MSc
2.º CICLO
FCUP
2018

U.PORTO

Antisense expression across the repeat insertion in *DAB1* and relevance for the pathogenic mechanism of SCA37

Ana Filipa Azevedo Castro

FC

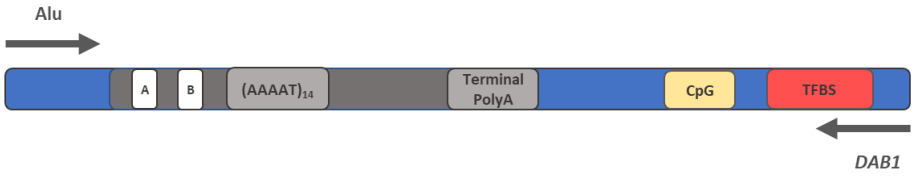
U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Antisense expression across the repeat insertion in *DAB1* and relevance for the pathogenic mechanism of SCA37

Ana Filipa Azevedo Castro

Master in Cell and Molecular Biology
Dissertation
Faculty of Science, University of Porto

2018



Antisense expression across the repeat insertion in *DAB1* and relevance for the pathogenic mechanism of SCA37

Ana Filipa Azevedo Castro

Dissertation

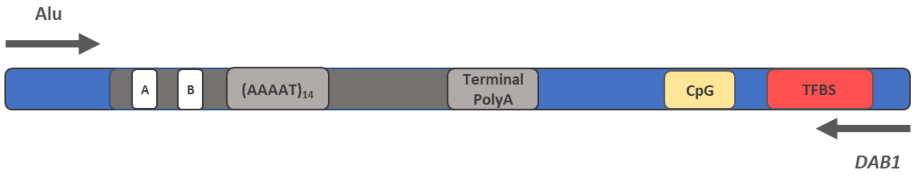
Master in Cell and Molecular Biology

Department of Biology, FCUP

2018

Supervisor

Isabel Silveira, PhD, IBMC, UP



Expressão da inserção repetitiva na cadeia de DNA contrária ao gene *DAB1* e a sua relevância para o mecanismo patogénico da SCA37

Ana Filipa Azevedo Castro

Dissertação

Mestrado em Biologia Celular e Molecular

Departamento de Biologia, FCUP

2018

Supervisor

Isabel Silveira, PhD, IBMC, UP

Ana Filipa Azevedo Castro
Master in Cell and Molecular Biology
Department of biology
Faculty of Sciences of University of Porto
Rua do Campo Alegre s/n
4169-007 Porto, Portugal
up201304512@fc.up.pt / afcastro@i3s.up.pt
Tel.: +351 220 408 800 / 910 367 791

Supervisor

Isabel Alexandra Azevedo Silveira, PhD
Group Genetics and Cognitive Dysfunction, Group Leader
IBMC – Institute for Molecular and Cell Biology
Instituto de Investigação e Inovação em Saúde
Room 214S2
Rua Alfredo Allen, 208
4200-135 Porto, Portugal
ISilveir@ibmc.up.pt
Tel.: +351 220 408 800

I, Ana Filipa Azevedo Castro, master student number 201304512 of Cell and Molecular Biology 2017/2018, hereby certify and declare on my honor the authenticity of this work. By signing this declaration, I acknowledge the consequences in case of copy of original works.

October 1, 2018

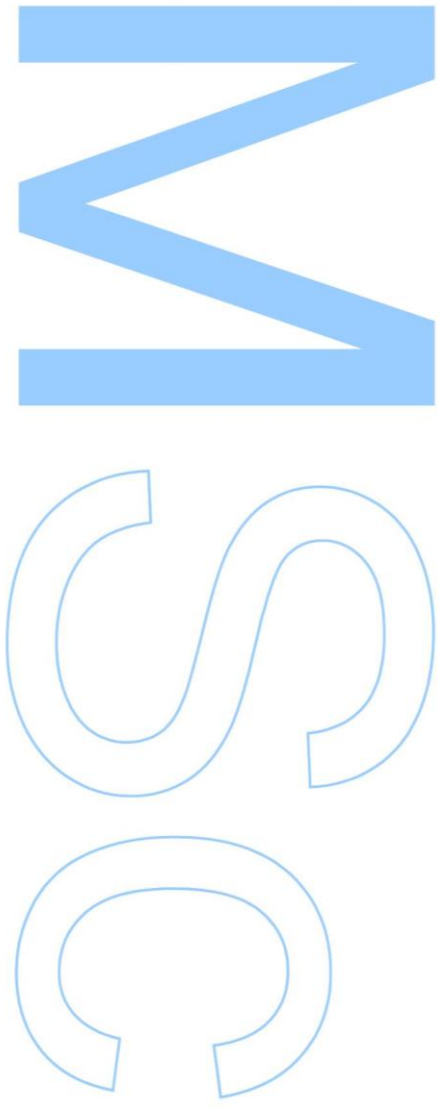
_____

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, _____/_____/_____



Acknowledgments

Este projeto foi realizado no grupo “Genetics of Cognitive Dysfunction”, no IBMC/i3S, no âmbito da dissertação para o Mestrado de Biologia Celular e Molecular da FCUP. A todas as entidades agradeço desde já pela oportunidade.

Para a realização deste projeto estiveram envolvidas várias pessoas às quais gostaria de agradecer pelo seu contributo, quer a nível científico, quer a nível pessoal.

Primeiro, gostaria de agradecer à Doutora Isabel Silveira, minha orientadora, pela sua total disponibilidade, permitindo a minha integração neste projeto do seu grupo. Pelo apoio, pelas opiniões e críticas construtivas e pelo conhecimento que me transmitiu, ajudando-me a solucionar os problemas que foram surgindo no seguimento do trabalho, o meu sincero obrigada.

Doutor José Bessa, o meu sincero agradecimento pela supervisão científica e pelo seu contributo, sem esta colaboração com o VDR, a realização deste projeto não seria possível. Primeiro, gostaria de agradecer por me aceitar no seu laboratório e no zebrário, depois pela sua disponibilidade, pelo seu apoio e conhecimento. O meu sincero obrigada pelo seu constante interesse que foi fundamental para este projeto chegar a este nível.

Joana, obrigada por tudo! Pelo conhecimento, pelo teu tempo, pela paciência… Realmente foste a “minha mãe de laboratório”, acompanhaste todo o trabalho, ensinaste-me tudo o que agora faço e ajudaste-me com todo o teu conhecimento científico. Também quero agradecer-te por todo o trabalho prévio que deu origem a este projeto. Obrigada, desejo-te o melhor mesmo!

Aos meninos do VDR, Renata, Fábio, Joana Teixeira, Marta, Joana Marques, Ana Gali, João, Ana Eufrásio, Diogo, Leonor e Isabel, por todo o apoio e dedicação, e por todos os bons momentos durante esta etapa, obrigada! Renata, pela tua total disponibilidade e por todas as vezes que paraste o teu trabalho para me ajudares, obrigada! Fábio, também me aturaste com as minhas dúvidas chatas (protocolos, reagentes,…), obrigada! Ana Eufrásio e Marta, obrigada pela vossa ajuda mesmo, sem vocês ainda não me teria entendido com o meu amigo confocal. Um agradecimento

muito especial à Joana Marques, levaste muitas vezes com as minhas dúvidas sobre técnicas e peixinhos, pela confiança e me teres emprestado as tuas linhas, e acima de tudo por todas as vezes que te chateei para marcar lupa, obrigada! Isabel, obrigada por todas aquelas palavras de força, aquelas que dizia às duas da tarde quando estava cansada de injetar e ainda não tinha almoçado, pela ajuda com todo o material, “vai almoçar querida, eu trato disso”, muito obrigada! Além disso, o tempo por vezes escasseia e sem a sua ajuda imprescindível, não teria peixinhos tão saudáveis, obrigada Isabel! Joana Teixeira, Ana Gali, João, Diogo e Leonor, não me esqueci de vocês, espero que saibam que também foram muito importantes mesmo, a vossa boa disposição, aquela música ou conversa durante as injeções foram imprescindíveis naqueles dias mais cansativos, obrigada!

Aos outsiders, Mafalda, Nuno e Ana André, não me esqueci de vocês! Quero que saibam que ~~foram~~ são muito importantes, obrigada pela vossa amizade. Por todas as conversas, a galhofa, o apoio, que permitiram manter a minha sanidade mental. Obrigada, meninos!!!

Gostaria ainda de agradecer aos meus pais e ao meu irmão pelo apoio e pelo investimento na minha educação académica, sem vocês não estaria agora a fazer o que gosto. Obrigada pela oportunidade e confiança.

Por último, mas não menos importantes, aqueles que me conhecem há anos e que sempre me apoiaram. Sim, estou a referir-me a vocês meninos, Ana, Ana Sofia, Bia, Daniela, Miguel e Tânia. Apesar de não estarem relacionados diretamente com este trabalho, vocês são sempre aqueles que me fazem rir quando volto a casa. Caminhos diferentes, tempo escasso, mas há sempre os nossos cafés mesmo quando estamos cheios de sono. Obrigada pela vossa amizade!

Abstract

Spinocerebellar ataxias (SCAs), whose cause is associated with mutations in short tandem repeats (STRs), are neurodegenerative diseases, characterized by a cerebellar atrophy and, consequently, a motor impairment. In 2017, an $(ATTTC)_n$ insertion within a normal $(ATTTT)_n$ sequence in a *DAB1* 5' UTR intron was identified as the causative mutation of SCA37. Although the main pathogenic mechanism remains unclear, an RNA-mediated toxicity may be implied in the SCA37 phenotype, as shown in a zebrafish model, demonstrating that $(AUUUC)_n$ insertion is deleterious *in vivo*. Sense and antisense transcription in noncoding repeat regions associated with neurodegenerative diseases interferes with normal cellular pathways by an RNA gain-of-function mechanism. Antisense transcripts may also play a role in SCA37 pathogenicity. Previous work of our team evidenced the existence of promoter activity in *DAB1* antisense strand, which was able to trigger transcription in zebrafish muscular and cerebellar tissues. In this project, we characterized the expression of two promoters in zebrafish embryos. Thus, we detected promoter expression in horizontal myosepta and muscular fibers, revealing that these promoters are tissue-specific. Moreover, we studied the potential of these promoters to interact with *cis*-regulatory elements as enhancers, demonstrating that antisense promoters are able to stablish a regulatory network with the midbrain-specific enhancer Z48, enlightening, more accurately, that antisense promoter activity may be regulated by brain-specific regulatory elements. As *in silico* analyses demonstrated, $(ATTTC)_n$ insertion may be located within a DNA regulatory region and, as regulatory elements may modulate the SCA37 toxicity, the existence of regulatory elements as enhancers were also investigated. However, no evidence of enhancer activity was detected in the repeat flanking region in *DAB1* orientation. Finally, to clarify the role of antisense transcripts in SCA37 toxicity, we verified that the $(GAAAU)_n$ RNA is not toxic for zebrafish embryos, suggesting that other regulatory role may be underlying its function. In summary, we characterized regulatory elements in the *DAB1* repeat flanking region to have insight in the SCA37 pathogenic mechanism.

Keywords Spinocerebellar ataxia type 37 (SCA37), neurodegeneration, antisense transcription, regulatory elements, cellular toxicity

Resumo

As ataxias espinocerebelosas (SCAs), cuja causa está associada a mutações em microssatélites, são doenças neurodegenerativas, caracterizadas por uma atrofia cerebelar e, consequentemente, uma perda motora. Em 2017, uma inserção $(ATTTC)_n$ numa sequência $(ATTTT)_n$ normal, localizada na 5' UTR da região intrónica do gene *DAB1*, foi identificada como causa da SCA37. Apesar do principal mecanismo patogénico ainda não ter sido esclarecido, uma toxicidade mediada pelo RNA pode estar implicada no fenótipo observado na doença, como demonstrado com o modelo peixe-zebra, no qual a inserção $(ATTTC)_n$ foi deletéria *in vivo*. A transcrição a partir de ambas as cadeias de DNA, em regiões repetitivas não-codificantes, associada a doenças neurodegenerativas, interfere com vias celulares normais através do ganho de função do RNA. Os transcritos da cadeia oposta ao gene *DAB1* também podem contribuir para um aumento ou diminuição da patogenicidade observada na SCA37. Trabalho prévio do nosso grupo evidenciou a existência de atividade promotora na cadeia oposta ao gene *DAB1*, que levou à transcrição nos tecidos muscular e cerebelar de peixe-zebra. Neste projeto, nós caracterizámos a expressão de dois promotores em embriões de peixe-zebra, detetando assim expressão dos promotores no miosepto horizontal e nas fibras musculares, revelando que esses promotores são específicos de determinado tecido. Além disso, estudámos o potencial desses promotores para interagir com elementos regulatórios em *cis* como *enhancers*, demonstrando que esses promotores na cadeia oposta ao gene *DAB1* conseguem interagir com o *enhancer* Z48, que é específico do mesencéfalo, revelando assim que estes promotores podem estar a ser regulados por elementos regulatórios específicos de cérebro. Como demonstrado através de análises *in silico*, a inserção $(ATTTC)_n$ pode estar localizada numa região regulatória do DNA e, como elementos regulatórios podem modular a toxicidade observada na SCA37, a existência de elementos regulatórios como *enhancers* foi também investigada. Contudo, nenhuma evidência de atividade de *enhancers* foi detetada na região flanqueante da sequência repetitiva $(ATTTT)_n$. Para elucidar a função desempenhada pelos transcritos da cadeia oposta ao gene *DAB1* na toxicidade observada na SCA37, verificámos que o RNA $(GAAAU)_n$ não é tóxico para embriões de peixe-zebra, sugerindo ter outra função regulatória. Em sumário, caracterizámos elementos regulatórios na região flanqueante ao trato repetitivo do gene *DAB1* para perceber o seu contributo para o mecanismo patogénico subjacente à SCA37.

Palavras-chave Ataxia espinocerebelosa tipo 37 (SCA37), neurodegeneração, transcrição bidirecional, elementos regulatórios, toxicidade celular

Table of contents

Introduction 1
Repeat diseases 1
Overview 1
Pathogenic pathways triggered by repeat mutations 3
Coding repetitive regions 3
Noncoding repetitive regions 3
Bidirectional transcription across repetitive tracts 5
Spinocerebellar ataxias 6
Spinocerebellar ataxia type 37 7
Finding gene regulatory DNA sequences 11
Regulatory DNA sequences in SCA37 15
Aims 17
Materials and methods 19
Zebrafish husbandry and manipulation 19
Cloning 19
Promoter activity characterization in *DAB1* antisense orientation 20
Promoter-enhancer interaction 20
Enhancer detection assay 21
In vitro Tol2 RNA synthesis 21
Microinjections 21
Imaging acquisition and analysis 22
Laser Scanning Confocal Microscopy 22
RNA toxicity assay 23
Results 25
Antisense transcription driven by upstream and downstream SCA37 repeat sequences 25
Promoters in the SCA37 region interact with a brain enhancer 29
No enhancer activity in the SCA37 repeat sequence 36
Antisense $(GAAAU)_n$ RNA is not toxic in zebrafish embryos 37
Discussion 39
Future perspectives 42
References 45
Appendix 1 53

Phenol chloroform DNA/RNA purification 53
Appendix 2 55
In vitro RNA synthesis 55
Appendix 3 57
Purification of RNA transcribed *in vitro* 57
A. Sephadex column preparation 57
B. RNA purification in Sephadex column 57
Appendix 4 59
Zebrafish microinjection 59
A. Courtship and spawning 59
B. Microinjection 60
Appendix 5 61
Appendix 6 63
Appendix 7 65
Appendix 8 67
Appendix 9 69

List of figures

Figure 1 Overview about the pathological pathways underlying the neurodegeneration in repeat expansion diseases ……..…..... 2

Figure 2 Schematic representation of $(ATTTC)_n$ insertion within a normal $(ATTTT)_n$ tract at the *DAB1* intergenic region ……..... 8

Figure 3 Pathological ins$(AUUUC)_n$ RNA leads to the formation of nuclear aggregates in SCA37 ……..... 9

Figure 4 RNA toxicity underlying SCA37 ……... 10

Figure 5 Genomic context of the repeat flanking regions in *DAB1* antisense strand or Alu-oriented strand ……..... 11

Figure 6 Gene regulation network ……... 13

Figure 7 Promoter activity assessment …….. 14

Figure 8 Zebrafish Enhancer Detection (ZED) vector …….. 15

Figure 9 Promoter activity across the repeat region of the *DAB1* antisense strand ... 16

Figure 10 Promoter activity across *DAB1* antisense strand ……... 26

Figure 11 Antisense transcription driven by the upstream SCA37 repeat sequence (Frag 1) 27

Figure 12 Antisense transcription driven by downstream SCA37 repeat sequence (Frag 3) 28

Figure 13 Schematic representation of promoter activity in *DAB1* antisense strand .. 29

Figure 14 Interaction between putative promoter sequences and midbrain-specific enhancer Z48 …….. 30

Figure 15 Characterization of EGFP expression driven by Frag T promoter-Z48 interaction in midbrain cells from 24 hpf to 72 hpf (A) and at 48 hpf (B) ……..... 31

Figure 16 Microinjected embryos expressing EGFP, at 48 hpf …….. 32

Figure 17 Characterization of EGFP expression triggered by Frag 1 promoter and Z48 enhancer interaction ……... 33

Figure 18 Analysis of Frag 2 and Z48 enhancer interaction ……... 34

Figure 19 EGFP expression driven by Frag 3 promoter and Z48 enhancer interaction ……..... 35

Figure 20 Evaluation of enhancer activity across Frag T in *DAB1* orientation ……. 36

Figure 21 No *in vivo* deleterious effect of the antisense ins$(GAAAU)_{54}$ RNA injection in zebrafish embryos ……... 38

Figure appendix 4 Apparatus of zebrafish breeding tank 59

Figure appendix 5 Characterization of F1 embryos expressing Frag 1 promoter 61

Figure appendix 6 Characterization of F1 embryos expressing Frag 3 promoter 63

Figure appendix 7 Evaluation of Frag T promoter-Z48 enhancer in F0 zebrafish embryos along the time 65

Figure appendix 8 Evaluation of Frag 1 promoter-Z48 enhancer in F0 zebrafish embryos 67

Figure appendix 9 Evaluation of Frag 3 promoter-Z48 enhancer in F0 zebrafish 69

List of abbreviations

ATXN8	ataxin 8
ATXNOS	ataxin 8 opposite strand
CAGE	Cap analysis gene expression
cDNA	complementary DNA
4C	Chromatin conformation capture circular
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
CRISPR-Cas9	Clustered Regularly Interspaced Short Palindromic Repeats-associated protein-9 nuclease
DAB1	*DAB1*, Reelin adaptor protein
DPF	Days postfertilization
DM	Myotonic dystrophy
DNA	Deoxyribonucleic acid
DRPLA	Dentatorubropallidoluysian atrophy
E3	Embryos medium
EGFP	Enhanced green fluorescence protein
FCUP	Faculty of Sciences of University of Porto
FMR1	Fragile X mental retardation 1
Frag	Fragment
FRDA	Friedreich's ataxia
FTD/ALS	Frontotemporal dementia/Amyotrophic lateral sclerosis
FXN	Frataxin
FXS	Fragile X Syndrome
FXTAS	Fragile X Tremor Ataxia Syndrome
HPF	Hours postfertilization
HD	Huntington's Disease
HDL2	Huntington's Disease Like-2
Hi-C	High-throughput chromosome conformation capture
i3S	Instituto de Investigação e Inovação para a Saúde
IBMC	Institute for Molecular and Cell Biology
IPATIMUP	Institute of Molecular Pathology and Immunology of the University of Porto
MBNL	Muscleblind-like

MJD	Machado-Joseph Disease
OPMD	Oculopharingeal muscular dystrophy
PBS	Phosphate-buffered saline
PolyA	Polyalanine
Poly-AP	Poly – Alanine-Proline
PolyG	Polyglicine
Poly-GP	Poly – Glycine-Proline
PolyP	Polyproline
Poly-PR	Poly – Proline-Arginine
PolyQ	Polyglutamine
PolyR	Polyarginine
1-phenil-2-thiourea	PTU
RACE	Rapid amplification of cDNA ends
RAN	Repeat-associated non-AUG
RNA	Ribonucleic acid
RNABP	RNA binding protein
RT-PCR	Reverse transcriptase-polymerase chain reaction
SBMA	Spinal bulbar muscular atrophy
SCA	Spinocerebellar ataxia
SINES	Short interspersed nuclear elements
SRSF	Serine/arginine-rich splicing factors
STR	Short tandem repeat
TFBS	Transcription factor binding site
TSS	Transcription start site
uORF	Upstream open reading frame
UTR	untranslated region
XLMR	X-linked mental retardation
WT	Wild-type
ZED	Zebrafish enhancer detection

Introduction

Repeat diseases

Overview

Approximately 3% of the human genome including exons, introns and 5' and 3' untranslated regions (UTRs), contains repetitive sequences of 1-6 nucleotides, called short tandem repeats (STRs)[1]. Several neurological and neuromuscular diseases are caused by repeat length expansion above a normal threshold or insertion of STRs, located in coding and noncoding regions[2].

In 1991, the fragile X syndrome (FXS) and spinal bulbar muscular atrophy (SBMA or Kennedy's disease) were the first repeat diseases reported caused by repeat expansions[1]. Thenceforth, many other neurological and neuromuscular diseases such as spinocerebellar ataxias (SCAs), myotonic dystrophy types 1 and 2 (DM1 and DM2), Huntington's disease (HD), dentatorubropallidoluysian atrophy (DRPLA), frontotemporal dementia (FTD)/amyotrophic lateral sclerosis (ALS), oculopharyngeal muscular dystrophy (OPMD), syndromic and non-syndromic X-linked mental retardation (XLMR) caused by trinucleotide, tetranucleotide, pentanucleotide or hexanucleotide repeat expansions have been found[3; 4].

Previously, only repeat expansion diseases had been associated with STR sequences, however, more recently, repeat insertions have been found pathogenic in SCA31, SCA37 and benign familial myoclonic epilepsy [5-7]. This new type of mutations consists in a pathogenic repetitive tract insertion, not found in normal individuals, within a normal polymorphic repetitive region. Both, repeat expansions and insertions, are dynamic mutations due to their ability to expand and contract upon intergenerational transmission. Somatic mosaicism among different tissues in the same individual has also been observed. Moreover, in the same family, from a generation to another, an increased severity of the disease may be observed since the repeat tend to expand, which causes an earlier onset of the disease, as occurs in SCAs 1, 2, 3 and 7, and, in DRPLA[8].

Repeat mutations may cause disease by three main mechanisms: gene loss-of-function, RNA gain-of-function and protein gain-of-function (**Figure 1**), that will be detailed in the following sections[9].

