Towards Trans-Splicing Gene Therapy for HD : Intronic Targets Identification in the Huntingtin Gene

Séverine Maire

NNT : 2018SACLS054

Towards *trans*-splicing gene therapy for Huntington's disease : *intronic target identification in the Huntingtin gene*

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Paris-Sud
au sein du Laboratoire des Maladies Neurodégénératives
(LMN / MIRCen / UMR CEA-CNRS 9199 / Institut François Jacob)

École doctorale n°568
Signalisations et réseaux intégratifs en biologie (Biosigne)
Spécialité de doctorat: Science de la vie et de la santé

Thèse présentée et soutenue à Fontenay-aux-roses, le 09/03/2018, par

Séverine Maire

Composition du Jury :

Mohammed Taouis — **Président**
Professeur, Université Paris-Sud
(Institut de Neurosciences Paris-Saclay)

Benjamin Dehay — **Rapporteur**
Chargé de recherche, Université de Bordeaux
(Institut des Maladies Neurodégénératives)

Philippe Ravassard — **Rapporteur**
Chargé de recherche, Université Pierre-et-Marie-Curie
(Institut du Cerveau et de la Moelle épinière)

Aurélie Goyenvalle — **Examinatrice**
Chargée de recherche, Université de Versailles Saint-Quentin-en-Yvelines
(Handicap neuromusculaire : Physiopathologie, Biothérapie et Pharmacologie appliquées)

Peter Vanhoutte — **Examinateur**
Directeur de recherche, Université Pierre-et-Marie-Curie
(Institut de biologie Paris-Seine)

Alexis Bemelmans — **Directeur de thèse**
Chercheur, CEA
(MIRCen – Molecular Imaging Research Center)

Thèse de doctorat

I do it for my dad and I do it for my mama

I do it for my shabs and I do it for my sistas

[…]

Plus I do it for the passion and I do it for the patience

No I'll never forget the pain from this fight.

RILÈS KACIMI (JANUARY 4, 1996 -)
rapper, songwriter, composer, and record producer

"I do it" (2016)

Acknowledgement

Je tiens tout d'abord à remercier l'ensemble des membres de mon jury pour avoir accepté d'évaluer mes travaux de thèse avec une mention particulière pour Philippe Ravassard et Peter Vanhouttte pour avoir pu se rendre disponible au pied levé pour ma soutenance.

Je souhaiterais également remercier Philippe Hantraye et Emmanuel Brouillet qui m'ont permis de réaliser mon stage de master 2 et ma thèse au sein de leurs installations.

Je tiens ensuite à remercier Alexis Bemelmans pour m'avoir confié ce projet de thèse, d'avoir eu assez confiance en moi pour m'accorder autant de liberté dans ce projet ce qui m'a permis d'aller au bout de toutes mes idées, même les plus perchées ! Merci de m'avoir donné l'opportunité de partir en congrès à l'étranger et notamment à la SfN. Ça a vraiment été des expériences formidables ! J'espère avoir d'autres occasions pour y retourner. Merci également de m'avoir permis d'avoir autant d'activités hors recherche (enseignement, membre du conseil de l'ED, inscription à la prépa Médiadix, etc). J'ai vraiment eu de la chance de pouvoir bénéficier de toutes ces opportunités qui ont enrichie mon parcours professionnel ; je n'aurais pas pu en profiter sans ton soutien. Je remercie également Marie-Claude Gaillard, qui m'a patiemment encadré depuis mon stage de master. Merci pour ton éternelle bienveillance et tes qualités humaines qui m'ont accompagné tout au long de ma thèse et qui m'ont permis de rester toujours motivée même dans les périodes plus difficiles !

Je remercie tous les collaborateurs qui m'ont aidé dans ce projet et leur plateforme associée. Merci à Jean Livet de l'institut de la vision pour son aide pour la caractérisation des spectres de la XFP. Merci aux différentes plateformes de cytométrie qui ont participé à ce projet de recherche, notamment à Charlène Lasgi de l'institut curie d'Orsay, Antonio Cosma et Sabrina Guenounou ainsi que Jan Baijer et Nathalie Deschamps du CEA de Fontenay-aux-roses. Un grand merci également à Karine Labadie, Caroline Menguy et Eric Mahieu du Genoscope d'Evry qui m'ont permis une avancée déterminante quant à la construction des banques et plus particulièrement pour l'étape de fragmentation des introns.

Je tiens également à remercier les chercheurs qui aux files des discussions m'ont permis de faire avancer ma réflexion sur mon projet. Je tiens notamment à remercier Frederic Ducongé et Ché Serguera dont les échanges m'ont toujours énormément apporté et m'ont guidé dans l'aboutissement d'étapes critiques du projet. Merci à Karine Cambon qui m'a souvent amené à me poser les bonnes questions. Avoir travaillé à tes cotés m'a permis d'apprendre beaucoup. Je tiens également à remercier Géraldine Liot qui m'a également beaucoup aidé, et qui m'a soutenu depuis le premier jour de notre rencontre. Un énorme merci pour toutes ces discussions qui m'ont fait avancer et merci d'avoir

toujours été là pour moi au quotidien, à répondre à toutes mes inquiétudes et mes questionnements professionnels et personnels. Merci d'avoir contribué à mon cheminement durant mon parcours de thèse !

Je tiens également à remercier l'ensemble de mon équipe et de mon laboratoire qui m'ont permis de vivre une aventure formidable pendant ces 3 ans de thèse. Merci à vous Julie, Gwen, Charlène et Noëlle qui contribuent à la bonne ambiance de l'équipe ! Merci de m'avoir toujours dépanné dans mes manipes et de m'avoir régulièrement proposé votre aide, notamment pour m'aider à entretenir mes cellules !!! Merci Noëlle pour toutes les discussions autour de nos transfections à parler sciences ou pas du tout ! Merci de m'avoir soutenu à mes débuts et d'avoir pris le temps d'être là pour moi dans mes moments de panique !

Merci à l'ensemble des doctorants (et post-doctorant !) qui ont constitué la « colonie MIRCen ». Je n'oublierais jamais toutes ces bons moments (les soirées à chanter du sniper en passant par émile et image et finissant systématiquement avec « les rois du monde » !, les vacances à la campagnes, à la mer, et de l'autre coté de la terre, nos coup de gueule et nos longues conversations autour d'un cosmo) en espérant que ce soit juste le début d'une longue série. Merci pour cette solidarité qui fait la force de MIRCen et qui nous permet d'avancer mieux, plus vite tous ensemble toujours dans la bonne humeur.

Pour finir, je remercie ma famille qui m'ont toujours donné les moyens de mener à bien tous mes projets. Merci d'avoir toujours été là pour moi, et de m'avoir soutenu pendant ces 9 ans de parcours universitaires !

Table of contents

List of abbreviations

3'SS	3' splice site
5'SS	5' splice site
AGEZ	AG exclusion zone
BD	binding domain
BDNF	brain-derived neurotrophic factor
BHK	baby hamster kidney
BP	branch point
CSS	cryptic splicing site
DEB	dystrophic epidermolysis bullosa
DNA	deoxyribonucleic Acid
DS	donor site
ESE	exon splicing enhancer
ESS	exon splicing silencer
FACS	Fluorescence activated cell sorting
HD	Huntington's disease
HEAT	Huntingtin, elongation factor 3, protein phosphatase 2A, and TOR 1
hnRNP	heterogenous nuclear ribonucleoprotein
HSP	heat shock protein
iPSC	induced pluripotent stem cells
ISE	intron splicing enhancer
ISS	intron splicing silencer
MaxEnt	maximum entropy
MFI	mean fluorescence intensity
miRNA	micro-RNA
mRNA	messenger RNA
ncRNA	non-coding RNA
NT	non-transfected
PolyQ	polyglutamine stretch
PPT	polypyrimidine tract
pre-mRNA	precursor mRNA
PTM	*trans*-splicing molecule
RISC	RNA-induced silencing complex

RNA	ribonucleic Acid
RP	retinitis pigmentosa
rRNA	ribosomal RNA
SELEX	systematic evolution of ligands by exponential enrichment
shRNA	small-hairpin RNA
siRNA	small interfering RNA
SmaRT	spliceosome-mediated RNA trans-splicing
snRNA	small nuclear RNA
snRNP	small nuclear ribonucleoprotein particle
SR	serine/arginine-rich
SRE	splicing regulatory element
tRNA	transfer RNA
TS	*trans*-splicing
XFP	chimeric protein
YFP	yellow fluorescent protein

Table of illustrations

Figures

Tables

Appendixes

1 Introduction

This thesis project is about a new gene therapy strategy for treating Huntington's disease. After 3 years, the project remains at an early stage of development and focuses on the molecular basis of *trans*-splicing. We only evaluated huntingtin introns using an artificial system adapted for molecular screening, thus excluding the pathological aspect of the Huntington's disease. In this context, Huntington's disease pathology was not the core of the thesis.

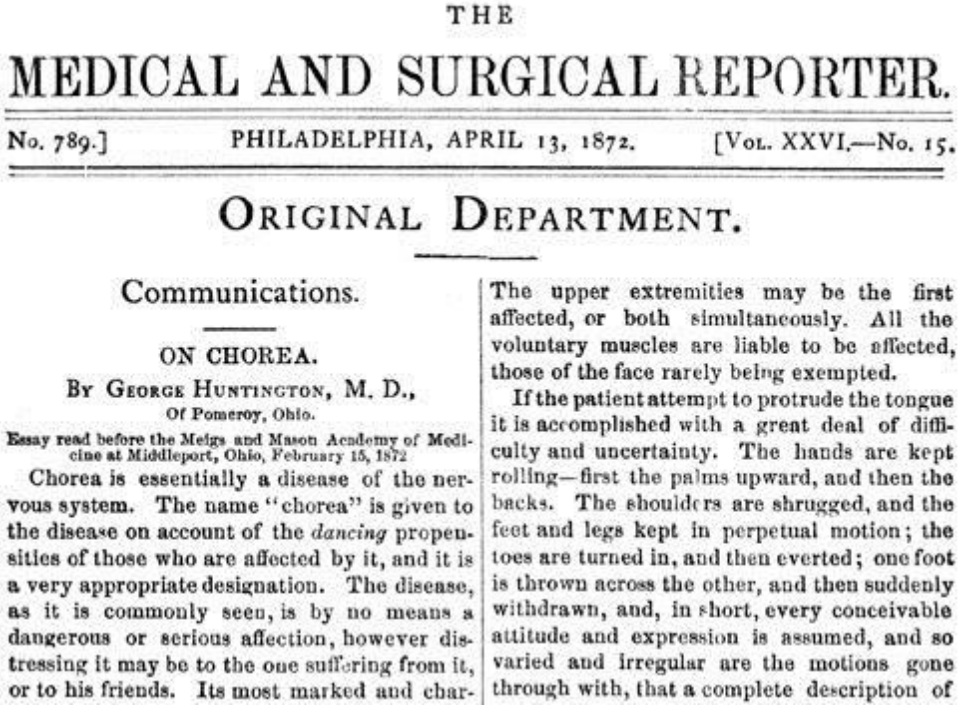
Thus, this introduction begins with only a short section about Huntington's disease, to provide contextual information about the purpose of this project with particular interest in therapies. Then, RNA biology will be presented to provide all the key elements for a comprehensive understanding of issues faced during the project time-course. The last part of the introduction will be dedicated to the *trans*-splicing process as an endogenous mechanism as well as a tool for gene therapy in genetic diseases such as Huntington's.

1.1 Huntington's disease

1.1.1 Physiopathology of HD

1.1.1.1 *History*

Huntington's disease (HD) was originally described in 1872 (Figure 1) by the American physician George Huntington (1850-1916). He depicted several aspects of HD including hereditary transmission and motor/cognitive issues. It is now well established that HD has an autosomal dominant genetic pathology, characterized by three kinds of symptom: motor dysfunction, cognitive decline and psychiatric disturbance (Figure 2).



THE

MEDICAL AND SURGICAL REPORTER.

No. 789.] PHILADELPHIA, APRIL 13, 1872. [Vol. XXVI.—No. 15.

ORIGINAL DEPARTMENT.

Communications.

ON CHOREA.

By George Huntington, M. D.,
Of Pomeroy, Ohio.

Essay read before the Meigs and Mason Academy of Medicine at Middleport, Ohio, February 15, 1872

Chorea is essentially a disease of the nervous system. The name "chorea" is given to the disease on account of the *dancing* propensities of those who are affected by it, and it is a very appropriate designation. The disease, as it is commonly seen, is by no means a dangerous or serious affection, however distressing it may be to the one suffering from it, or to his friends. Its most marked and characteristic feature is the involuntary character of the movements. The upper extremities may be the first affected, or both simultaneously. All the voluntary muscles are liable to be affected, those of the face rarely being exempted.

If the patient attempt to protrude the tongue it is accomplished with a great deal of difficulty and uncertainty. The hands are kept rolling—first the palms upward, and then the backs. The shoulders are shrugged, and the feet and legs kept in perpetual motion; the toes are turned in, and then everted; one foot is thrown across the other, and then suddenly withdrawn, and, in short, every conceivable attitude and expression is assumed, and so varied and irregular are the motions gone through with, that a complete description of

*Figure 1: **Beginning of George Huntington's original 1872 paper "On chorea"** in the Medical and Surgical Reporter. Extracted from Bates et al. 2005[1]*

It has been demonstrated that the pathology begins before clinical symptoms appear, so exhibiting a prodromal period. Some signs that can be motor (decreased movement speed and reaction times), cognitive (emotion recognition and learning impairments), or psychiatric (irritability, mood disorder) are often indicators of HD development. Day-to-day, these could be unnoticed or assumed to be personality changes due to aging, but in an HD family context the prodromal period is detected more easily and is the first warning of the beginning of the HD clinical phase.

Progressively, symptom severity increases as described in Figure 2 with mean disease onset around 40 years old and death about 15-20 years later. At late HD stages, HD patients lose all their autonomy and need strong healthcare for basic daily subsistence.

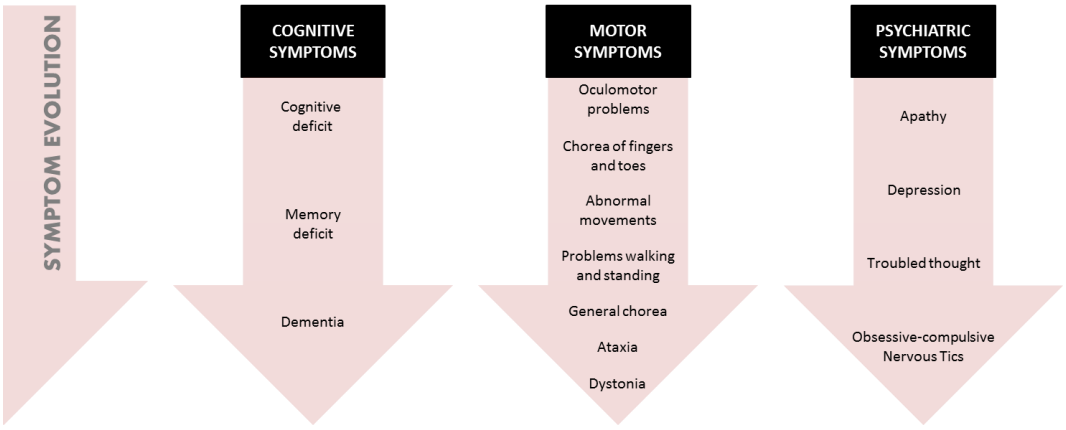


*Figure 2: **Clinical symptoms.** There are three categories of symptom in HD pathology: cognitive, motor and psychiatric. These symptoms evolve slowly, with onset around 40 years old and disease progression that usually lasts around 20 years, with complete loss of autonomy in late stages.*

1.1.1.2 *Prevalence*

HD is considered a rare disease with around 2.71 cases per 100,000 people worldwide[2], but this prevalence varies according to location (Figure 3). Disparities depend on ethnicity, with a high occurrence in populations of West European descent (>5 per 100 000) as shown by past immigration patterns. In contrast, very low occurrences are observed in China and Japan, with fewer than 0.5 per 100,000. According to Inserm, in France, 18,000 people carry the HD mutation, including 12,000 in the asymptomatic period[3].

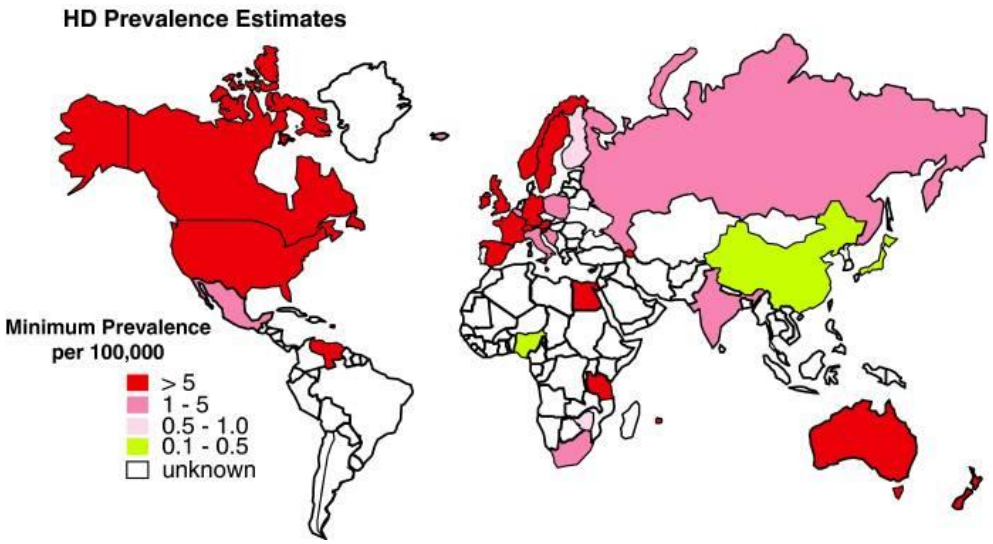


*Figure 3: **Worldwide estimates of the prevalence of HD**. Overall, the prevalence of HD is much higher in European populations than in East Asia. Note that prevalence studies occurring before the discovery of the HD gene in 1993 could underestimate the true prevalence of HD by as much as 14–24%[4,5]. In particular, many of the studies in Africa have small sample sizes and the HD diagnosis has not been confirmed by molecular testing. As HD phenocopy disorders are relatively common in Africa[6], these studies could have significantly overestimated the HD prevalence in these regions. Currently, the Maracaibo region of Venezuela has the highest reported worldwide prevalence of HD (700 per 100,000). Venezuela was colonized by the Spanish in the 16th century, and the origins of HD in Venezuela can be traced back to Europe[7]. Also see Harper et al.[8], Conneally et al.[9] and Al Jader et al.[10] for earlier reviews on the worldwide prevalence of HD. Cartography adapted from Warby et al 2011[11].*

1.1.1.3 *Neuropathology*

As a neurodegenerative disease, HD's main feature is the preferential and early vulnerability of a particular region in brain: the striatum. In humans, it is composed of two distinct structures, the caudate nucleus and the putamen. These two nuclei are a part of the basal ganglia, consisting of several nuclei depicted in Figure 4. They play a key role in initiation and coordination of normal voluntary movements with major involvement of striatal nuclei and an area of frontal cortex[12]. Thus, progression of striatal atrophy is related to worsening motor symptoms.

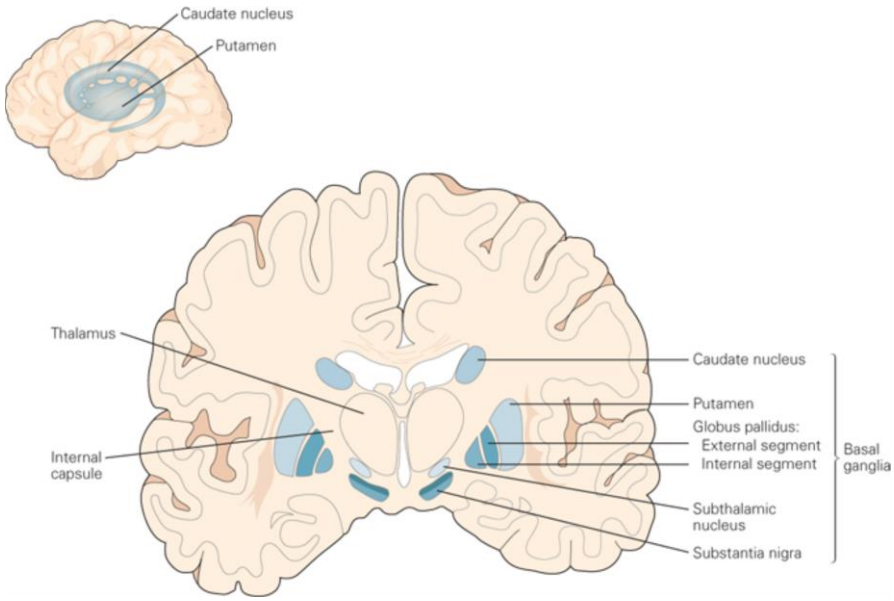


*Figure 4: **The basal ganglia and surrounding structures**. The nuclei of the basal ganglia are identified on the right in this coronal section. Adapted from Nieuwenhuys, Voogd, and Van Huijzen 1981[13]*

GABAergic medium spiny neurons of the striatum are the most vulnerable neuronal population in HD, with degeneration starting in the caudate and then in the putamen[14,15]. Since medium spiny neurons represent 95% of the striatum neuronal population, their degeneration is the direct cause of striatal atrophy. Thus, the striatum is the first and the most affected region compared to others like the neocortex, which also shows signs of degeneration during the course of the disease. Dilation of the lateral ventricles is also observed, in particular by MRI in advanced cases, as shown in Figure 5.

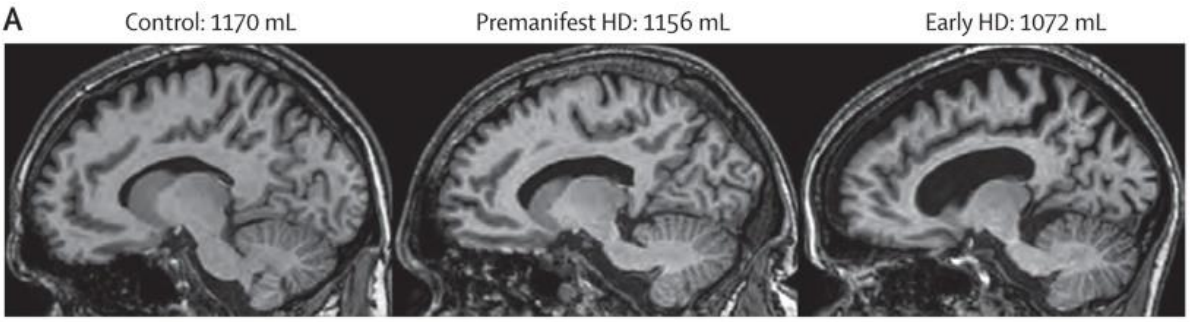


*Figure 5: **MRI sagittal sections of control, premanifest and early Huntington's disease human brains**, showing regional atrophy in HD. Pictures show 3T volumetric MRI scans of the whole brain in a 50-year-old control (left panel), a 55-year-old individual with preHD (middle panel), and a 49-year-old with early Huntington's disease (right panel). Brain volumes are corrected for intracranial volume. Extracted from Tabrizi et 2009[16]*

1.1.2 Huntingtin

HD is a genetic disease and results from a mutation in the Huntingtin gene (*HTT*).