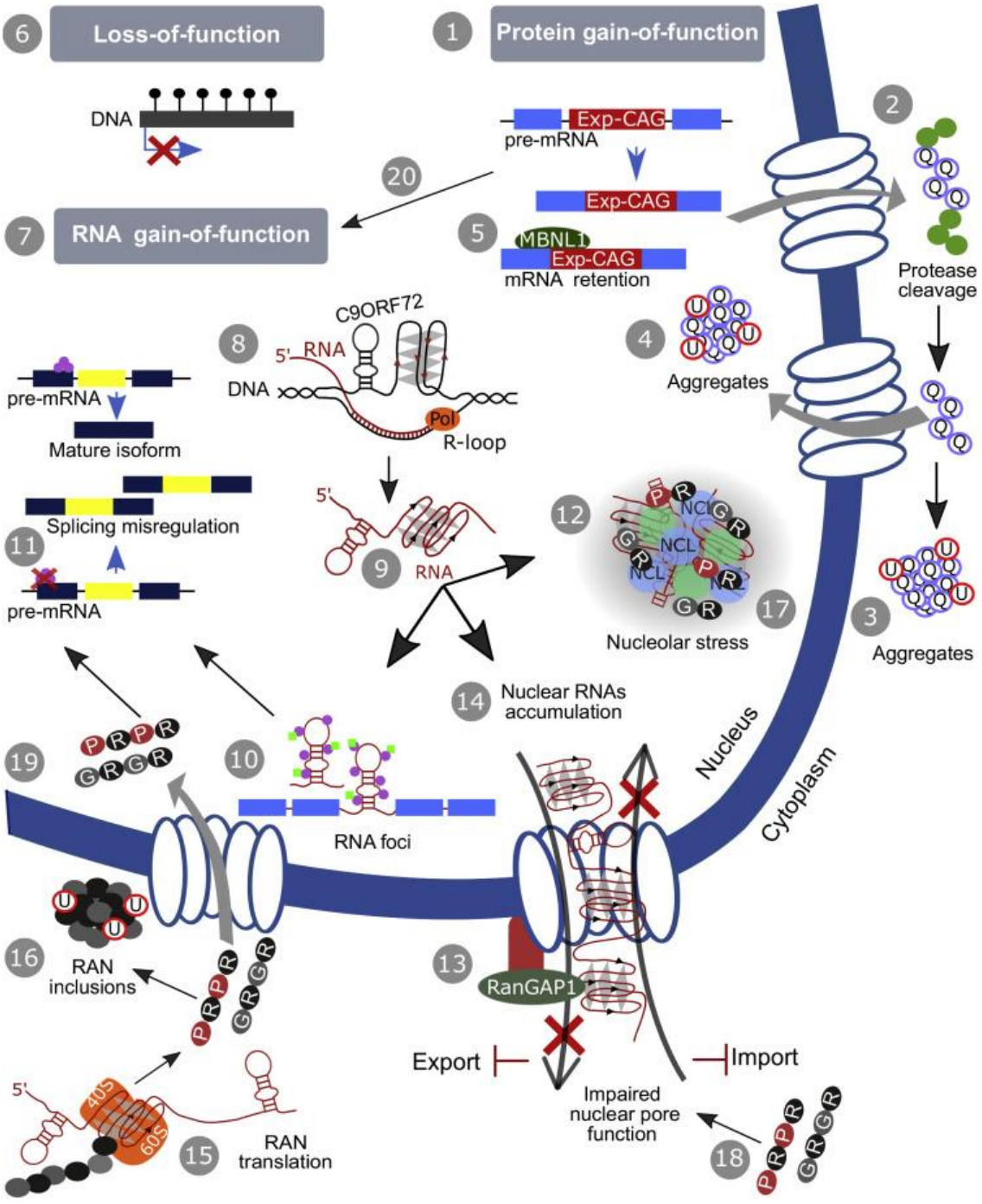


Figure 1 Overview about the pathological pathways underlying the neurodegeneration in repeat expansion diseases. In coding regions, the protein gain-of-function is the principal mechanism, responsible for the cellular toxicity. After gene transcription, the expanded CAG RNA are translated into proteins containing expanded polyQ stretches. These mutant proteins form aggregates in the cell, compromising the native function of these proteins. In noncoding regions, the two major mechanisms are gene loss-of-function and RNA gain-of-function. Some repeats lead to epigenetic alterations and gene silencing. These repeat RNAs form RNA foci, which recruit several RNABP, leading to a disruption of biological pathways in which these RNABP are involved. Furthermore, these repeat RNAs may be translated by RAN-translation, contributing to cellular toxicity. Adapted from Loureiro et al., 2016.

Pathogenic pathways triggered by repeat mutations

Coding repetitive regions

Protein gain-of-function

Pathogenic repeats in DNA coding regions include expansions of trinucleotide CAG and GCN repeats, which are translated into proteins containing polyglutamine (polyQ) or polyalanine (polyA) repetitive tracts. The trinucleotide repeat expansions encode long tracts of amino acids that compromise the native function of the protein, and those abnormal proteins have the ability to form aggregates in the cytoplasm and in the nucleus of the cell. This pathogenic pathway occurs by a protein gain-of-function process and is detected in all pathogenic repeat expansions located in coding DNA regions, like those that cause SCA 1, 2, 3, 6, 7 and 17, in addition to other neuronal diseases such as SMBA, HD, DRPLA, and OPMD[4; 9; 10].

Noncoding repetitive regions

In noncoding regions, several expanded repetitive sequences causing disease have been reported associated with trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide repeats[3]. These repeats may be pathogenic either when the transcription is inhibited, and the gene is silenced causing gene loss-of-function, or when the gene is transcribed, and the RNA acquires a function that leads to cellular toxicity by RNA gain-of-function.

Gene loss-of-function

In diseases resulting from a gene loss-of-function mechanism, the gene expression may be decreased or even silenced by epigenetic modifications, leading to reduction or even absence of protein product, as occurs in FXS and Friedreich's ataxia (FRDA)[1]. In FXS, $(CGG)_n$ expansion above 200 units, in the 5' UTR of fragile X mental retardation 1 (*FMR1*) gene, leads to CpG island methylation and hence to gene silencing and absence of the gene protein FMRP[4; 11; 12]. In FRDA, the GAA expansion in intron 1

of frataxin (*FXN*) gene is responsible for histone modifications in flanking repeat regions, which induces heterochromatin formation and *FXN* transcription inhibition. Consequently, a decrease of *FXN* expression leads to a reduction of the frataxin protein[11].

RNA gain-of-function

In several noncoding gene regions, the mutant repeats are transcribed producing an expanded RNA, which contributes to toxicity in the cell. Those RNAs cause toxicity in the cell by an RNA gain-of-function mechanism. The repetitive RNA is responsible for (1) RNA foci formation, (2) disruption of splicing in other cellular mRNAs originated by sequestration of RNA-binding proteins (RNABPs), mainly splicing factors, in RNA foci, (3) abortive transcripts resulting from R-loops formation, (4) nucleocytoplasmic transport impairment, and (5) noncanonical translation[3]. In noncoding repeat diseases, wherein the gene is transcribed, the repetitive sequences of RNA form secondary structures, which sequester several RNABPs like splicing or regulatory factors that modulate splicing, leading to nuclear aggregates formation. Thus, the expanded transcripts may disturb the normal function of essential proteins by affecting their accessibility in the cell[1; 13]. Many splicing factors recruited by secondary and quaternary repetitive RNA structures, in foci, have been described for diseases like DM types 1 and 2, SCA31, SCA36, Fragile X Tremor Ataxia Syndrome (FXTAS) and FTD/ALS[1; 5; 14-16]. For instance, in both DM1 and 2, the muscleblind-like (MBNL) proteins are retained by the expanded CUG/CCUG through hairpin structures, leading to splicing misregulation of several MBNL gene targets[1; 12; 17; 18]. In turn, sequestration of serine/arginine-rich splicing factors (SRSFs) by mutant RNAs may also lead to a splicing dysregulation, as described in SCA31 and SCA36[5; 15]. Furthermore, the formation of RNA/RNABP aggregates increases the cellular toxicity, interfering with the RNA processing (splicing, nuclear export, localisation) and translation[3; 12; 19]. Some proteins of the nuclear pore complex bind to these RNA foci, compromising the nucleocytoplasmic transport, as referenced in FTD/ALS, which is caused by a hexanucleotide expansion in *C9ORF72*[1; 3; 13; 19-21].

The abnormal repeat RNA may acquire unusual conformations derived from mismatch paired nucleotides with DNA or RNA, in repetitive stretches[22]. One such conformations are the R-loops (DNA::RNA hybrid) formed during transcription[3; 23], other are the hairpins; triple-stranded and slipped-DNAs structures can contribute to repeat instability, as in FRDA and FXS, or to transcription abortion[2].

The expanded RNAs may escape to the cytoplasm, inducing their noncanonical translation[1]. Repeat-associated non-AUG (RAN) translation is a noncanonical translation mechanism observed in several neurological diseases caused by repeat expansions[24; 25]. The secondary structures of the repetitive RNA are able to recruit ribosomes and initiate translation without an AUG-initiation codon across the three reading frames[1]. Thus, the repeat expanded RNA stretches is translated into homopolymeric, dipeptide, tetrapeptides and pentapeptide proteins, depending on the repeat, and these translational products aggregate in the cytoplasm and are transported to the nucleus, increasing the neuronal toxicity[24-29]. RAN translation was firstly reported in 2011, for SCA8. The stable RNA hairpins formed by the CAG repeats from the ataxin 8 (*ATXN8*) recruit the translational machinery and trigger the initiation of translation, leading to the production of polyQ proteins. These homopolymeric proteins accumulate particularly in the nuclei of the Purkinje cells, leading to an increase of toxicity in cerebellar tissues[24; 30].

Some upstream open reading frames (uORFs) in the 5' UTR may trigger RAN translation. These sequences are able to recruit ribosomes through near-cognate codons (e.g. UUG or CUG) and initiate the translation without an AUG-initiation codon. Thus, the RAN translation may be initiated upstream of the repeat, leading to different repetitive proteins and, consequently, different outcomes in repeat diseases, as reported in FXTAS and FTD/ALS[24; 25; 27; 31].

Bidirectional transcription across repetitive tracts

Bidirectional transcription has been identified in several repeat diseases located in coding and noncoding regions[16; 32-36]. The repeats may be transcribed from both strands such as in SCA7, SCA8, HD, HDL2, FXTAS, DM1 and FTD/ALS[35;34 36; 39]. Sense and antisense transcripts contribute to the disease phenotype in SCA7, SCA8, FXTAS and FTD/ALS[16; 35-37].

In SCA8, both protein and RNA gain-of-function are associated with bidirectional transcription. An expanded CAG in *ATXN8* gene orientation encodes a toxic polyQ protein and, in the *ataxin8 opposite strand* (*ATXNOS*) an untranslated CUG RNA is transcribed, leading to RNA foci formation in cerebellar cortex[36].

In FXTAS and FTD/ALS, both sense and antisense transcripts were identified, contributing to RNA gain-of-function. In FXTAS, which is caused by a premutation (55-200 CGG repeats) in the 5' UTR of the *FMR1*, the RAN translation occurs from both sense and antisense transcripts across different reading frames. The repetitive RNA from

the expanded sense transcripts is translated into polyglycine (polyG) proteins, whereas from the antisense transcripts, proteins containing polyproline (polyP), polyarginine (polyR) and polyalanine (polyA) tracts are synthesized. These proteins are responsible for the formation of intranuclear inclusions in neuronal tissues[16; 24; 25; 27]. In FTD/ALS, the repetitive RNA resulting from the expression of *C9ORF72* antisense strand also forms RNA foci and may be translated into different dipeptide-repeat proteins: poly-AP (poly – Alanine-Proline), poly-PR (poly – Proline-Arginine) and poly-GP (poly – Glycine-Proline)[26; 38; 39]. Studies have demonstrated that these dipeptide-repeat proteins interfere with the cellular processes and form neuronal inclusions, increasing the cellular toxicity[35; 39; 40].

The identification of repetitive loci bidirectional transcribed have revealed that not only the repeat expression in gene-oriented strand drives to pathogenicity, but also the expression in the antisense strand may interfere with different cellular pathways, leading to an increase in toxicity.

Spinocerebellar ataxias

SCAs are a group of autosomal dominant inherited neurodegenerative diseases, mainly characterized by cerebellar atrophy due to Purkinje cell loss[8; 41; 42]. The degeneration may extend to the brainstem and spinal cord[8; 9].

Up to date, 44 types of SCAs have been genetically identified[6; 43]. They are caused by classical mutations such as duplications, deletions or point mutations and dynamic mutations like repeat expansions and insertions that can be located in both coding and noncoding DNA regions[42; 43]. The SCAs associated with STRs are either caused by CAG repeat expansions in coding regions, such as SCA 1, 2, 3, 6, 7 and 17, or by trinucleotide, pentanucleotide and hexanucleotide repeats in noncoding regions. In 2009, the first repeat insertion mutation was discovered associated with the SCA31, which is caused by a $(TGGAA)_n$ insertion within a normal polymorphic $(TAAAA)_n$ in an intron of brain expressed, associated with Nedd4 (*BEAN*) gene[5]. In 2017, the second repeat insertion was found in SCA37 by the laboratory led by Dr Isabel Silveira, wherein this project is being performed. SCA37 is caused by an $(ATTTC)_n$ insertion in the middle of an $(ATTTT)_n$ located in an intronic region of the *DAB1, reelin adaptor protein* (*DAB1*) gene[6].

The estimated prevalence of SCAs varies according to the geographic region from 1.6 to 5.6 per 100 000 inhabitants[6]. In Portugal and worldwide, the most common SCA is SCA3/Machado-Joseph Disease, followed by SCA2, SCA37 and DRPLA[44].

Spinocerebellar ataxia type 37

The SCA37 is a progressive pure cerebellar ataxia characterized by dysarthria as the first symptom, whose onset ranges from the late teens to the 6th decade.

SCA37 is caused by an $(ATTTC)_n$ insertion in a noncoding region of *DAB1* gene. The mutation associated with SCA37 was described in 2017, by the research group led by Dr Isabel Silveira, in six families from southern Portugal[6]. The published work reported the genetic mapping of the disease, the discovery of the mutation and an RNA gain-of-function involvement in the pathology. In parallel with this work, in 2013, a Spanish group reported the clinical description and genetic mapping of SCA37 to chromosome 1p32, in a Spanish kindred[45]. This last study found that affected individuals are characterized by a pure cerebellar ataxia and abnormal vertical eye movements. Although abnormal vertical eye movements have not been detected in Portuguese SCA37 families, the remaining clinical features overlap in all families.

SCA37 is caused by an $(ATTTC)_n$ insertion within an $(ATTTT)_n$ sequence in a 5' UTR intron of *DAB1* (DAB1, reelin adaptor protein) on chromosome 1p32. The $(ATTTT)_n$ tract observed in nonpathological alleles can vary from 7 to 400 repetitive units, whereas the pathogenic pentanucleotide alleles have a configuration $[(ATTTT)_{60-79}(ATTTC)_{31-75}(ATTTT)_{58-90}]$ (**Figure 2**). The $(ATTTC)_n$ insertion shows intergenerational instability, with a tendency to increase in size when transmitted to the next generation. When the father is the transmitting parent, this instability is increased compared with when the mother is the transmitting parent. The SCA37 pentanucleotide is transcribed and the pathogenic mechanism involved in the disease is mediated by a toxic RNA, as shown by Seixas et al[6].

In SCA37, the mutant transcripts form RNA aggregates in the nucleus of human cell lines (**Figure 3**). Like expanded repeat RNAs reported in other diseases[1; 13], the $(AUUUC)_n$ RNA may have the ability to bind RNABP, forming pathological RNA foci that dysregulate normal cellular pathways.

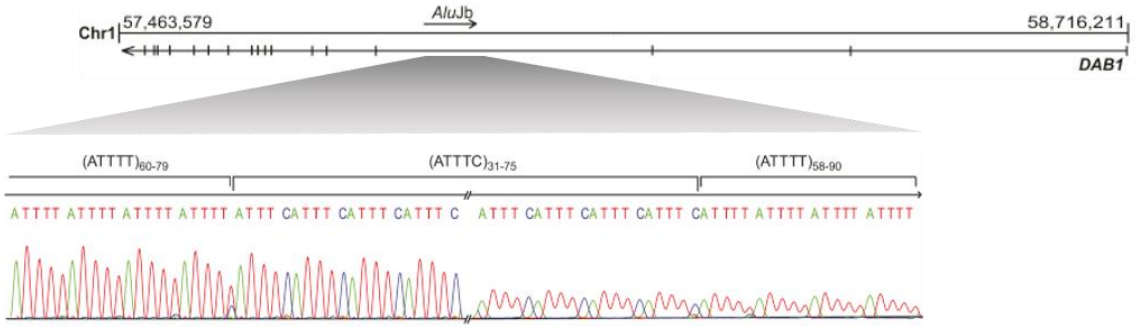


Figure 2 Schematic representation of $(ATTTC)_n$ insertion within a normal $(ATTTT)_n$ tract at the *DAB1* intergenic region. Pathological insertion in the *DAB1* antisense strand is in the middle poly(A) of an AluJb sequence, on chromosome 1. SCA37 affected individuals have the pathogenic $(ATTTC)_n$ allele that ranges from 31 to 75 pentanucleotide repeats, flanked by normal $(ATTTT)_n$, as represented in electropherogram. Adapted from Seixas et al., 2017.

Danio rerio, commonly known as zebrafish, is a teleost fish from the Southeast Asia, that became one of the most important vertebrate model organisms for developmental studies. The zebrafish model has also been widely used in studies related to functional genomics and human neuronal diseases due to its great homology with human. Zebrafish, whose complete genomic sequence is already known, have a rapid development and large broods all over the year[46-48]. Several technical approaches have been developed, enabling an easy zebrafish manipulation and its inexpensive husbandry in small laboratorial areas. Furthermore, the transparency of the embryos is a critical feature for live-imaging analyses[48; 49]. Given these unique features, molecular research in neurodegenerative diseases such as Parkinson's disease, Alzheimer's disease, ALS, and spinocerebellar ataxias like SCA3/MJD and SCA13 have used zebrafish models[50-52]. Due to the advantages in usage of this animal model and the published reproducible results in other neurodegenerative diseases, zebrafish was used to perform functional studies in SCA37.

When the mutant RNA insertion was microinjected into one to two-cell stage zebrafish embryos, an increased toxicity was also observed, leading to an abnormal development and, an increased lethality rate at 24 hours postfertilization (hpf) compared with embryos injected with RNA from the nonpathological alleles[6] (**Figure 4**). As described for other repeat diseases, repeat RNA may lead to a toxic cellular environment or may be translated by noncanonical mechanisms such as RAN-translation, being toxic for the cells[1; 12].

The SCA37 mutation is located in the *DAB1* gene, which is a neurodevelopmental gene, involved in early neuronal migration. DAB1 protein is a cytoplasmic adaptor of the reelin signalling pathway, essential for neuronal cell migration and brain development. Four *DAB1* transcriptional variants have been identified spanning the repetitive region in the sense strand. The gene expression is higher in human fetal brain than in adult cerebellum[6]. A recent study demonstrated an increase of *DAB1* RNA and protein expression in post-mortem SCA37 cerebella compared with controls[53].

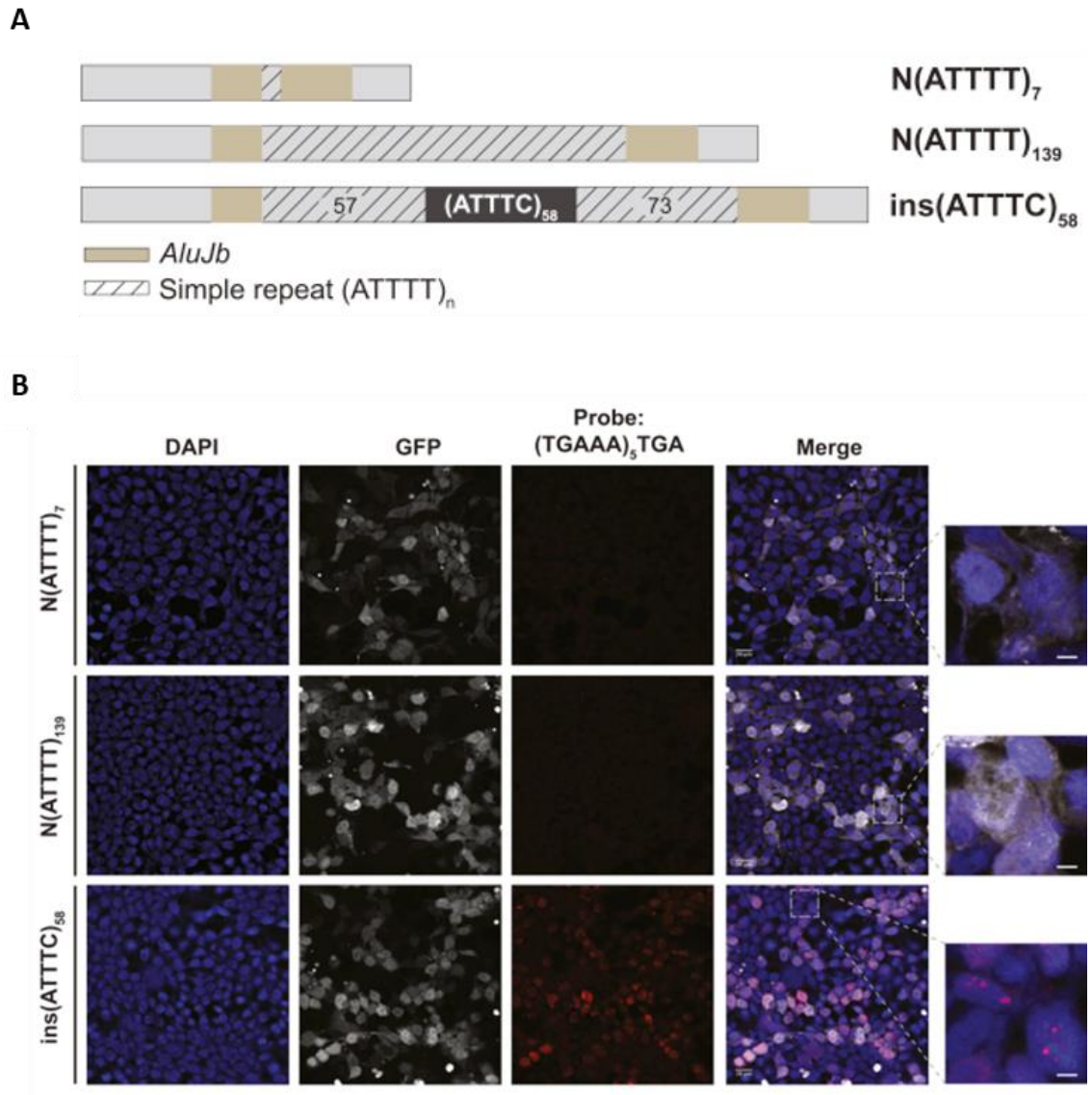


Figure 3 Pathological ins(AUUUC)$_n$ RNA leads to the formation of nuclear aggregates in SCA37. (**A**) Schematic representation of normal N(ATTTT)$_7$ and N(ATTTT)$_{139}$ alleles and the pathological ins(ATTTC)$_n$ allele, used to understand the RNA toxicity in this neurological disease. (**B**) Overexpression of the pathological ins(ATTTC)$_n$ in transfected human embryonic cell line HEK293T demonstrated the ability of this mutant RNA to form nuclear aggregates, contrarily to the normal alelles, which suggests that RNA may have an important role for the SCA37 pathogenicity. Adapted from Seixas et al., 2017.

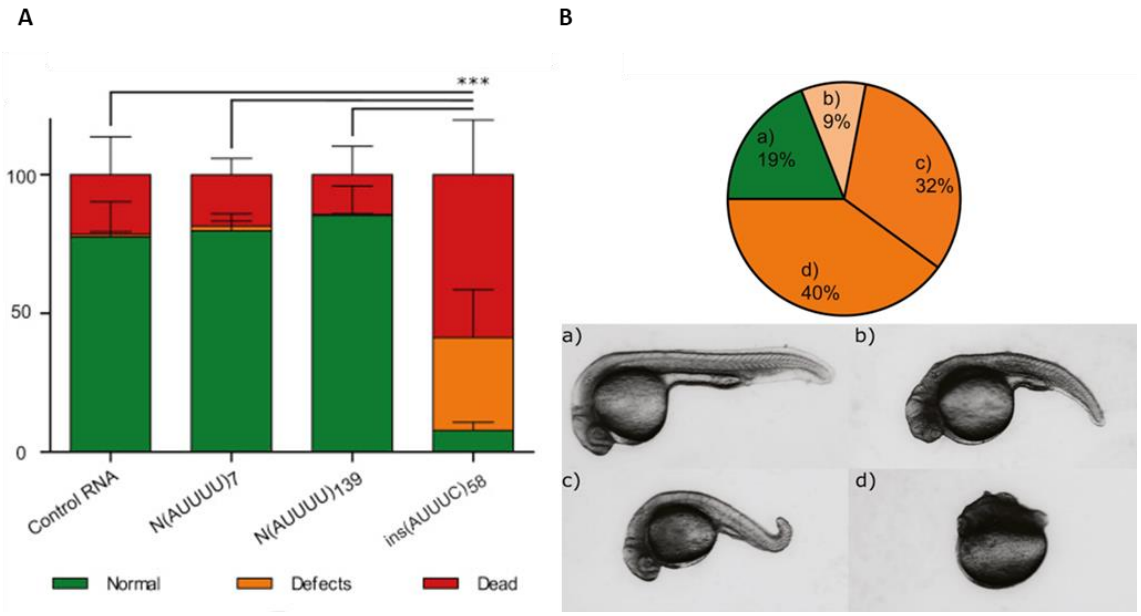


Figure 4 RNA toxicity underlying SCA37. (**A**) Zebrafish embryos that were injected with the pathological ins(AUUUC)$_{58}$ RNA demonstrated an increased lethality rate when compared with those injected with the normal N(AUUUU)$_{7}$ and N(AUUUU)$_{139}$ RNAs or the Cas9 RNA control ($p < 0.001$; χ^2 test for the lethality rate). (**B**) Pathological ins(AUUUC)$_{58}$ injection also led to an increase of developmental defects, as observed in embryos (b), (c) and (d) when compared with wild-type phenotype (a). Adapted from Seixas et al., 2017.

In the *DAB1* opposite strand, the (ATTTT/AAAAT)$_{n,}$ is located in the middle polyA of an Alu element (**Figure 5**). In other repeat diseases such as FRDA, DM2, SCA10 and SCA31, the pathogenic repeats are also within Alu elements[5; 54-56]. Alu elements are one of the most abundant transposable elements in the human genome, belonging to the short interspersed nuclear elements (SINEs). They are primate specific retroelements, composed by two different monomeric sequences separated by a short polyA, having also a longer polyA at the 3' end. Considering the phylogenetic age, the Alu elements are mainly classified into three subfamilies with the AluJ being the most ancestral; this is followed by the AluS and finally the AluY is the youngest subfamily of Alus. Only some elements of the AluY subfamily are actually transposable active in the human genome[57-60].

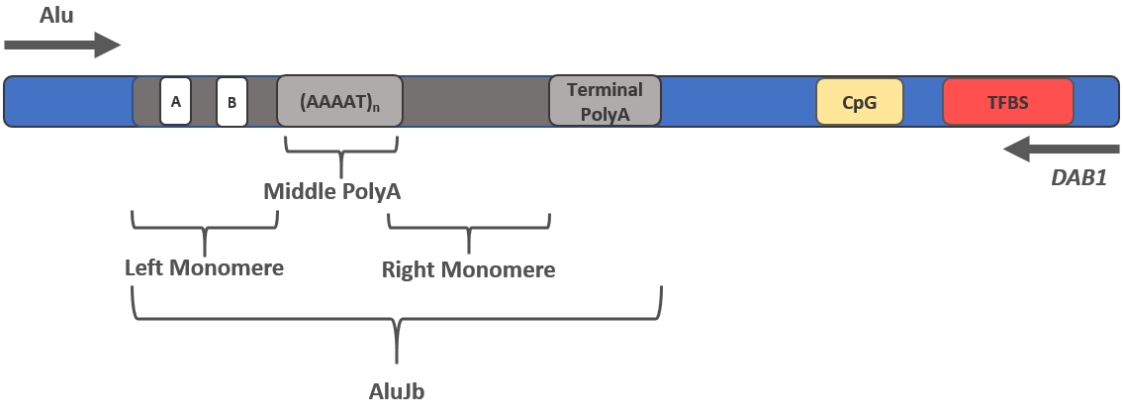


Figure 5 Genomic context of the repeat flanking regions in *DAB1* antisense strand or Alu-oriented strand. Alu elements are composed by two monomers separated by a middle polyA, containing a longer polyA at 3' end. Boxes A and B, which are present in the left monomer, are two internal promoter regions. In Alu-oriented strand, a polymorphic $(AAAAT)_n$ region is located in the middle polyA of an AluJb element. Preliminary *in silico* analysis in UCSC genome browser revealed a CpG island and a TFBS at downstream of the repetitive region

The genomic context of the SCA37 pentanucleotide is complex in the *DAB1* opposite strand, the pentanucleotide is located within an Alu element which is upstream of a CpG island and a transcription factor binding site (TFBS) (**Figure 5**). Preliminary *in silico* analysis of the repeat genomic context, using the UCSC genome browser, raised the hypothesis that the region where the repeat is located might have a function at the transcriptional regulatory level. Genomic elements, like promoters, enhancers, silencers and insulators in the SCA37 repeat flanking regions may influence the levels of expression of the toxic insertion $(ATTTC)_n$, having an impact in the molecular pathways leading to the neurodegenerative SCA37 phenotype.

Finding gene regulatory DNA sequences

Genome wide sequencing has been determinant to increase the extensive data about the complexity behind gene regulation in eukaryotes, particularly in the human genome[61; 62]. In the last decades, several unknown DNA regions have been studied regarding their genomic function[63]. Currently, there is a concept that gene expression is highly regulated by DNA regulatory elements (promoters, enhancers, silencers,

insulators) and epigenetic modifications (DNA methylation, histone modifications), responsible to coordinate complex networks underlying the spatiotemporal transcription[62]. Promoters and enhancers are considered as two important DNA regulatory elements, allowing a differential and tissue-specific gene expression[64]. Promoter sequences, that are usually nearby to transcription start sites (TSSs), are able to initiate transcription by RNA polymerase II recruitment, whereas *cis*-regulatory elements, as enhancers, are distant elements that interact with promoters to regulate the transcriptional rates through chromatin loops[62; 64] (**Figure 6**). Several high-throughput molecular techniques have been developed to study DNA-DNA interactions such as 4C (chromatin conformation capture circular), ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) and Hi-C (high-throughput chromosome conformation capture), enabling the detection of DNA long-range interactions[65; 66].

Gene noncoding regions have been described as regulatory regions that are able to define the chromatin conformation and the transcription initiation and elongation, recruiting transcription factors and RNA polymerase II. Thus, the identification of these regulatory regions is mandatory to understand their biological roles and the implications of their DNA variations in disease[67].

To find sequences with promoter activity, a traditionally used assay is dual-luciferase activity. However, this method does not allow the study of expression *in vivo*, at different time points, lacking the description of the tissue specificity of the regulatory modules[68; 69]. Zebrafish transgenesis, using a *Tol2* transposon system to integrate a vector, wherein the sequence of interest is cloned upstream of an enhanced green fluorescent protein (EGFP) sequence, without a promoter (**Figure 7**), offers several advantages. This approach is based on the assessment of EGFP expression *in vivo*, in different tissues using a fluorescence stereoscope. Furthermore, the zebrafish is not euthanized, and thus, the activity can be studied at different time points during animal development and adulthood. Moreover, this strategy allows the identification of TSSs by 5'- rapid amplification of cDNA ends (5'-RACE), finding transcripts originated from the activity of specific promoters. To assess the ability of the promoter to interact with an enhancer, a brain-specific enhancer like the Z48 enhancer, known as Z54272 enhancer[70-73], can be used. Thus, a vector harbouring EGFP and Z48 sequences downstream of a putative promoter sequence allows indirect visualization of promoter and Z48 enhancer interaction by EGFP expression in zebrafish midbrain.

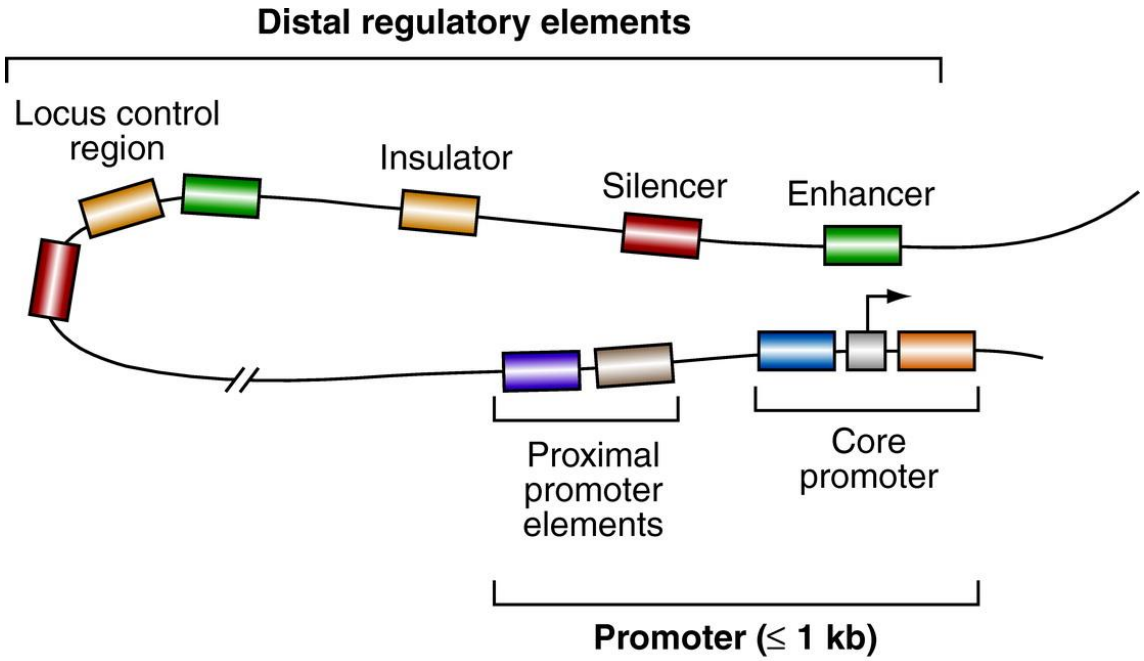


Figure 6 Gene regulation network. Transcription factors and RNA polymerase II interact with a promoter, which may be composed by core promoter and nearby proximal promoter elements, to activate transcription. Transcription is modulated by other regulatory elements, such as insulators, silencers and enhancers, that are able to increase or decrease the transcriptional levels. Promoter and interactions with their distal regulatory elements through DNA loops are the key for gene expression. Adapted from Maston G.A. et al., 2006.

To find enhancers there is a simple method that is the zebrafish enhancer detection (ZED) vector[72]. This ZED vector is composed by a transgenesis control cassette and an enhancer detection cassette (**Figure 8)** that allows detection of *cis*-regulatory sequences[72]. In this system, the success of fragment integration in the zebrafish genome is controlled by the RFP expression, while the enhanced activity is controlled by the EGFP expression, whose gene is associated with a minimal promoter. Therefore, EGFP expression increment is related to the presence of an enhancer sequence.

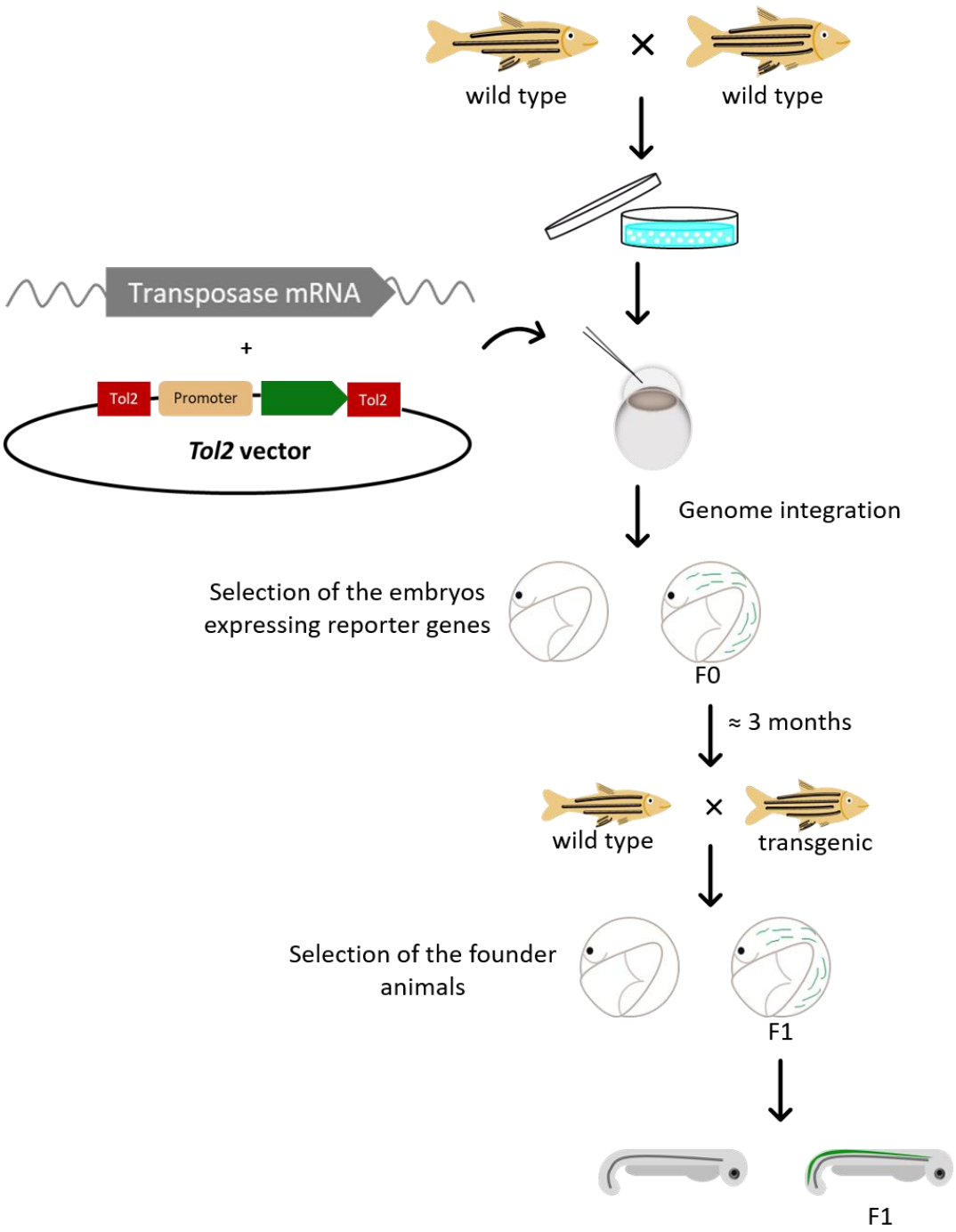


Figure 7 Promoter activity assessment. Zebrafish transgenesis are performed using *Tol2* transposon system. *Tol2* element, currently used to create zebrafish transgenic lines, is a DNA transposon identified in medaka fish (*Oryzias latipes*) genome[78-80]. *Tol2* RNA synthesised *in vitro* is co-microinjected with a *Tol2* vector containing *Tol2* recognition sites flanking the interest DNA. After translation of transposase RNA by zebrafish, *Tol2* protein recognizes these recognition sites, integrating the DNA randomly in the fish genome. Microinjected animals may be selected by fluorescence expression and the transgenic are raised to adulthood. These mosaic zebrafishes aging three months are then backcrossed with wild-type animals to identify transgenic animals whose integration occurred in the germline (founder animals) to characterize the expression in F1 embryos.

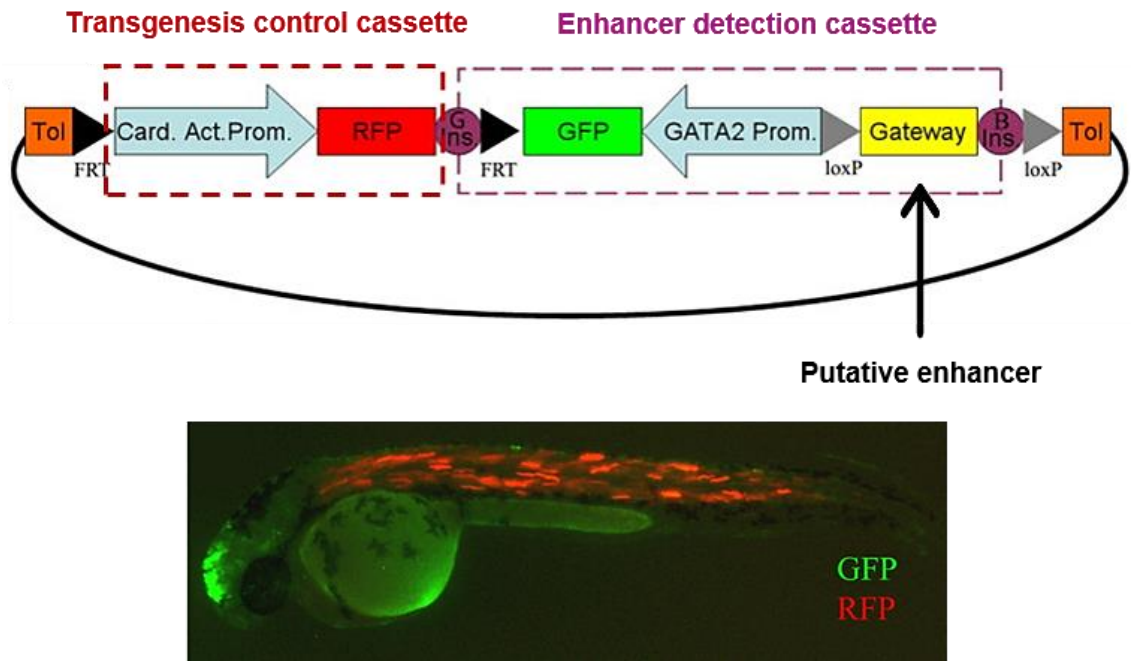


Figure 8 Zebrafish Enhancer Detection (ZED) vector. The vector contains two cassettes for the transgenesis control (red) and the enhancer detection (violet), flanked by *Tol2* recognition sites (orange). The transgenesis control cassette contains a cardiac actin promoter (pale blue), that triggers the expression of *red fluorescent protein* (*RFP*) (red) in the cells of the embryo with the integrated DNA sequence. The enhancer detection cassette is composed by a gateway site (yellow) for the insertion of the fragment of interest, to evaluate its enhancer activity by a gata2 minimal promoter (pale blue) with an EGFP (green) downstream. Adapted from Bessa et. al, 2009.

Regulatory DNA sequences in SCA37

Previous work of our team on regulatory regions in the SCA37 locus detected promoter activity across the repeat region in the *DAB1* antisense strand[74]. Promoter activity driven by antisense intronic region where the repeat is located (Total fragment or Frag T) was detected in muscular and cerebellar tissues of F1 zebrafish embryos (**Figure 9**, top panel). The characterization of promoter elements in this regulatory region was assessed by fragmentation of the region in 3 sequences (Frag 1-3), resulting in EGFP expression (**Figure 9**, lower panel). Frag 1 sequence contains a genomic region upstream of Alu element and its left monomer, Frag 3 comprises the genomic region downstream of Alu where a CpG island and a TFBS is located. Functional genomic assays were performed with an $(ATTTT)_{14}$ normal allele.

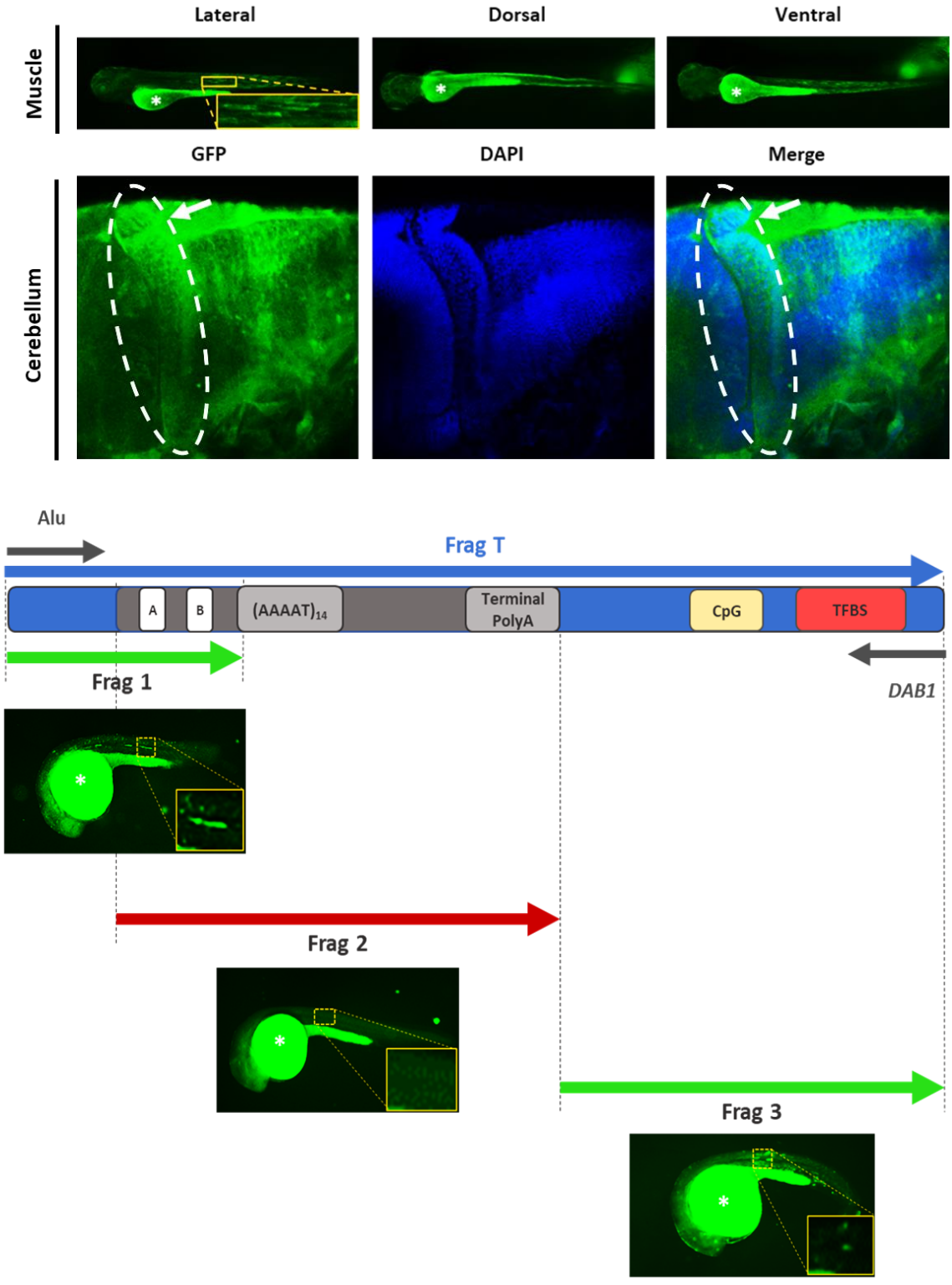


Figure 9 Promoter activity across the repeat region of the *DAB1* antisense strand. Antisense sequence of repeat flanking region (Frag T) triggers EGFP expression in muscle fibers and cerebellum of zebrafish embryos, at 72 hpf (**top panel**). After Frag T was fragmentation, promoter activity was detected upstream of the repeat (Frag 1) and at downstream of the Alu element (Frag 3) (**lower panel**). Asterisks represent regions of autofluorescence (embryos yolk).

Aims

In several repeat diseases such as SCA8, HD, FXTAS, DM1 and FTD/ALS, both sense and antisense strands containing the repetitive tract are transcribed and these RNAs may disrupt essential biological pathways[3; 32-34; 36; 39]. Preliminary results using zebrafish embryos, showed promoter activity in two regions flanking the repeat in *DAB1* antisense strand. Thus, it is important to understand in which tissues and stages the antisense transcripts are expressed, as well as to identify the TSS(s) in human cerebellum. Furthermore, since the repeat is located close to putative regulatory elements, it is crucial to know if this complex genomic region has a regulatory function. Therefore, this project aims to:

a) Describe the antisense transcription across the *SCA37* normal repeat in different tissues and stages of the transgenic zebrafish
b) Clarify the genomic regulation of antisense transcription
c) Identify the TSS(s) using zebrafish transgenic animals
d) Investigate the presence of enhancers in the repeat flanking region

The characterization of the antisense transcriptional network performed in this work is the basis for the investigation of potential pathogenic pathways led by the antisense $(GAAAU)_n$ RNA in SCA37.