1.1.2.1 *HTT description and HD mutation*

HTT, also called IT15 (for interesting transcript 15), was discovered in 1993 by a collective of researchers named the Huntington Disease Collaborative Research Group[17]. The gene is localized on the short arm of the chromosome 4 (4p16.3) in humans (Figure 6). *HTT* is highly conserved across vertebrates[18], in particular in mammals in which:

- the first 17 amino acids are commonly conserved
- a CAG repeat region in exon 1, encoding for a polyglutamine stretch (PolyQ), is also systematically found with a varying number of repeats.

The HD mutation is precisely located within these CAG repeats, and corresponds to an expansion of the number of repeats. In healthy subjects, this does not go beyond 36, whereas in HD subjects it is superior to 40 (Figure 6). This repeat number is the cut-off for which the development of HD is systematic in patients.

Between 36 and 40, a high variability is observed among subjects. In this case, symptoms start in later life and so HD is less likely to be diagnosed because of an extensive prodromal period. To summarize, HD is a pure genetic disorder, originating from the CAG repeat number in exon 1 of the *HTT* gene.

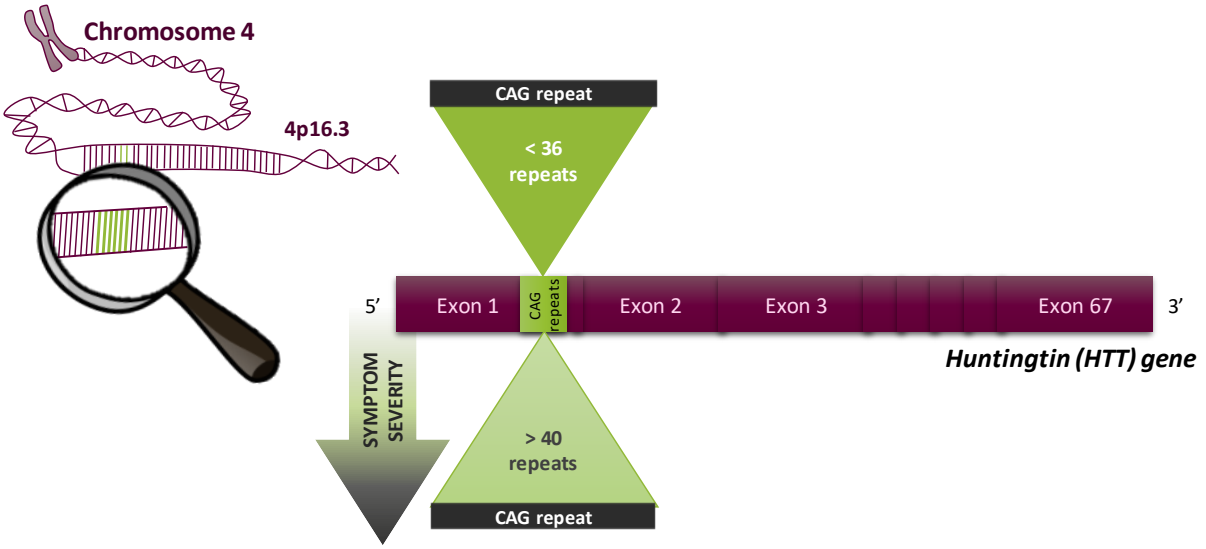


*Figure 6: **Schematic representation of the HD mutation in the HTT gene.** Purple boxes represent exon sequences. The green box corresponds to the CAG repeat motif in exon 1. The upper black box represents repeat number in healthy cases and the lower black box in pathological cases.*

HTT is a large gene, composed of 67 exons encoding a 348 kDa protein named Huntingtin (HTT). Besides the fact that HTT is ubiquitously expressed, its expression level depends on developmental stage and cell type, with a higher level in the brain as depicted below in Figure 7 and Figure 8, respectively.

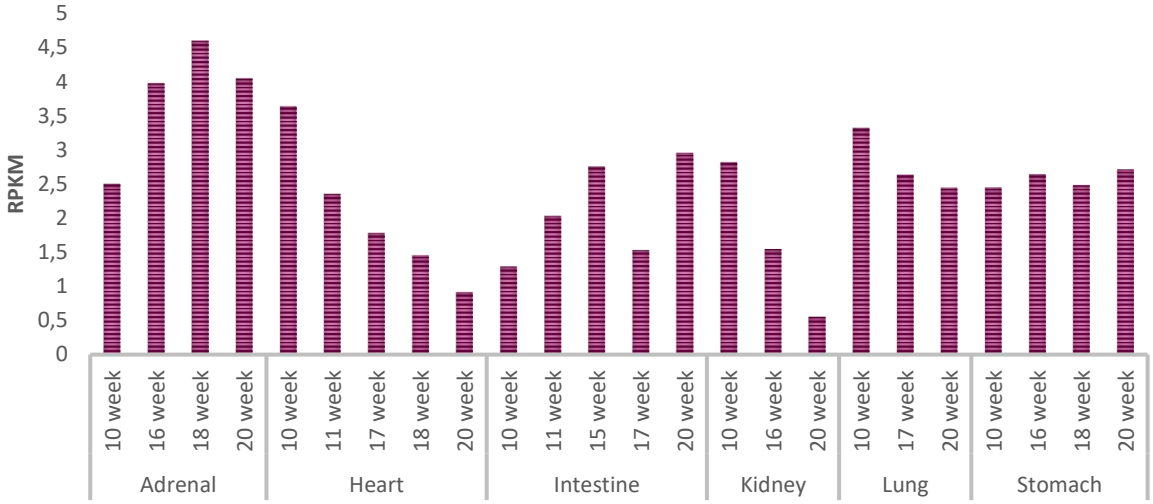


Figure 7: ***Variation of HTT expression level in different tissues across development.*** *35 human fetal samples from 6 tissues (3 - 7 replicates per tissue) collected between 10 and 20 weeks gestation time were sequenced using Illumina TruSeq Stranded Total RNA. BioProject: PRJNA270632 Project title: Tissue-specific circular RNA induction during human fetal development. Adapted from Szabo et al 2015[19], datasets available online from the NCBI at https://www.ncbi.nlm.nih.gov/gene/3064 (HTT ID:3064)*

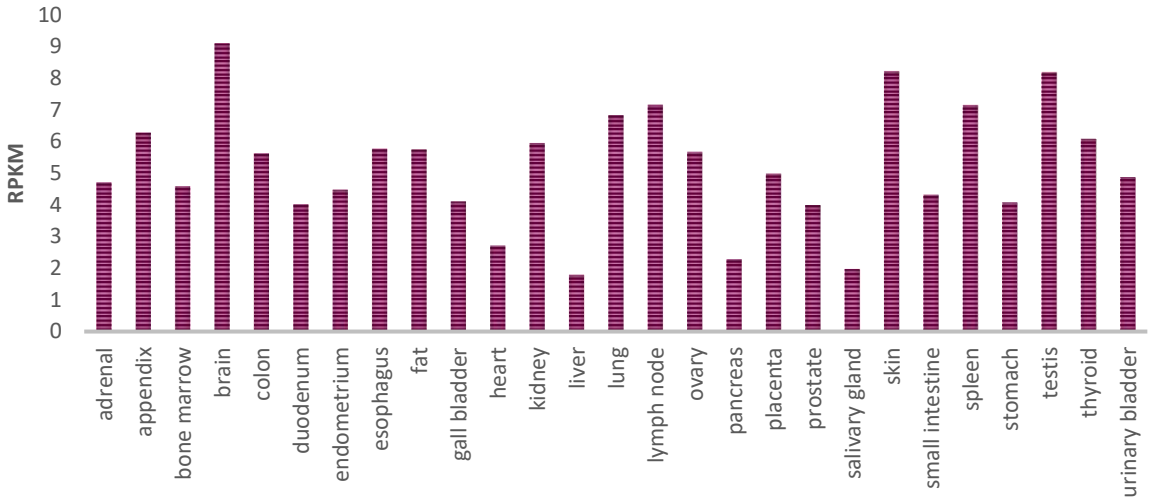


Figure 8: ***Variation of HTT expression level among tissues.*** *RNA-seq was performed on tissue samples from 95 human individuals representing 27 different tissues in order to determine tissue-specificity of all protein-coding genes. BioProject: PRJEB4337 entitled HPA RNA-seq normal tissues. Adapted from Fagerberg et al, 2014[20], datasets available online from the NCBI at https://www.ncbi.nlm.nih.gov/gene/3064 (HTT ID:3064)*

Moreover, HTT is a multifunctional protein that exhibits several subdomains (listed next in purple) and post-translational modification sites (listed next in green) with distinct functions, suggesting a larger spectrum of actions in cells (Figure 9).

- **Polyglutamine domain (PolyQ)**: binds to transcription factors thanks to its capacity to form a polar zipper structure[21].
- **Polyproline domain (PolyP)**: stabilizes the polyglutamine tract[22]
- **Three HEAT (Huntingtin, Elongation factor 3, protein phosphatase 2A, and TOR 1) repeats:** act on protein-protein interactions[23,24].
- **NES domain:** nuclear export signal.
- **NLS domain:** nuclear localization signal.
- **Caspase and calpain cleavage sites:** involved in proteolysis of HTT itself[25].
- **Ubiquitination sites:** control of stability, function and intracellular localization resulting in maintenance of HTT homeostasis in cells[26].
- **Sumoylation sites**: regulate HTT localization, activity and stability[22].
- **Phosphorylation sites**: regulate HTT clearance[27], aggregate formation[28] and vesicle transport[29,30].
- **Palmitoylation sites**: regulate vesicular trafficking and synaptic vesicle function[31]
- **Acetylation sites**: involved in macroautophagy[32], protein-protein interaction, stabilization and turn-over[33].

More detailed information is given in the review by Zuccato, Valenza, and Cattaneo (2010)[34] and the book by Bates, Tabrizi and Jones (2014)[35].

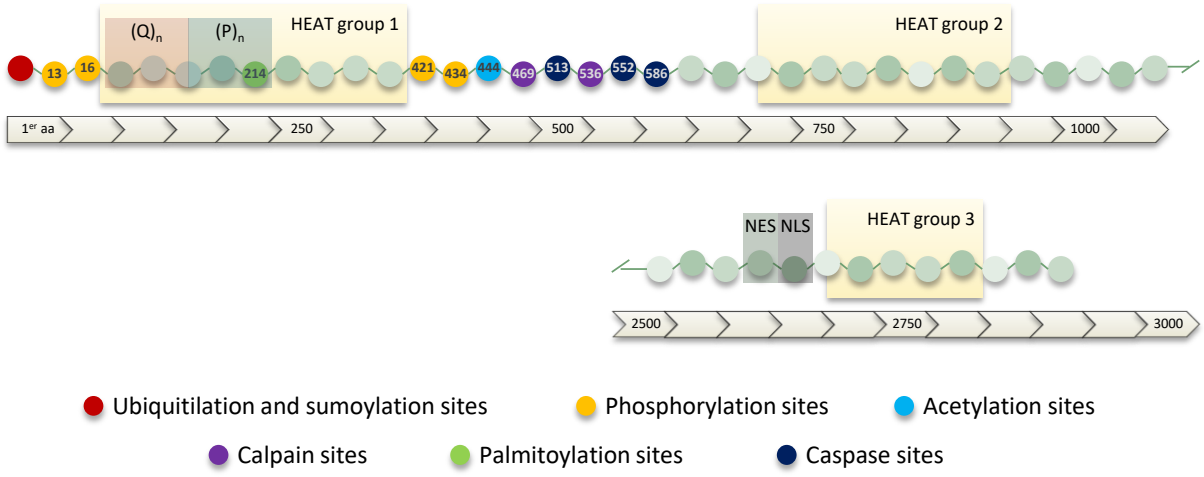


*Figure 9: **Schematic representation of HTT protein functional domains.** Linked bubbles represent amino acid chains, with the chevrons underneath giving a scale of relative position. Colored boxes represent functional domains: $(Q)_n$ for polyglutamine, $(P)_n$ for polyproline, HEAT for Huntingtin, Elongation factor 3, protein phosphatase 2A, and TOR 1, NES for Nuclear Export Signal, NLS for Nuclear Localization Signal. Adapted from Zuccato et al (2010)[34]*

CAG repeats code for the polyQ domains, and in the case of an HD mutation, lead to an elongated stretch of glutamine in the N-terminal part of mutant HTT protein (mHTT)[36–39]. It is possible to measure the precise CAG repeat length in an individual, and it has been demonstrated that CAG repeats can increase through CAG tract instability during spermatogenesis, thus limiting these events to paternal transmission[40]. This instability is probably associated with hairpin loop formation and is likely to be driven by CAG repeat length. When CAG tract has pathologic length with particular *HTT* polymorphisms into the adjacent CCG repeats (PolyP domain), this is likely to exhibit greater instability[41].

In addition to CAG tract instability, the repeat length is also correlated with the age of onset and the severity of disease. As depicted in Figure 10, the larger the CAG repeats are, the earlier the HD symptoms appear and thus the faster the pathology progresses. This aspect is observed in HD patients as well as in cells in tissue cultures carrying the HD mutation, with pathological changes such as mHTT aggregation or vesicular transport impairment dependent on CAG repeat length, suggesting an intrinsic dysregulation tightly linked to mHTT expression.

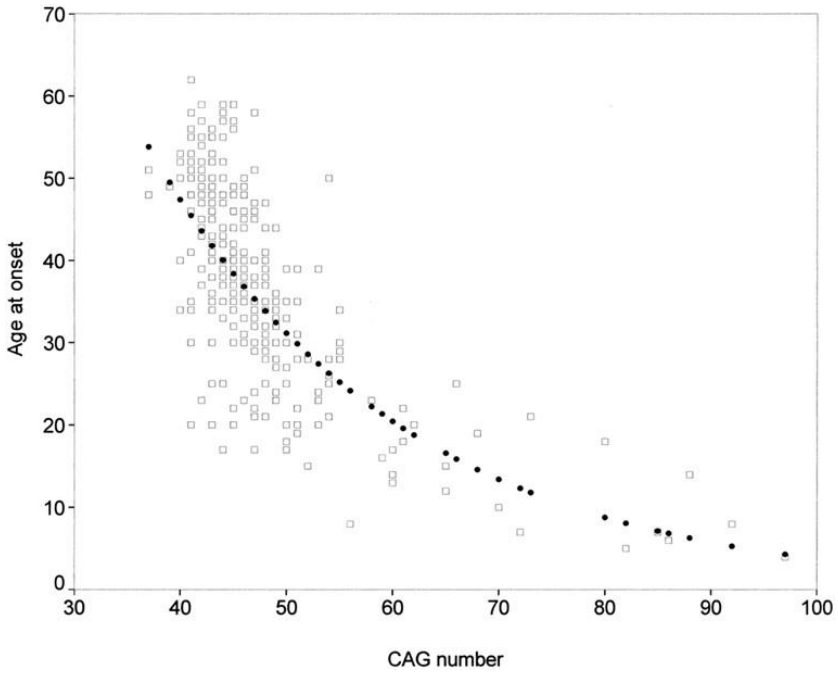


Figure 10: ***Scatterplot of age at onset versus CAG repeat number*** *in 319 HD patients born after 1930 (open squares). The expected values under an exponential regression are also reported as filled circles. Extracted from Squitieri et al., 2000*[42]

Indeed, this elongation results in changes in protein conformation, with a misfolding of mHTT, thus causing simultaneously a toxic gain of function[43,44] and loss of normal function[45], with dramatic consequences in neurons. The normal functions of HTT remain unclear but it is now known that it is involved in:

- vesicle transport
- gene transcription regulation
- protein synthesis regulation
- RNA trafficking
- autophagy
- embryonic development
- brain development

Some of these functions are impaired in HD, resulting from the loss of HTT function. For instance, when mHTT is expressed, vesicle-mediated BDNF transport is reduced[29,30], mitotic spindle orientation is impaired[46], and the autophagy process is inhibited. These pathological features, as with many other cellular impairments, impact neuron survival[47].

1.1.2.2 *Toxic N-terminal fragment of mHTT*

However, toxic gain of function is known to be mediated by N-terminal fragments of mHTT (N-mHTT). N-mHTTs are generated after cleavage at caspase sites (① on Figure 11) but recent studies also suggest direct production of an mHTT fragment resulting from alternative splicing[48] (② on Figure 11). N-mHTT contains the polyQ domain and it has been demonstrated that N-mHTT is sufficient to induce the main HD characteristics in animal models, suggesting that N-mHTT is more toxic than the full-length protein[49]. N-mHTTs are found in soluble form and in aggregate form with intranuclear and cytoplasmic inclusions. They interact and enclose physiological partners, thus preventing normal signaling (⑥ on Figure 11) such as:

- transcriptional alterations
- metabolic dysfunctions
- proteasome impairment
- stress response abnormalities

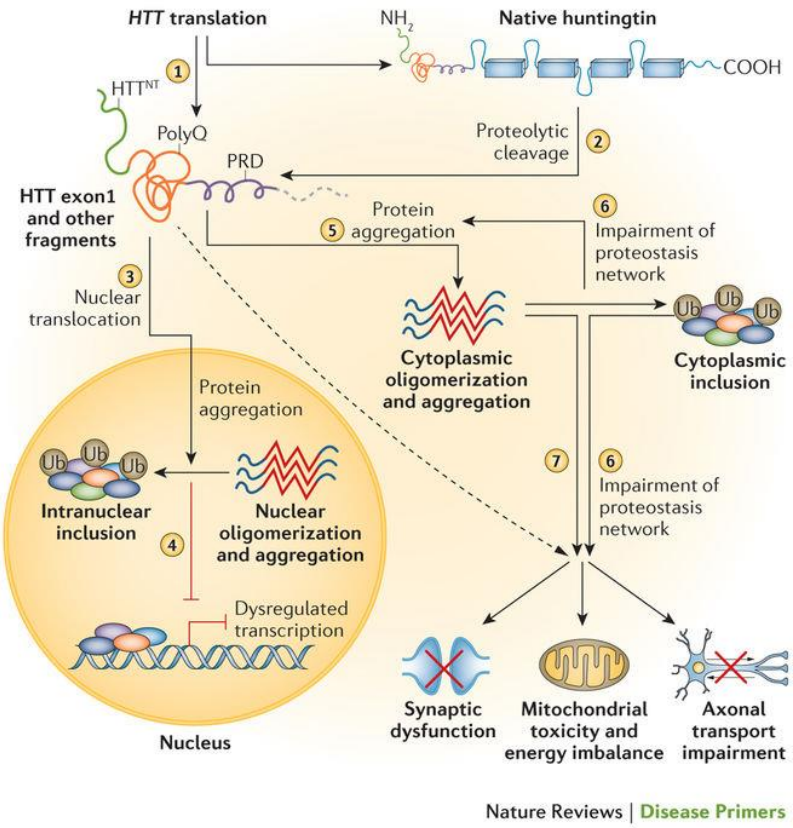


*Figure 11: **Schematic representation of pathogenic cellular mechanisms in Huntington's disease**. (1) HTT is translated to produce the full-length huntingtin protein as well as an N-terminal exon1 HTT fragment (the result of aberrant splicing). The length of the polyQ tract in these proteins depends on the extent of somatic instability. (2) Full-length native huntingtin is cleaved through proteolysis to generate additional protein fragments. (3) Protein fragments enter the nucleus. (4) Fragments are retained in the nucleus through self-association, oligomerization, and aggregation—leading to the formation of inclusions, a process that causes transcriptional dysregulation through the sequestration of other proteins as well as other incompletely defined mechanisms. (5) Huntingtin fragments oligomerize and aggregate in the cytoplasm. (6) The aggregation of huntingtin is exacerbated through the disease-related impairment of the proteostasis network, which also leads to global cellular impairments. (7) The aberrant forms of huntingtin result in additional global cellular impairments including synaptic dysfunction, mitochondrial toxicity, and a decreased rate of axonal transport. Abbreviations: polyQ, polyglutamine; PRD, proline-rich domain; Ub, ubiquitin. Extracted from Bates et al 2005[1]*

Regarding the plethora of cellular impairments that have been identified, it seems complicated to determine a precise target to reverse defects generated by HD mutation. However, without knowing all these mechanisms in detail, different therapeutic possibilities have been hypothesized. Beyond the symptomatic treatment of psychiatric or motor symptoms, some of these strategies, known as disease modifiers, aim to promote the survival of GABAergic neurons in the striatum. Moreover, since the disease is purely genetic, it is also tempting to intervene even further upstream in the pathogenic process by acting directly on the mutation. In the next part of this introduction, we will focus on the different therapeutic strategies that have been tested or are still under development for HD.

1.2 Therapeutic approaches for HD

1.2.1 HD models

1.2.1.1 *Animal models*

The development of efficient therapeutic strategies involves adequate animal models to mimic the disease. A wide variety of HD models have been developed, including vertebrate and invertebrate models like *Drosophila melanogaster*[50] and *Caenorhabditis elegans*[51,52], since *HTT* is highly conserved across species. These models have allowed easier manipulation compared to vertebrates and thus enhanced the understanding of HTT normal function and some aspects of HD.

Large animal models of HD have also been studied including transgenic HD rhesus monkeys[53], and pigs[54] expressing N-mHTT. In this case, mHTT aggregates are found extensively in neurons and in particular in nuclei and neuronal processes, equivalent to HD patient brains. Other models with larger HTT fragments have likewise been studied in pigs[55] and sheep[56], demonstrating a milder phenotype. Overall issues with these kinds of model are related to fertility and life-span which are both highly reduced.

For therapeutic development, the most common animal models are rodents because they have a similar nervous system organization to humans, and they are small animals with easy and fast reproduction as well. Regarding disease relevance, most transgenic mouse models demonstrate HD features such as aggregation, nuclear inclusion, and motor deficits, with more or less a degree of severity as listed in Table 1.

A new rat model named BACHD, expressing the full length human mHTT gene, including exons, introns and 97 CAG repeats, was generated by the team of Hoa Huu Phuc Nguyen in 2012[57]. This model exhibits motor deficits and also presents cognitive deficiency, cortex degeneration as well as mHTT aggregates [57,58]. Recently, gene expression impairments were also demonstrated, but only 39% of gene expression changes were similar to human HD samples, though differences are possibly due to the severe progression of post-mortem human samples. This model will be further detailed in the perspective chapter.

In parallel to transgenic animal models, other strategies are based on:

- **Excitotoxicity:** injection of toxins to reproduce neurodegeneration in the animal brain. Quinolinic, ibotenate, kainic and 3-nitropropionic acids have demonstrated their capacity to generate HD models when injected by inducing striatal lesions[59–63].
- **Gene transfer**: the whole *mHTT* gene cannot be introduced into a viral vector since it would exceed vector capacity. However, the use of viral vectors expressing N-mHTT with varying CAG repeats (from 44 to 82) has demonstrated its relevance using lentiviral mediated delivery in rodents[64,65]. In these models, animals develop HD striatal neuropathology including aggregates and degeneration. Similar studies with N-mHTT (82 CAG repeats) have been done in non-human primates[66], with HD behavioral characteristics such as spontaneous dyskinesia appearing 15 weeks after injection.