Materials and methods

Zebrafish husbandry and manipulation

Zebrafish husbandry conditions and manipulation procedures were carried out according to National guidelines of animals housing and care (Decreto-Lei nº 113/2013), and European Union legislation for zebrafish experimentation (Directive 2010/63/EU). Zebrafish were grown and maintained in the licenced Zebrafish facility of Vertebrate Development and Regeneration group at i3S led by Dr José Bessa.

Zebrafish embryos were maintained in an incubator at 28.5°C in petri dishes containing embryos medium (E3) (5 mM NaCl, 0.17 mM KCl, 0.33 mM CaCl2, 0.33 mM MgSO4, 0.00015% methylene blue) up to 6 days postfertilization (dpf). Animals for which was expected to detect fluorescent proteins expression, E3 medium was supplemented with 1x of 1-phenil-2-thiourea (PTU)[75], that inhibits animal pigmentation for image acquisition purpose. At 6 dpf, embryos were placed in Zebrafish facility nursery at 24°C on a 14hr/10hr light dark circadian cycle. All the larvae transferred to nursery were bleached from 10 to 28 hpf with a 0.03% sodium hypochlorite solution (Sigma-Aldrich). At 10 dpf, the recirculating water system was turned on. After one month in nursery, zebrafish were transferred to larger aquaria (3.5 L). All the animals were fed three times a day with GEMMA Micro 150 (Skretting) until reach one month and, after that, they were fed with Zebrafeed 400-600 (Sparos).

Cloning

Total fragment (Frag T) with the upstream (Frag 1), $(ATTTT)_{14}$ tract (Frag 2) and downstream (Frag 3) sequences in both *DAB1* and Alu orientations (**Figure 9**) were previously cloned in pCR™8/GW/TOPO® TA Cloning plasmids (Invitrogen™). This vector, which comprises the attL1 and attL2 sites, is an entry clone vector commonly used to Gateway™ recombination[76; 77]. Then, the sequences cloned in this vector were recombined with the different destination vectors used in this work (1GWC2EGFP, miniTOL_MCS(pUC19)-GW GFP-Z48 and ZED vector) by Gateway™ technology using Gateway™ LR Clonase™ II Enzyme mix (Invitrogen™).

Plasmids were transformed in NZYStar Competent Cells (NZYTech) and DNA plasmids extraction were performed using ZR Plasmid Miniprep™-Classic (Zymo Research), according to the manufacture's recommendations. Sequences cloned in the vectors described were confirmed by Sanger sequencing.

Promoter activity characterization in *DAB1* antisense orientation

To characterize the activity of each Frag 1 and Frag 3 promoters across *DAB1* antisense strand (**Figure 6**), Frag 1 and Frag 3 were cloned in 1GWC2EGFP plasmid in *DAB1* antisense orientation, upstream of an EGFP sequence without promoter. Then, 1GWC2EGFP plasmids were microinjected in zebrafish embryos with a *Tol2* transposon system.

Zebrafish embryos injected with 1GWC2EGFP, in which the fragment was integrated (mosaics for the DNA sequence injected), were raised in Zebrafish nursery until reach the adulthood. Transgenic animals were then backcrossed, to select the zebrafish that integrated the transgenesis in germline (founders). The offspring of these animals was analysed and, those with EGFP positive animals in the offspring were selected as founders. Then, EGFP expression was analysed and characterized in these F1 embryos, from 24 hpf up to 7 dpf.

Promoter-enhancer interaction

To study the ability of the promoter sequences to interact with enhancers, the entry vectors containing Frag T and the fragmented Frag 1, Frag 2 and Frag 3 sequences in *DAB1* antisense orientation were recombined to miniTOL_MCS(pUC19)-GW GFP-Z48 destination vector that contains the midbrain-specific enhancer Z48. To guarantee that EGFP expression was actually triggered by the interaction of the cloned sequences with the Z48 enhancer, an empty miniTOL_MCS(pUC19)-GW GFP-Z48 was used as negative control. Thus, EGFP expression was then analysed from 24 hpf up to its vanishing in F0 embryos.

Enhancer detection assay

To detect enhancer elements across the repeat flanking region, Frag T was previously cloned in ZED vector[72]. In this vector, Frag T was cloned upstream of a minimal promoter that drives the EGFP expression (**Figure 9**). If the cloned sequences contain enhancers, an increase in EGFP expression is detected. ZED vector comprises a cardiac actin promoter that drives the RFP expression, at 48 hpf, in heart and muscles, which functions as transgenesis control cassette (**Figure 9**), enabling to select transgenic animals. An empty ZED vector was also used as negative control. Thus, EGFP expression was compared in F0 embryos over time, to detect its enhancement in the presence of an enhancer element.

In vitro Tol2 RNA synthesis

To synthesise *Tol2* RNA, PCS2FA plasmid, which contains the *Tol2* sequence under a SP6 promoter, was linearized with 60U Anza™*NotI* (Invitrogen™) overnight at 37°C. Linearized DNA was then purified with phenol chloroform to eliminate RNAses (detailed protocol in **Appendix 1**). *Tol2* RNA was synthesised with SP6 RNA polymerase (Thermo Fisher Scientific) (detailed protocol in **Appendix 2**). RNA was purified in a sephadex column (detailed protocol in **Appendix 3**) and then, re-purified with phenol chloroform.

Microinjections

All the 1GWC2EGFP, ZED vector and miniTOL_MCS(pUC19)-GW GFP-Z48 vectors comprise two *Tol2* arms that are recognized by Tol2 transposase[78-80]. This recognition leads to the transposition of the sequences of interest randomly in zebrafish genome. Previously to microinjection procedure, 1GWC2EGFP, ZED vector and miniTOL_MCS(pUC19)-GW GFP-Z48 plasmids were purified with phenol chloroform (detailed protocol in **Appendix 1**), and then co-injected in one to two-cell stage zebrafish embryos together with *Tol2* RNA (detailed protocol in **Appendix 4**).

Zebrafish animals from wild-type AB line were crossed to obtain embryos for microinjection procedures[79; 80]. To promoter activity assay, microinjected embryos were raised to generate a transgenic line, whereas to enhancer assay and promoter and Z48 enhancer interaction assay, three independent microinjections (replicates), in which the fragments flanked by *Tol2* arms were microinjected in at least 200 embryos, were performed at different days.

Imaging acquisition and analysis

Fluorescence expression was analysed in Leica M205 FA stereomicroscope from Leica Microsystems, using the Leica Applications Suite (LAS) software. Imaging acquisition was performed using Orcha Flash 4.0 LT system. A 300 mgL^{-1} tricaine solution (ethyl 3-aminobenzoate methanesulfonate or MS222) (Sigma-Aldrich) was used as anesthetic in imaging acquisition. Imaging edition was posteriorly performed in ImageJ program.

Laser Scanning Confocal Microscopy

To specifically localize the promoter activity as well as its interaction with Z48 enhancer, EGFP expression was analysed by confocal imaging. F1 zebrafish embryos of 1GWC2EGFP vector (promoter activity assay) and microinjected embryos with miniTOL_MCS(pUC19)-GW GFP-Z48 (promoter-enhancer interaction analysis) were fixed overnight at 72 hpf and 48 hpf, respectively. Embryos were fixed in 4% formaldehyde in Phosphate-buffered saline (PBS), overnight at 4ºC. Then, larvae were washed in 0.1% Triton in PBS for 10min, repeating the procedure 5x. Embryos were then permeabilized with 0.5% Triton in PBS for 30 min, and then, stained with 0.001% DAPI overnight at 4ºC. Imaging acquisition was performed using the SP5 II TCS Laser Scanning Confocal Microscope from Leica Microsystems with a 20x glycerol objective.

RNA toxicity assay

T7-pCDH-CMV-MCS-EF1-GFP-T2A-Puro plasmids comprising the normal $N(AAAAT)_7$ and $N(AAAAT)_{139}$ alleles and the pathological insertion $[(TAAAA)_{56}(CTTTA)_{54}(TAAAA)_{44}]$ under a T7 promoter were linearized. T7-pCDH-CMV-MCS-EF1-GFP-T2A-Puro plasmids were linearized with 60U Anza™ *EcoR*I (Invitrogen™) overnight at 37°C, a restriction enzyme that cuts the DNA sequence downstream of the repeats of interest. Then, plasmids were purified with phenol chloroform (detailed protocol in **Appendix 1**) and $N(UAAAA)_7$, $N(UAAAA)_{139}$ and $ins(GAAAU)_{54}$ RNAs were *in vitro* transcribed using T7 RNA polymerase (Thermo Fisher Scientific) (detailed protocol in **Appendix 2**). RNAs were purified with a sephadex column (detailed protocol in **Appendix 3**) and then re-purified with phenol chloroform (detailed protocol in **Appendix 1**). Cas 9 RNA, used as control in this experiment, was synthesized using the same protocol.

Approximately 5 nL of $N(AAAAU)_7$, $N(AAAAU)_{139}$, $ins(GAAAU)_{54}$ and Cas 9 RNAs were microinjected in one to two-cell stage embryos, at a concentration of 100 $ng\mu L^{-1}$. At least 200 embryos were microinjected per condition and microinjection procedures were performed in triplicated. To analyse the developmental defects and lethality of the embryos, they were observed at 6 hpf, 24 hpf and 48 hpf. Embryos lethality was statistically analysed with a χ^2 test.

Results

Antisense transcription driven by upstream and downstream SCA37 repeat sequences

Previous work of our team showed the existence of two putative promoters in *DAB1* antisense orientation, flanking the repetitive SCA37 region[74] (**Figure 9**). Now, to characterize the activity of these promoter sequences in different tissues and developmental stages, we analysed stable F1 transgenic lines for the upstream (Frag 1) and downstream (Frag 3) repeat sequences (**Figure 10A**). Their promoter activity was compared with the promoter activity of the total fragment (Frag T) containing the repeat with upstream and downstream sequences, in F1 zebrafish lines. An empty 1GWC2EGFP vector was used as a negative control to guarantee that EGFP expression was entirely driven by promoter sequences (**Figure 10B**). The offspring of two founder animals, with transgenesis integration in germline, was characterized per condition (Frag 1 and Frag 3) from 24 hpf to 7 dpf.

Regarding Frag 1, the promoter expression in stable lines was only detected at 48 hpf (**Figure 11A**). In F1 embryos expressing Frag 1, EGFP expression was noticed in muscular tissue, particularly in horizontal myosepta from 48 hpf to 7 dpf, at least (**Figure 11 A** and **Figure appendix 5**). Regarding promoter activity in embryos brain, no strong EGFP signal was clearly found in F1 embryos, as confirmed using confocal microscopy at 72 hpf. However, some EGFP signal driven by Frag 1 in brain structures may occur at other developmental stages **(Figure 11B)**.

Concerning Frag 3 promoter activity, this sequence was able to trigger EGFP transcription from 24 hpf to 7 dpf in F1 zebrafish embryos (**Figure 12A** and **Figure appendix 6**). EGFP expression was detected in muscle fibers along the body of the embryos (**Figure 12A**). A slight difference in EGFP was detected between F1(1) and F1(2) founder animals at 24 hpf (**Figure 12A**). In F1(1), EGFP was found in notochord and muscular tissue, whereas in F1(2) EGFP was only expressed in muscular tissue. From 48 hpf to 7 dpf, EGFP signal was detected in muscular tissue in embryos of both F1 lines (**Figure 12A** and **Figure appendix 6**). At 72 hpf, the EGFP signal was also evaluated in brain of F1 embryos driven by Frag 3 promoter, by confocal microscopy. However, no evidence of EGFP expression driven by the Frag 3 promoter was noticed (**Figure 12B**).

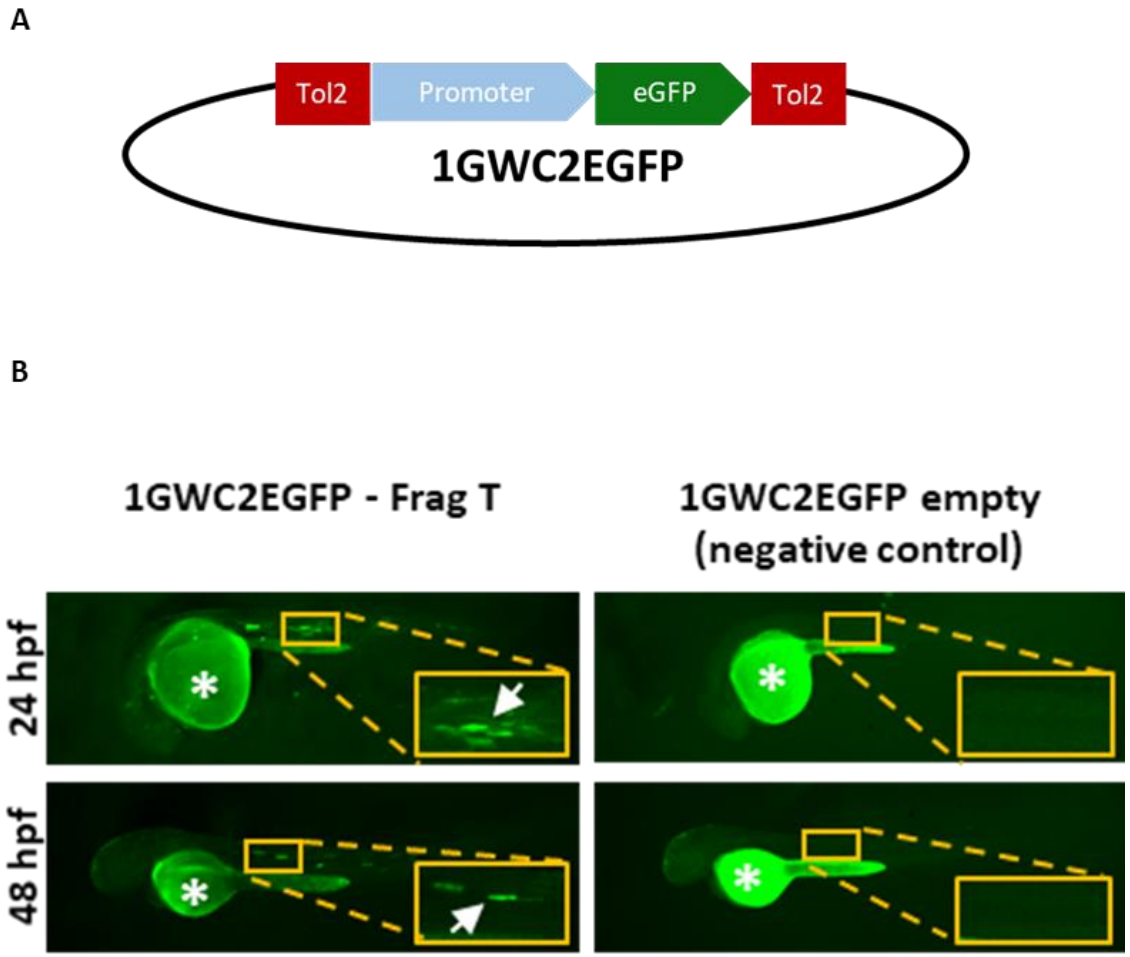


Figure 10 Promoter activity across *DAB1* antisense strand. (**A**) Scheme of 1GWC2EGFP vector used; each putative promoter sequence was cloned upstream of an EGFP sequence without promoter. (**B**) An empty 1GWC2EGFP vector was microinjected in zebrafish embryos and compared with a positive control, Frag T, to guarantee that EGFP transcription is only driven by the tested sequences; asterisks indicate autofluorescence regions such as the yolk of the embryos.

Summarising, we detected a tissue-specific promoter activity of Frag 1 and Frag 3 in horizontal myosepta and muscle fibers, respectively. In transgenic zebrafish embryos containing Frag T, we observe EGFP expression in these both tissues, meaning that Frag T expression results from a cumulative effect of the promoter activity of Frag 1 and Frag 3 (**Figure 13**). Contrarily to promoter activity detected in muscular tissue, when we investigated the promoter activity in zebrafish embryos cerebellum, we found that EGFP is expressed in cerebellar tissue of Frag T transgenic embryos, but not in the cerebellum of Frag 1 and Frag 3 transgenic embryos (**Figure 13**). These results suggest that there is synergic activity between Frag 1 and Frag 3 promoters, leading to the cerebellar expression in Frag T.

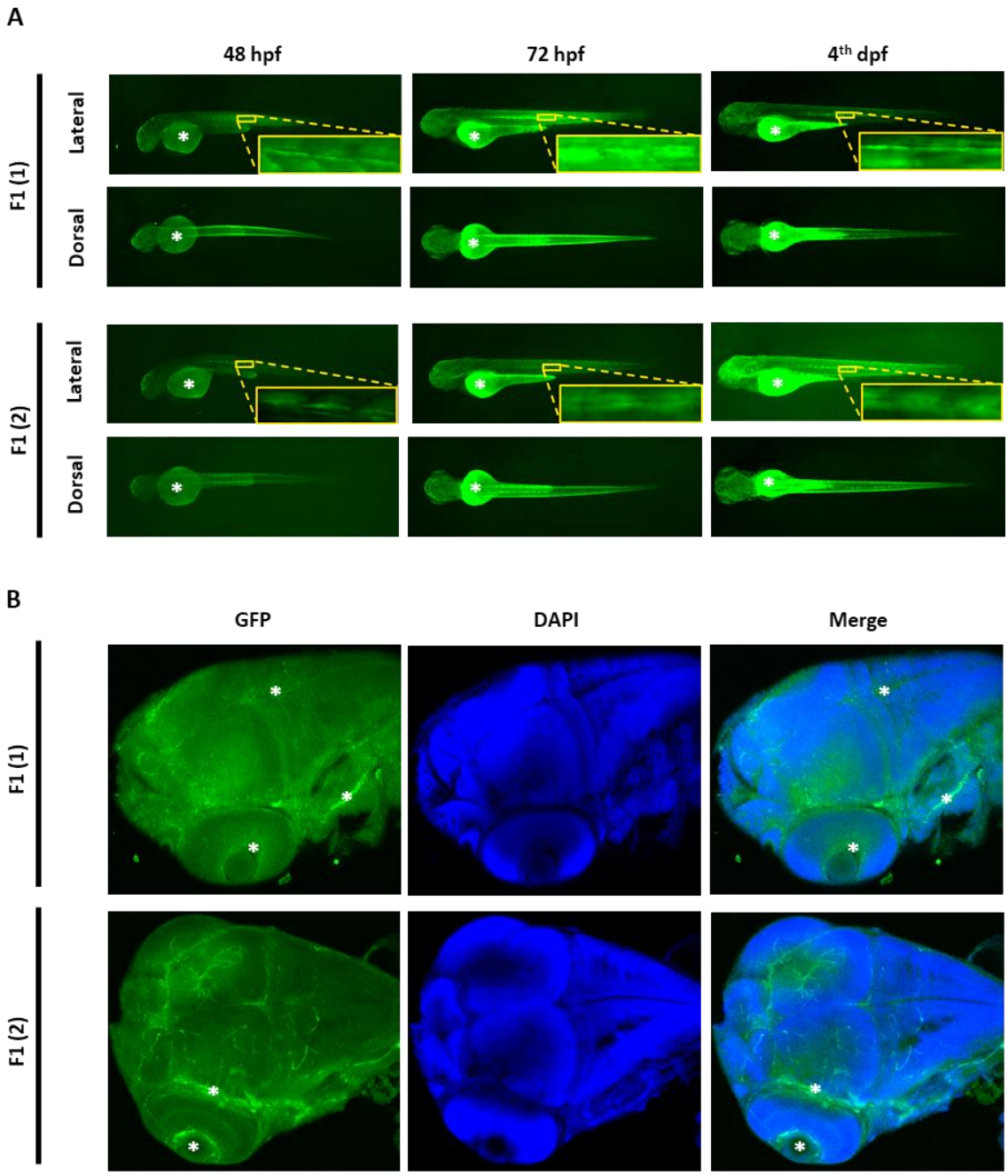


Figure 11 Antisense transcription driven by the upstream SCA37 repeat sequence (Frag 1). (**A**) At 48 hpf, EGFP was firstly observed in the horizontal myosepta of F1 zebrafish embryos (1 and 2). (**B**) Z projection of confocal images of brain structures of Frag 1 animals (F1 generation); Z-stacks with 10 and 20 focal planes. Asterisks indicate autofluorescence regions such as larvae yolk and blood cells.

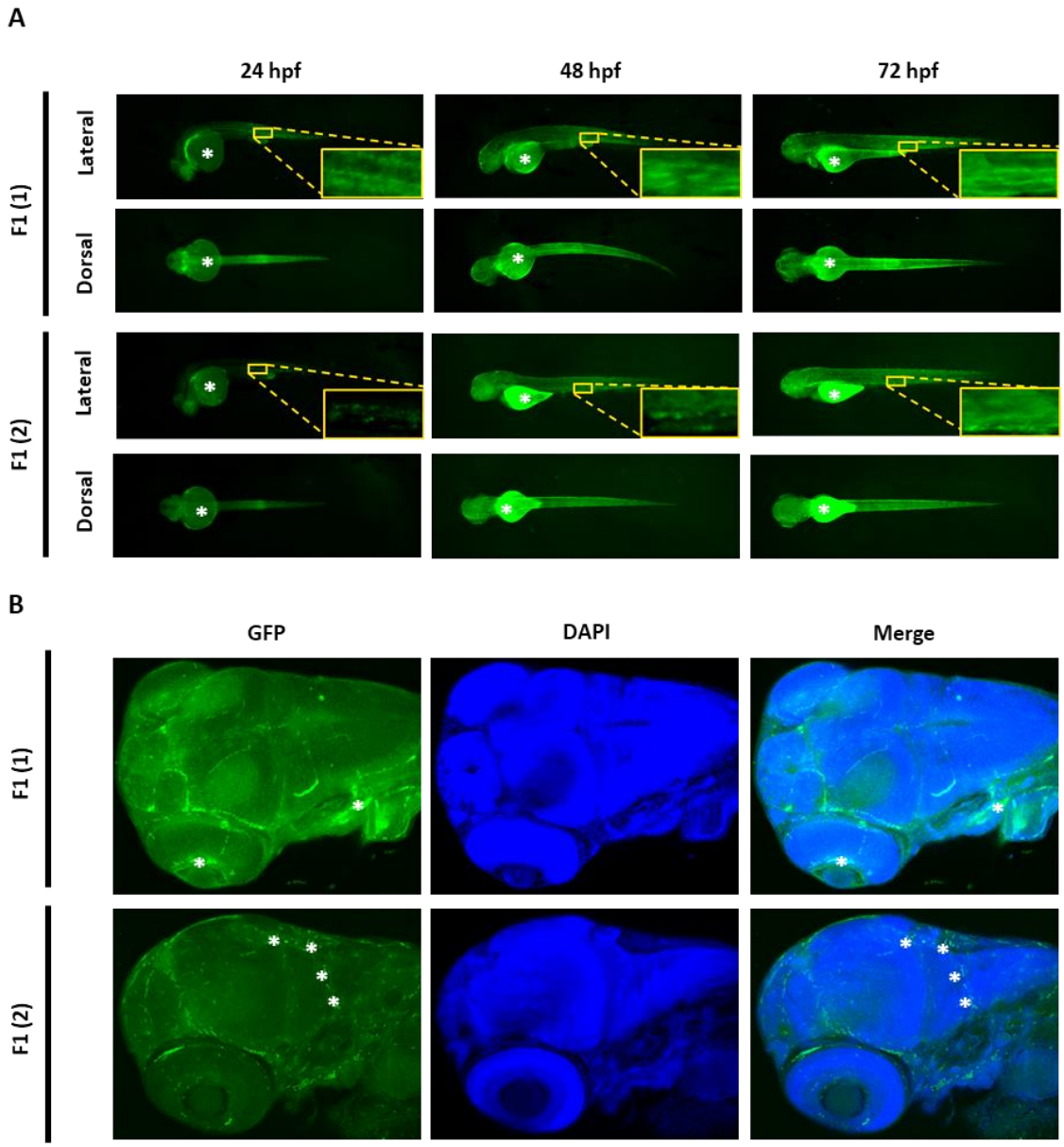


Figure 12 Antisense transcription driven by downstream SCA37 repeat sequence (Frag 3). (**A**) At 24 hpf, the F1(1) demonstrated an EGFP expression in notochord that was not detected in F1(2) embryos; F1(2) shows EGFP expression in muscular tissue. At 48 hpf, EGFP was observed along the muscle fibers of the embryos in both founders. (**B**) Founders 1 and 2, Z-stacks with 13 and 9 focal planes in brain tissue, respectively, were acquired. Asterisks indicate autofluorescence regions such as larvae yolk and blood cells.