To conclude, a lot of HD animal models are available, but no one reproduces all HD features as observed in patients due to limited timeframe of animal model to develop HD phenotypes. Few animal models express the full length mHTT gene, thus reducing the construct validity and the number of potential targets at DNA and RNA levels that could be tested.

As mentioned at the beginning of this introduction, this project is at an early stage, and before assessing our gene therapy strategy *in vivo* in a relevant animal model (See in Perspectives page 151), we used an *in vitro* HD model, more suited to validate a preliminary proof of concept.

Table 1: ***Transgenic mouse models of HD***

Model	Transgene	Promotor	Lifespan	Disease severity	Neuropathological features	References
R6/1	Human *HTT* exon 1 R6/1: 116Q	Human *HTT*	8 months	Motor deficit at 5 months	Nuclear inclusions and HTT aggregates at 5 months	[67]
R6/2	Human *HTT* exon 1 R6/2: 144Q	Human *HTT*	13 weeks	Juvenile-onset HD Motor deficit at 5-6 weeks	Nuclear inclusions and HTT aggregates at 4 weeks	
N171-82Q	N-terminal human *HTT* (first 171 amino acids) 82Q	Mouse prion	4-6 months	Motor deficit at 12 weeks	Age-dependent aggregates Neuronal loss at 20 weeks	[68,69]
BACHD	Full length chimeric human *HTT* 97Q	Human *HTT*	Normal	Adult onset HD Motor deficit at 24 weeks	Striatal and cortical atrophy (~28% and ~32% respectively) at 48 weeks	[70]
Conditional BACHD	Inducible using Cre/LoxP system					[71]
YAC128	Full length chimeric human *HTT* 128Q	Human *HTT*	Normal	Adult onset HD Motor deficit at 24 weeks	Striatal and cortical atrophy (~12% and ~8% respectively) at 48 weeks	[72,73]
GFAP-HD	N-terminal human *HTT* 160Q	Human GFAP	Earlier than WT	Motor function deficit		[74]
Hu97/18 **Hu128/21**	No mouse *HTT* Double human HTT allele					[75,76]

Table 2: ***Knock-in mouse models***

Model	Transgene	Promotor	Lifespan	Disease severity	Neuropathological features	References
Knock-in mouse models	Endogen mouse *HTT* with expanded CAG repeats 111, 140, 150 and 190Q	Mouse HTT	Normal	Late onset Variable motor function deficit	HTT aggregates in striatal neurons Nuclear inclusion CAG repeats dependent	[77–81]

1.2.1.2 *Cellular models*

Parallel to the use of animal models (limited by their biological relevance and ethics considerations given animal experimentation issues), *in vitro* experiments are, up to a certain point, fully suitable for studying the physiopathology of the disease, as well as for assessment of therapeutic benefit.

The easiest cellular models to work with are **cell lines** made immortal with continuous cell division. They have limited relevance in terms of HD mimicking, but are suitable for primary biological process validation. One of the most common cell lines is HEK293, frequently used in neuroscience field and screening studies. This cell line is derived by transformation of primary cultures of human embryonic kidney, but some data suggests neuronal lineage, with positive immunostaining for neurofilament (NF) subunits [82] and comparable levels of HTT expression compared to other brain cell lines[83].

Primary cultured cells are rodent cells put in culture at an early stage of development exhibiting particular characteristics of a given cell type. This is a useful strategy to work with endogenous murine or human mHTT. Moreover, it allows discrimination of specific cell type effects on the disease by specific culture of neurons[84,85], astrocytes[86] or microglia[87] for instance, and is suitable for gene therapy assessment[88]. However, it is more complicated to implement, since these cells are more vulnerable to death and it requires particular skills to isolate specific cell types during dissection and to optimize cell culture conditions.

Induced pluripotent stem cells (iPSC) are cells extracted from HD patient fibroblasts and reprogrammed into self-renewing pluripotent cells and differentiated into specific neuronal lineages such as, for example, medium spiny neurons [89] (Figure 12). These cells carry complete human genomes with *in situ* HD mutations. They are able to exhibit particular aspects of medium spiny neurons such as DARP-32 as well as cellular impairments mediated by HD mutations like gene expression[90] and proteasome activity[91] alterations. Still, protocols to generate iPSCs are complex with specific transcription factor treatments including Oct3/4, Sox2, Klf4, and c-Myc[92], and differentiation can be very challenging due to HD mutation effects on neurogenesis. However, it is possible to suspend differentiation in an intermediate state between iPSC and a given cell type. These intermediate cells are called progenitors and they can be stored to facilitate experiment time courses. In this condition, these cells are also appropriate for screening uses as well as for therapy purposes [93].

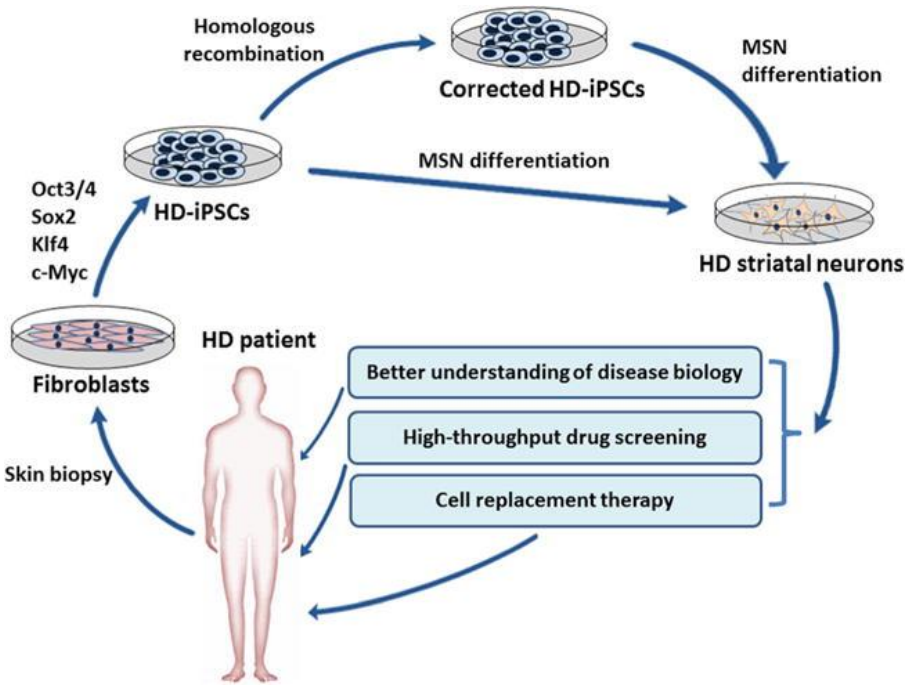


*Figure 12: **The generation and application of iPSCs in HD research.** HD patient-specific iPSCs can be obtained by reprogramming of skin fibroblasts. Established iPSCs can be used as a tool for better understanding the molecular basis of HD. iPSC technology can also be coupled with high-throughput screening that provides a more efficacious platform to assess novel drug candidates aimed at stopping or slowing the disease process. Moreover, HD-iPSCs can be differentiated into specific cell types predominantly affected in the disease (striatal MSNs). Emerging gene therapies make the genetic correction of HD-iPSCs feasible, paving the way for autologous transplantation strategies of healthy iPSCs or iPSC-derived neural cells. Extracted from Lui et al 2016[89]*

We have listed many models available for HD that have been extensively used to identify potential therapeutic targets. However, no curative treatment is available for patients to date and the next part reviews previously-explored strategies for HD therapeutic development.

1.2.2 Symptomatic treatments

From the beginning of contemporary therapeutic strategy, most active compounds were provided by the chemical industry. This traditional process of drug discovery is based on the identification of active compounds, from large libraries of chemicals, with a given biological activity. Then, their potential therapeutic efficacy is assessed in different diseases, exhibiting similar symptoms, and related to the biological activity criteria defined for the screening.

In the context of HD, two kinds of drugs have emerged depending on whether they are disease modifiers or not. Non-disease modifying drugs aim to improve patient quality of life by targeting essentially psychiatric and motor symptoms. However, only a few are specifically dedicated to HD, including tetrabenazine and its derivatives, acting on chorea. Other classical anti-psychotic drugs are also prescribed but at the discretion of the clinician (Table 3). Most of these drugs have not been

assessed in rationally-designed clinical studies (double-blind, placebo-controlled), but can be useful pain relief for HD patients, who face a complete lack of alternative treatment.

*Table 3 : **Most frequently prescribed drugs for chorea in Huntington's disease (HD)**. Adapted from Coppen et al, 2017[94]*

Drug	Adverse effects
Tetrabenazine	Depression, somnolence, parkinsonism, insomnia, akathisia, anxiety, nausea
Olanzapine	Sedation, weight gain, dry mouth, Parkinsonism
Risperidone	Parkinsonism, akathisia, sedation, hyperprolactinemia
Tiapride[a]	Sedation, parkinsonism
Quetiapine	Weight gain, dry mouth, parkinsonism, sedation, akathisia
Aripiprazole	Sedation, parkinsonism, akathisia, cardiac arrhythmias
Clozapine	Orthostatic hypotension, sedation, weight gain, increased seizure risk, agranulocytosis
Haloperidol	Tardive dyskinesia, sedation, parkinsonism, akathisia, tachycardia
Sulpiride	Sedation, parkinsonism, hyperprolactinemia, akathisia, weight gain
Amantadine	Insomnia, hallucinations, anxiety, agitation, cardiac arrhythmias, dry mouth

Drugs listed in this table are the most commonly prescribed drugs for the treatment of chorea based on international surveys among HD experts and registered HD patients (Priller et al.[95], Orth et al. [96], Burgunder et al.[97]). Most commonly reported adverse effects are described (Brown et al. [98], Videnovic[99])

[a]Tiapride is only available in European countries

In parallel, disease modifier medications, acting on particular aspects of HD cellular impairments, and relying on a better comprehension of HD physiopathology, raised promising results in animal models. Indeed, from a pre-clinical point of view, most cellular processes impaired by HD can be targeted, but from a clinical point of view, seventy-four clinical studies targeting HD around the world are referenced by the US library of medicine (clinicaltrials.gov) to assess drugs for HD at different stages, including twelve on going at the time of writing these words. Despite the large range of therapeutic targets assessed and as many associated drugs, most of these trials failed to reach therapeutic efficacy.

The main reasons for failure were limited number of HD patients available, short treatment duration, study design, and the outcome measurement. Moreover, pharmaceutical companies are less motivated to conduct clinical trials to assess drugs without patent rights[94], thus reducing pools of potential medication.

1.2.3 Advanced therapy medicinal products

The pharmaceutical industry is undergoing deep change. The end of patent validity, associated with biotechnology progress, motivates therapeutic strategy changes favoring advanced therapy medicinal products, covering:

- Somatic cell therapy medicinal products
- Tissue engineered products
- Gene therapy medicinal products
- Combined advanced therapy medicinal products.

In an HD context, where pathology is complex with multiple cellular defects, concomitantly with no available traditional drugs, these emerging medication categories could be an important source of hope to develop a curative treatment.

1.2.3.1 *Cell-replacement therapy*

For a somatic cell therapy medicinal product, the most promising strategy is cell-replacement therapy, since one of the main characteristics of HD is the specific loss of striatal medium spiny neurons. In particular, loss of striatal projection neurons occurs at the earliest pathological stages[100,101]. Providing normal functioning cells expressing normal HTT is thus likely to produce an important therapeutic benefit.

The main challenge of this strategy is to obtain relevant cells to introduce into the brain. Two main sources have been used:

- Fetal cells
- iPSCs

Fetal cells were the first to be used, with very promising results in rodents and non-human primates, whereas results in HD patients were more variable. Although many patients exhibited motor improvement and better striatal metabolism, this was not durable, with progressive deterioration 4 to 6 years after surgery[102]. So, the use of fetal cells is limited by the graft quality, inducing varying degrees of vascularization, which is believed to impact therapy success. Moreover, the amount of available fetal tissue is restricted, quality control is limited because of tissue viability, and there is a high risk of graft rejection[103] and tau pathology development[104].

In this context, the use of iPSCs has been favored since it is feasible to amplify iPSCs *in vitro* and to generate cell line banks using GMP (Good Manufacturing Practice) protocols. Additionally, technological advances now allow the production of HD patient-derived iPSCs, reducing immunogenicity problems if autograft is performed. From fibroblasts, cells can be corrected and reprogrammed as either neuronal stem cells (NSC) or mesenchymal stem cells. Although a very promising strategy, improvements still need to be made to improve the differentiation protocol, regulate graft overgrowth, and ensure accurate progenitor cell fate once injected into the brain[103].

1.2.3.2 *Gene therapy*

Gene therapy can be divided into three strategies:

- Disease modification
- Gene silencing
- Genome editing

The first category is very close to a symptomatic strategy, since it acts as a disease-slowing treatment by impacting particular aspects of the disease. A principal technique is to supplement the brain with trophic factors[105] which are reduced in HD, promoting neuronal loss.

BDNF[106–108] (Brain-Derived Neurotrophic factor) is a potential therapeutic agent. When overexpressed, it induces improvement of HD phenotypes in mice, but it is also associated with adverse effects such as weight loss and seizure activity. Absolute level BDNF is critical for adequate therapeutic benefit which also depends on side effect and delivery.

The use of **CNTF**[109] (Ciliary neurotrophic factor) has been developed up to the clinical trial stage, but is limited by the reliability of the delivery system. Indeed, CNTF is introduced in Baby Hamster Kidney (BHK) cell capsules exhibiting variable CNTF release; 6 months after, only 40% of capsules are still active, while in other cases, capsule content is deteriorated[110,111].

GDNF (Glial-cell-line-derived neurotrophic factor) and **NGF** (nerve growth factor) are also therapeutic candidates. Multiple studies have been undertaken with fibroblast-mediated expression[112,113] with promising results in rat models.

In parallel, **cholesterol metabolism** has also been targeted since its synthesis pathway is very specific to the brain and is known to be reduced in neurodegeneration. In mice, overexpression of the enzyme CYP46A1 via an AAV vector restores normal cholesterol levels, inducing motor deficit improvements and striatal atrophy decrease[114].

Since HD genetically is characterized by complete penetrance, a curative treatment is unlikely to emerge from these kinds of strategies. Gene silencing and editing, however, could be more consistent since they target the source of pathology instead of targeting specific altered mechanisms that are poorly understood.

It has been demonstrated that HD disease pathology is mediated by mHTT, so a rational strategy is to reduce expression of this toxic protein by targeting messenger RNA. This process is called "gene silencing" and is the most promising approach to date. Two kinds of technologies have been developed including RNAi (interfering RNA) and ASO (antisense oligonucleotides), both depicted in Figure 13. The most common RNAi strategy uses siRNA (short interfering RNA), and shRNA (small-hairpin RNA) derived from miRNA (microRNA) cellular machinery (See Non-coding RNAs page 52). ASOs are modified DNA molecules, using RNAse H degradation or merely protein translation inhibition mediated by the RNA binding.

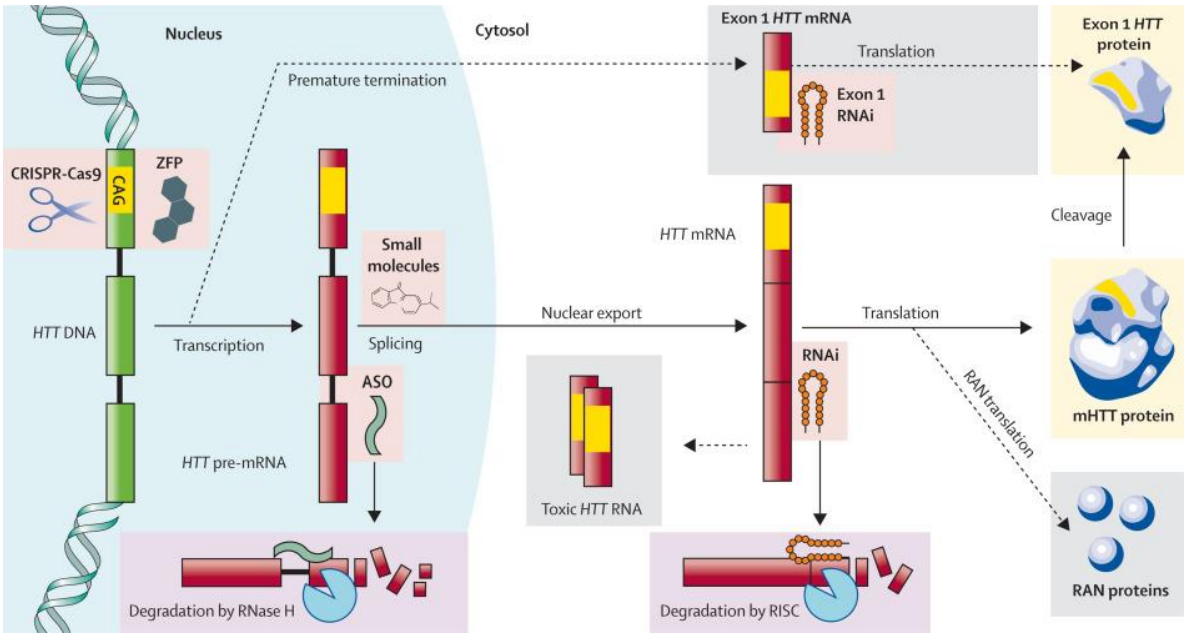


Figure 13: ***Production of huntingtin protein (HTT) and targeted molecular therapies in development to reduce its expression.*** *Yellow sections of the DNA, RNA, and protein represent the pathogenic expanded CAG tract and its polyglutamine product. Therapeutic approaches are indicated with pink boxes. Yellow boxes indicate the most widely accepted toxic species. Dotted arrows and grey boxes indicate proposed non-traditional mechanisms for the production of toxic fragments. The chief mechanisms of action of ASOs and RNAi compounds are shown at the bottom of the figure in purple boxes. CRISPR-Cas9=clustered regularly interspaced short palindromic repeats–CRISPR-associated system. ZFP=zinc finger protein. ASO=antisense oligonucleotide. RNAi=RNA interference compound. RISC=RNA-induced silencing complex. RAN=repeat-associated non-ATG. Extracted from Wild & Tabrizi, 2017*[115]

The challenge of these technologies is to perform allele-specific silencing, since most HD patients are heterozygous. Specific allele targeting, likely to be safer, has been developed, but non-allele specific silencing also exhibited an efficient decrease of mHTT, with limited side effects. The non-allele specific strategy is now in a clinical trial as IONIS-HTT_{RX} (ClinicalTrials.gov; NCT02519036, Appendix 7). This is the first therapy targeting the cause of HD to reach that far in the therapy development process. In a brief communication to the Huntington's community, Roche (Indianapolis, IN), involved in the trial, claimed that the highest dose of IONIS-HTT_{RX} induces the largest reduction of mHTT and its safety and its tolerability appear to be validated[116]. Extensive results will be available in first half of 2018.

Gene silencing also aims to reduce mHTT in removing the HD mutation at DNA level. Two major approaches have been developed including (Figure 13):

- ZFP (Zinc Finger Protein) method
- CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology

ZFPs are transcription factors used to repress gene expression through the recognition of specific DNA motifs by a modified endonuclease. In an HD context, it leads to a selective decrease of mHTT, and HD symptom improvements in mice[117]. Recently, the CRISPR system, an easy and powerful tool to generate driven genomic DNA breaks, has emerged and rapidly spread in use. Indeed, CRISPR is a bacterial mechanism, modified from its original goal of protection against viral DNA. It requires two components: a guide RNA (gRNA) and an endonuclease, the most commonly used being Cas9 (CRISPR associated protein 9) from *Streptococcus Pyogenes*.

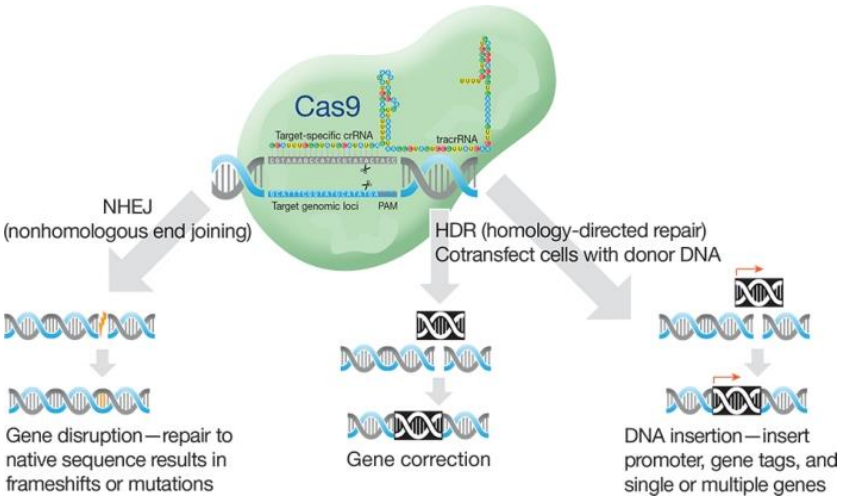


*Figure 14: **A CRISPR-Cas9 targeted double-strand break**. Cas9-mediated cleavage occurs on both strands of the DNA, three base pairs upstream of the NGG proto-spacer adjacent motif (PAM) sequence on the 3' end of the target sequence. The specificity is supplied by the guide RNA (gRNA), and changing the target only requires a change in the design of the sequence encoding the gRNA. After the gRNA unit has guided the Cas9 nuclease to a specific genomic locus, the Cas9 protein induces a double-strand break at the specific genomic target sequence. Adapted from https://www.thermofisher.com.* [118]

The cleavage specificity is driven by the gRNA which recognizes a particular sequence called PAM for Protospacer-Adjacent motif, upstream to the target sequence. Then, Cas9 is recruited, the target sequence is cleaved, and different processes take place, depicted in Figure 14:

- non-homologous end joining (NHEJ)
- homology-directed repair (HDR)

We will focus on NHEJ mechanisms since no HDR data are yet available in an HD context.

To inactivate mHTT, several strategies have been developed including the use of two gRNAs to remove a large region of *HTT* DNA, showing very promising results [119–121]. However, both mutant and normal alleles are targeted and so both HTT protein levels are reduced. Another issue of CRISPR-Cas9 is a potential side effect induced by off-target DNA breaks. When CRISPR-Cas9 is introduced in the brain, it is continuously expressed. Once the removal of the mutated portion is achieved, Cas9 endonuclease is still active leading to potential adverse effects from off target cleavage. So the main concern now is to solve this issue. A recent study led by the team of Nicole Déglon validated a new strategy to reduce over-CRISPR-Cas9 activation using two gRNAs depicted in the bottom of Figure 15:

- One targeting a sequence on the *HTT* gene, close to the translation start site (sgHTT)
- One targeting the ATG of Cas9 (sgCa9).

Both sgRNAs are designed to block translation, but each one is under control of a different RNA Pol III promoter to regulate its time-course expression, with a weak 7SK promoter for sgCas9 and a stronger H1 promoter for sgHTT. Using this system, Cas9 expression is transient, with a peak of expression around 1 week after treatment, which is synchronized to the expression level of sgHTT, thus allowing efficient DNA breaks in the *HTT* gene and a significant decrease of huntingtin level (Figure 15, top). In parallel, a lower level of sgCas9 is enough to drastically eliminate Cas9 as early as 2 weeks after treatment, thus quickly reducing potential off-target breaks.