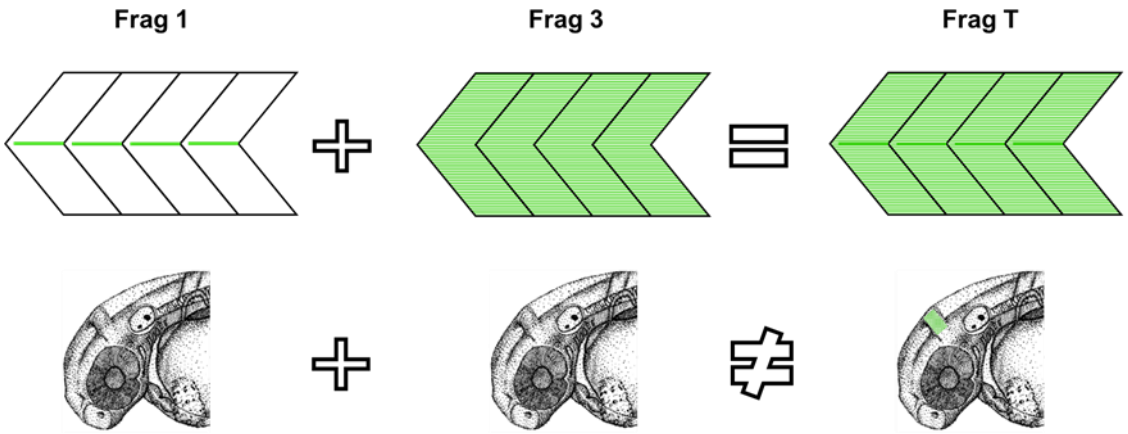


Figure 13 Schematic representation of promoter activity in *DAB1* antisense strand. Promoter activity in muscular tissue observed in Frag T embryos results from the activity of Frag 1 and Frag 3 promoters in horizontal myosepta and muscle fibers, respectively. In cerebellar tissue, no EGFP expression was detected in Frag 1 and Frag 3 embryos, however EGFP expression was observed in Frag T embryos.

Promoters in the SCA37 region interact with a brain enhancer

To investigate if the promoters identified in the SCA37 repeat flanking region, in Alu orientation, were able to interact with brain-specific enhancers, Frag T and its fragmented sequences, Frag 1, Frag 2 and Frag 3, were cloned upstream EGFP and Z48 enhancer sequences (**Figure 14A**). We tested if these promoter sequences were able to interact with the strong midbrain-specific enhancer Z48 that would drive EGFP expression to midbrain cells (**Figure 14B**). An empty miniTOL_MCS(pUC19)-GW GFP-Z48 plasmid was also microinjected in zebrafish embryos as negative control, to guarantee that the EGFP expression in the midbrain of embryos injected with vectors containing promoter sequence and Z48 enhancer is the result of interaction between these sequences (**Figure 14C**). Each miniTOL_MCS(pUC19)-GW GFP-Z48 plasmid containing the promoters sequences (Frag T, Frag 1, Frag 2, Frag 3) was independently microinjected three times, in at least 200 embryos.

Promoter and Z48 enhancer interaction was initially analysed in Frag T (**Figure 15**). EGFP signal in midbrain cells was observed at 24 hpf, weakening at 6 dpf (**Figure 15A** and **Figure appendix 7**). As the Frag T, Frag 1 and Frag 3 have promoter activity in muscular cells, this expression was used as a transgenesis control in injected embryos. EGFP expression was detected widely in midbrain in 39.7% of microinjected embryos (**Figure 15A** and **Figure 16**). To confirm that promoter-Z48 enhancer interaction drives EGFP expression that is located in midbrain cells, 48 hpf embryos were also fixed and visualised using a confocal microscope (**Figure 15B**).

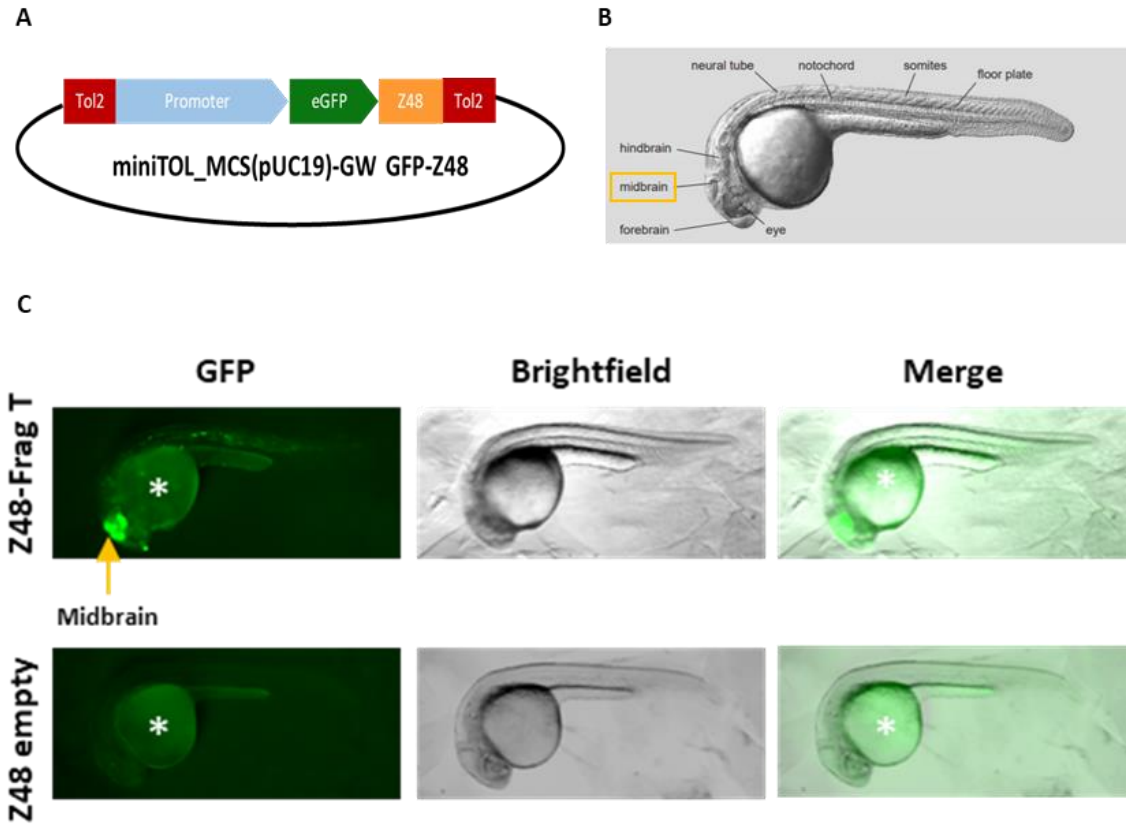


Figure 14 Interaction between putative promoter sequences and midbrain-specific enhancer Z48. (**A**) miniTOL_MCS(pUC19)-GW GFP-Z48 vector, in which each putative promoter sequence (Frag T, Frag 1, Frag 2 and Frag 3) was cloned upstream of an EGFP to evaluate their ability to recognize brain-specific enhancers. (**B**) Schematic representation of midbrain position in a 29 hpf zebrafish embryo. (**C**) Top image shows EGFP expression in midbrain of a positive interaction between Z48 and Frag T. Lower image is an empty miniTOL_MCS(pUC19)-GW GFP-Z48 vector used as negative control, to guarantee that EGFP transcription is driven only in presence of the tested promoters, eliminating artefacts inherent to vector; asterisks indicate autofluorescence regions (embryos yolk). Adapted from Haffter P. et al., 1996.

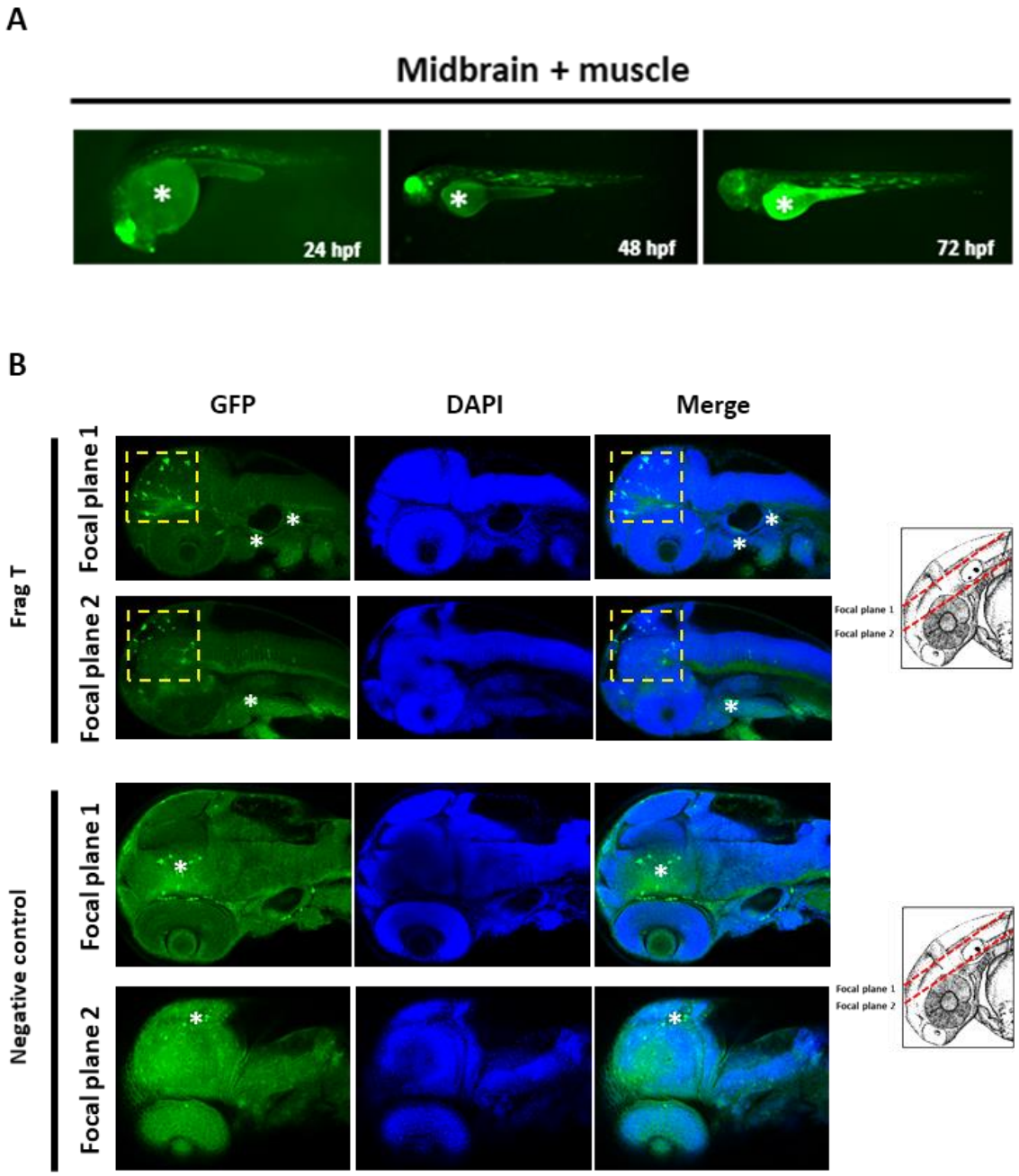


Figure 15 Characterization of EGFP expression driven by Frag T promoter-Z48 interaction in midbrain cells from 24 hpf to 72 hpf (**A**) and at 48 hpf (**B**). (**A**) EGFP expression was detected in whole midbrain of zebrafish embryos, however, promoter activity of Frag T maintained its expression in muscular tissue. (**B**) Two focal planes enabled the visualization of EGFP signal (delimited by yellow box) in whole midbrain through confocal microscopy. As negative control, an empty miniTOL_MCS(pUC19)-GW GFP-Z48 plasmid was used. Asterisks indicate autofluorescence regions such as embryos yolk and blood cells. Illustration from Institute of Neuroscience, University of Oregon.

Promoter-Z48 enhancer activity was then evaluated in Frag 1, Frag 2 and Frag 3. At 24 hpf, Frag 1 promoter-Z48 enhancer interaction was unable to drive EGFP expression in midbrain cells, while Frag T and Frag 3 promoters-Z48 interaction were able to drive expression at this time point. At 48 hpf, Frag 1 promoter was able to interact with Z48 enhancer in 16.0% of embryos, leading to a faint EGFP expression in a superficial subpopulation of midbrain cells, 15.1% of the embryos injected with Frag 1-Z48 exhibited EGFP expression only in muscle tissue but not in midbrain (**Figure 16**, **Figure 17A** and **Figure appendix 8**). To localize the EGFP expression in midbrain cells of embryos expressing Frag 1, embryos were fixed at 48 hpf and analysed through confocal microscopy (**Figure 17B**). Confocal microscopy enabled to ensure the localisation of EGFP expression superficially in midbrain cells.

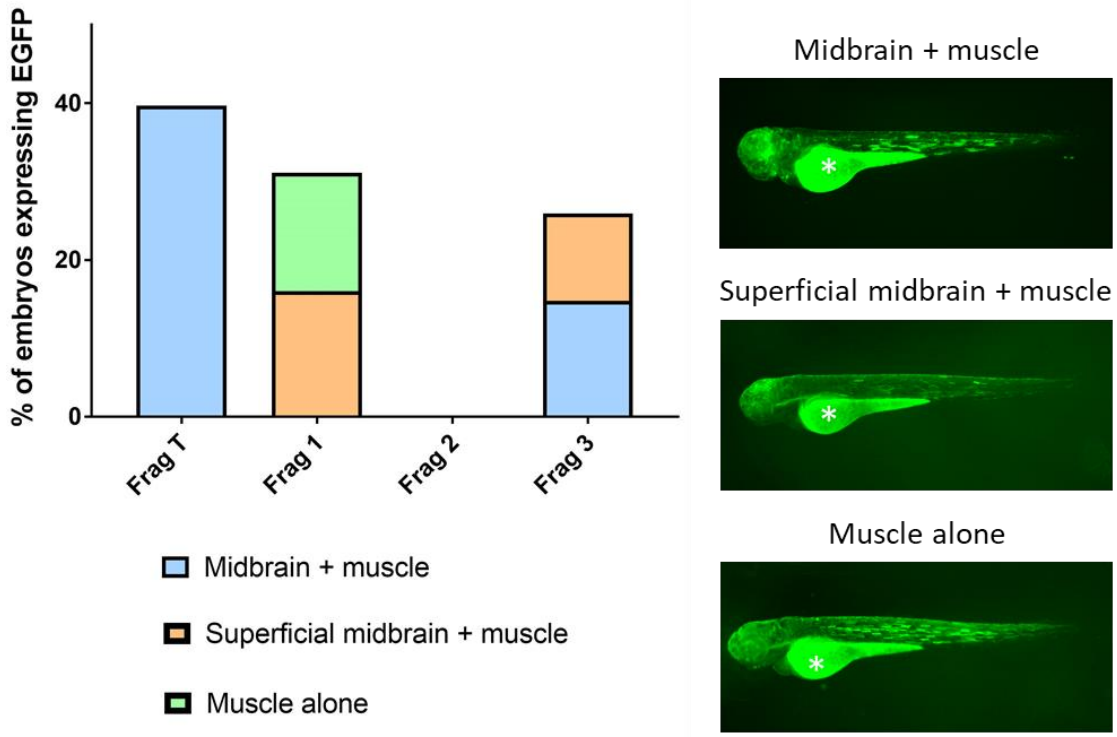


Figure 16 Microinjected embryos expressing EGFP, at 48 hpf. Positive embryos for transgenesis maintained EGFP expression in muscle tissue. Frag T promoter and Z48 enhancer interaction led to a widespread EGFP expression in midbrain cells in 39.7% of transgenic embryos; the interaction between Frag 1 promoter and Z48 was only detected superficially in midbrain cells in 16% of transgenic embryos; in 14.8% of transgenic Frag 3 promoter animals, EGFP was widespread in midbrain, while 11.1% the interaction occurred only in a subpopulation of midbrain cells, similarly to Frag 1 promoter; asterisks indicate autofluorescence regions such as the embryos yolk.

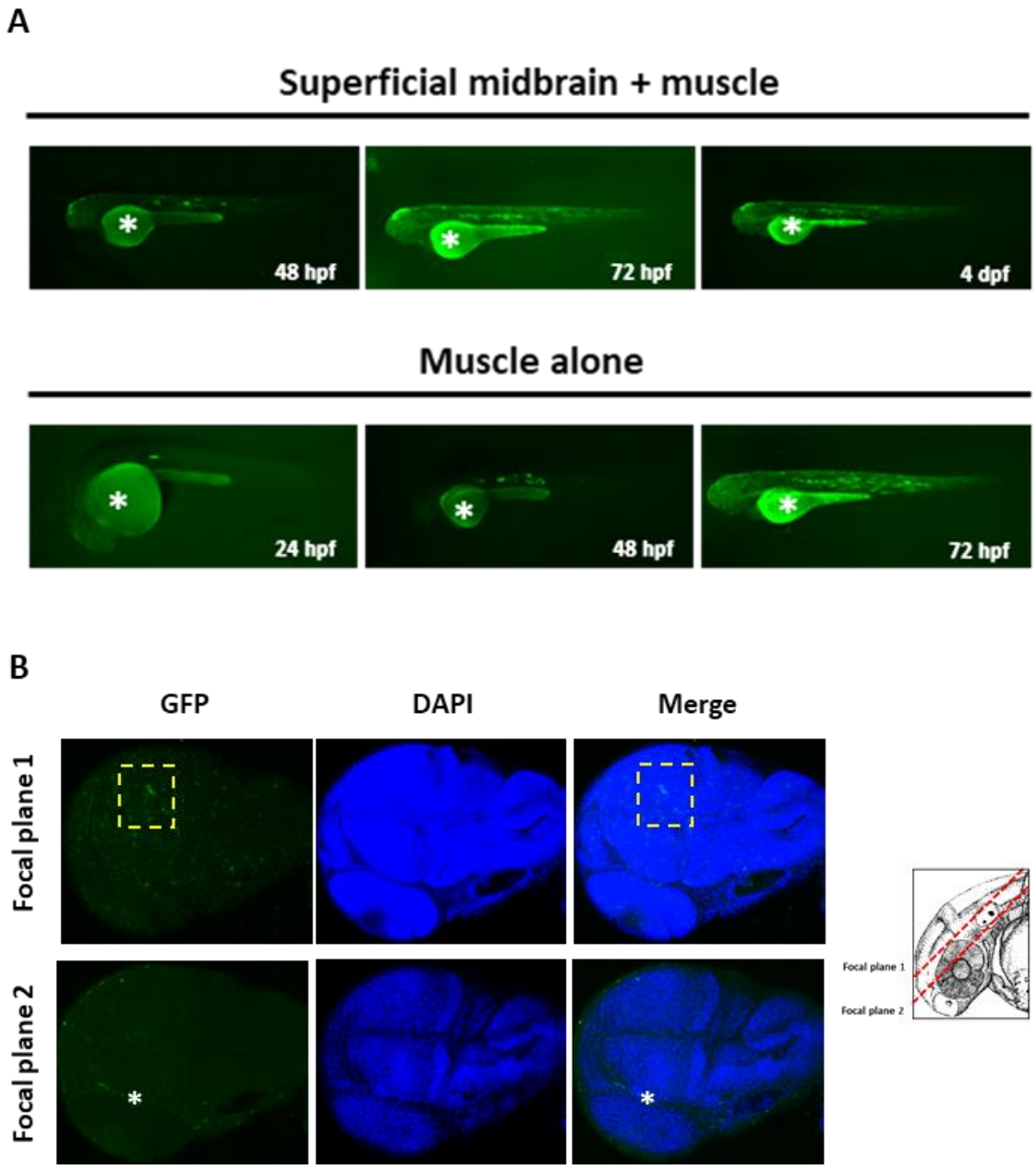


Figure 17 Characterization of EGFP expression triggered by Frag 1 promoter and Z48 enhancer interaction. (**A**) From 48 hpf up to 4 dpf, EGFP expression driven by Frag 1 promoter-Z48 interaction was detected superficially in midbrain of zebrafish embryos. However, some transgenic embryos were unable to induce EGFP transcription in midbrain at the same developmental stage. At 24 hpf, EGFP was only detected in muscular tissue of transgenic embryos. (**B**) Two focal planes enabled the visualization of EGFP signal (delimited by yellow box) superficially in midbrain cells of 48 hpf transgenic embryos, using confocal microscopy. Asterisk indicate autofluorescence regions. Illustration from Institute of Neuroscience, University of Oregon.

Although promoter activity was not detected in Frag 2 in preliminary assays, the interaction of this antisense sequence with Z48 was also tested, as weak promoters may only drive transcription when interacting with an enhancer. However, EGFP expression was not detected in Frag 2-Z48 enhancer interaction in zebrafish embryos (**Figure 16** and **Figure 18**).

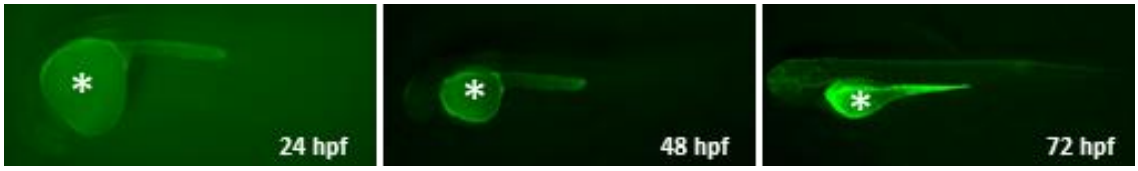


Figure 18 Analysis of Frag 2 and Z48 enhancer interaction. Given the hypothesis that Frag 2 had a week promoter, able to interact with an enhancer, Z48 enhancer was used to test its activity, at least in midbrain cells. EGFP expression was not detected in zebrafish embryos; asterisks indicate autofluorescence regions such as embryos yolk cells.

Interaction between Frag 3 promoter and Z48 enhancer was also analysed. EGFP expression was detected in midbrain of 25.9% of microinjected zebrafish embryos at 24 hpf. Whereas 14.8% of these embryos expressed EGFP widely in midbrain, its expression in 11.1% of them was restricted to a subpopulation of cells (**Figure 16**, **Figure 19A** and **Figure appendix 9**). Frag 3 promoter-Z48 enhancer interaction was also analysed using confocal microscopy to assess the promoter-enhancer interaction in brain tissue (**Figure 19B**).

Although the promoter and Z48 enhancer interaction verified in Frag 1 and Frag 3, EGFP expression driven by Frag T in midbrain cells seems to be much stronger than in fragmented sequences, suggesting that Frag T expression has a synergic interaction of different promoters (**Figure 16**).

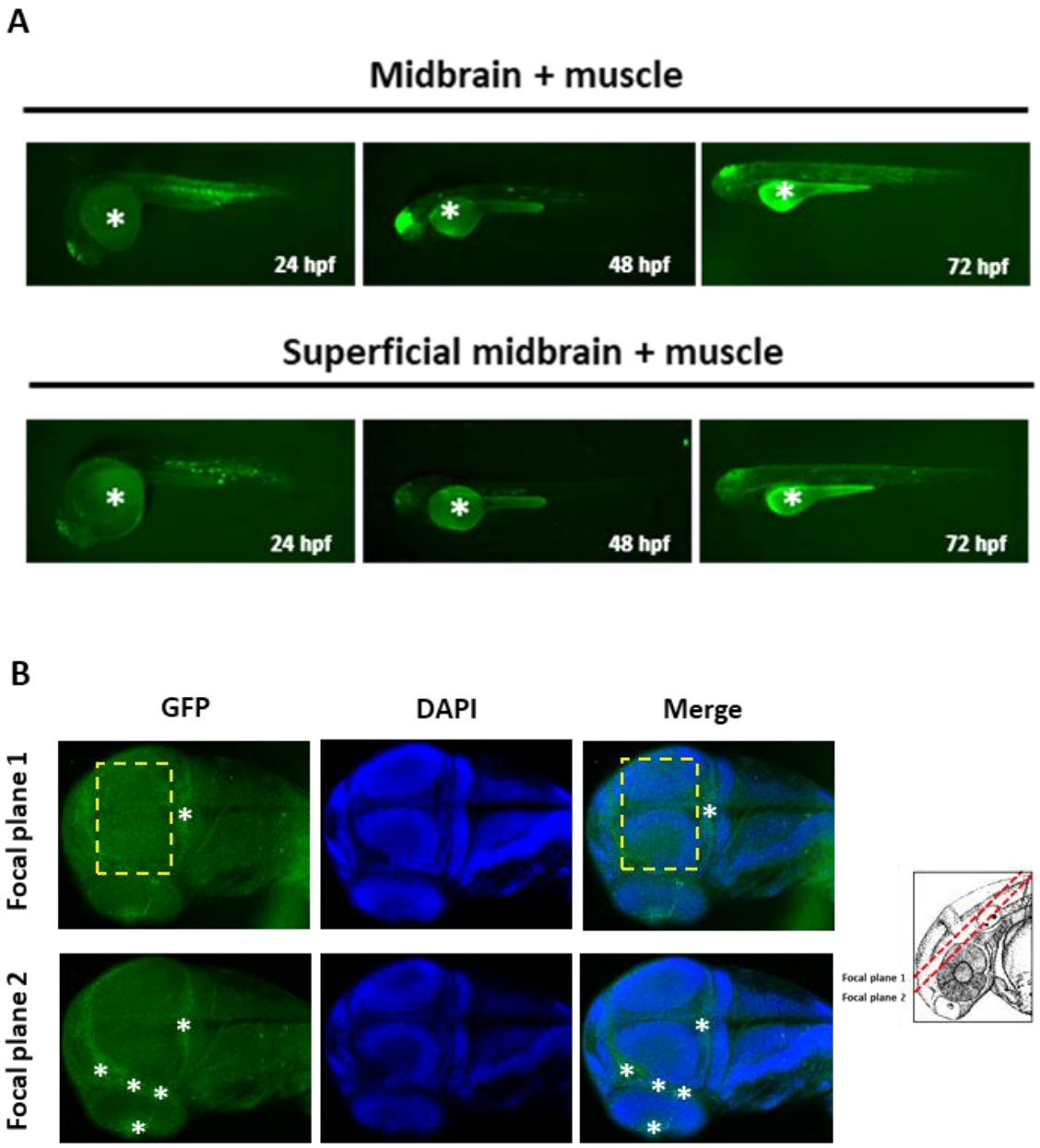


Figure 19 EGFP expression driven by Frag 3 promoter and Z48 enhancer interaction. (**A**) Transgenic animals with Frag 3 promoter and Z48 enhancer were able to trigger EGFP signal in whole midbrain at 24 hpf. EGFP expression was detected superficially in midbrain of some zebrafish embryos. (**B**) Two focal planes allowed visualization of EGFP signal (delimited by the yellow box) in midbrain, using confocal microscopy. Asterisks indicate autofluorescence regions such as embryos yolk and blood cells. Illustration from Institute of Neuroscience, University of Oregon.

No enhancer activity in the SCA37 repeat sequence

To evaluate Frag T enhancer activity, the sequence in *DAB1* orientation was cloned in the ZED vector (**Figure 9**). In *DAB1* antisense orientation, Frag T showed promoter activity, thus the enhancer assay using ZED vector tool was performed in sense orientation. As the ZED vector contains a minimal promoter driving the EGFP expression, an empty ZED vector was also microinjected and used as negative control. EGFP expression was then compared in transgenic animals.

At 48 hpf, 72 hpf and 4 dpf, no EGFP signal enhancement was observed in animals injected with ZED vector containing the Frag T when compared with animals injected with an empty ZED vector (**Figure 20**).

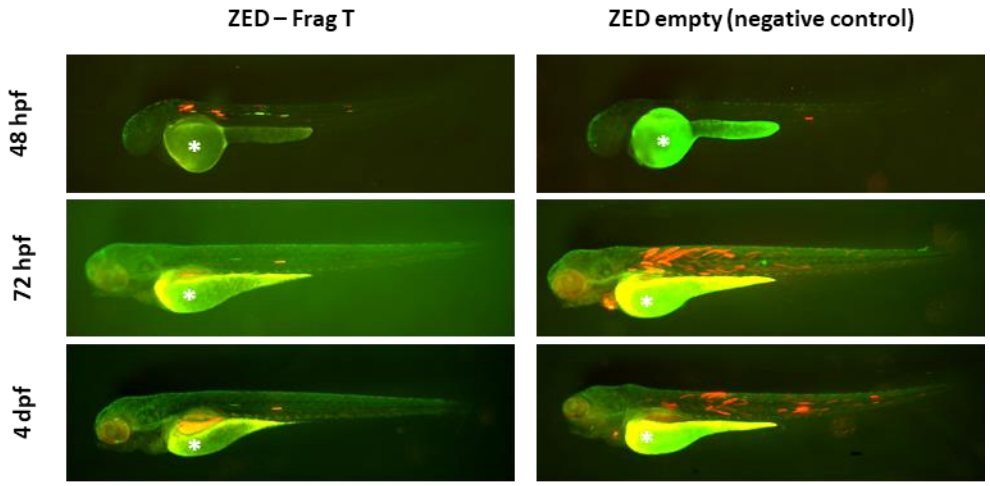


Figure 20 Evaluation of enhancer activity across Frag T in *DAB1* orientation. Transgenic embryos injected with ZED-Frag T and ZED empty vector (left and right panels, respectively) were compared for EGFP expression enhancement; RFP expression is a marker of transgenesis integration; asterisks indicate autofluorescence regions such as the yolk of zebrafish larvae.

Antisense $(GAAAU)_n$ RNA is not toxic in zebrafish embryos

To investigate if the antisense repeat insertion RNA contributes to the pathogenicity driven by the SCA37 insertion, normal antisense RNAs $N(AAAAU)_7$ and $N(AAAAU)_{139}$, and the pathogenic antisense insertion RNA $ins(GAAAU)_{54}$ flanked by AAAAUs and AluJb monomers, were microinjected in one to two-cell stage zebrafish embryos. Animals viability and morphology was analysed at 24 hpf and 48 hpf. As control, Cas 9 RNA was used to normalize the lethality and phenotypic abnormalities underlying the microinjection procedure.

At 24 hpf, Cas9 control RNA showed 27.91% lethality, $N(AAAAU)_7$ 22.44%, $N(AAAAU)_{139}$ 17.64%, and $ins(GAAAU)_{54}$ 25.43% (average of three replicas with at least 200 embryos each). The lethality rate between $ins(GAAAU)_{54}$ and the other conditions did not reach statistically differences (**Figure 21**). The values of developmental abnormalities and dead animals were significantly different between Cas9 control RNA and normal $N(AAAAU)_{139}$ RNA (average of three replicas with at least 200 embryos each; χ^2 test $p < 0.0001$); normal $N(AAAAU)_7$ RNA and normal $N(AAAAU)_{139}$ RNA (average of three replicas with at least 200 embryos each; χ^2 test $p < 0.05$); normal $N(AAAAU)_{139}$ RNA and pathogenic $ins(GAAAU)_{54}$ RNA (average of three replicas with at least 200 embryos each; χ^2 test $p < 0.001$) (**Figure 21**). At 48 hpf, the lethality rate remained statistically non-significant between $ins(GAAAU)_{54}$ and each of other condition, excepting $N(AAAAU)_{139}$ (χ^2 test $p < 0.001$) (data not shown).

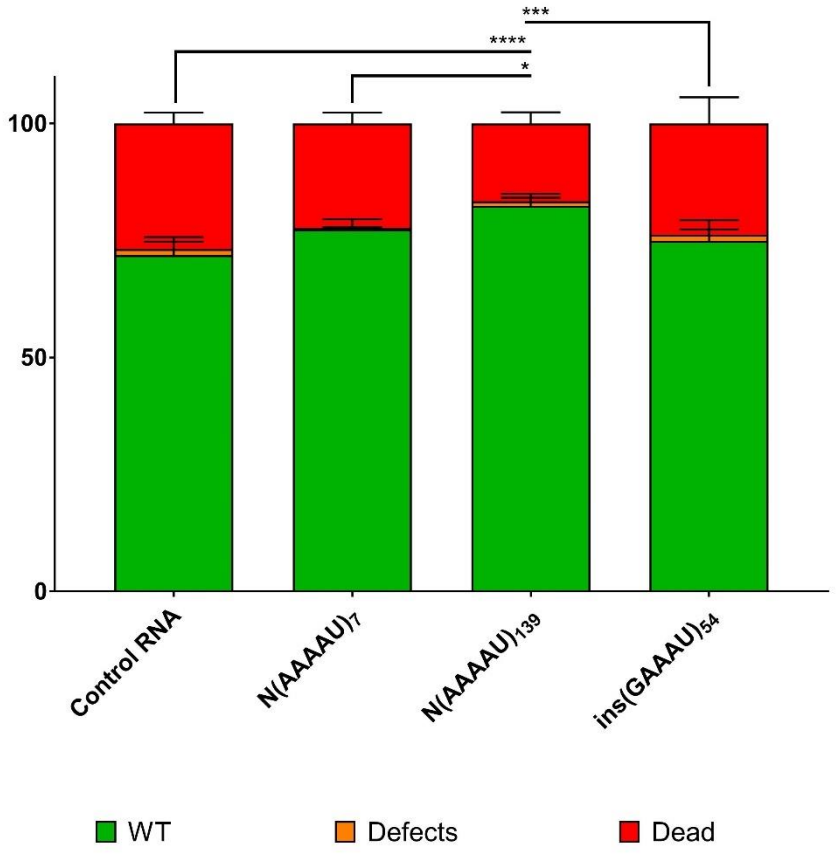


Figure 21 No *in vivo* deleterious effect of the antisense ins(GAAAU)$_{54}$ RNA injection in zebrafish embryos. Percentage of embryos with wild-type phenotype (WT), developmental defects (defects) and lethality (dead) at 24 hpf, after Cas9 control, N(AAAAU)$_7$, N(AAAAU)$_{139}$ and ins(GAAAU)$_{54}$ RNA injections (average of three replicas with at least 200 embryos each; χ^2 test for the lethality rate). Non-significative differences were observed between Cas9 control RNA and ins(GAAAU)$_{54}$, N(AAAAU)$_7$ and ins(GAAAU)$_{54}$. Significative differences were observed between N(AAAAU)$_{139}$ and the other RNA conditions (average of three replicas with at least 200 embryos each; χ^2 test for the lethality rate; ****$p < 0.0001$; ***$p < 0.001$; *$p < 0.05$).

Discussion

In several repeat diseases as SCA8, FTD/ALS and FXTAS, the repeat bidirectional transcription has been reported as contributing for the disease pathogenicity[16; 35; 36]. In this work, we showed that the *DAB1* $(ATTTC)_n$ insertion may be transcribed in antisense orientation, a finding that may have implications in SCA37 pathogenicity.

Previous work has shown promoter activity in *DAB1* antisense strand in the repeat flanking region, upstream and downstream the repeat[74]. In this work, we characterized the activity of these promoter regions in F1 zebrafish embryos. We found that the upstream repeat (Frag 1) promoter is active in muscular tissue, particularly in the horizontal myosepta, a cluster of differentiated structures of myomere, responsible to separate dorsal and ventral muscular masses[47; 81], while the downstream repeat (Frag 3) promoter is active in muscular tissue, showing that these two promoters are tissue-specific and their activity does not overlap in muscular tissue. Horizontal myosepta, which is important for the migration of the cells of posterior lateral line (sensorial organ containing the neuromasts, responsible to feel water oscillations), develops parallelly to the notochord[82; 83]. Regarding the Frag 3 F1 generation, we found a different pattern of promoter activity at 24 hpf in the offspring of two founders. In the offspring of one Frag 3 founder, we detected EGFP expression in notochord, but we did not find the same expression pattern in the offspring of the second founder. As we did not see a corroborative result regarding EGFP expression in notochord in the offspring of both founders, we neglected this result. This may be explained by the genomic context where sequence integration occurred, being possible that Frag 3 sequence had been integrated in a region susceptible to regulatory elements that act in notochord-specific genes, leading to EGFP expression in this zebrafish structure. Interestingly, at muscular level, from 48 hpf to 7 dpf, the promoter activity found in F1 zebrafish from founders of Frag 1 and Frag 3 overlaps the promoter activity of Frag T, where promoter activity is detected in both muscular tissue and horizontal myosepta. However, contrarily to Frag T where cerebellar EGFP expression was detected, no cerebellar expression was detected in F1 of Frag 1 and Frag 3 promoters, at 72 hpf. This result shows that Frag 1 and Frag 3 alone do not have promoter activity in cerebellum of zebrafish animals. However, when these promoters are in the same sequence (Frag T), they are able to drive EGFP expression in this brain region. This suggests that Frag T has a more robust promoter activity than Frag 1 and Frag 3 promoters separately. Frag T sequence may comprise regulatory

elements required for promoter activation that were disrupted during Frag T fragmentation in Frag 1, 2 and 3. Overall, the promoter activity found in transgenic zebrafish embryos for the $(AAAAT)_n$ flanking region indicates that these promoters may also be active in human cerebellum in *DAB1* opposite strand. Thus, in the future, it is mandatory the identification of Frag 1 and Frag 3 TSSs in zebrafish transgenic animals in order to enable the identification of the RNAs transcribed under these promoters. If the activity of Frag 1 and Frag 3 promoters is detected in human cerebellum, a *DAB1* antisense transcription role may be suggested in SCA37 disease. Especially, Frag 1 that seems to drive repeat expression in antisense orientation, a mechanism already associated with repeat diseases[84].

Promoter activity was evaluated using a vector containing the putative promoter upstream the EGFP sequence in zebrafish stable transgenic lines. Thus, promoter activity was assessed by detecting EGFP expression under a fluorescence stereoscope. This approach may have a limitation, as background of images may mask EGFP expression and give a false indication of fluorescence location on the embryos tissues. To overcome this issue, an *in situ* hybridization of whole embryo using a labelled probe complementary to the transcript could be performed to clarify the expression location of Frag 1 and Frag 3 promoters.

The promoters found in this work are able to interact with enhancers cloned in their vicinity, suggesting that they could function in a similar way in *DAB1* genomic landscape. Both Frag 1 and Frag 3 are able to interact with the midbrain-specific enhancer Z48 leading to EGFP expression in this region[73]. This indicates that *cis*-regulatory elements as cerebellar-specific enhancers may be involved in *DAB1* antisense expression regulation in different tissues or developmental stages. Interestingly, when comparing the Frag T, Frag 1 and Frag 3 promoters interaction with Z48 enhancer, we observed different patterns of expression. Frag T-Z48 enhancer interaction led to a widespread promoter activation in midbrain. Frag 3-Z48 interaction led to promoter activation in whole midbrain in only 14.8% of the embryos, whereas, in the remaining 11.1% of the embryos, the Frag 3-Z48 interaction only led to promoter activity in superficial cells of zebrafish midbrain. A weaker promoter-Z48 interaction was found with Frag 1, where 16.0% of the embryos showed Frag 1-Z48 interaction in superficial midbrain cells and in the remaining 15.1% no interaction was detected. Similar to the results obtained with promoter assay, no evidence of cumulative effect was observed in interaction between Z48 enhancer and Frag 1 and Frag 3 promoters, which explains the Frag T-Z48 activity detected in midbrain cells, showing that this interaction is stronger with the whole sequence than the separated Frag 1 and Frag 3, suggesting that the Frag T contains important regions for promoter-Z48 interaction that are not

present in Frag 1 or Frag 3 sequences, or the Frag T promoter activity is stronger by itself. Thus, these antisense promoters may interact with *cis*-regulatory elements belonging to *DAB1* regulatory landscape, that may regulate the antisense transcription in space and time.

To investigate if Frag 2 had also promoter activity when activated by an enhancer, we tested the interaction between Frag 2 and Z48, however, we did not detect any EGFP expression, supporting the result that Frag 2 does not contain a promoter sequence and there is no promoter activity across the AluJb element or it requires upstream sequences. Alu elements are able to interact with promoters at long-range distances as enhancers[85], which may be an hypothesis for Frag 2 function. Thus, it is possible that this AluJb element may have a regulatory role in both *DAB1* gene regulation and antisense transcriptional regulation. To investigate the presence of an enhancer element in the repeat flanking region, we cloned the Frag T in *DAB1* orientation in a ZED vector[72]. Only fragments in sense orientation were tested using this molecular tool, because promoter activity was already detected across this region in antisense orientation. As ZED vector comprises a minimal promoter that may be enhanced by a promoter or an enhancer, the outcome of ZED vector does not allow distinction between promoter and enhancer elements. Using this strategy, we did not detect enhancer activity in *DAB1*-oriented Frag T sequence in zebrafish embryos, indicating that an enhancer is not present in the repeat flanking region.

The RNA-mediated toxicity is a common pathogenic pathway to noncoding repeat diseases[5; 15; 39]. Previously, we have shown that the $(AUUUC)_n$ RNA is highly toxic *in vivo* in injected zebrafish embryos[6]. As in diseases like FTD/ALS[86], the antisense repeat mitigates the toxicity of the sense repeat and we detected one promoter (in Frag 1 sequence) immediately upstream of the repetitive region, we hypothesized that this promoter would drive the repeat expression in antisense orientation and the antisense repeat might be toxic itself. Thus, we investigated the toxicity of the SCA37 antisense repeat *in vivo*. Normal and mutant antisense RNAs were microinjected in one to two-cell stage of zebrafish embryos, however, non-significant differences were found between ins$(GAAAU)_{54}$ and Cas9 control or between ins$(GAAAU)_{54}$ and normal N$(AAAAU)_7$ RNAs, suggesting the $(GAAAT)_n$ insertion is not toxic *in vivo*. Unexpectedly, the lethality rate of N$(AAAAU)_{139}$ was significantly lower than in the other conditions ($p<0.001$), suggesting that the longer normal RNA N$(AAAAU)_{139}$ may have a protective effect when compared with the other tested RNAs. Perhaps, in an affected individual with a large normal allele, the large normal RNA may interfere with the SCA37 phenotype, protecting from toxicity. However, clinical investigation together with co-injection of N$(AAAAU)_{139}$ with $(GAAAU)_{54}$ RNA need to be performed to clarify this hypothesis.

In summary, in this study we identified two promoters in *DAB1* antisense orientation that are able to interact with brain-specific enhancers. One promoter is located upstream the repeat and may be able to drive antisense repeat expression in human cerebellum and the other promoter is located downstream the repeat. Although the antisense repeat is not toxic *in vivo*, their expression may have other regulatory functions, such as the interference with the toxic sense repeats, modulating the SCA37 pathogenicity.

Future perspectives

The present work enabled to shed light on a putative role of *DAB1* antisense promoters in SCA37 pathogenicity, however further work needs to be performed to understand 1) how the antisense promoters identified in this study are expressed in human cerebellum, and 2) what is their role in SCA37 pathogenicity. Now it is mandatory to map the TSSs of the identified promoters (Frag 1 and Frag 3) in zebrafish embryos to later investigate their expression in human cerebellum. For that, after RNA extraction from stable transgenic zebrafish lines containing Frag T, we will perform a 5' RACE using a gene specific primer that anneal on EGFP, to identify the TSSs of the transcripts from Frag 1 and Frag 3 promoters. Later, to characterize the full-length transcripts driven by promoters in Frag 1 and Frag 3 in human cerebellum, the brain region most affected by this SCA, we will perform a 3' RACE using human cerebellum RNA and specific primers that anneal in the sequence detected in zebrafish 5' RACE. As Frag 1 and Frag 3 showed specific expression in horizontal myosepta and muscle fibers, respectively, we will also perform the characterization of these promoters in RNA from human muscle.

The characterization of promoter activity from *DAB1* antisense strand was here investigated using the short normal allele $(AAAAT)_{14}$. As the two promoters characterized in this project are located upstream and downstream of the AAAAT repeat (**Figure 9**), differences in repeat length or nucleotide composition may modify promoters activity due to chromatin rearrangements. Thus, it is necessary to perform the characterization of these antisense promoters containing the large normal allele $(AAAAT)_{139}$ and the pathogenic $(GAAAT)_n$ allele to understand if different repeat size or repeat nucleotide composition changes transcript expression. For that, we will clone into 1GWC2EGFP vectors the larger normal allele and the pathogenic allele upstream of an EGFP sequence to assess their promoter activity, injecting these vectors in one to two-cell

stage zebrafish embryos. If differences in promoters expression are detected, other methods such as ChIA-PET or Hi-C may be used to understand the chromatin interactions in this region[66].

In FXS and FRDA, the repeat expansions lead to a gene loss-of-of-function by gene expression silencing[1]. FXS is caused by CGG repeat expansion, above 200 units, in the 5' UTR of the *FMR1*. This expansion is methylated, and the methylation extends to the gene promoter, leading to *FMR1* silencing[87]. In FRDA, the expanded GAA repeat in the first intron of *FXN* gene leads to the histone modifications in the repeat flanking region with consequent heterochromatin formation, impeding the gene transcription initiation and elongation[11]. Given the genomic context of SCA37 locus (**Figure 5**), in which a CpG island is located at downstream of the ATTTT repeat, we hypothesized that the pathogenic insertion may lead to the CpG island methylation and/or histone modifications, and, consequently, to a silencing of Frag 3 and/or Frag 1 promoter. Thus, CpG island methylation and histone modifications may lead to heterochromatin formation, which might have an impact in *DAB1* expression. If differences in promoters activity are detected using the pathogenic insertion allele, we may perform a chromosome conformation capture-on-chip (4C) to verify the DNA conformation in this region. According to our results, Frag 1 and Frag 3 promoters may have a synergic activity that might be compromised due to these chromatin modifications. Furthermore, the different nucleotide composition may lead to a different genomic regulation by *cis*- and *trans*- regulatory elements, as enhancers or transcription factors binding. Thus, the different nucleotide content may interfere with DNA chromatin dynamics, that may enable a different interaction and binding of regulatory elements, which may lead to a transcription dysregulation.

The SCA37 pathogenic mechanism includes an RNA-mediated toxicity by the repeat in *DAB1*-oriented strand, as demonstrated by the $(AUUUC)_{58}$ RNA injection in zebrafish embryos[6]. In FTD/ALS and FXTAS, the bidirectional transcription has been reported, contributing for disease pathogenicity[16; 35]. Given that we found promoter activity across the *DAB1* antisense strand triggered by a promoter upstream of the repeat, we hypothesised that the antisense transcripts containing the $(GAAAU)_n$ RNA may also play a role in the disease toxicity[74]. In this work, our results revealed that the $(GAAAU)_{54}$ RNA insertion is not toxic for zebrafish embryos, however, we do not exclude a role for antisense repeat transcription in SCA37 disease. The antisense RNA may act at transcriptional level, decreasing or increasing the transcription of the sense strand due to heterochromatin formation or recruitment of transcriptional machinery, respectively[88; 89]. Hypothetically, the antisense RNA insertion might bind to the sense RNA insertion, reducing or increasing the sequestration of RNABP and, consequently, decreasing or

increasing the cellular toxicity. In future, to unravel the role of both strands in SCA37 pathogenicity, we will perform the co-injection of $(AUUUC)_{58}$ and $(GAAAU)_{54}$ RNAs in zebrafish embryos. One hypothesis that could explain the significant decrease of lethality and malformations in zebrafish embryos after the $(AAAAU)_{139}$ RNA injection is the possibility of acting as a long-noncoding RNA that may interfere with apoptosis pathway, since these RNA molecules may interact with mRNAs and/or proteins involved in this pathway. For example, long noncoding MALAT1 was already reported to inhibit the caspase-8 activity, contributing for apoptosis suppression[90]. To test this hypothesis, some apoptosis markers should be analysed by immunofluorescence and Western blot to investigate the differences of protein localisation and expression between $(AAAAU)_{139}$ RNA and an RNA control, after zebrafish embryos injections. Given the reduction of lethality verified in zebrafish embryos injected with $(AAAAU)_{139}$ (**Figure 21**), we hypothesised that this long normal allele may have a protective effect, that may decrease the cellular toxicity when present together with the pathogenic insertion allele. The co-injection of $(AUUUC)_{58}$ and $(AAAAU)_{139}$ RNAs will also be performed to understand the protective effect of $(AAAAU)_{139}$ allele in an affected individual with these both alleles.