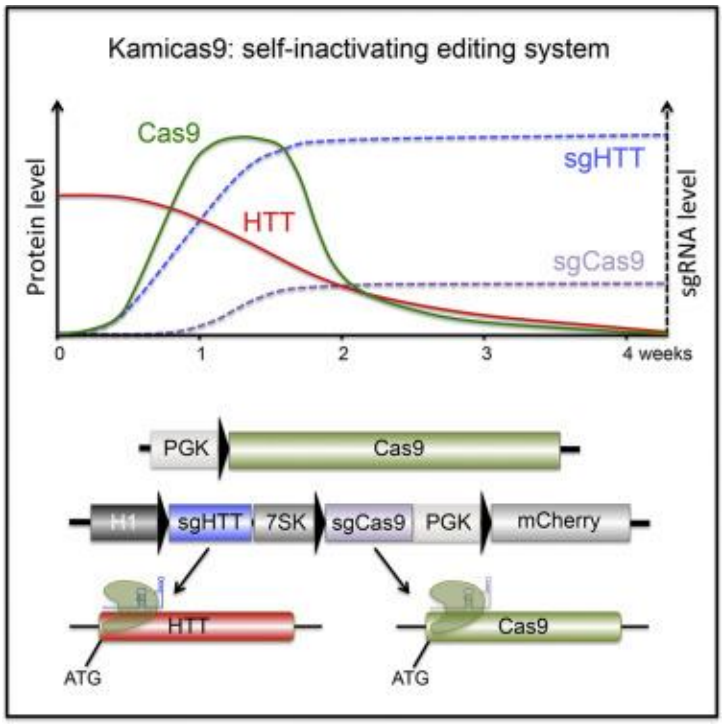


*Figure 15: **Graphical abstract of Merienne et al** (2017)[122]. The upper graph represents variation of protein and sgRNA levels after treatment. Underneath are pictograms of the different constructs in their Cas9 system and their targets.*

Another strategy to reduce off-target breaks was developed by the team of Jennifer Doudna, who ran for CRISPR system discovery in collaboration with Emmanuelle Charpentier in 2012[123]. Instead of injecting a plasmid encoding Cas9 endonuclease and sgRNA, they propose to inject in the brain Cas9 ribonucleoprotein complexes, which are preassembled, and once in cells, are quickly degraded because they are very short-lived.

This strategy allows a transient expression of the CRISPR-Cas9 system, eliminating off-target effects and reducing immune responses induced by persistent Cas9 expression. Moreover, it avoids the use of viral vectors, also reducing elicited immune responses. This strategy has produced promising results *in vivo* in transgenic mice expressing a fluorescent reporter system to detect precise DNA break by Cas9 ribonucleoproteins, inducing tdTomato positive cells when cleavage is performed. They demonstrated efficient DNA breaks when injected in the striatum, giving substantial hope for application in HD pathology[124].

Combining strategies can also be considered. Late stage HD will probably require grafted cells as well as ASO in order to replace neuronal loss and to prevent cell to cell transmission of mHTT from remaining cells.

To conclude this first part, although no cure is available for patients to date, a lot of new promising strategies have emerged, though most of them failed in clinical trial stages. Thus other innovative treatments are still needed to be developed.

To summarize the state of the art, some aspects of HD remain to be better characterized, but the cause of pathology is clearly defined and targeting the HD mutation directly at the RNA level appears to be the most successful strategy to date. In this context, a *trans*-splicing strategy could be optimal for HD.

In the next chapter, the basis of molecular biology will be reviewed with a particular interest in RNA biology. RNAs are now better understood with the discovery of functions impacting cellular function. In this context, we will focus on RNA processing, leading us to splicing. The final part will be dedicated to *trans*-splicing, which is a particular maturation process with potential applications in gene therapy.

1.3 RNA *trans*-splicing

Before introducing *trans*-splicing, it is needed to provide more details about RNA for a better understanding of biological context. In the next part, short presentation about RNA specificities will lead us across their variety, their complexity, and their wide range of functions in cells.

1.3.1 RNA generalities

1.3.1.1 *The central dogma*

Francis Crick mentioned for the first time in 1956 [125,126,] the idea that genetic information is transmitted from DNA to RNA to protein (Figure 16). Ever since, this model has been named the central dogma of molecular biology, providing an explanation about the transit of genetic information inside the cell.

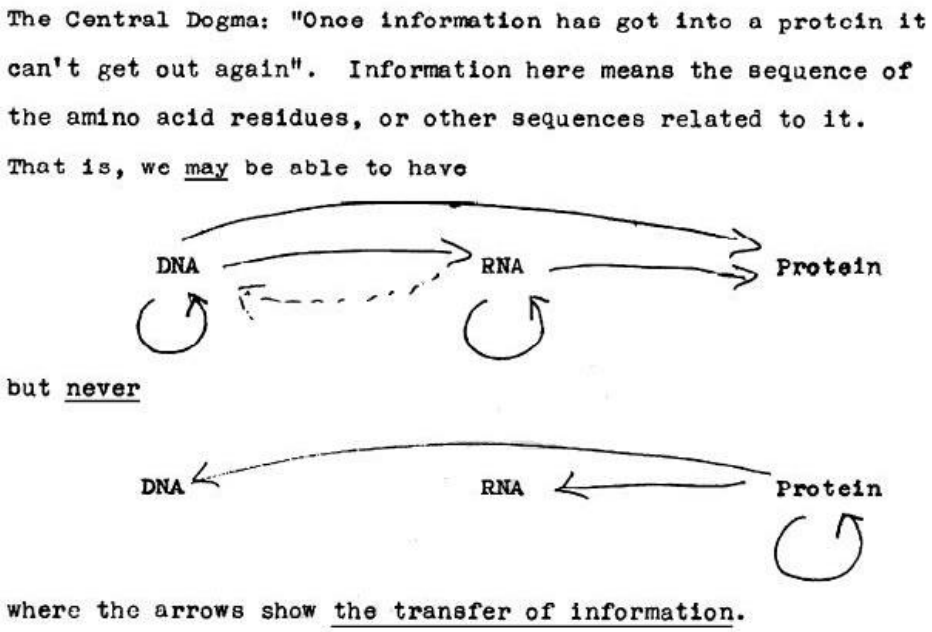


Figure 16: ***Handwritten central dogma representation of F. Crick*** *from an unpublished note made in 1956* [127]

Our comprehension of the central dogma has evolved with better understanding of cell machinery and gene expression mechanisms (Figure 17). These processes, mostly occurring during genetic information transfers, are known to be regulated through a variety of mechanisms targeting DNA, RNA and proteins such as epigenetic processes, post-transcriptional and post-translational modifications.

The central dogma has been challenged by many discoveries, highlighting more complex forms of genetic information transfer. However, none of these processes strikes down the main exclusion principle that states: "**Information flows between different forms of nucleic acids are allowed, but information cannot be recovered from amino acids to nucleic acids**". Reverse translation seems to be prohibited, mainly because protein folding is likely to prevent a possible "read" of the amino acid sequence, and denaturation of protein would lead to protein degradation. Regarding advances in

central dogma comprehension, it appears that RNA is a powerful intermediate between DNA and protein, sharing both the property of being a genetic information carrier and a folding capacity allowing multiple functions in cells as depicted in next section.

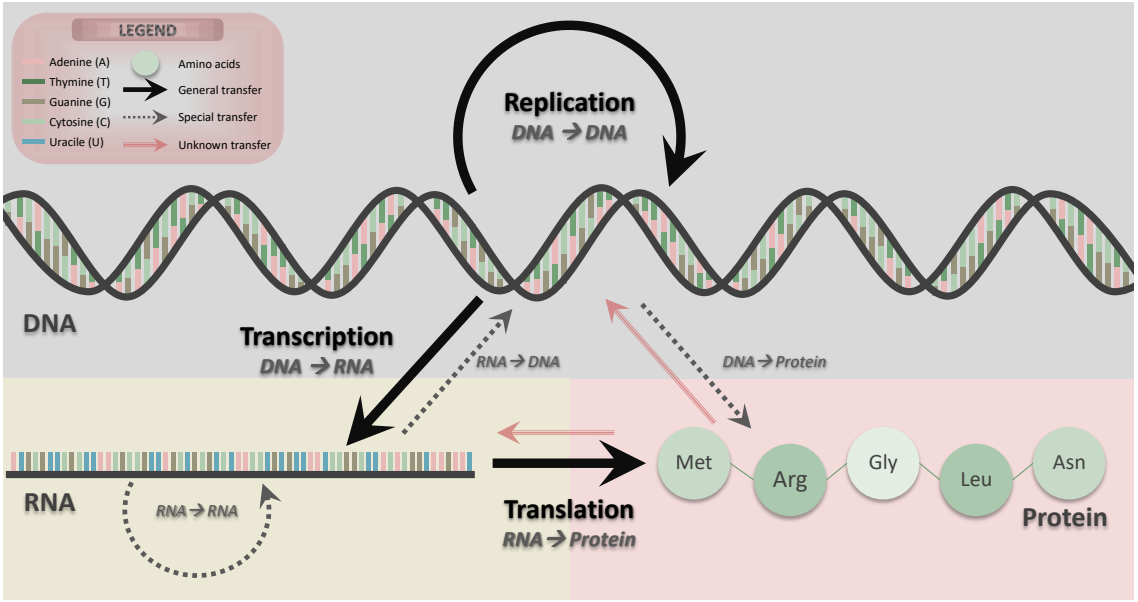


*Figure 17: **Schematic representation of the central dogma.** The central dogma assumes that genetic information is stored in DNA, which can duplicate its content during replication. This information is then transferred to RNA during transcription which is the first step of gene expression. Transcription, another kind of copying process, allows the production of one RNA strand from one strand of complementary DNA. RNA is designed to be used as a template to generate proteins during translation. Since the genetic code in DNA and RNA is driven by nucleotide triplets called codons (Appendix 1)[128], translation allows the conversion of this information into amino acid sequences to produce proteins. These two transfers of information are the main kinds operation in cells and each step is mandatorily conserved across all living species. However, under specific conditions, special transfers may occur such as (i) Reverse transcription, with the production of DNA molecules from RNA, (ii) RNA replication with duplication of RNA molecules and (iii) Direct production of protein from DNA. These processes take place in case of viral gene expression or in laboratory experiments with controlled environments. Other transfers are summarized in an unknown transfers category and to date no evidence demonstrates the existence of these reactions. Adapted from Crick et al [126].*

1.3.1.2 *Structure and functions*

As depicted in Appendix 1, DNA and RNA share some features in terms of both content (genetic code) and structure (nucleotides). However, RNA sets itself apart from DNA with slight differences, including a capacity to be folded into multiple shapes thanks to its non-canonical pairing properties (possible because they are mainly single stranded). These abilities allow RNA to exhibit catalytic properties making them much closer to protein. This discovery was awarded with a Nobel prize in 1989 shared by Thomas Cech and Sidney Altman. They both demonstrated that RNA possesses functional features but in different cellular processes: self-splice RNAs for Cech (See Spliceosome-mediated splicing page 59) and biosynthesis of tRNAs by ribonuclease P for Altman (See Non-coding RNAs page 52).

These discoveries have led to a new hypothesis suggesting that one single molecule type (RNA) can couple the capacity to encode its own reproductive machinery information and the capacity to perform chemical reactions for its own needs. This observation has led to the hypothesis that RNAs could have

been the basis of the first form of life [129–132]. This is called the RNA world hypothesis, considering that RNAs were the first genetic information holder, long before the development of the contemporary cellular system.

As mentioned before, key components in this hypothesis are that RNAs form a wide variety of molecules in terms of size and related secondary structure (Appendix 1). Indeed, RNAs can be folded into highly elaborated structures, suitable for chemical reactions essential to life. So RNAs allow all forms of genetic information transmission, and their folding capacity plays an essential role in their catalytic functions.

Folding seems to be the key and could be included into a more comprehensive vision of the central dogma (Figure 18). According to this point of view, RNA molecules can be divided into two categories:

- non-coding RNAs (ncRNAs) with functional properties
- coding RNAs carrying genetic information called messenger RNAs (mRNAs)

1.3.1.3 *Non-coding RNAs*

NcRNA functional properties make ncRNAs much closer to proteins, sharing folding ability but with an additional capacity to perform reversal folding (Figure 18). NcRNAs are synthetized during transcription like mRNAs, but when that exhibit catalytic properties, they are called ribozymes and they display a wide diversity of functions in cells:

- Protein assembly with
  - **transfer RNAs** (tRNAs): making the link between mRNA and protein. They translate genetic information and provide amino acids to the ribosome for their incorporation into newly synthesized nascent polypeptide chains.
  - **ribosomal RNAs** (rRNAs): carrying out the catalysis property of ribosomes (a ribonucleoprotein complex involved in protein synthesis). As depicted in Figure 19, a ribosome contains two rRNA subnits: a small one (40S) to decode the mRNA and a large one (60S) to catalyze peptide linkage[133].
- RNA modification with
  - **spliceosomal RNAs**: involved in the splicing process, further detailed in section 1.3.2.2 Spliceosome-mediated splicing.
  - **ribonuclease/RNAse P**: exhibiting ribozyme activity with a catalytic RNA subunit involved in 5' end maturation in precursor tRNAs[134]. Another category of RNAse P constituted of protein only also exists.

- Gene expression regulation with
  - **Riboregulators**: antisense RNAs acting at the level of translation initiation by blocking AUG motifs on mRNA under environmental stimuli[135]
  - **Riboswitches**: a large ncRNA category of partly double stranded RNA characterized by its capacity to regulate cellular functions by inducing structural rearrangement in response to intracellular physical signals, such as metabolites and ions[136]. They bind to small metabolites which, in turn, mediate gene activity regulation.[137]

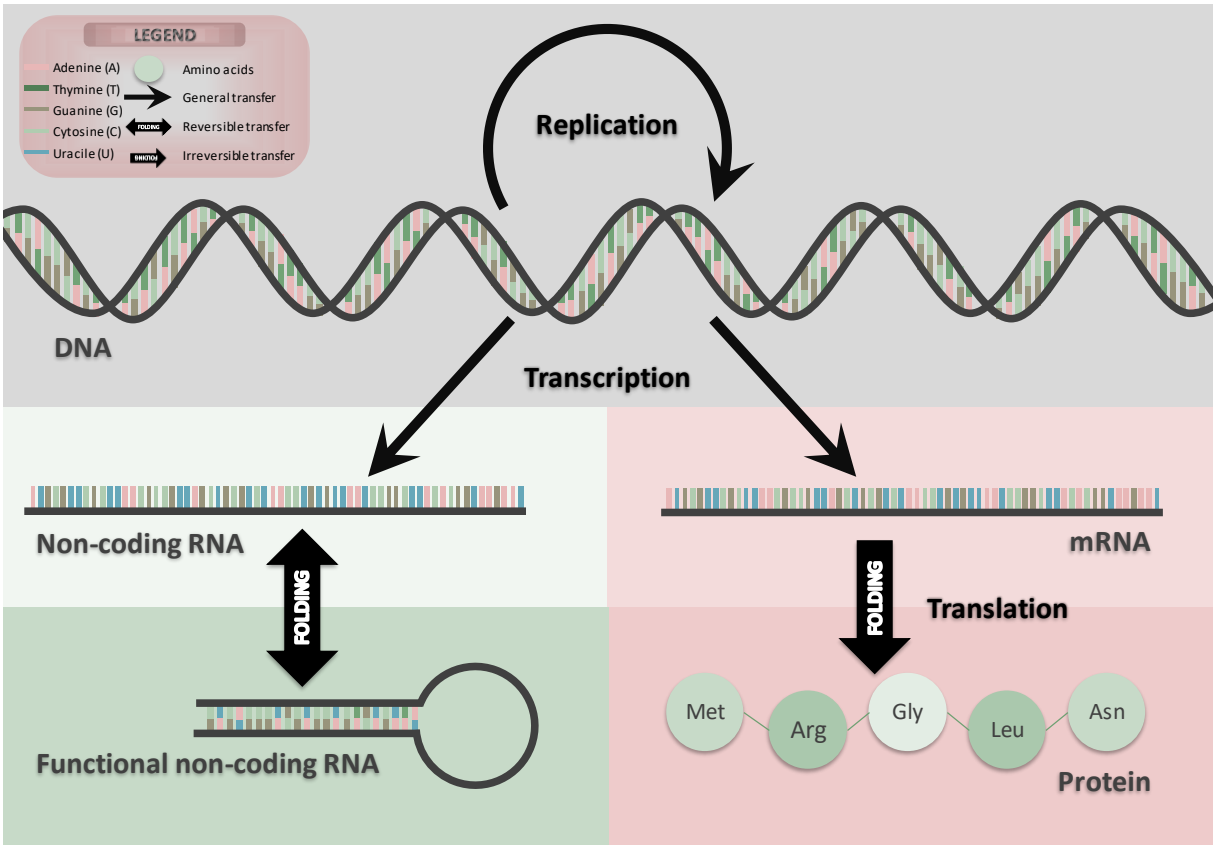


*Figure 18: **Updated version of the central dogma.** Transcribed RNA can have two different fates: whether to be used for protein synthesis as mRNA or whether to have specific activity in cells after folding. This representation highlights the importance of RNA in information transfer. Adapted from Koonin et al [138]*

Other ncRNAs without catalytic properties also play an essential role in cells. In this case, this is not RNA folding but their pairing capacity which mediates activity in cells, mainly in acting on signaling. Indeed, double strand RNAs are likely to initiate RNA cleavage by the activation of Dicer ribonuclease. This mechanism is called the RNA-induced silencing complex (RISC) and is involved in many processes such as:

- development with micro-RNAs (miRNAs)
- genome protection with small interfering RNA (siRNAs).

Taken all together, RNAs are actively involved in major cellular processes because of their structural flexibility and their capacity of quick adaptation which make them striking molecule to target and/or to use (see textbox Selex page 66).

In this context, the main limit is prediction of RNA structure based on its sequence, which is a necessary to go further. Different prediction approaches have been imagined, both experimental and computational. However, due to cellular complexity limitations preventing the stability of RNA conformation, none of these approaches has given results reliable enough to be used with confidence *in vivo*. This drawback is also a consequence of **the versatility of RNA**. Indeed, across evolution, nature developed other forms such as protein for a more sustainable life system since RNAs are not likely to allow the development of more complex systems. Regarding ncRNA's potential, they are assumed to be powerful targets to be used in biological research, complementary to mRNAs, essential to produce protein.

1.3.1.4 *Messenger RNAs*

mRNAs are synthesized during transcription by the enzyme RNA polymerase in three stages: initiation, elongation and termination:

- **initiation**: identification of the promotor *cis*-acting element such as a "TATA box" and unwinding of the two strands of the DNA double helix
- **elongation**: repeated selection of the correct ribonucleoside triphosphate (RNA nucleotide) and catalysis of the formation of a phosphodiester bond. Elongation is unidirectional along DNA from 5' to 3' extremities.
- **termination**: identification of the termination signal defining the transcript end.

Although mRNAs are universally expressed in all kinds of living cells with similar basic biochemistry of RNA synthesis, the machinery to generate RNAs and proteins depends on the cell type. Differences in terms of space and time are found between prokaryotes and eukaryotes due to their cell organization depicted in Figure 19. Main differences involve regulation, polymerase types and last but not the least post-transcriptional processing.

Indeed, eukaryote mRNAs are first generated as mRNAs (pre-mRNAs) in the nucleus during transcription thanks to RNA polymerase II**.** To generate mature mRNAs, pre-mRNAs need to undergo several maturations steps before being exported outside the nucleus including:

- **Capping**: addition of a 7-methyl guanosine (cap) at the 5' extremity for:
    - regulating nuclear export
    - preventing transcript degradation
    - promoting translation by ribosome recognition
- **Polyadenylation**: addition of a poly-A-tail at the 3' end to:
    - protect mRNA from degradation
    - increase mRNA stability
    - end transcription
- **Splicing**: removal of introns and ligation of exons. This process will be described in more detail in the next chapter.

Pre-mRNAs are single strand nucleic acids carrying the same sequence as the coding strand of DNA. In the eukaryote genome, it is known that the coding information is fragmented into scattered sequences called exons. So, exons code information conserved into the mature mRNAs. However, at DNA and pre-mRNAs levels, exons are spaced apart by introns which are non-coding sequences located in between (See Intron specifications page 57 and intron text box page 66).

Exons and introns are together involved in splicing, a maturation process which is at the heart of this project, and described extensively in the next section by introducing intron signaling complexity, molecular mechanisms of splicing and their regulation pathways.

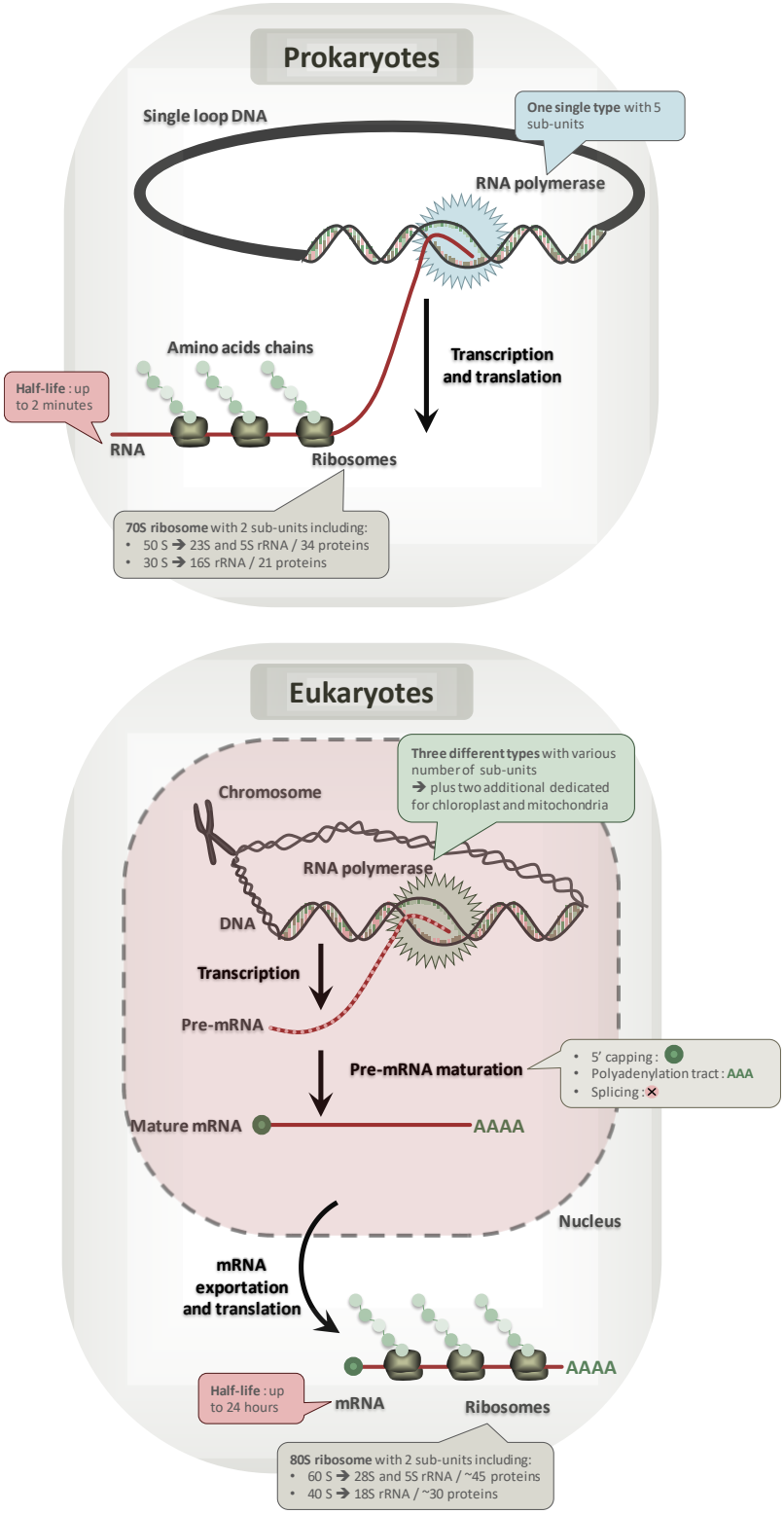


Figure 19: ***Transcription and translation in prokaryotes and eukaryotes.*** *In prokaryotes: intron-free mRNAs are mainly polycistronic, meaning that one single mRNA is the template for several polypeptide chains. This can be up to 20 different proteins. In eukaryotes: from one single gene transcribed as pre-mRNAs, cells can generate different isoforms of mature mRNAs during splicing, thus producing different proteins. This process is called alternative splicing. Following maturation, one mRNA encodes for only one polypeptide. For this reason, eukaryotic mRNAs are monocistronic. Adapted from Berg et al 2002*[139]

1.3.2 Splicing

1.3.2.1 *Intron specifications*

Intron sequences are easily identified in the genome since there are known sequence motifs that define precisely intron sequences and boundaries. They are marked out by the donor site (DS) at the 5' end and the acceptor site (AS) at the 3' end, most of the time characterized by "GT" and "AG" motifs respectively (Figure 20 and Table 4). The DS and AS are part of much larger splice consensus sequences, called 5' splice site (5'SS) and 3' splice site (3'SS). These splice sites correspond to the nucleotides that surround exon-intron boundaries. Some of these nucleotides are highly conserved such as DS and AS, and others appear with known occurrence based on their location. Exon-intron boundary consensus motifs are summarized in Table 4, which represents a matrix of splice site occurrence.

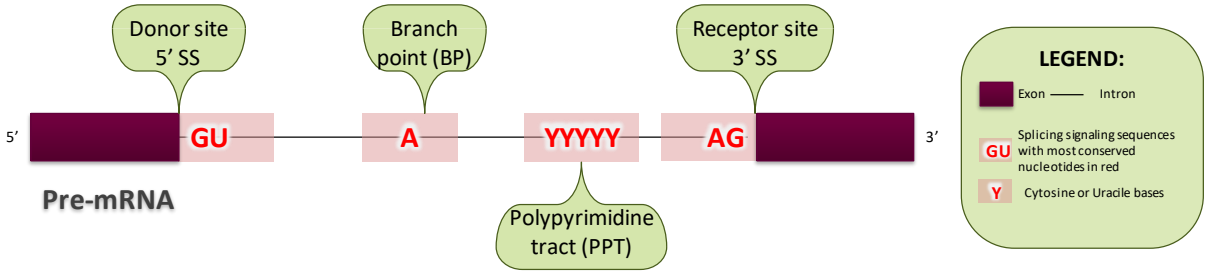


Figure 20: ***Schematic representation of intron splicing sequences.***

The two additional core splicing sequences are the branch point (BP) and the polypyrimidine tract (PPT) as depicted in Figure 20. These signals can be predicted by using bioinformatics tools to define intron composition and boundaries. The region between BP and 3'SS is defined by the absence of AG dinucleotides (r2 in Figure 21)[140,141]. This particularity is used to predict BP which can be located at an unusual distant location. For this purpose, an extended region named the AG exclusion zone (AGEZ) is characterized by a sequence spanning from the 12th nucleotide upstream through the first AG_{a3} before r2 to the 3'SS AG as depicted in Figure 21.

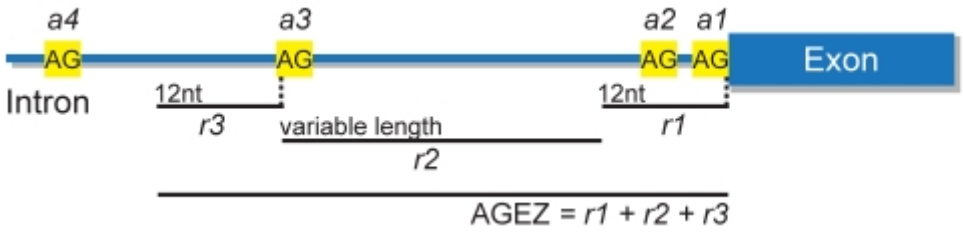


Figure 21: ***AG exclusion zone (AGEZ)****.BP location relative to the 3'SS (a1) is dependent on the presence/absence of additional AG dinucleotides in the intron. The most common situation is the absence of AGs in the region between the BP and the 3'SS. However, these can occur either at locations close to the 3'SS (so a2 in r1) where they may compete with the 3'SS signal, or very close to the BP (a3 in r3), where they are bypassed possibly due to steric constraints. Any AG occurring in r2 is likely to be recognized as 3'SS. Therefore, BPs are usually located inside the region defined as r1+r2+r3 – the AG exclusion zone (AGEZ). Adapted from Corvelo al (2010)*[141]

Table 4: ***Percentage probabilities of nucleotide occurrence in humans according to their relative position to both 5' and 3'SS****. For 5'SS tables, negative numbers are located in exons and positive numbers are located in introns. For 3'SS, negative numbers are located in introns and positive numbers are located in exons. Isochores are large DNA regions (> 300 kb) characterized by specific GC composition variation mapped in mammalian chromosomes. High and low GC isochore cutoffs are, respectively, superior and inferior to 50% of GC in a sequence. Adapted from Zhang et al, 1998[142].*

5' Splice Sites - high GC isochores

	-3	**-2**	**-1**	**1**	**2**	**3**	**4**	**5**	**6**
A	32	56	8	0	0	38	70	5	13
C	38	15	4	0	0	4	9	5	21
G	19	15	80	100	0	56	14	86	25
T	11	14	8	0	100	2	7	4	41

5' Splice Sites - low GC isochores

	-3	***-2***	***-1***	***1***	***2***	***3***	***4***	***5***	***6***
A	38	62	12	0	0	71	73	11	21
C	31	10	4	0	0	2	6	6	10
G	18	12	77	100	0	24	8	75	14
T	13	16	7	0	100	3	13	8	55

3' Splice Sites - high GC isochores

	-15	**-14**	**-13**	**-12**	**-11**	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**1**	**2**
A	10	8	7	8	6	6	4	8	8	7	6	19	2	100	0	21	19
C	41	42	41	40	38	43	42	46	49	54	45	38	82	0	0	13	21
G	15	14	14	13	13	12	13	14	10	8	8	26	0	0	100	58	29
T	34	36	38	39	43	39	41	32	33	31	41	17	16	0	0	8	31

3' Splice Sites - low GC isochores

	-15	**-14**	**-13**	**-12**	**-11**	**-10**	**-9**	**-8**	**-7**	**-6**	**-5**	**-4**	**-3**	**-2**	**-1**	**1**	**2**
A	15	14	13	11	10	10	11	12	13	11	10	26	7	100	0	26	24
C	24	21	20	22	21	22	25	28	28	25	22	25	55	0	0	11	15
G	10	12	10	9	10	9	10	10	8	5	5	15	1	0	100	50	20
T	51	53	57	58	59	59	54	50	51	59	63	33	37	0	0	13	41

1.3.2.2 *Spliceosome-mediated splicing*

The main goal of splicing is to produce mature mRNA encoding meaningful information for the translation step. This process is ATP-dependent and occurs concomitantly to transcription, meaning that the pre-mRNA is spliced as soon as it is synthesized. Thus, for long pre-mRNAs, splicing of the beginning of the pre-mRNA occurs while it is still not entirely synthesized. Splicing requires the synchronized action of multiple splicing factors, which as a whole are called the spliceosome.

The spliceosome is the major molecular complex catalyzing the splicing reaction. It is formed from 5 subunits called small nuclear ribonucleoprotein particles (snRNPs). Each snRNP involves RNA-protein interaction and is composed of[143]:

- a small nuclear RNA (snRNA) specifying the subunit type
- up to 100 different proteins

The subunit composition of the spliceosome changes during the time course of the splicing process in a stepwise manner. Each subunit acts in cooperation with the others and has a defined role. The most common association involves U1, U2, U4, U5 and U6 snRNPs, and this composition is likely to be highly conserved across cell types. However, other snRNPs are involved in other particular splicing processes representing 0.35% of human introns[144]. They are called minor spliceosomes and are known to target splice site sequences of rare pre-mRNA isoforms with functional analogous snRNPs.

The splicing process takes place in two steps that are detailed in the next section:

- Assembly
- Catalysis

1.3.2.2.1 Assembly

Firstly, some snRNPs and other splicing factors that recognize specific motifs are required to trigger splicing. It begins with a U1 interaction at the 5'SS, U2AF1 at PPT and U2AF2 at the 3'SS, which are respectively U2 auxiliary factor 1 and 2[145] (Figure 22, Complex E or Early complex).

Then, U2 interacts with the BP (Complex A or pre-spliceosome)[146]. These first recognitions initiate other snRNP recruitments called the tri-snRNP for U4, U6 and U5 (Complex B or pre-catalytic spliceosome)[147], inducing changes into the pre-mRNA conformations resulting in spatial proximity between 5'SS and 3'SS. Then it switches towards complex B* with the release of U1 and U4.

So, this first stage of assembly corresponds to a cascade of snRNP recruitments that induces pre-mRNA folding in a shape suitable for catalysis (Complex C), and thus making the initiation of the first cleavage ready to be performed.

1.3.2.2.2 Catalysis

From a chemical point of view, the first enzymatic reaction is a *trans*-esterification driven by the adenosine nucleotide located in the BP and in particular by its ribose 2'OH free end which is bulged [148]. This 2'OH-adenosine performs a nucleophilic attack on the 5'SS phosphate (of the first intron nucleotide which is most commonly a guanosine) producing a 3',5'-phosphodiester break on the 5'SS and a 2',5'-phosphodiester bond between the 5' intron end and the BP.

This first cleavage produces two splicing intermediates (Figure 22):

- the upstream exon with a free 3'OH end
- the so-called lariat composed of the loop-shape intron followed by the downstream exon.

The second cleavage is initiated by the free 3'OH from the upstream exon which carries out, in turn, a nucleophilic attack on the 3'SS phosphate, resulting in the release of the lariat intron and the ligation of both exons together.

Spliceosome-mediated splicing is extensively described in Suné-Pou et al (2017)[149].

1.3.2.2.3 Self-spliced intron

Self-spliced introns also exist and perform intron excision and exon ligation as well, without involvement of any external factors[150,151]. In these cases, introns perform the same chemical reaction with the two *trans*-esterification steps and the lariat formation.

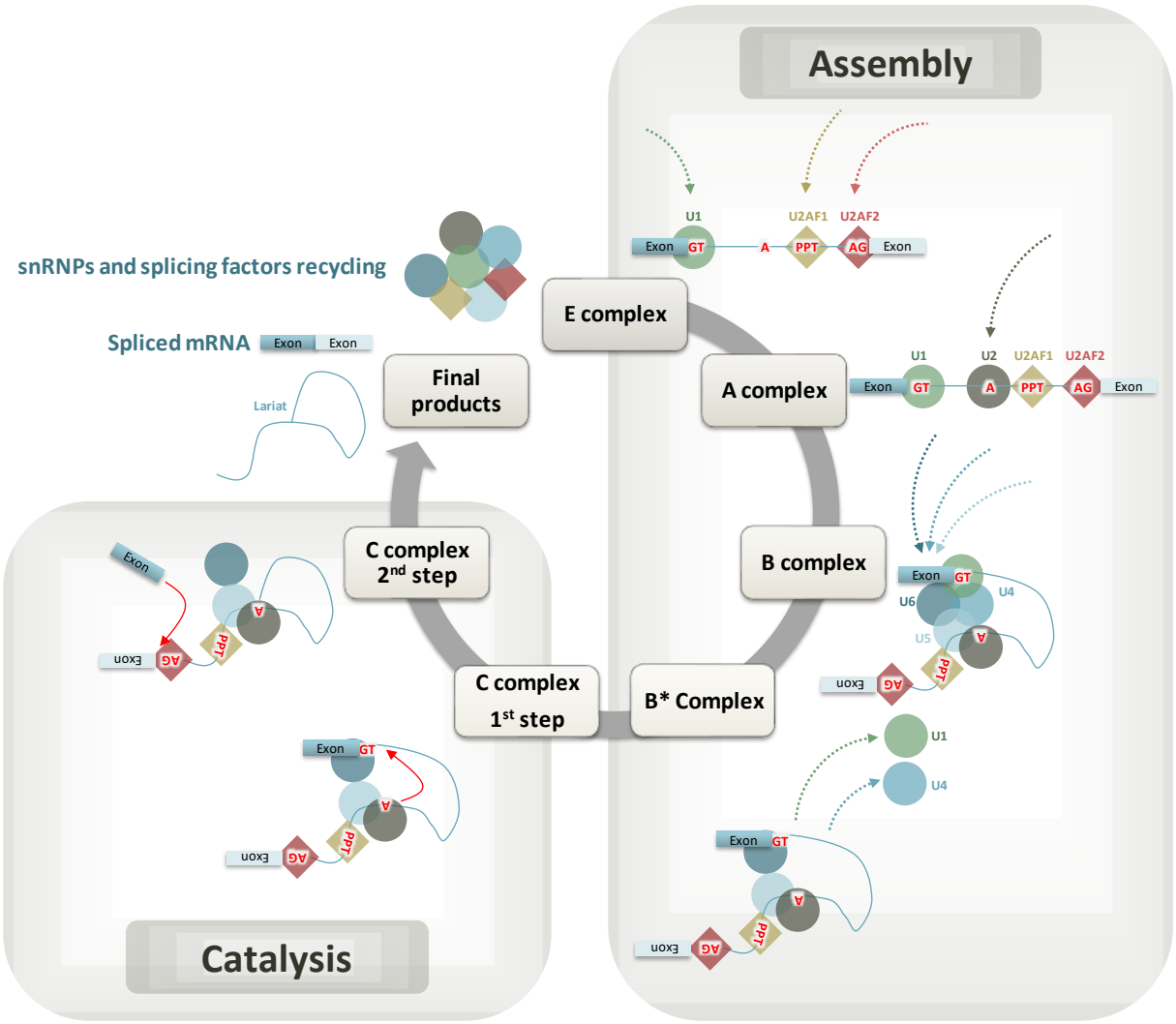


Figure 22: ***Schematic representation of the spliceosome assembly and pre-mRNA splicing.*** *In the first step of the splicing process, the 5' splice site (GU, 5'SS) is bound by the U1 snRNP, and U2AFs cooperatively to the polypyrimidine tract (PPT), and the 3' splice site (AG, 3' SS) to assemble complex E. The binding of the U2 snRNP to the branch point (A) results in the pre-spliceosomal complex A. Subsequent steps lead to the binding of the U4/U5–U6 tri-snRNP and the formation of complex B. Complex C is assembled after rearrangements that detach the U1 and U4 snRNPs to generate complex B*. Complex C is responsible for the two trans-esterification reactions at the SS. Additional rearrangements result in the excision of the intron, which is removed as a lariat RNA, and ligation of the exons. The U2, U5, and U6 snRNPs are then released from the complex and recycled for subsequent rounds of splicing. Adapted from* [149]

1.3.2.3 *Splicing regulation*

Spliceosome-mediated splicing involves the synchronization of components that are highly regulated in terms of time course. These different sub-units are all potential targets to modulate gene expression during mRNA maturation. Indeed, splicing is required for most mRNA processing and its modulation contributes to producing different isoforms of mature mRNA from one single gene.

1.3.2.3.1 Splicing enhancers and silencers

To perform splicing, essential splicing factors are recruited on pre-mRNAs. It is believed that this recruitment takes place through pairing interactions, and these interactions are modulated by *trans*-acting splicing factors also called splicing regulatory elements. SnRNP pairing occurs on the core splice signals including 5'SS, 3'SS, PPT and BP whereas splicing regulatory elements recognize other less characterized sequences, divided into 4 categories according to their locations and activity (Figure 23):

- Exonic splicing enhancers (ESEs)
- Exonic splicing silencers (ESSs)
- Intronic splicing enhancers (ISEs)
- Intronic splicing silencers (ISSs)

These sequences are *cis*-regulatory, known to activate or inhibit spliceosome assembly through splicing regulatory element recognition. Splicing regulatory elements are *trans*-regulatory elements mainly constituted by:

- SR proteins for serine/arginine-rich proteins
- hnRNP for heterogenous nuclear ribonucleoprotein particles

SR proteins are known to enhance spliceosome activity through ESE and ISE recognition whereas hnRNPs are likely to prevent spliceosome stabilization by interacting with ESS and ISS[152].

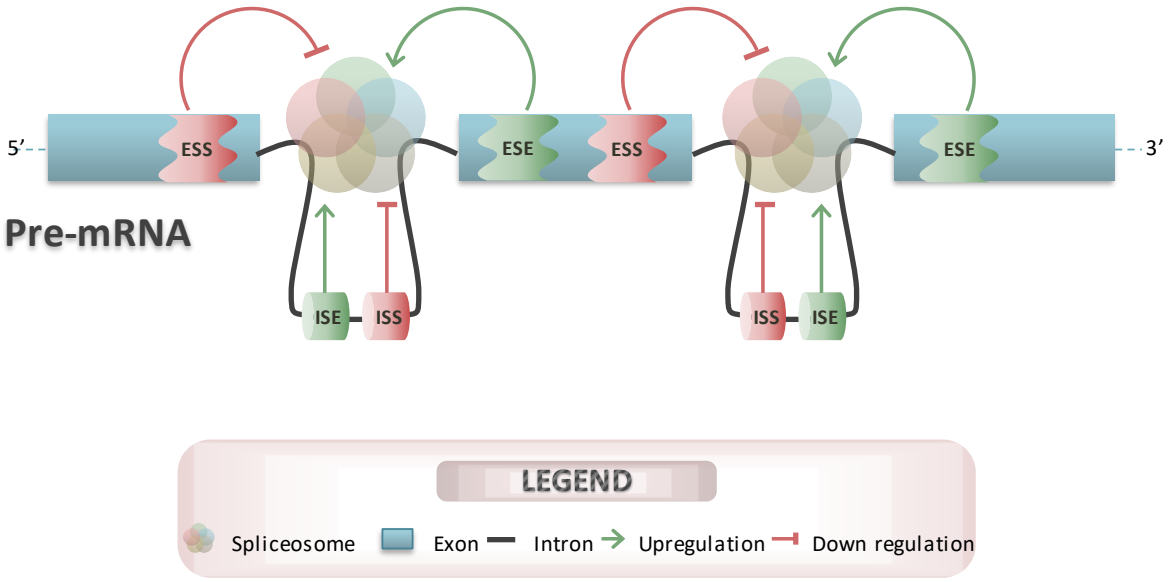


*Figure 23: **Enhancer and silencer elements in exons and in introns** (ESE: exonic splicing enhancers; ESI: exonic splicing silencers; ISE: intronic splicing enhancers; ISI: intronic splicing silencers) contribute to splicing regulation.*

1.3.2.3.2 Exon and intron definition

As evoked above, splicing regulatory elements are believed to act on splicing regulation, and in particular, on the stabilization of the spliceosome. This stabilization depends on the organization of the pre-mRNA transcripts with respect to the relative size of exons and introns. For instance, 90% of the mammalian genome is characterized by short exons (50 to 250 bp) with relatively large introns in between (hundreds to thousands of bp) [153,154]. In this context, splicing regulatory element interactions with the spliceosome are believed to take place across exons (Figure 24, right). This process is called "exon definition", in which the stabilization of snRNAs and other unknown factors is performed across exons in downstream 5'SS and in upstream 3'SS.

In contrast, the opposite model, called "intron definition", is a process dedicated to short introns (fewer than 250 bp, Figure 24 left) in which interactions between splicing regulatory elements and spliceosome sub-units occur primarily across introns.

Intron definition is believed to take place in invertebrates, fungi and plants, whereas exon definition has been mostly observed in humans and mice. Thus, the relative size of intron seems to induce different activation pathways to perform efficient splicing. In exon definition, the U2AF2/1 complex in 3'SS interacts with the U1 subunit located on the 5'SS of the next intron, thus interaction occurs across exons if introns are too large. In intron definition, the U1 subunit and the U2AF1/2 complex are in the same intron sequence and so interaction is through introns.

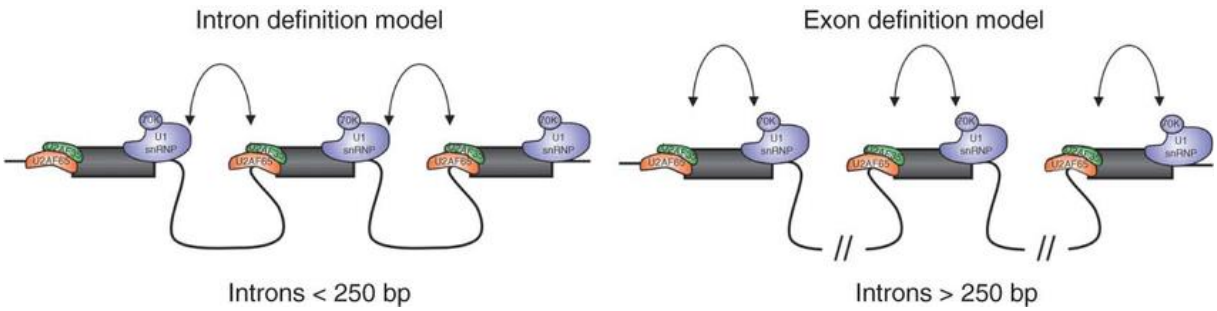


*Figure 24: **Exon and Intron definition models.** The left panel depicts the Intron definition model according to which pairing between the splice sites takes place across an intron when long exons are separated by short (<250 bp) introns. On the other hand the right panel shows the Exon definition model where the splice site communication occurs across exons when they are separated by long (>250 bp) introns. Adapted from De Conti et al (2013)[153]*

1.3.2.3.3 Alternative splicing and exon skipping

There are a plethora of splicing regulation pathways and more than 90% of the human genome involves alternative splicing[155,156]. Alternative splicing is the formation of different mRNA isoforms from a single pre-mRNA, thanks to a different arrangement of the exons in the mature mRNAs. Alternative splicing may follow different patterns of exon rearrangement summarized in Figure 25. HTT being involved in many cellular processes is expressed as different isoforms listed in Appendix 6.