Therefore, in the future, we will focus our work in the understanding of the genomic regulation behind SCA37 locus and the pathways that lead to an RNA-mediated toxicity, aiming to understand the contribution of antisense expression for the cellular toxicity in human cerebellum, that will give insight in the pathogenic mechanism underlying the SCA37.

References

1. Rohilla, K.J., and Gagnon, K.T. (2017). RNA biology of disease-associated microsatellite repeat expansions. Acta Neuropathologica Communications 5, 63.
2. Schmidt, M.H.M., and Pearson, C.E. (2016). Disease-associated repeat instability and mismatch repair. DNA Repair 38, 117-126.
3. Loureiro, J.R., Oliveira, C.L., and Silveira, I. (2016). Unstable repeat expansions in neurodegenerative diseases: nucleocytoplasmic transport emerges on the scene. Neurobiology of Aging 39, 174-183.
4. Nelson, D.L., Orr, H.T., and Warren, S.T. (2013). The Unstable Repeats - Three Evolving Faces of Neurological Disease. Neuron 77, 825-843.
5. Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T.*, et al.* (2009). Spinocerebellar Ataxia Type 31 Is Associated with "Inserted" Penta-Nucleotide Repeats Containing (TGGAA)n. The American Journal of Human Genetics 85, 544-557.
6. Seixas, A.I., Loureiro, J.R., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C.L.*, et al.* (2017). A Pentanucleotide ATTTC Repeat Insertion in the Non-coding Region of DAB1, Mapping to SCA37, Causes Spinocerebellar Ataxia. The American Journal of Human Genetics 101, 87-103.
7. Ishiura, H., Doi, K., Mitsui, J., Yoshimura, J., Matsukawa, M.K., Fujiyama, A.*, et al.* (2018). Expansions of intronic TTTCA and TTTTA repeats in benign adult familial myoclonic epilepsy. Nature Genetics 50, 581-590.
8. Paulson, H.L. (2009). The Spinocerebellar Ataxias. Journal of Neuro-Ophthalmology 29, 227-237.
9. Paulson, H.L., Shakkottai, V.G., Clark, H.B., and Orr, H.T. (2017). Polyglutamine spinocerebellar ataxias - from genes to potential treatments. Nature Reviews Neuroscience 18, 613-626.
10. Messaed, C., and Rouleau, G.A. (2009). Molecular mechanisms underlying polyalanine diseases. Neurobiology of Disease 34, 397-405.
11. Evans-Galea, M.V., Hannan, A.J., Carrodus, N., Delatycki, M.B., and Saffery, R. (2013). Epigenetic modifications in trinucleotide repeat diseases. Trends in Molecular Medicine 19, 655-663.
12. Anthony, K., and Gallo, J.-M. (2010). Aberrant RNA processing events in neurological disorders. Brain Research 1338, 67-77.

13. Zhang, N., and Ashizawa, T. (2017). RNA toxicity and foci formation in microsatellite expansion diseases. Current Opinion in Genetics & Development 44, 17-29.
14. García-Murias, M., Quintáns, B., Arias, M., Seixas, A.I., Cacheiro, P., Tarrío, R.*, et al.* (2012). 'Costa da Morte' ataxia is spinocerebellar ataxia 36: clinical and genetic characterization. Brain 135, 1423-1435.
15. Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y.*, et al.* (2011). Expansion of Intronic GGCCTG Hexanucleotide Repeat in NOP56 Causes SCA36, a Type of Spinocerebellar Ataxia Accompanied by Motor Neuron Involvement. American Journal of Human Genetics 89, 121-130.
16. Kong, H.E., Zhao, J., Xu, S., Jin, P., and Jin, Y. (2017). Fragile X-Associated Tremor/Ataxia Syndrome: From Molecular Pathogenesis to Development of Therapeutics. Frontiers in Cellular Neuroscience 11, 128.
17. Thornton, C.A., Wang, E., and Carrell, E.M. (2017). Myotonic dystrophy: approach to therapy. Current Opinion in Genetics & Development 44, 135-140.
18. Scotti, M.M., and Swanson, M.S. (2016). RNA mis-splicing in disease. Nature Reviews Genetics 17, 19-32.
19. Zhang, K., Donnelly, C.J., Haeusler, A.R., Grima, J.C., Machamer, J.B., Steinwald, P.*, et al.* (2015). The C9ORF72 repeat expansion disrupts nucleocytoplasmic transport. Nature 525, 56-61.
20. DeJesus-Hernandez, M., Mackenzie, I.R., Boeve, B.F., Boxer, A.L., Baker, M., Rutherford, N.J.*, et al.* (2011). Expanded GGGGCC hexanucleotide repeat in non-coding region of C9ORF72 causes chromosome 9p-linked frontotemporal dementia and amyotrophic lateral sclerosis. Neuron 72, 245-256.
21. Renton, A.E., Majounie, E., Waite, A., Simón-Sánchez, J., Rollinson, S., Gibbs, J.R.*, et al.* (2011). A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72, 257-268.
22. Jain, A., and Vale, R.D. (2017). RNA phase transitions in repeat expansion disorders. Nature 546, 243.
23. Kumar, V., kashav, T., Islam, A., Ahmad, F., and Hassan, M.I. (2016). Structural insight into C9orf72 hexanucleotide repeat expansions: Towards new therapeutic targets in FTD-ALS. Neurochemistry International 100, 11-20.
24. Green, K.M., Linsalata, A.E., and Todd, P.K. (2016). RAN translation—what makes it run? Brain research 1647, 30-42.
25. Cleary, J.D., and Ranum, L.P.W. (2017). New developments in RAN translation: insights from multiple diseases. Current Opinion in Genetics & Development 44, 125-134.

26. Gitler, A.D., and Tsuiji, H. (2016). There has been an awakening: Emerging mechanisms of C9orf72 mutations in FTD/ALS. Brain Research 1647, 19-29.
27. Kearse, M.G., Green, K.M., Krans, A., Rodriguez, C.M., Linsalata, A.E., Goldstrohm, A.C.*, et al.* (2016). CGG Repeat associated non-AUG translation utilizes a cap-dependent, scanning mechanism of initiation to produce toxic proteins. Molecular cell 62, 314-322.
28. Jackson, G.R. (2017). SCA31 Flies Perform in a Balancing Act between RAN Translation and RNA-Binding Proteins. Neuron 94, 4-5.
29. Zu, T., Cleary, J.D., Liu, Y., Bañez-Coronel, M., Bubenik, J.L., Ayhan, F.*, et al.* (2017). RAN Translation Regulated by Muscleblind Proteins in Myotonic Dystrophy Type 2. Neuron 95, 1292-1305.e1295.
30. Zu, T., Gibbens, B., Doty, N.S., Gomes-Pereira, M., Huguet, A., Stone, M.D.*, et al.* (2011). Non-ATG–initiated translation directed by microsatellite expansions. Proceedings of the National Academy of Sciences of the United States of America 108, 260-265.
31. Green, K.M., Glineburg, M.R., Kearse, M.G., Flores, B.N., Linsalata, A.E., Fedak, S.J.*, et al.* (2017). RAN translation at C9orf72-associated repeat expansions is selectively enhanced by the integrated stress response. Nature Communications 8, 2005.
32. Batra, R., Charizanis, K., and Swanson, M.S. (2010). Partners in crime: bidirectional transcription in unstable microsatellite disease. Human Molecular Genetics 19, R77-R82.
33. Budworth, H., and McMurray, C.T. (2013). Bidirectional transcription of trinucleotide repeats: Roles for excision repair. DNA Repair 12, 672-684.
34. Wilburn, B., Rudnicki, Dobrila D., Zhao, J., Weitz, Tara M., Cheng, Y., Gu, X.*, et al.* (2011). An Antisense CAG Repeat Transcript at JPH3 Locus Mediates Expanded Polyglutamine Protein Toxicity in Huntington's Disease-like 2 Mice. Neuron 70, 427-440.
35. Mori, K., Arzberger, T., Grässer, F.A., Gijselinck, I., May, S., Rentzsch, K.*, et al.* (2013). Bidirectional transcripts of the expanded C9orf72 hexanucleotide repeat are translated into aggregating dipeptide repeat proteins. Acta Neuropathologica 126, 881-893.
36. Ikeda, Y., Daughters, R.S., and Ranum, L.P.W. (2008). Bidirectional expression of the SCA8 expansion mutation: One mutation, two genes. The Cerebellum 7, 150-158.

37. Sopher, B.L., Ladd, P.D., Pineda, V.V., Libby, R.T., Sunkin, S.M., Hurley, J.B.*, et al.* (2011). CTCF regulates ataxin-7 expression through promotion of a convergently transcribed, antisense non-coding RNA. Neuron 70, 1071-1084.

38. Jiang, J., and Cleveland, D.W. (2016). Bidirectional Transcriptional Inhibition as Therapy for ALS/FTD Caused by Repeat Expansion in C9orf72. Neuron 92, 1160-1163.

39. Zu, T., Liu, Y., Bañez-Coronel, M., Reid, T., Pletnikova, O., Lewis, J.*, et al.* (2013). 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 110, E4968-E4977.

40. Yin, S., Lopez-Gonzalez, R., Kunz, R.C., Gangopadhyay, J., Borufka, C., Gygi, S.P.*, et al.* (2017). Evidence that C9ORF72 Dipeptide Repeat Proteins Associate with U2 snRNP to Cause Mis-splicing in ALS/FTD Patients. Cell Reports 19, 2244-2256.

41. Lo, R.Y., Figueroa, K.P., Pulst, S.M., Perlman, S., Wilmot, G., Gomez, C.*, et al.* (2016). Depression and clinical progression in spinocerebellar ataxias. Parkinsonism & Related Disorders 22, 87-92.

42. van der Stijl, R., Withoff, S., and Verbeek, D.S. (2017). Spinocerebellar ataxia: miRNAs expose biological pathways underlying pervasive Purkinje cell degeneration. Neurobiology of Disease 108, 148-158.

43. Sequeiros, J., Martins, S., and Silveira, I. (2012). Epidemiology and population genetics of degenerative ataxias. In Handbook of Clinical Neurology, S.H. Subramony and A. Dürr, eds. (Elsevier), pp 227-251.

44. Coutinho, P., Ruano, L., Loureiro, J.L., Cruz, V.T., Barros, J., Tuna, A.*, et al.* (2013). Hereditary ataxia and spastic paraplegia in portugal: A population-based prevalence study. Journal of Anerican Medical Association Neurology 70, 746-755.

45. Serrano-Munuera, C., Corral-Juan, M., Stevanin, G., San Nicolás, H., Roig, C., Corral, J.*, et al.* (2013). New subtype of spinocerebellar ataxia with altered vertical eye movements mapping to chromosome 1p32. Journal of the American Medical Association Neurology 70, 764-771.

46. Nasiadka, A., and Clark, M.D. (2012). Zebrafish Breeding in the Laboratory Environment. Institute Laboratory Animal Research Journal 53, 161-168.

47. Kimmel, C.B., Ballard, W.W., Kimmel, S.R., Ullmann, B., and Schilling, T.F. (1995). Stages of embryonic development of the zebrafish. Developmental Dynamics 203, 253-310.

48. Stewart, A.M., Braubach, O., Spitsbergen, J., Gerlai, R., and Kalueff, A.V. (2014). Zebrafish models for translational neuroscience research: from tank to bedside. Trends in Neurosciences 37, 264-278.
49. Kalueff, A.V., Stewart, A.M., and Gerlai, R. (2014). Zebrafish as an emerging model for studying complex brain disorders. Trends in Pharmacological Sciences 35, 63-75.
50. Martín-Jiménez, R., Campanella, M. & Russell. (2015). New Zebrafish Models of Neurodegeneration. Current Neurology and Neuroscience Reports 15.
51. Liu, H., Li, X., Ning, G., Zhu, S., Ma, X., Liu, X.*, et al.* (2016). The Machado–Joseph Disease Deubiquitinase Ataxin-3 Regulates the Stability and Apoptotic Function of p53. Public Library of Science Biology 14, e2000733.
52. Issa, F.A., Mazzochi, C., Mock, A.F., and Papazian, D.M. (2011). Spinocerebellar Ataxia Type 13 Mutant Potassium Channel Alters Neuronal Excitability and Causes Locomotor Deficits in Zebrafish. The Journal of Neuroscience 31, 6831-6841.
53. Corral-Juan, M., Serrano-Munuera, C., Rábano, A., Cota-González, D., Segarra-Roca, A., Ispierto, L.*, et al.* (2018). Clinical, genetic and neuropathological characterization of spinocerebellar ataxia type 37. Brain 141, 1981-1997.
54. Clark, R.M., Dalgliesh, G.L., Endres, D., Gomez, M., Taylor, J., and Bidichandani, S.I. (2004). Expansion of GAA triplet repeats in the human genome: unique origin of the FRDA mutation at the center of an Alu. Genomics 83, 373-383.
55. Kurosaki, T., Ueda, S., Ishida, T., Abe, K., Ohno, K., and Matsuura, T. (2012). The Unstable CCTG Repeat Responsible for Myotonic Dystrophy Type 2 Originates from an AluSx Element Insertion into an Early Primate Genome. Public Library of Science ONE 7, e38379.
56. Kurosaki, T., Matsuura, T., Ohno, K., and Ueda, S. (2009). Alu-Mediated Acquisition of Unstable ATTCT Pentanucleotide Repeats in the Human ATXN10 Gene. Molecular Biology and Evolution 26, 2573-2579.
57. Chen, L.-L., and Yang, L. (2017). ALUternative Regulation for Gene Expression. Trends in Cell Biology 27, 480-490.
58. Ade, C., Roy-Engel, A.M., and Deininger, P.L. (2013). Alu elements: An intrinsic source of human genome instability. Current opinion in virology 3, 639-645.
59. Deininger, P. (2011). Alu elements: know the SINEs. Genome Biology 12, 236-236.
60. Ugarkovic, D. (2011). Long Non-Coding RNAs.(Springer Berlin Heidelberg).
61. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G.*, et al.* (2001). The Sequence of the Human Genome. Science 291, 1304.

62. Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional Regulatory Elements in the Human Genome. Annual Review of Genomics and Human Genetics 7, 29-59.
63. Mercer, T.R., and Mattick, J.S. (2013). Understanding the regulatory and transcriptional complexity of the genome through structure. Genome Research 23, 1081-1088.
64. Kim, T.-K., and Shiekhattar, R. (2015). Architectural and functional commonalities between enhancers and promoters. Cell 162, 948-959.
65. Ron, G., Globerson, Y., Moran, D., and Kaplan, T. (2017). Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains. Nature Communications 8, 2237.
66. Li, Y., Chen, C.-y., Kaye, A.M., and Wasserman, W.W. (2015). The identification of cis-regulatory elements: A review from a machine learning perspective. Biosystems 138, 6-17.
67. Spielmann, M., and Mundlos, S. (2016). Looking beyond the genes: the role of non-coding variants in human disease. Human Molecular Genetics 25, R157-R165.
68. A. Sherf, B., L. Navarro, S., Hannah, R., and V. Wood, K. (1996). Dual-Luciferase TM Reporter Assay: An Advanced Co-Reporter Technology Integrating Firefly and Renilla Luciferase Assays.
69. Alcaraz-Pérez, F., Mulero, V., and Cayuela, M. (2008). Application of the dual-luciferase reporter assay to the analysis of promoter activity in Zebrafish embryos. BMC Biotechnology 8, 81.
70. de la Calle-Mustienes, E., Feijóo, C.G., Manzanares, M., Tena, J.J., Rodríguez-Seguel, E., Letizia, A.*, et al.* (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. Genome Research 15, 1061-1072.
71. Allende, M., Manzanares, M., Tena, J., Feijoo, C., and Gómez-Skarmeta, J. (2006). Cracking the genome's second code: Enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. Methods 39, 212-219.
72. Bessa, J., Tena, J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A.*, et al.* (2009). Zebrafish Enhancer Detection (ZED) Vector: A New Tool to Facilitate Transgenesis and the Functional Analysis of cis-Regulatory Regions in Zebrafish. Developmental Dynamics 238, 2409-2417.
73. Liu, C., Song, G., Mao, L., Long, Y., Li, Q., and Cui, Z. (2015). Generation of an Enhancer-Trapping Vector for Insertional Mutagenesis in Zebrafish. Public Library of Science ONE 10, e0139612.

74. Loureiro, J.R., Seixas, A.I., Costa, C., Ordóñez-Ugalde, A., Marcelino, H., Oliveira, C.L.*, et al.* (2017). A spinocerebellar ataxia mapping to the SCA37 locus is caused by a pentanucleotide ATTTC repeat insertion in the noncoding region of the DAB1 gene. In American Society Human Genetics Meeting. (Orlando, USA.
75. Karlsson, J., von Hofsten, J., and Olsson, P.-E. (2001). Generating Transparent Zebrafish: A Refined Method to Improve Detection of Gene Expression During Embryonic Development. Marine Biotechnology 3, 522-527.
76. Katzen, F. (2007). Gateway® recombinational cloning: a biological operating system. Expert Opinion on Drug Discovery 2, 571-589.
77. Hartley, J.L., Temple, G.F., and Brasch, M.A. (2000). DNA Cloning Using In Vitro Site-Specific Recombination. Genome Research 10, 1788-1795.
78. Ni, J., Wangensteen, K.J., Nelsen, D., Balciunas, D., Skuster, K.J., Urban, M.D.*, et al.* (2016). Active recombinant Tol2 transposase for gene transfer and gene discovery applications. Mobile DNA 7, 6.
79. Kawakami, K. (2007). Tol2: a versatile gene transfer vector in vertebrates. Genome Biology 8, S7-S7.
80. Clark, K.J., Urban, M.D., Skuster, K.J., and Ekker, S.C. (2011). Transgenic Zebrafish Using Transposable Elements. Methods in Cell Biology 104, 137-149.
81. (2002). Pattern Formation in Zebrafish.(Springer-Verlag Berlin Heidelberg).
82. Gompel, N., Cubedo, N., Thisse, C., Thisse, B., Dambly-Chaudière, C., and Ghysen, A. (2001). Pattern formation in the lateral line of zebrafish. Mechanisms of Development 105, 69-77.
83. López-Schier, H., Starr, C.J., Kappler, J.A., Kollmar, R., and Hudspeth, A.J. (2004). Directional Cell Migration Establishes the Axes of Planar Polarity in the Posterior Lateral-Line Organ of the Zebrafish. Developmental Cell 7, 401-412.
84. Swinnen, B., Bento-Abreu, A., Gendron, T.F., Boeynaems, S., Bogaert, E., Nuyts, R.*, et al.* (2018). A zebrafish model for C9orf72 ALS reveals RNA toxicity as a pathogenic mechanism. Acta Neuropathologica 135, 427-443.
85. Su, M., Han, D., Boyd-Kirkup, J., Yu, X., and Han, J.-Dong J. (2014). Evolution of Alu Elements toward Enhancers. Cell Reports 7, 376-385.
86. Donnelly, C.J., Zhang, P.-W., Pham, J.T., Heusler, A.R., Mistry, N.A., Vidensky, S.*, et al.* (2013). RNA Toxicity from the ALS/FTD C9ORF72 Expansion Is Mitigated by Antisense Intervention. Neuron 80, 415-428.
87. Okray, Z., de Esch, C.E.F., Van Esch, H., Devriendt, K., Claeys, A., Yan, J.*, et al.* (2015). A novel fragile X syndrome mutation reveals a conserved role for the carboxy-terminus in FMRP localization and function. EMBO molecular medicine 7, 423-437.

88. Dang, Y., Cheng, J., Sun, X., Zhou, Z., and Liu, Y. (2016). Antisense transcription licenses nascent transcripts to mediate transcriptional gene silencing. Genes & development 30, 2417-2432.
89. Cho, D.H., Thienes, C.P., Mahoney, S.E., Analau, E., Filippova, G.N., and Tapscott, S.J. (2005). Antisense Transcription and Heterochromatin at the DM1 CTG Repeats Are Constrained by CTCF. Molecular Cell 20, 483-489.
90. Su, Y., Wu, H., Pavlosky, A., Zou, L.-L., Deng, X., Zhang, Z.-X.*, et al.* (2016). Regulatory non-coding RNA: new instruments in the orchestration of cell death. Cell death & disease 7, e2333-e2333.

Appendix 1

Phenol chloroform DNA/RNA purification

For RNA purification all the material used in the procedure is RNAse free and all the steps were performed at 4ºC.

1. If necessary, bring the DNA/RNA volume up to 100 µL with nuclease free HyClone HyPure water (GE Healthcare Life Sciences)
2. Add 100 µL of phenol chloroform isoamyl alcohol (Fisher Scientific) at 4°C
3. Mix vigorously and centrifuge at high speed (13 000 rpm) for 5 min
4. Transfer the upper phase (aqueous phase) to a new tube
5. Add 100 µL of chloroform (Fisher Scientific) at room temperature
6. Mix vigorously and centrifuge at high speed (13 000 rpm) for 5 min
7. Place the upper phase (aqueous phase) to a new tube RNAse free and add 10 µL of sodium acetate (3 M, pH 5.2) (Merch, Milipore) at room temperature to each 100 µL of upper phase collected
8. Add 2 volumes of 100% ethanol (Merch, Milipore) at -20°C
9. Mix gently, inverting the tube 2-4 times, and precipitate the DNA at -80°C for at least 1 hr
10. Centrifuge at high speed (13 000 rpm) for 15 min
11. Discard the ethanol and wash the pellet with 70 % ethanol at -20°C
12. Centrifuge at high speed (13 000 rpm) for 5 min
13. Discard the supernatant and dry the pellet
14. Resuspend the pellet in HyClone HyPure water (GE Healthcare Life Sciences) and place on ice
15. Quantify the DNA/RNA in NanoDrop™ 1000 Spectrophotometer (Thermo Fisher Scientific)

Appendix 2

In vitro RNA synthesis

For *in vitro* RNA synthesis, all the material used in the procedure is RNAse free.