The most common and predictable form of alternative splicing is exon skipping[157]. In this case, an exon sequence is removed with surrounding intron sequences during splicing (See textbox dedicated to Tau page 67, as an example). It is believed that exon skipping takes place in introns characterized by "weak" signaling of their 5'SS and 3'SS strength[158] (see MaxEnt in Materials and methods page 84 and in Intron 9 and 5'SS competition page 139).

Some weak splice sites can be considered as alternative splicing cryptic splice sites, since they are not supposed to be spliced. These sites have a low activity and they are disadvantaged by their position. However, once a cryptic splice site is activated through a specific phosphorylation state of *trans*-splice regulators, it can result in a high rate of mis-splicing, which in some cases can be pathological[159].

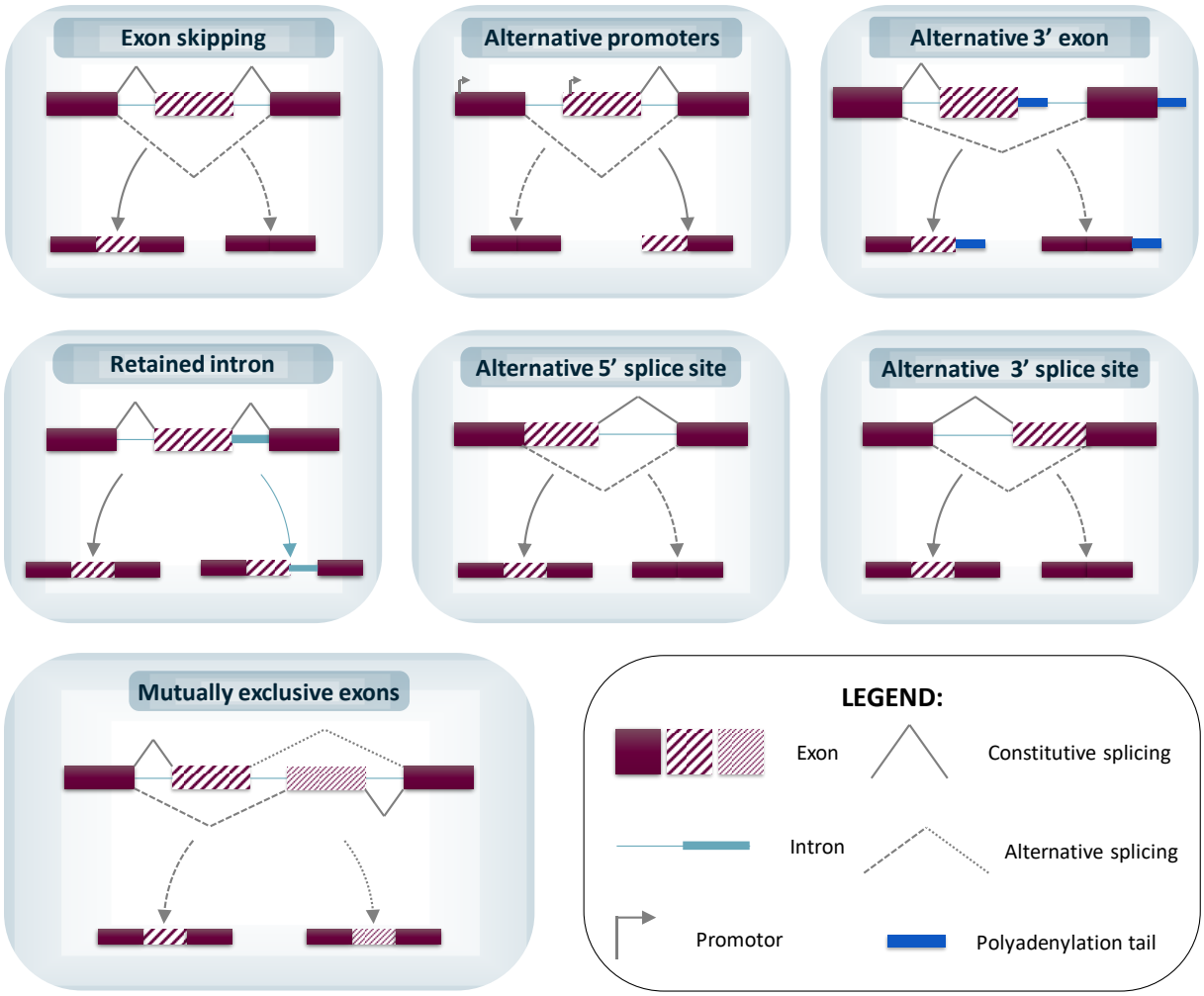


*Figure 25: **Different patterns of alternative splicing**. The v-shaped lines show the different ways in which the exons can be joined in a final mRNA. The most common is exon skipping from the final mRNA. Mutually exclusive are a pair of consecutive cassette exons where only one of the exons is included in the mRNA. Alternative 5' splice sites are consecutive 'donor sites' that change the length of an exon at its 3' end. Conversely, alternative 3' splice sites are consecutive acceptor sites that change the 5' end of the exon. Alternative promoters and alternative 3' exons create different first exons and different last exons on the mRNA, respectively. Retained introns can be excised as a typical intron or remain in the final mRNA. A single gene can have multiple positions and patterns of alternative splicing to create a family of many different mRNAs and proteins through the inclusion or skipping of various alternatively spliced RNA segments. Adapted from Vuong et al (2016).*[157]

To go further…

SELEX

RNA evolution accelerator

Since punctual mutation in RNA sequences occurs commonly in cells, RNA biological functions evolve faster. Indeed, when changes occur in sequences, it impacts RNA shape and possibly improves or prevents some RNA biological functions. This aspect is studied in our laboratory with a particular type of RNA selection method called SELEX (Systematic Evolution of Ligands by EXponential enrichment). From oligonucleotide libraries, SELEX allows a discrimination of oligonucleotides according to their affinity for a given ligand. This selection is based on multiple cycles of oligonucleotide amplification, mutation, and selection steps. Products of SELEX are called aptamers and are the synthetic version of riboswitches. SELEX illustrates the power of RNAs and how RNAs could be used as biological tools.

INTRON

From DNA trash to splicing band leader

When introns were discovered, they were first believed to have no function and they were considered as remaining sequences discarded by bacteria because of their high rate of division. This hypothesis remained since it was consistent with the weak representation of introns in unicellular eukaryote organisms. It has been demonstrated now that introns are specific to eukaryotic cells and are highly involved in gene expression modulation by promoting or preventing expression of surrounding exons into mature mRNAs.

Tau

The champion of splicing regulation

Tau is encoded by the MAPT gene and is involved in microtubule stabilization in axons. Tau is known to be expressed in neurons as various isoforms arising from AS events. These isoforms are co-expressed in a regulated ratio differing according to brain region and developmental stage. Each isoform is likely to have determined roles in terms of microtubule-binding capacity and modulates axonal transport based on cell needs. Interestingly, the proportion of these isoforms in cells is linked with the vulnerability of brain region to undergo neurodegenerative processes such as Alzheimer's disease. Tau protein is a perfect example to illustrate the complexity of splicing regulators. Indeed, due to its high involvement in several neurodegenerative processes, splicing interactors have been well-characterized to address the isoform ratio. A high diversity of potential interactors emerges including cis-elements, SR proteins, micro-RNAs and kinases (**Erreur ! Résultat incorrect pour une table.**).

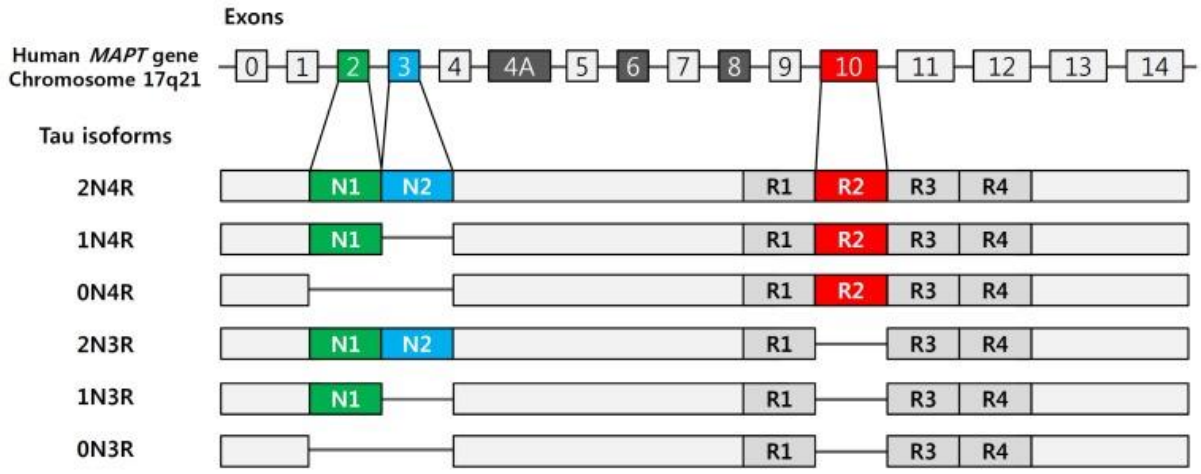


*Figure 26: **Tau protein isoforms in the human brain**, Tau protein isoforms in the human brain. Six tau isoforms are present in the human brain through different combinations of the splicing of exons 2, 3, and/or 10. The aspects of the N-terminal projection domain, N1 (green) and N2 (blue), are produced from exons 2 and 3, respectively. Exon 10 encodes the second aspect of the microtubule-binding repeat domain, R2 (red). Depending on the presence of the R2 domain, tau proteins become either 3R or 4R tau. Adapted from Park et al (2016)[160]*

1.3.3 Endogenous *trans*-splicing

The splicing introduced above is *cis*-splicing, as opposed to *trans*-splicing (TS). The main difference between these two kinds is the number of pre-mRNA molecules involved. In *cis*-splicing, only one single pre-mRNA molecule is processed. This is the most common form of splicing but, a special form of splicing can occur between different pre-mRNA molecules. In this case, exons from distinct pre-mRNAs can be combined into a single mature mRNA. In *trans*-spliced pre-mRNAs, introns are still removed but exons are then put together, thus producing a single chimeric mature mRNA (Figure 27).

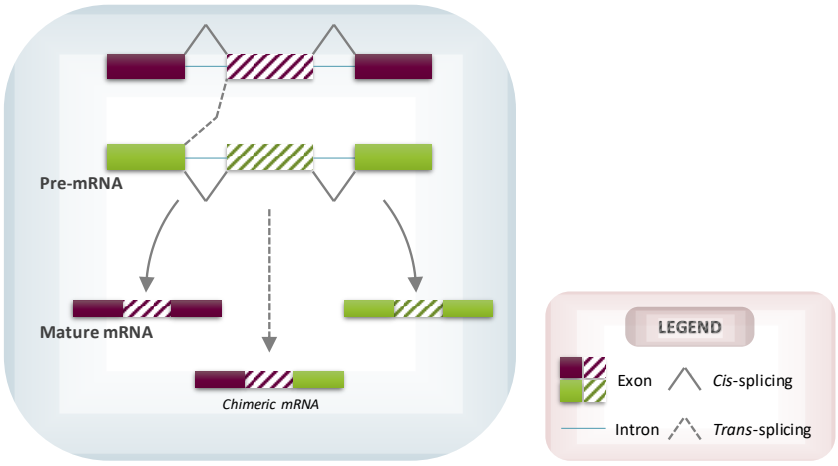


*Figure 27: **Schematic representation of cis- and trans-splicing between two pre-mRNAs**. Two different pre-mRNAs are represented in purple and in green. The v-shaped lines represent potential splicing processes, resulting in mature mRNA.*

TS is a widespread mechanism in species covering vertebrates, invertebrates[161,162] and even in prokaryote with group II introns, known to be mobile genetic elements with catalytic property[163]. Different kinds of TS have been identified, as depicted in Figure 28. This process was first described in 1986 in trypanosoma mRNA processing. Two studies from Murphy et al[164] and Sutton et al [165] identified an additional sequence in mRNA provided by a “splice leader” (SL) or “mini-exon” (Figure 28B). It is now known that SL is common in lower eukaryotes with RNA specialized for TS, and occurs during splicing in a two-step mechanism[166]. Other mechanisms include intergenic TS with exons from other genes generating chimeric mRNA and intragenic TS involving pre-mRNA from the same genomic locus. This last-mentioned TS type mediates exon repetition, sense/antisense fusion and exon scrambling. In rodents, studies have shown that TS mediates the expression of different protein isoforms such as immunoglobulin[167] which results from intergenic TS, or carnitine octanoyltransferase[168] which results from intragenic TS.

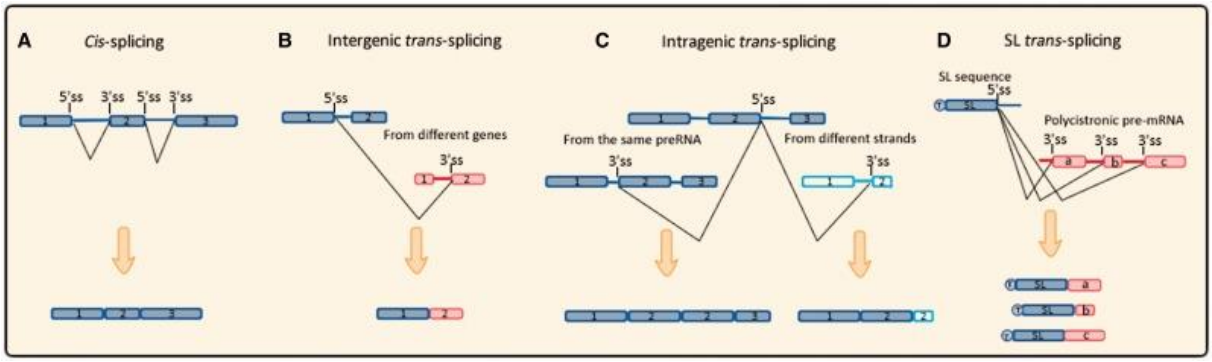


Figure 28: ***Schematic diagram of different types of pre-RNA splicing events.*** *(A) Cis-splicing. After excision of introns, exons of the same pre-mRNA are joined together to form a linear molecule. (B) Intergenic trans-splicing. Transcripts from different genes or even different chromosomes could be spliced and generate a non-linear chimeric molecule. (C) Intragenic trans-splicing. Boxes with vertical lines represent exons transcribed from the other strand. In the same gene, the splicing reaction occurs between two identical transcripts, or, transcripts from different strands, leading to exon-duplication and sense–antisense fusion. (D) SL trans-splicing. Red boxes represent structural genes, while T represents the TMG cap on the Spliced-leader (SL) mini-exons. SL exon is produced from tandem repeated SL gene cluster, splicing reaction occurs between SL exon and distinct structural genes of a polycistronic pre-mRNA to generate an array of mature "capped" transcripts. Extracted from review Lei et al*[169]

TS event frequency is dependent on species as depicted in Figure 29 (Bottom). In lower organisms, TS is a common process like in *C. elegans* with more than 70% of genes *trans*-spliced. It can achieve up to 100% of genes when a splice leader is involved such as in the dinoflagellate branch (*Protozoa*) whereas in mammals, where other TS types occur, TS is a very rare event. It is believed SL TS is found in organisms using operons to drive their gene expression, but the operon disappears and turns to *cis*-splicing in vertebrates. Remaining TS events are limited to some processes with a lower frequency.

Table 5 describes genes with identified chimeras in humans that are associated with function. TS appears to be involved in cancer, but the precise connection remains to be elucidated since some chimeric RNAs are found in cancer and in normal cells. However, some studies suggest that chimeric mRNA is involved in chromosomal translocation[170] and could be a good indicator of tumorigenesis as biomarkers for diagnosis[171]. Emerging role of TS in regulation of pluripotency capacity in human embryonic stem cell has been demonstrated in recent studies, suggesting a potential contribution in gene expression[172,173]. Other functions of TS are suggested by Table 5 including gene expression regulation, signal transduction, cell survival and growth.

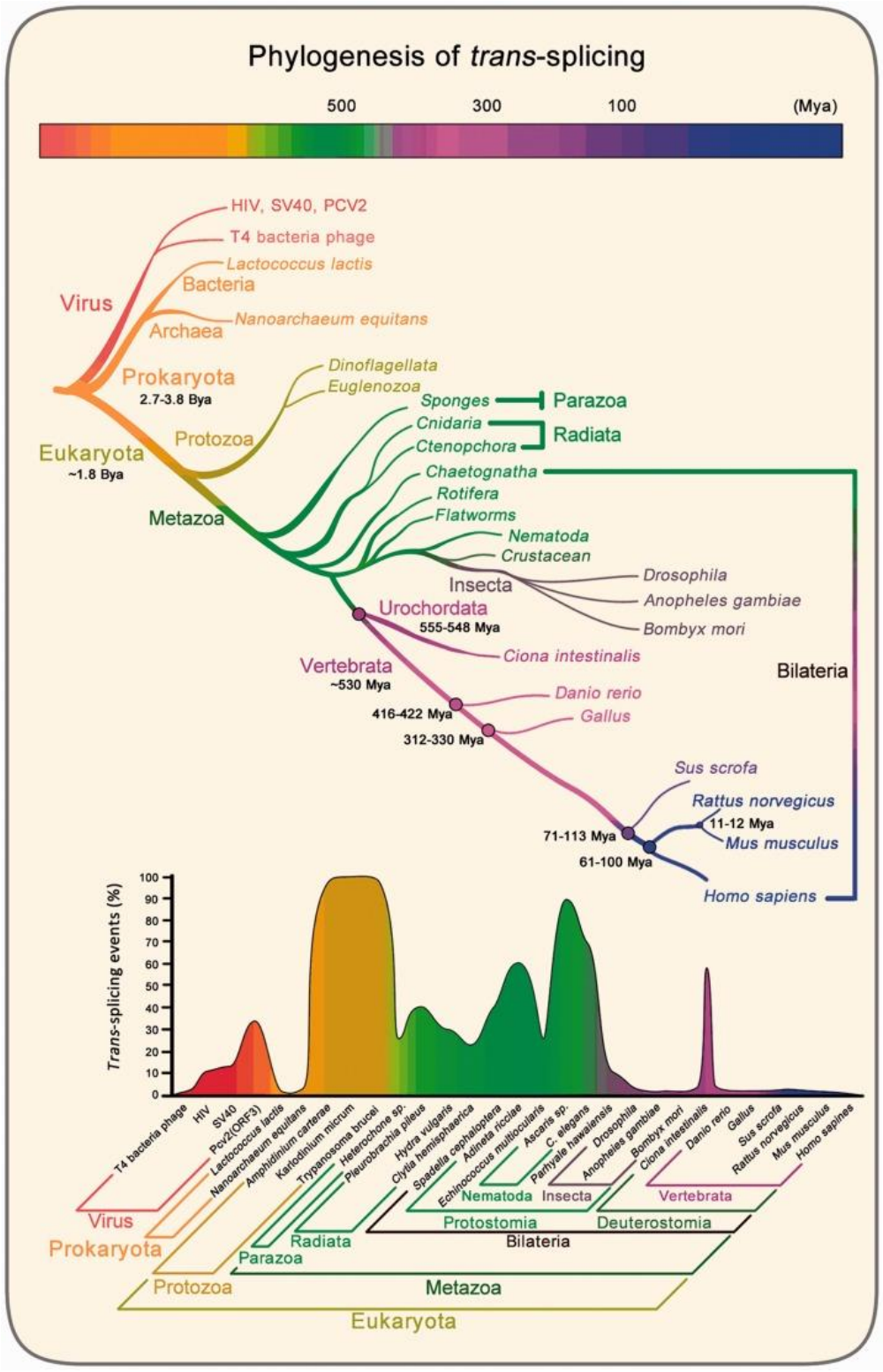


Figure 29: ***Phylogenetic analysis of trans-splicing events.*** *Evolutionary tree and time scale refer to Benton et al*[174]*. Ba, billion years ago; Ma, million years ago. In the lower panel, the percentage of trans-splicing events and trans-spliced gene numbers are relative to the total amount of gene numbers for a species. Extracted from Lei et al 2017*[175]

Table 5: ***Typical trans-splicing chimeras in Homo sapiens (Human).*** *Extracted from Lei et al*[169]

Involved genes or chimeras	Function description	References	Experiments verified
ATAC-1- Exon Xa/Xb	Gene expression regulation	(Yu et al. 1999)	RT-PCR; RNase protection assay; Northern blot
ATAC-1-Ampr	Gain antibiotic resistance	(Hu et al. 2013)	RT-PCR; in vitro trans-splicing
CAMK2G-SRP72	Unknownn	(Breen et al. 1997)	PCR; Genetic mapping; Western blot; (?)
CDC2L2	Transcriptional regulation	(Jehan et al. 2007)	FISH; RT-PCR
C-myb	Proto-oncogene	(Vellard et al. 1991)	RT-PCR; Sequencing
CoAA-RBM4	Regulate stem/progenitor cell differentiation	(Brooks et al. 2009)	RT-PCR; in vitro trans-splicing
CYCLIND1-TROP2	Cell growth	(Guerra et al. 2008)	RT-PCR; Northern blot; RNase protection assay (?)
CYP3A4, 5, 7, 43	Catalytic activity	(Finta et al. 2002)	RT-PCR; Northern blot; RNase protection assay
hER alpha	Gene expression regulation	(Flouriot et al. 2002)	RT-PCR; Southern blot
JAZF1-JJAZ1	Anti-apoptotic protein	(Li et al. 2008)	RT-PCR; Southern blot; in vitro trans-splicing
PJA2-FER	Cancer biomarker	(Kawakami et al. 2013)	RT-PCR
PAX3-FOXO1	Cancer biomarker	(Yuan, Qin, et al. 2013)	RT-PCR; FISH
RGS12	G-protein signaling	(Chatterjee et al. 2000)	RT-PCR
Sp1	Transcription factor	(Takahara et al. 2000)	RT-PCR; Southern blot; RNase protection assay
ZC3HAV1L-CHMP1A	Genome rearrangement	(Fang et al. 2012)	RT-PCR (?)
AF4, AF9, ELL, ENL, MLL, ETV6, NUP98, RUNX1, EWSR1	DNA repair and chromosomal translocation	(Kowarz et al. 2012; Kowarz et al. 2011)	RT-PCR (?)
TMEM79-SMG5	Cancer biomarker	(Kannan et al. 2011)	PCR; Sequencing (?)
tsRMST	Pluripotency maintenance of hESCs	(Wu et al. 2014)	RT-PCR; RNase protection assay
TSNAX-DISC1	G1/S transition and endometrial carcinoma (EC) development	(Li et al. 2014)	PCR; RNA-Seq (?)

1.3.4 Application of *trans*-splicing in gene therapy

TS is an unusual process in vertebrates while being a promising therapeutic strategy for genetic disease. If there is a mutation in a given exon, it is feasible to remove the mutated exon by TS. The concept is to replace the mutated exon by a mutation-free version when exons are ligated together during TS. Indeed, TS can be used for therapeutic purposes, since it can modify mRNA sequences with another pre-mRNA carrying corrected exons. The first proof of concept emerged in 1999 with the use of an artificial pre-mRNA[176].

1.3.4.1 *Pre-trans-splicing molecule*

To perform TS for a therapeutic purpose, it is necessary to use exogenous pre-*trans*-splicing molecules (PTMs). A PTM is composed of (Figure 30):

- **A replacement cDNA**: corresponding to the mutation-free exon sequence. This coding sequence will be introduced into the final mature mRNA, and normal protein will be expressed with recovered functions.
- **An artificial intron:** exhibiting splicing signals, including 5'SS, BP and PTT, to perform and promote TS over *cis*-splicing. This sequence represents one of the main challenges in terms of TS performance because its donor site competes directly with the donor site of the endogenous pre-mRNA.
- **A binding domain (BD):** corresponding to an anti-parallel sequence, hybridizing with a specific pre-mRNA sequence. The BD is the other very important component since it drives the whole molecule to the specific location to perform TS, thus allowing a spatial proximity between the two RNA molecules. BD hybridization can also hide important signaling sequences on endogenous pre-mRNA as well as inducing folding changes. Taken together, BD is likely to disturb normal splicing signaling to favor TS. Thus, depending on the matching sequence, PTMs will be more or less efficient at promoting TS activity. So BD is believed to be the key element determining TS efficiency.

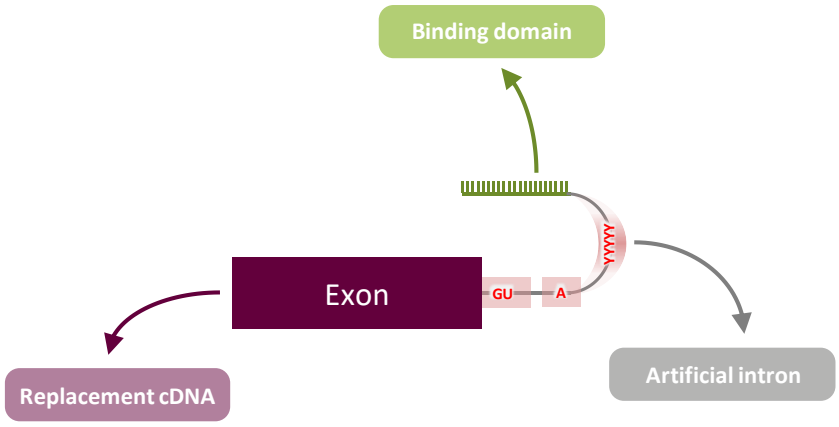


Figure 30: ***Schematic representation of a PTM.***

PTMs are able to promote specific exon exchange via a technology called spliceosome-mediated RNA *trans*-splicing (SmaRT). This technology has already been developed in several different diseases. However, these strategies are not equal in terms of TS efficiency, as depicted in Figure 31 and in a review that we have published in 2016 enclosed at the end of this manuscript[177].

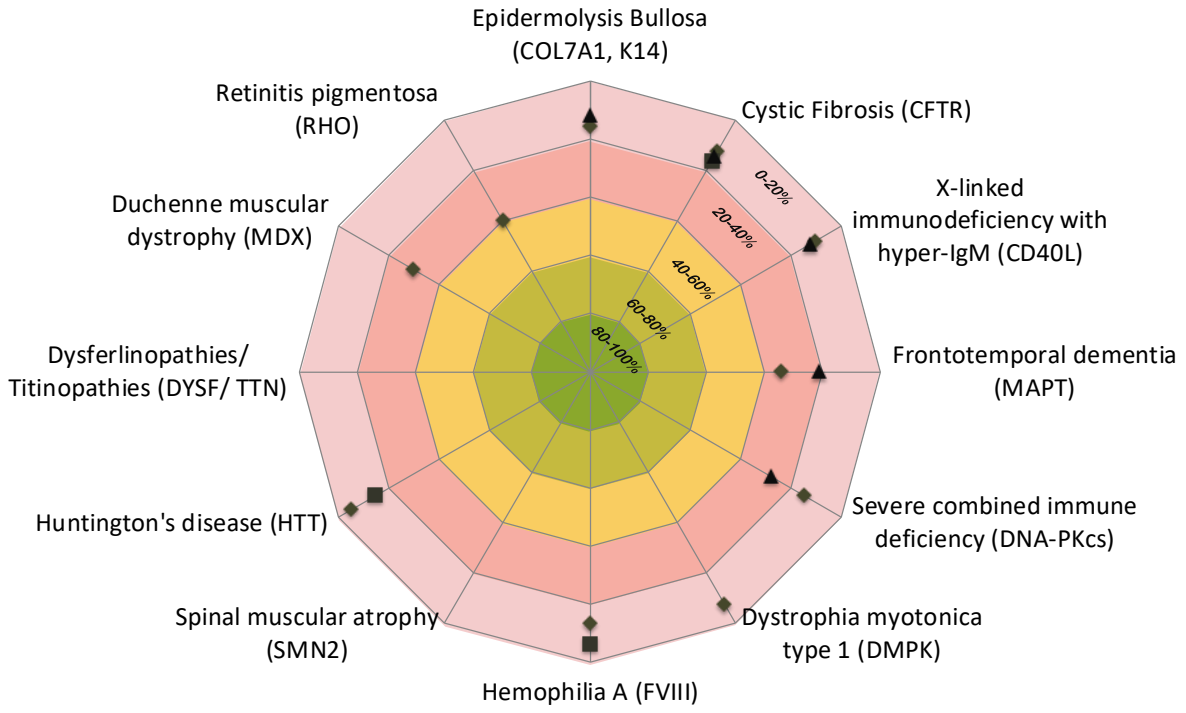


Figure 31 : ***Summary of the main publications reporting the implementation of a trans-splicing strategy for gene therapy of genetic diseases****. Each axis of the depicted polygon is dedicated to a different genetic pathology with a corresponding gene. The TS rate for each study is represented by a grey symbol and they are positioned in proportion to their score along the axes, with center corresponding to 100% and extremities to 0%. Intermediate % are represented by assorted colors. Adapted from Berger, Maire et al. (2016)[177]*

The therapeutic benefit provided by TS depends on the type of target mutation and whether the mutation is dominant or recessive. Indeed, TS is easier to develop for recessive diseases, because a small portion of normal protein can be enough to observe a beneficial effect. In this case, we can deal with different kind of situations, among them the following:

- the absence of gene expression
- a mutated protein with a complete loss of function
- a mutated protein with a partial loss of function

In the two first categories, the use of TS can be very efficient since it would reproduce the heterozygous case where the mutation-free allele compensates, completely or partially, the loss of function. Indeed, the severity of the pathology could be reduced, or even eliminated, using TS since conversion of a small fraction of mutant pre-mRNA could be enough to obtain significant therapeutic benefits (See for instance muscular dystrophy diseases in textbox page 77).

For example, dystrophic epidermolysis bullosa (DEB) has been extensively studied in TS experiments by the team of Eva Muraue. DEB is a recessive disease due to mutations in COL7A1 leading to defects or absence of type VII collagen protein with promising TS treatments. They have demonstrated that around 2% of corrected COL7A1 mRNA is enough to obtain 53% of type VII collagen recovery compared to controls (100%) associated with correction of the DEB phenotype[178].

Therapeutic benefit is much more challenging in cases of dominant diseases. In this kind of disease, the mutations most of the time lead to a toxic gain of function. In this case, there is expression of a mutated protein which exhibits pathological activity in the affected cells. At some point, a small portion of mutated protein can be sufficient to develop the disease. Thus, to apply a TS strategy in this case, it requires a significant rate of corrected proteins to reach a therapeutic threshold. Therapeutic benefit would result in a major reduction of mutated protein level. It is thus necessary to achieve a high correction of the mutant mRNA.

To illustrate, Retinitis pigmentosa (RP) is heterogenous genetic disease leading to visual impairment resulting from degeneration of cone photoreceptors. This disease is caused by multiple kind of mutation in the rhodopsin gene and a TS study has shown that 40% of mRNA correction in a severe dominant mutation context (P347L) is not sufficient to reduce degeneration[179].

1.3.4.2 *Advantages of trans-splicing*

Compared to previously described therapeutic strategies, TS exhibits several advantages:

- **Correction** of the mutant protein via a natural pre-mRNA process
- **Silencing** of mutant mRNA expression
- **Expression level preservation** in targeting endogenous mRNA, and thus holding the expression level on standard basis
- **Reduced side effects** since mRNA is short-lived, potential mis-targeting is transient
- **Suitable for large genes** with partial vectorization of the transgene

When TS occurs, it simultaneously induces mRNA correction and mutation elimination in exchanging exons. Since TS acts directly on pre-mRNAs, this exon exchange will not induce overexpression or complete knock-down of a given gene. Only their own levels of expression will define the available amount of proteins to correct.

Moreover, pre-mRNAs are transitory targets. So, potential off-targeting on mRNA will be transient, unlike for CRISPR-Cas9, for which the strategy is to target DNA with sustained modifications. Moreover, CRISPR-Cas9 is derived from a bacterial process and could have negative effects on inflammation processes when expressed in the brain, whereas TS already takes place in cells as a rare but innate process.

So, TS is very promising, in particular for very large genes. Because of vector limitations, large genes are complicated to target by gene therapy since their size often overwhelms the packaging of currently available gene transfer vectors such as AAV (adeno associated virus). Using TS, only a small portion is removed, thus solving this vector issue. Moreover, some genetic pathologies like RP or DEB have multiple genetic origins with different mutations located in the same gene. With TS, there is no need to develop specific molecules for each mutation; one PTM is sufficient to correct a large range of mutations.

Moreover, SmaRT has the capacity to repair all mutations in the coding sequence location. If mutations are located in the 5’ part of a given gene, the strategy is to perform 5’ replacement. In this case, the artificial intron is designed to join the 3’ end of its own replacement cDNA exon to the 5’ end of a given endogenous pre-mRNA exon. In the same manner, 3’ replacement and internal replacement can be achieved as depicted in Figure 32.

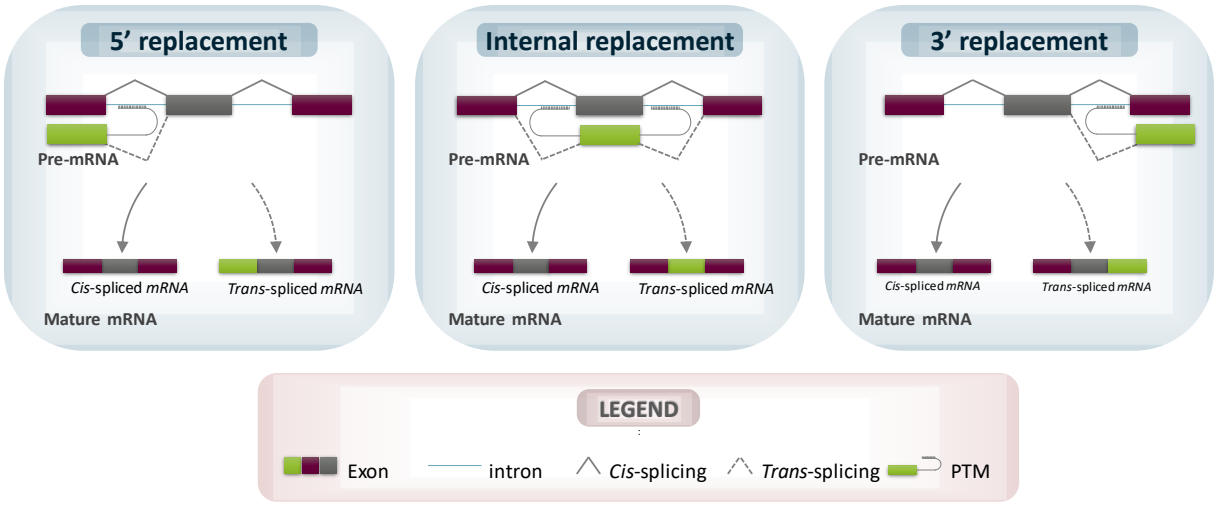


*Figure 32: **Different modes of action of trans-splicing**. Trans-splicing occurs between two splice sites located on two different pre-mRNAs. In SMaRT technology, an exogenous RNA, called the pre-mRNA trans-splicing molecule (PTM), is used to replace one or several exons of an endogenous pre-mRNA. Depending on the orientation of the PTM, it is possible to replace 3’ exon(s) (left), 5’ exon(s) (right), or even internal exon(s) (centre).*

To conclude, TS is a very promising strategy, exhibiting several advantages which could be useful for treating HD pathology.

To go further…

TS for DMD and BMD

A two tier system

DMD and BMD are muscle disorders characterized by a mutated dystrophin gene with different severity levels. Differences between these diseases are whether the dystrophin mutation disrupts the open reading frame (DMD) or stays within the open reading frame (BMD) of the gene. So DMD is much more severe, with an earlier age of onset and a higher rate of progression than BMD, while impairing the same tissue including skeletal muscles. Using a TS strategy, all mutations can be corrected with the same PTM but with a higher level of recovery in case of DMD than BMD[180].

Proof of TS feasibility has already been demonstrated on several genes, including *HTT*[181]. Recent studies led by Rindt et al [181,182] showed the possibility of correcting *HTT* mutations by targeting *HTT* intron 1 using different PTMs with different BDs, named "tethers" (Figure 33). They demonstrated improved TS activity when the branch point was targeted (Tether #4 and #9) reaching from 10% to 15% of TS rate. However, this quantification was performed using end-point PCR and no precision was provided about methodology of quantification, thus implying that band intensity was used, which is not particularly accurate. Despite this lack of precise quantification of the TS rate, this study provided the first evidence of a possible phenotype improvement such as reduced cell death after BDNF withdrawal and ATP level recovery in HD iPSC lines. These results confirmed that TS could be a promising approach in HD.

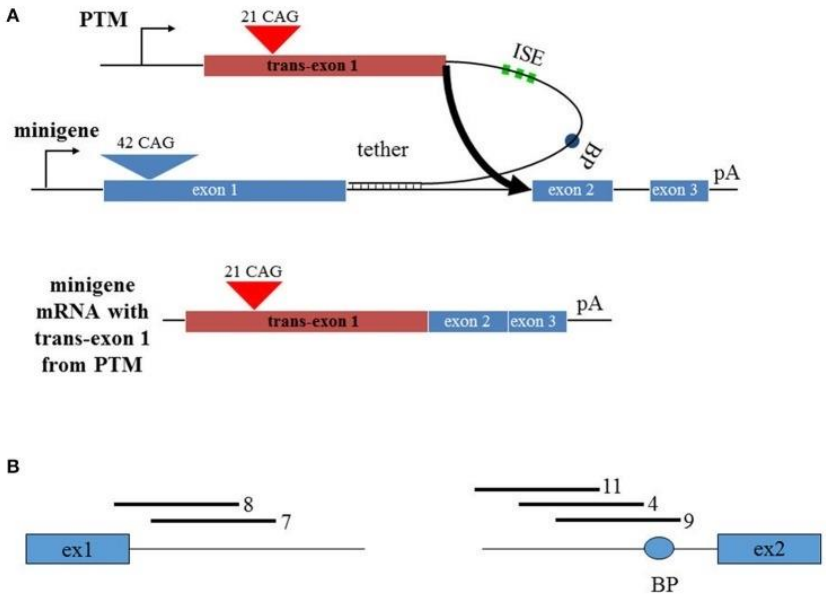


Figure 33: ***Trans-splicing in HTT intron 1.*** *(A) Schematic of PTM targeting. The PTM is driven by the CMV promoter and contains 21 CAG repeats within HTT exon 1. The PTM contains intronic splice enhancers (ISE), a branch point (BP), and a "tether" that targets the PTM to specific sequences within the HTT or mini-gene pre-mRNA. The HTT mini-gene is driven by the CMV promoter and consists of exon1-2-3 with intervening introns and a polyA signal. (B) Relative position of the PTM tethering elements for the PTMs. T7 and T8 target the 5' end of intron 1; T4, T9, and T11 target the 3' end of intron 1. Adapted from Rindt et al (2017)* [182]

The main issue is that differential TS rates are observed across TS studies giving differential benefits, which depend partly on the variety of PTM assessed. Indeed, it is known that TS efficiency relies on several parameters and BD seems to be the key element to trigger since its binding on the target intron can have multiple actions in the splicing process. It is known that the length, sequence composition and localization of BD all impact TS efficiency[183,184], but no clear evidence has emerged to elaborate a theory for rationale design of BD.

BD design faces constraints like the fact that PTMs targets pre-mRNAs, single stranded molecules which can be folded into complex conformations. Some intron regions in the pre-mRNA target can be

hidden, and thus unavailable for PTM hybridization. Unfortunately, RNA conformation is poorly predictable by bioinformatics tools for large RNA sequences. Adding to the complexity, splice sites such as ISE and ISS are sparse along the intron sequence, and these splice signals can be dysregulated by PTM hybridization. These regulatory sequences are furthermore complicated to localize. Thus prediction of efficient PTM targeting along an intron is very challenging.

Most of the TS studies that have been published to date have opted for different strategies to choose BD sequences driven by splicing hypotheses such as blocking U2 binding[185] or U1 binding[158,181] or targeting 5' and 3'SS[186,187], BP/PPT[158,182], exon/intron junction[188,189]. Other studies have designed their BDs randomly[184,190] and sometimes BD strategy is not even precisely mentioned[191–194]. From these studies, no convergent hypothesis has emerged, probably because a low number of PTMs was evaluated.

Indeed, in most studies only a small number of PTMs are assessed. Sometimes, target intron screening is implemented[184,190], but the most extensive study led by the team of Johann W. Bauer involved 10 different PTMs to screen intron 9 of the plectin gene, and 22 different PTMs to screen intron 30 with a high variation of TS activity across these introns[184].

Consequently, in this project, we decided to implement a "high content" screening of at least one hundred different PTMs for each *HTT* target intron to identify very efficient PTMs for therapeutic purposes. This project is structured around four main axes:

Axis 1: Construction of molecular tools appropriate for TS "high content" screening.

- Develop a fluorescent reporter system to implement a high content screening strategy for PTM libraries.

Axis 2: Evaluation of splicing activity in the fluorescent reporter system.

- Validate the splicing capacity of selected *HTT* introns in an artificial environment

Axis 3: High content screening of PTM libraries

- Identify hotspot regions on target *HTT* introns using mTurquoise2 fluorescence

Axis 4: Validation of TS rate at mRNA levels for the best PTM derived from the screen

- Quantify precisely the ratio of corrected mRNA

3.1 Bioinformatics analysis

3.1.1 Sequence alignment

Multiple sequence alignment of GFP, YFP, XFP and mTurquoise2 was performed using Clustal Omega version 1.2.4 (https://www.ebi.ac.uk; Tools > Multiple Sequence Alignment > Clustal Omega) and then colorized in Microsoft Word 2013.

Sequence alignment with amino acid equivalence was performed with the Blast tool of Serial cloner version 2.6.

Sequences are listed in Appendix 3.

3.1.2 MaxEnt scoring

MaxEnt is a probabilistic model of splice motifs including 5'SS and 3'SS. Developed by Yeo and Burge [195], MaxEnt scores were calculated using 9-mers comparisons with different splicing site models. These comparisons account for non-adjacent as well as adjacent nucleotide positions introduced in weight matrix models and analyzed in a probabilistic manner.

The MaxEnt score dedicated to 5'SS is generated by an online tool named: "MaxEntScan::score5ss for human 5' splice sites" (http://genes.mit.edu/burgelab/MaxEnt/XMaxEntscan_scoreseq.html).

3.2 Hybrid mini-gene cloning

3.2.1 GFP-derived protein cloning

The N-terminal part of YFP (yellow fluorescent protein) and C-terminal part of mTurquoise2 sequences were extracted from plasmids and ClaI restriction sites added at the extremities. ClaI was inserted in the middle and at either side of YFP and mTurquoise2 sequences using site-directed mutagenesis inducing a silent mutation (QuickChange kit, Agilent – see primers' pair #1 in Appendix 4).

3.2.2 Intron cloning

Depicted in Figure 34, *HTT* target introns were PCR amplified from genomic DNA of HEK293T cells, with primers specific for *HTT* intron sequences containing ClaI linkers (see primers #2, #3, #4, #5 in Appendix 4). Thanks to ClaI compatible ends, *HTT* introns were introduced into XFP. However, these constructions lead to two additional amino acids in *cis*-spliced products, resulting from an additional ClaI site at the YFP-mTurquoise2 boundary. It appears that this minor change completely inhibits fluorescence; therefore afterwards the ClaI site located upstream of the intron was removed by site-directed mutagenesis (QuickChange kit, Agilent - see primers' pairs #6, #7, #8, #9 in Appendix 4).

3.2.3 Lenti-virus expression cassette

ClaI-deleted hybrid mini-genes were inserted in a lentiviral shuttle plasmid by Gateway recombination (Life Technologies) under the transcriptional control of the short form of the ubiquitous promoter elongation factor 1 alpha (EF1). Transgene expression cassettes were inserted in reverse orientation to the lentiviral genome to avoid intron splicing during vector production.

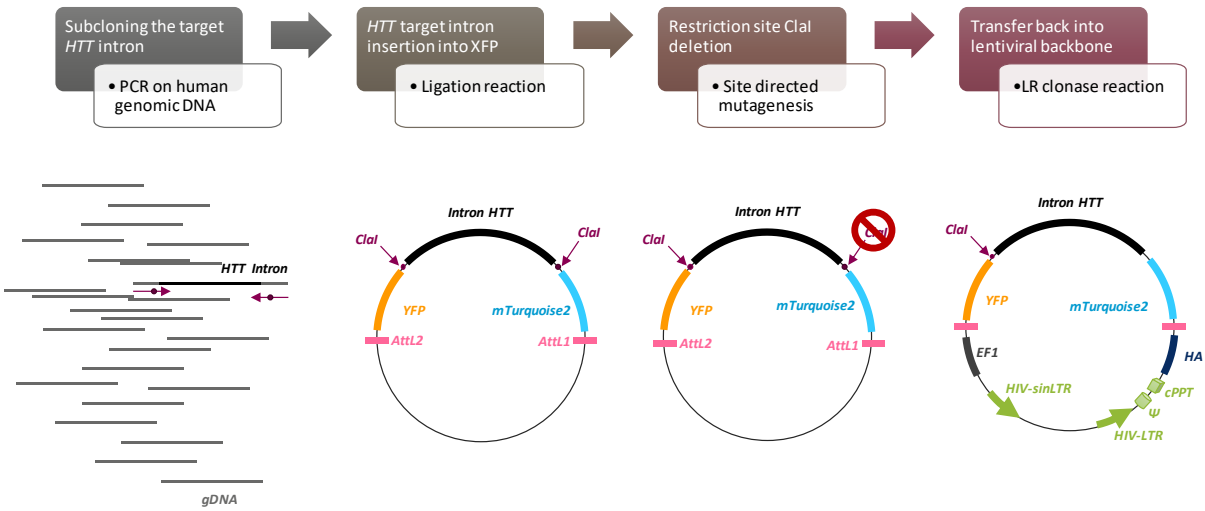


Figure 34: ***Hybrid mini*-gene construction *steps***

3.3 PTM libraries

3.3.1 PTM backbone

The PTM backbone is a pPur plasmid with ampicilin and puromycin resistance genes. PTM expression was under the control of the ubiquitous cytomegalovirus (CMV) promoter. It is composed of the N-terminal part of mTurquoise2, an artificial intron, and at the end, an AhdI restriction site was used to linearize the PTM backbone. TA cloning was performed after T-tailing of the vector backbone using Taq polymerase with dTTP at 72°C for 2 hours on an AhdI-linearized PTM backbone.