1. Linearize >3 µg of DNA plasmid and confirm the digestion, using 0.3 µg of DNA, by 0.8 % agarose gel electrophoresis
2. Quantify the plasmid DNA using the NanoDrop™ 1000 Spectrophotometer (Thermo Scientific)
3. Purify the DNA sample according to phenol chloroform DNA/RNA extraction (**Appendix 1**)
4. Prepare a 50 µL mix at final volume for RNA synthesis as followed:
    a. Add 1x transcription buffer (Thermo Scientific), 5 mM DTT (Invitrogen™) and 1 mM NTPs (New England Biolabs) and incubate at 37°C for 5 min
    b. Add 2-2.5 mM Cap Analog [m7G(5')ppp(5')G] (New England Biolabs) and incubate at 37°C for 1 min
    c. Add the purified DNA sample and incubate at 37°C for 1 min
    d. Add 120U NZY Ribonuclease Inhibitor (NZYTech) and incubate at 37°C for 1 min
    e. Add 40U RNA polymerase (Thermo Scientific) and incubate at 37°C for 1 hr
    f. Add 20U RNA polymerase and incubate at 37°C for 1 hr
    g. Add 1U DNAse I (Thermo Scientific) and incubate at 37°C for 30 min
5. Purify the transcribed RNA according to purification of RNA transcribed *in vitro* (**Appendix 3**)

Appendix 3

Purification of RNA transcribed *in vitro*

A. Sephadex column preparation

1. Remove the plugger of a 1 mL syringe and plug the syringe with sterile glass wool (0.1 mm)
2. Fill the syringe with illustra Sephadex G-50 Fine DNA Grade (GE Healthcare Life Sciences) and place it into a 15 mL disposable tube
3. Centrifuge at 4 000 rpm for 5 min at 4°C
4. Discard the flow-through and, if necessary, compact the column with more sephadex (up to 0,6 mL)
5. Discard the flow-through and place a new 1 mL tube at the bottom of the syringe to collect the RNA
6. Add 50 µL of HyClone HyPure water (GE Healthcare Life Sciences) to sephadex column and centrifuge at 4 000 rpm for 5 min at 4°C
7. Verify if the collected volume is 50 µL, if not repeat the centrifugation again

B. RNA purification in Sephadex column

1. Bring the RNA volume to 50 µL with HyClone HyPure water (GE Healthcare Life Sciences)
2. Place the RNA sample to sephadex column (A) and centrifuge at 4 000 rpm for 5 min at 4°C
3. Place the purified RNA sample into a new tube
4. Repeat the centrifugation and collect the purified RNA sample to the same tube
5. Quantify the RNA in NanoDrop™ 1000 Spectrophotometer (Thermo Fisher Scientific)

Appendix 4

Zebrafish microinjection

A. Courtship and spawning

1. At the end of the day prior to microinjection procedure, prepare 1 L breeding tank (**Figure 1 Appendix 4**) and place 3 females and 2 males in each side of the tank
2. At first hours of light cycle, place the breeding tank above a lamp, removing the divider of the tank and a small amount of water, to allow animals courtship
3. Collect the fertilized eggs in water facility system

Figure appendix 4 Apparatus of zebrafish breeding tank. Prior to microinjection procedure, 1L external tank containing an internal tank with a perforated bottom, a device and a lid was seated with water facility system. After the spawning, animals were placed on other external tank using the internal tank and the fertilized eggs were easily collected from the bottom of the external tank. Adapted from Techniplast.

B. Microinjection

1. Prepare the needles from glass capillaries using a puller (Narishige)
2. Cut off the needle's tip with forceps under a stereoscope
3. Prepare the injection mix as followed: 50 ng of the interest vector + 50 ng of *Tol2* RNA (ratio 1:1) + 10% phenol red (Sigma-Aldrich) (dye used to visualize the injection procedure)
4. Load the needle with microinjection mix and calibrate the microinjector (Narishige) pressure to inject the embryos with 5 nL approximately
5. Positioning the embryos with a help of a microscope slide and microinject approximately 5 nL into the yolk of one to two-cell stage embryos
6. Transfer the injected embryos to a new petri dish with E3 medium with PTU and incubate them at 28°C

Appendix 5

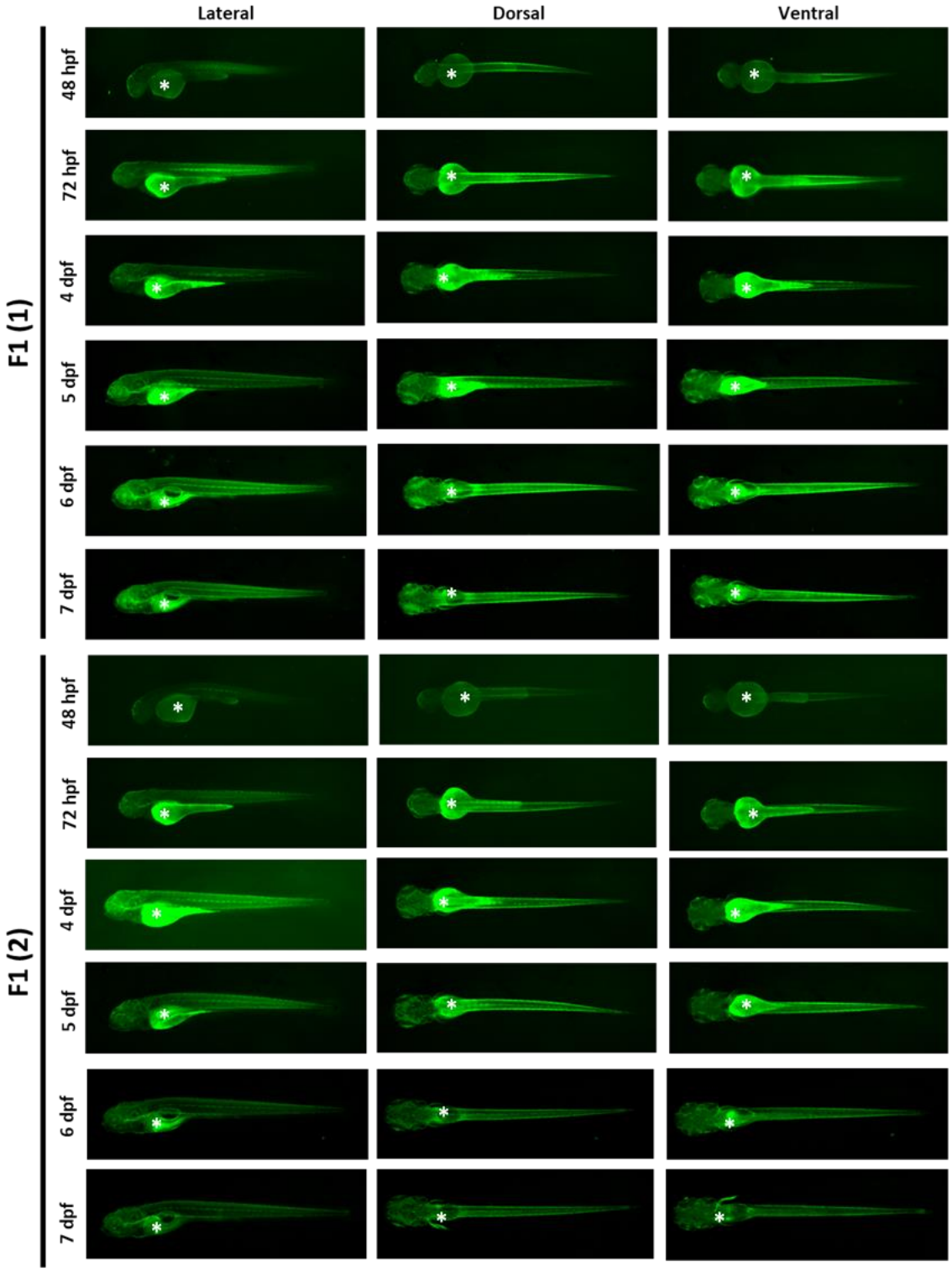


Figure appendix 5 Characterization of F1 embryos expressing Frag 1 promoter. The offspring of two founders was analysed from 48 hpf up to 7 dpf. EGFP expression was detected in muscular tissue, specifically in horizontal myosepta of the zebrafish offspring of F1(1) and F1(2); asterisks indicate autofluorescence regions.

Appendix 6

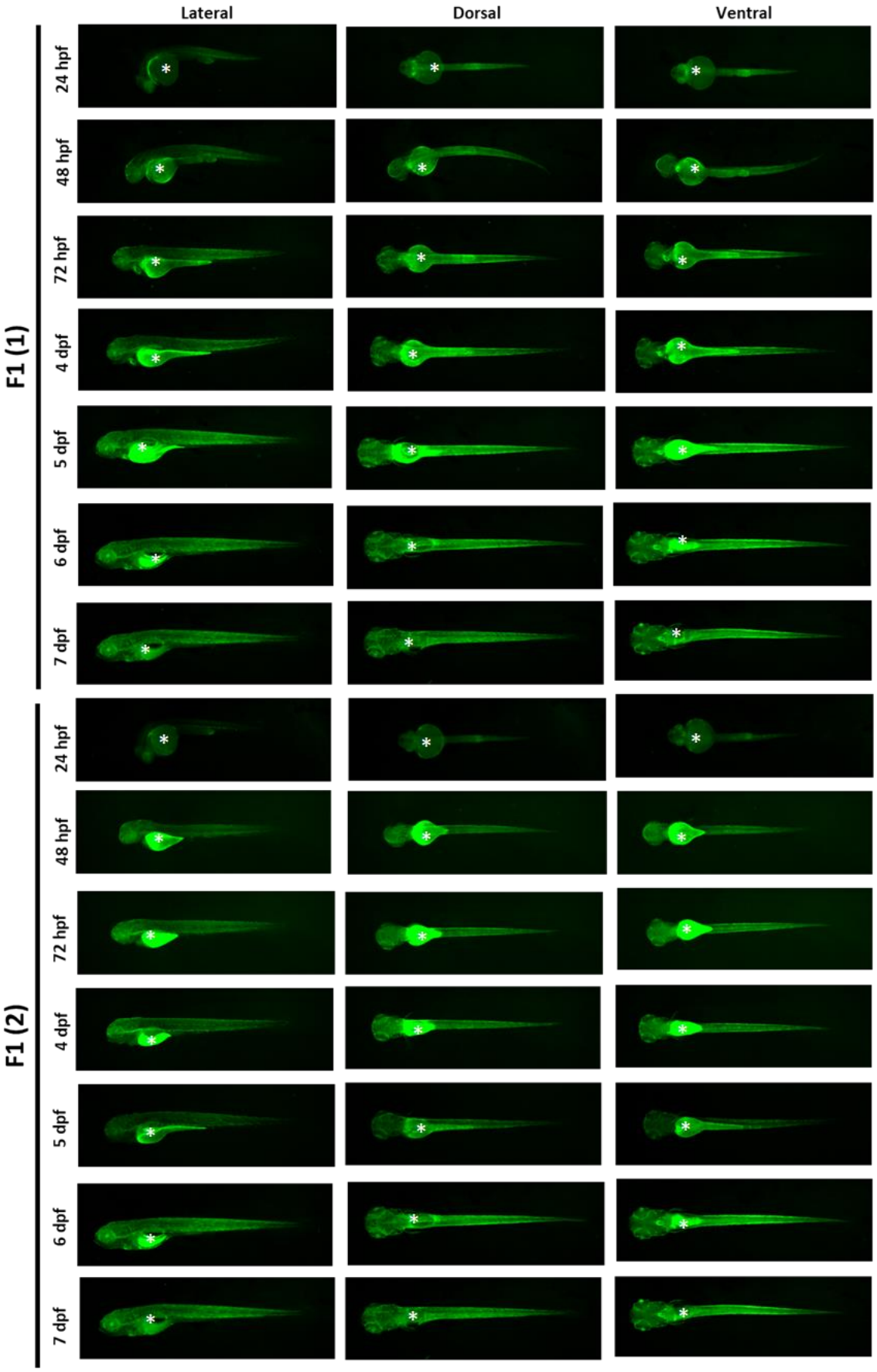

Lateral
Dorsal
Ventral
F1 (1)
24 hpf
48 hpf
72 hpf
4 dpf
5 dpf
6 dpf
7 dpf
F1 (2)
24 hpf
48 hpf
72 hpf
4 dpf
5 dpf
6 dpf
7 dpf

Figure appendix 6 Characterization of F1 embryos expressing Frag 3 promoter. The offspring of two founders was analysed from 24 hpf up to 7 dpf. EGFP expression was detected in muscle fibers along the body of zebrafish embryos. At 24 hpf, the offspring of F1(1) exhibit an EGFP expression in notochord, contrarily to the offspring of F1(2). Asterisks indicate autofluorescence regions.

Appendix 7

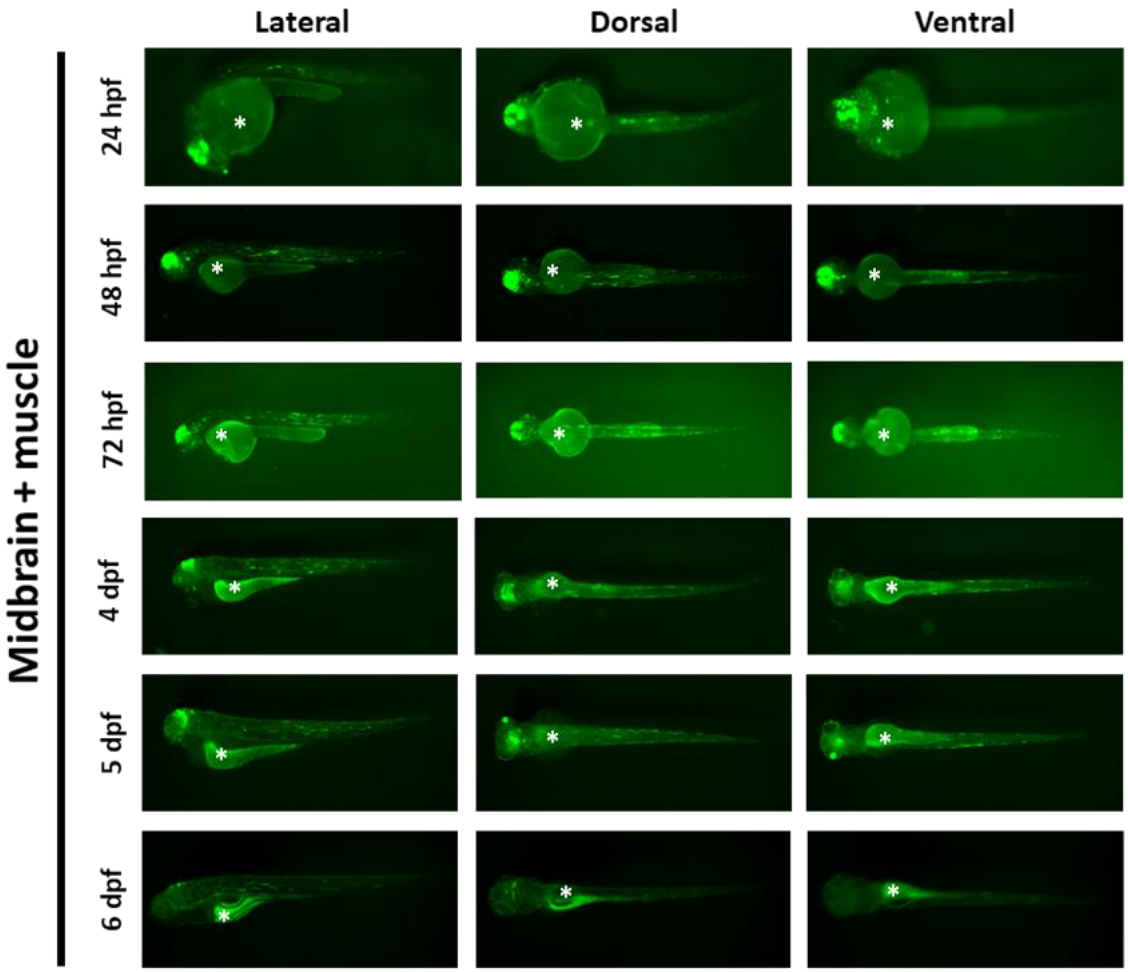


Figure appendix 7 Evaluation of Frag T promoter-Z48 enhancer in F0 zebrafish embryos along the time. In Frag T, promoter-enhancer interaction was observed from 24 hpf up to 6 dpf. EGFP expression was detected in whole midbrain cells. Although midbrain expression, EGFP was also detected in muscle fibers along the body of zebrafish embryos, as a result of promoter specificity for these regions. Asterisks indicate autofluorescence regions.

Appendix 8

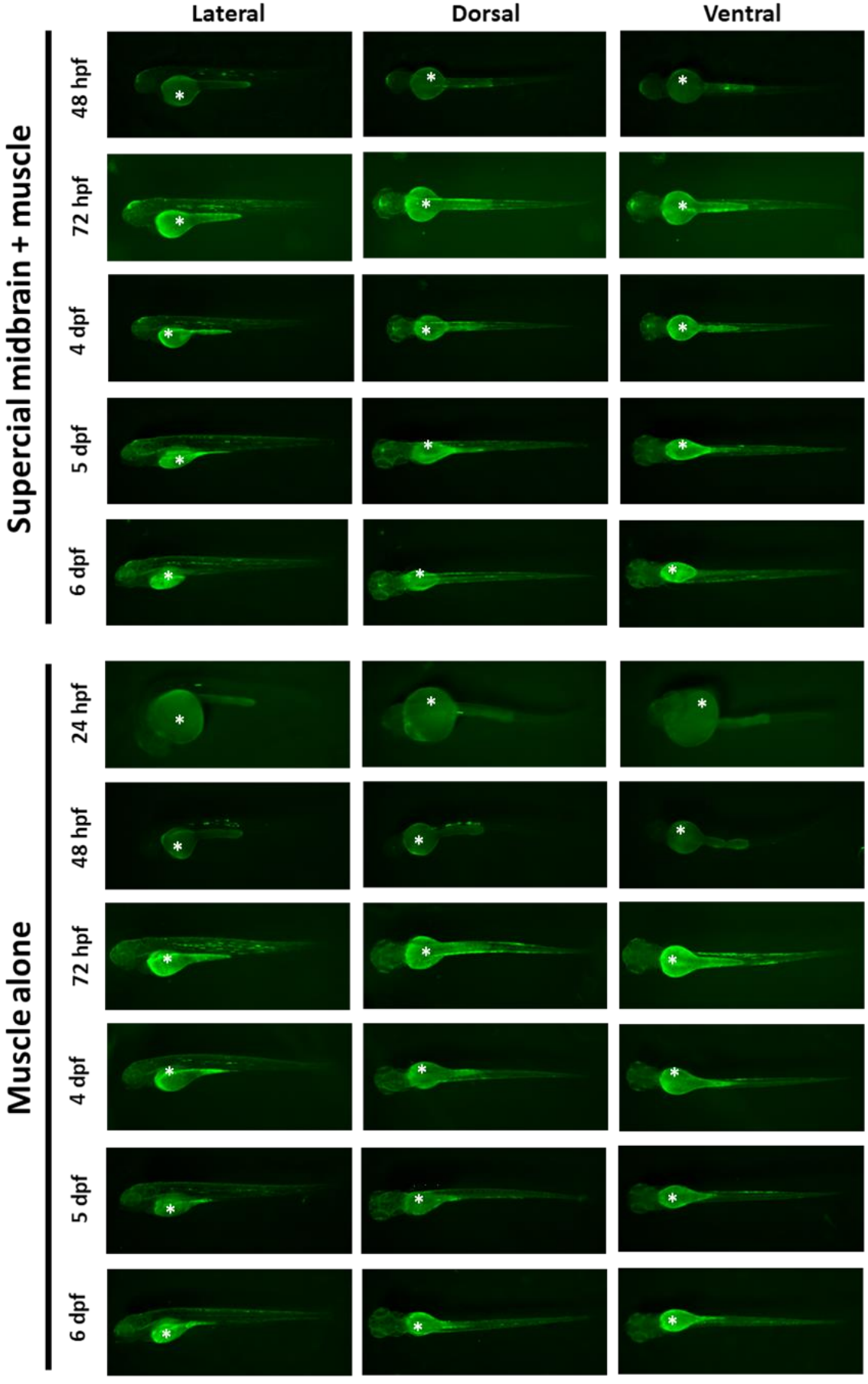

Lateral
Dorsal
Ventral
Supercial midbrain + muscle
48 hpf
72 hpf
4 dpf
5 dpf
6 dpf
Muscle alone
24 hpf
48 hpf
72 hpf
4 dpf
5 dpf
6 dpf

Figure appendix 8 Evaluation of Frag 1 promoter-Z48 enhancer in F0 zebrafish embryos. In Frag 1, promoter-enhancer interaction was observed from 48 hpf up to 6 dpf. EGFP expression was detected in superficial midbrain cells in some zebrafish embryos. In some transgenic embryos, EGFP in midbrain was not detected. EGFP expression was also observed in muscle fibers, as a result of promoter specificity for these regions. Asterisks indicate autofluorescence regions.

Appendix 9

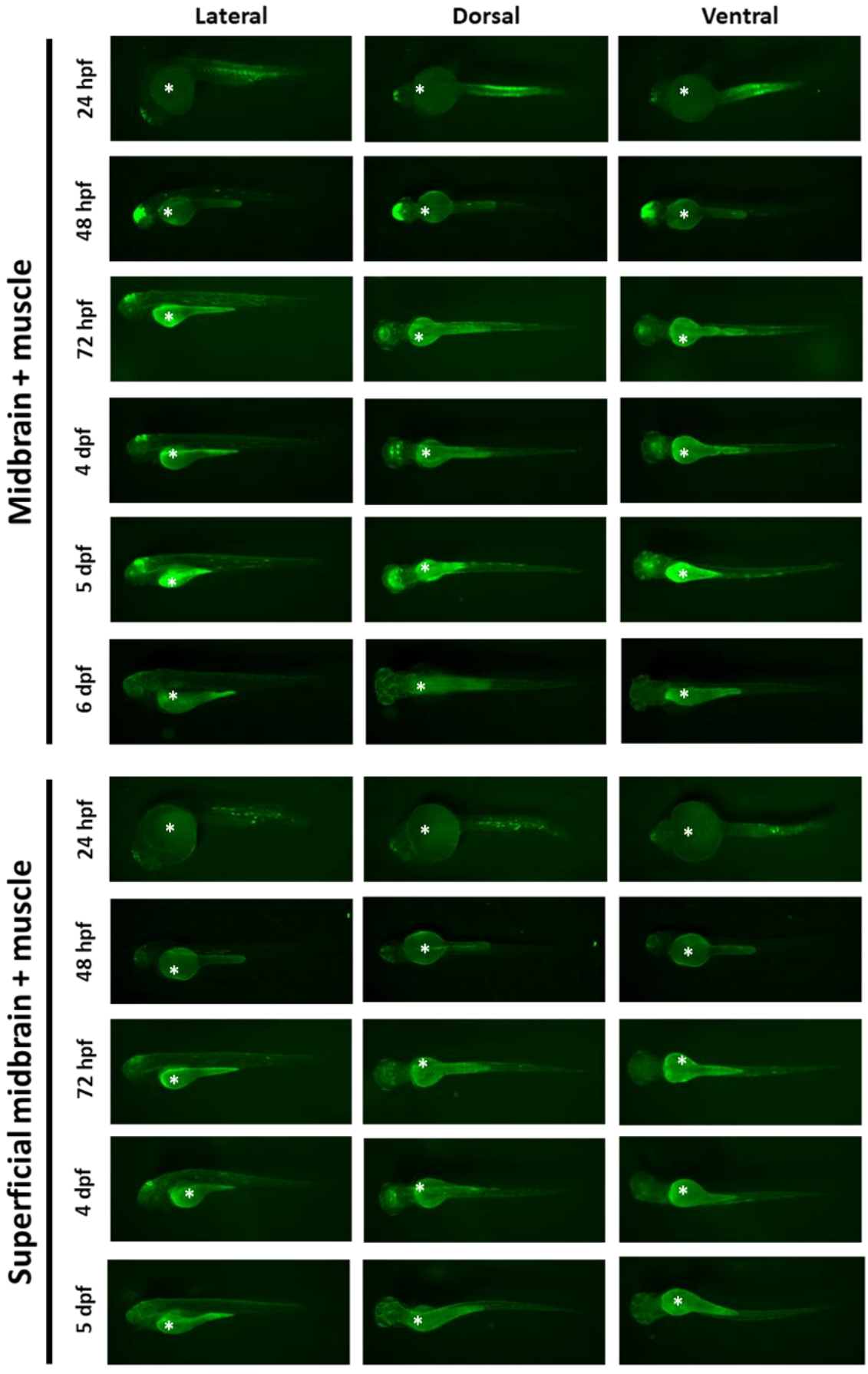

Lateral
Dorsal
Ventral
Midbrain + muscle
24 hpf
48 hpf
72 hpf
4 dpf
5 dpf
6 dpf
Superficial midbrain + muscle
24 hpf
48 hpf
72 hpf
4 dpf
5 dpf

Figure appendix 9 Evaluation of Frag 3 promoter-Z48 enhancer in F0 zebrafish. In Frag 3, promoter-enhancer interaction was observed from 24 hpf up to 6 dpf, and EGFP expression was detected spread in whole midbrain cells in some zebrafish embryos. However, in some embryos, EGFP expression was only detected in superficial cells of midbrain. Although midbrain expression, EGFP was detected in muscle fibers in all transgenic embryos as a result of promoter specificity for these regions. Asterisks indicate autofluorescence regions.