3.3.2 Intron 3 and 9

HTT introns were purified on an agarose gel after ClaI restriction from hybrid mini-gene intermediate constructs. As shown in Figure 35, purified introns were fragmented by ultrasonic wave mediated sonication on Covaris following the manufacturer's instructions (E-serie) to obtain 200bp long fragments (in collaboration with Genoscope at Evry). Sonicated products were purified (NucleoSpin Gel and PCR Clean-up, Macherey-Nagel) to remove fragments inferior to 20bp. Since mechanical fragmentation generates mostly over-hang ends, fragments were repaired and then A-tailed, to be

suitable for the subsequent TA cloning reaction (NEBNext® Ultra™ End Repair/dA-Tailing Module, New England Biolabs).

Random A-tailed *HTT* fragments were introduced in T-tailed vectors, and positive clones were selected on ampicillin supplemented medium. Ligations were validated by colony PCR to make sure of BD insertion, and by Sanger sequencing to identify anti-sense intron sequences. PTM DNA mini-preparations were provided by Eurofins after sequencing.

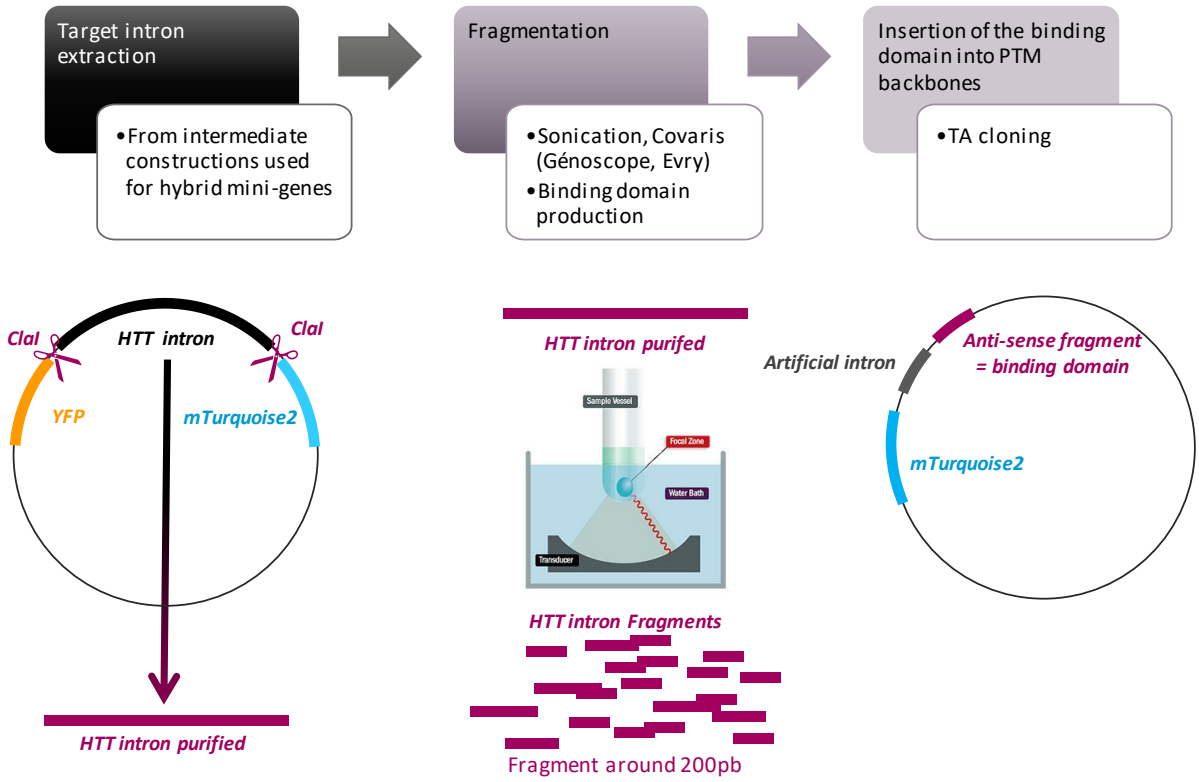


Figure 35: ***PTM cloning steps.***

3.3.3 Intron 20

For intron 20, BDs were generated by PCR from HEK293T genomic DNA (see primers' pair #10 in Appendix 4) using a DNA polymerase suitable for TA cloning by adding an A-tail at the 3' extremities of PCR products. The BD was then inserted in the same manner (TA cloning, colony PCR and Sanger sequencing).

3.3.4 PTM_0

PTMs without BD were also generated from empty PTM vector clones. This construct was used as a universal negative control for transfection.

3.3.5 Binding domain analysis

The first analysis with a bioanalyzer was systematically performed since sonication can vary from one sample to another, especially for short fragments. For subsequent sub-cloning in the PTM backbone, we selected only samples that presented a high diversity of fragments distributed on a regular gaussian shape centered on 200bp, as assessed by analysis on a DNA 100 bioanalyzer chip (Figure 50 p109).

Following transformation of the library in DH10ß highly competent cells, PCR colony screening was performed to make sure BD was inserted before sequencing. After having 100% of positive clones several times in a row, suggesting very efficient TA cloning, we excluded this additional validation, and we decided to include all BD lengths in libraries which were further analysed through sequencing.

3.4 Cell culture and transfection

HEK293T cells were grown in Dulbecco's modified Eagle's medium (DMEM, Gibco) supplemented with 10% heat-inactivated fetal bovine serum (Gibco), 100 U/mL penicillin and 100µg/mL streptomycin (Gibco). HEK-293T cells were incubated at 37°C with 5% CO_2.

For calcium phosphate transfection, HEK293T cells were plated in 24-well plates at a density of 100,000 cells/well. Cell concentrations were calculated using a Scepter 2.0 Cell Counter (Merck Millipore). The following day, transfections were performed with 500ng of plasmid DNA per well using calcium phosphate precipitation with HEPES buffer solution (Sigma).

3.4.1 PTM screening conditions

Co-transfection of the hybrid mini-gene and individual PTMs was performed in different ratios depending on XFP fluorescence:

- PTMi3 ➔ 1:10
- PTMi9 ➔ 1:1
- PTMi20 ➔ 1:1

Each PTM was tested in two wells from a common transfection mix. Each plate included at least:

- One non-transfected (NT) well
- One negative control well, using PTM_0
    - One positive control mixing full mTurquoise2 plasmid with PTM0. For each condition, we replaced the DNA amount dedicated to hybrid mini-gene by plasmid encoding for mTurquoise2 in order to mimic 100% of TS.

3.4.2 Fluorescence activated cell sorting (FACS)

Transfection was performed in 10cm^2 petri plates at a density of 3,000,000 cells/well. Cell concentrations were calculated using a Scepter 2.0 Cell Counter (Merck Millipore). The following day, transfections were performed with 16µg of plasmid DNA per well using calcium phosphate precipitation with HEPES buffer solution (Sigma). Before FACS, transfected cells were dissociated using trypsin and collected in PBS to be sorted.

3.5 Emission spectra

Emission spectra were acquired on transfected HEK293T cells fixed with 1% PFA 48h after transfection. We used two different excitation wavelengths, 405nm and 440nm. Mean fluorescence intensities were acquired from 450 nm to 595 nm in 5 nm steps. For each fluorescent protein, 8 regions of interest (ROI) were selected, each corresponding to brightly fluorescent cells. Mean fluorescence intensities were averaged and normalized to maximum signal of total ROIs as 100%.

3.6 Cytometry profile of XFP and mTurquoise2

Fluorescence profiles were assessed on transfected HEK293T cells 48h after transfection using Partec cytometers (Cyflow® Space) using the following configuration:

Channel	Lasers	Detectors
mTurquoise2	405 nm	455/50 nm
GFP	488 nm	536/40 nm

3.7 Epifluorescence acquisition

Fluorescence emitted from living transfected cells was acquired using a 10x objective on an inverted epifluorescence microscope (Zeiss), fitted with a temperature and CO_2 controlled chamber, and driven by the ExploraNova software suite. An XY automatic acquisition method was implemented on a motorized stage to have 16 pictures per well, in the manner shown in Figure 36 (Morpho Strider, Exploranova). Every 4 pictures, an auto-focus was performed on the bright field channel and mTurquoise2 fluorescence was observed on the CFP channel of the microscope mounted with an excitation filter of 436/20nm and an emission filter 480/40 nm. Acquisition parameters were always set at 800 ms exposure with a gain of 200.

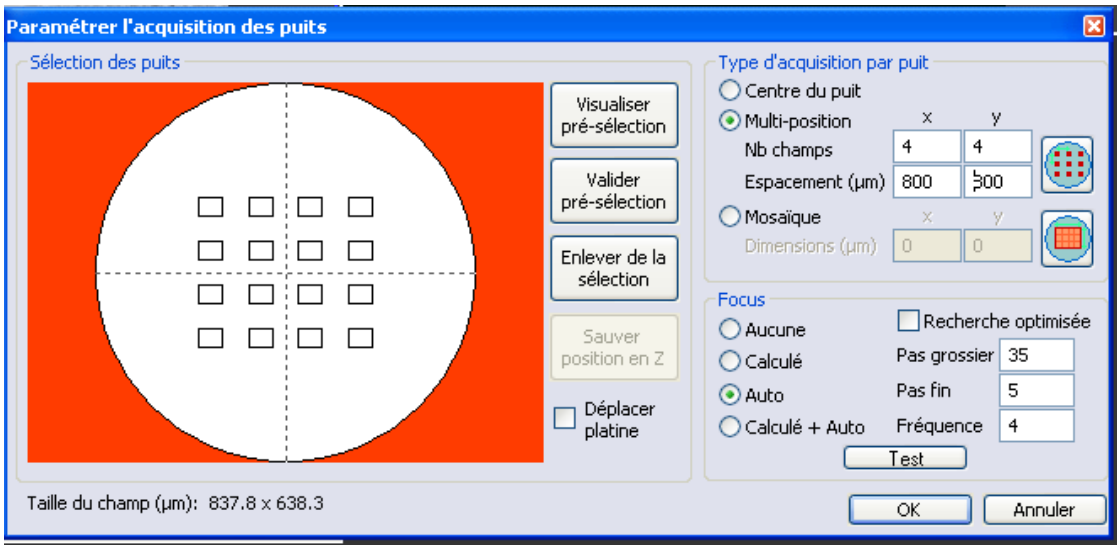


Figure 36: ***Screenshot of acquisition procedure.***

3.8 Epifluorescence data analysis

Pictures with non-specific fluorescence were excluded. Remaining ones were processed with ImageJ using a custom-made macro to automatically apply a given threshold based on NT noise, being negative controls with auto fluorescence background. From selected regions of interest determined by the threshold, the ImageJ macro also included automatic measurement of area and mean gray value (Figure 37).

```
Dialog.create("Avant de commencer");
Dialog.addNumber("combien d'image:", 400);
Dialog.show();
nbf=Dialog.getNumber();


run("Clear Results");

waitForUser("Ouvrir la première photo CFP");

for (i=0; i<=nbf-1; i++) {
setThreshold(37, 255, "over/under");
run("Create Selection");

run("Measure");
resetThreshold();
run("Select None");
run("Open Next");
}
run("Close All");

waitForUser("Enregistrer les résultats de CFP")
```

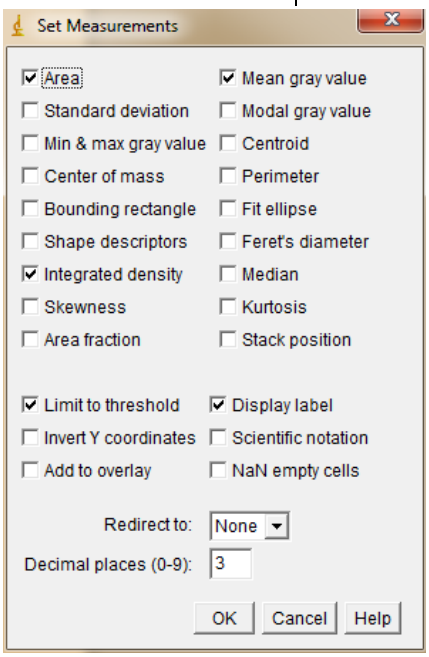


Figure 37: ***ImageJ Macro.***

3.8.1 PTM screening data processing

For each picture, we calculated the mean gray value and integrated density (1) which results in the multiplication of an ROI's area and mean gray value. Integrated values were transformed to a logarithmic scale (2) to obtain a normal distribution of datasets. Z-score transformation was then performed to reduce variability across plates and experiments (3). The mean gray value was also analyzed and transformed following steps (2) and (3) and is mentioned as MFI for "mean fluorescence intensity".

(1)

$$x_{ij} = Area \times Mean\ Gray\ Value$$

With $\boldsymbol{x_{ij}}$ = integrated density value of well i from plate j

(2)

$$x'_{ij} = \log_{e^1} x_{ij}$$

With $\boldsymbol{x'_{ij}}$ = log-transformation of $\boldsymbol{x_{ij}}$

(3)

$$x''_{ij} = \frac{x'_{ij} - \mu_j}{\sigma_j}$$

With $\boldsymbol{x''_{ij}}$ = final normalized output value

$\boldsymbol{\mu_j}$ = mean of x'_{ij} from plate j

$\boldsymbol{\sigma_j}$ = standard deviation of x'_{ij} from plate j

3.9 Quantification of mRNA TS rate

TS events being very rare, we needed to isolate cells exhibiting TS activity to detect TS products. Positive mTurquoise2 cells were selected using FACS in collaboration with Nathalie Déchamps and Jan Baijer from CEA-DSV-IRCM. From all events detected in samples by cytometer, the first gate was determined to isolate cell population according to FSC and SSC channels. The second gate determined singlet population (Figure 38, Appendix 5). The final gate was determined on PTM_0 samples to discriminate negative from positive cells using the following configurations.

Channel	Lasers	Detectors
mTurquoise2	408 nm	450/30 nm
GFP	488 nm	530/40 nm

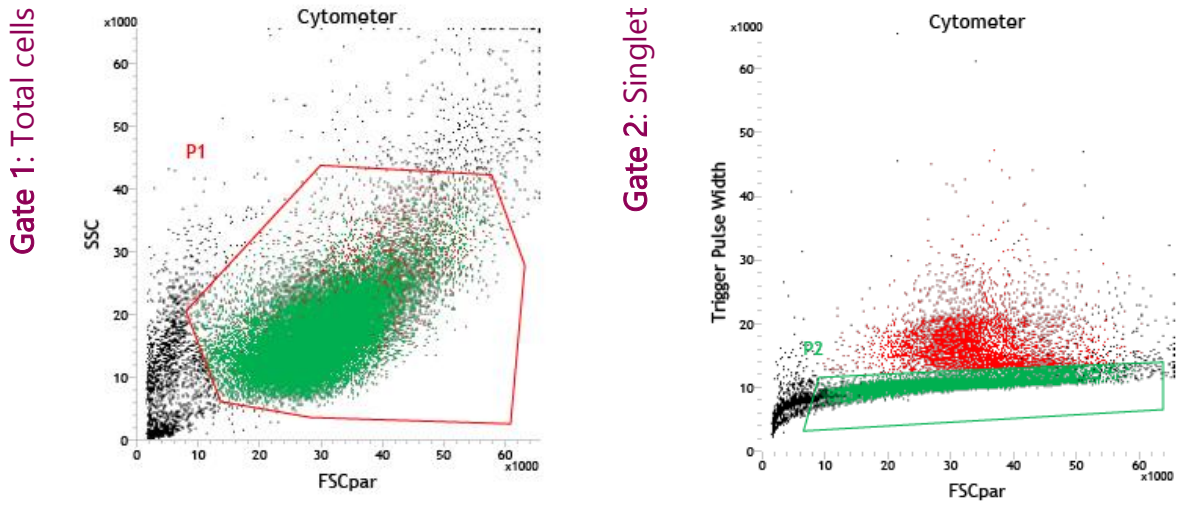


Figure 38: ***Gating methods to isolate and sort mTurquoise2 cells.*** *More details in Appendix 5*

After sorting, cells were collected and lysed in Trizol for RNA extraction. After the addition of chloroform, samples were centrifuged at 12,000 g for 15 min at RT. The aqueous phase was collected, and 1 volume of 70% ethanol was added. RNA samples were purified on a column (RNeasy micro kit, Qiagen), then reverse-transcribed into cDNA (SuperScript IV VILO Master Mix with ezDNase, Life Technologies). Both spliced products, XFP and mTurquoise2 sequences were amplified by end-point PCR (Hotfire, Solis BioDyne, see primers' pair #11 in Appendix 4). PstI and DrdI digestions were performed with respectively 20 units and 5 units with 250ng of purified PCR products at 37°C for 45 min followed by 10 min at 65°C. TS was estimated using the area under the curve of non-digested peaks for PstI and digested peaks for DrdI from electropherogram profiles obtained after migration on a DNA 1000 bioanalyzer chip (2100 expert system, Agilent Technologies).

3.10 Intron retention detection by quantitative PCR

Cells were transfected with each hybrid mini-gene construct and total mRNAs were extracted on a column (Nucleospin RNA II extraction kit, Macherey-Nagel) and reverse-transcribed into cDNA (SuperScript IV VILO Master Mix with ezDNase, Life Technologies). Then qPCR was performed to amplify exon/intron boundaries and expression levels were normalized to PPIA expression (see primers' pairs #12, #13, #14, #15 in Appendix 4)

Intron 3	109 bp	93 bp
Intron 9	163 bp	191 bp
Intron 17	143 bp	181 bp
Intron 20	348 bp	308 bp

Figure 39: ***Schematic representation of qPCR strategy to assess intron retention and amplicon length.***

3.11 Statistical analysis

3.11.1 Correlation matrix

Pearson correlation coefficient "r" was calculated with the formula (1):

$$(1)\quad r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

x = individual sample value from datasets "X"

$\bar{x}$ = mean sample value from datasets "X"

y = individual sample value from datasets "Y"

$\bar{y}$ = mean sample value from datasets "Y"

To determine p-values, **t-values** are required and are calculated with formula (2)

$$(2)\quad t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

r = Pearson correlation coefficient

n = number of observations

The p-value was then determined using the Excel function LOI.STUDENT.BILATERALE(t,deg_freedom) with n – 2 degrees of freedom.

4.1 A Fluorescent reporter system allows easy detection of TS

4.1.1 Fluorescent reporter system

To implement PTM screening, we developed a fluorescent reporter system based on GFP-derived protein. It is composed of two different constructs depicted in Figure 40.

On one hand, we built a hybrid mini-gene (see Figure 34 in Materials and methods page 85) containing the N-terminal part of YFP and the C-terminal part of mTurquoise2 protein. We introduced in between these two artificial exons, a given *HTT* intron to make a hybrid mini-gene with *cis*-splicing activity.

On the other hand, we generated libraries of PTMs (see Figure 35 in Materials and methods page 86) encoding for the N-terminal part of mTurquoise2 protein, thus leading to the full mTurquoise2 ORF (open reading frame) when the hybrid mini-gene is *trans*-spliced.

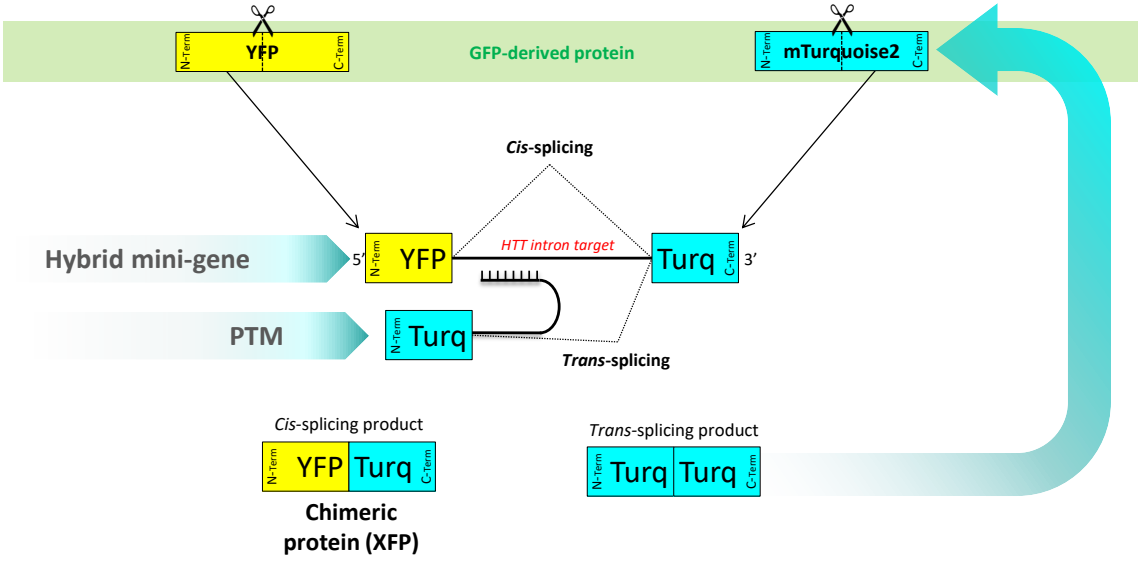


*Figure 40: **Schematic representation of the fluorescent reporter system.** YFP and mTurquoise2 are derived from GFP. The hybrid mini-gene has 2 artificial exons: within 5' the sequence encoding for the half YFP sequence and in 3' the half of mTurquoise2 sequence. The target HTT intron is introduced in between these exons, thus providing a splicing property to the hybrid mini-gene. The cis-spliced product of the hybrid mini-gene leads to a chimeric protein (XFP) which results in the association of the N-terminal sequence of YFP and C-terminal sequence of mTurquoise2. In this fluorescent reporter system, the PTM is designed to carry the complementary 5' part of the mTurquoise2 sequence. When PTM induces TS, it leads to the association of the two halves of mTurquoise2, thus resulting in the recovery of standard mTurquoise2 protein (and fluorescence).*

Since YFP and mTurquoise2 are derived from GFP with high sequence homology (Figure 41A), *cis*-spliced mRNA leads to a chimeric protein (XFP) which is able to emit fluorescence (Figure 42B). However, small changes in amino acid sequence result in drastic changes in fluorescence properties. Indeed, YFP and mTurquoise2 share 97.49% of their nucleotide sequence (699/717). As depicted in Figure 41B, nucleotide changes are not continuous and induce a small shift resulting in 5 differential

amino acids. XFP corresponds to the first 390 nucleotides of YFP, it has two differential amino acids compared to YFP and 3 compared to mTurquoise2. Interestingly, these minor changes in terms of amino acids lead to important differences of their fluorescence properties.

A

```
GFP             ATGGTGAGCAAGGGCGAGGAGCTGTTCACCGGGGTGGTGCCCATCCTGGTCGAGCTGGAC
YFP             atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggac
XFP             atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggac
mTurquoise2     atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggac
                ************************************************************

GFP             GGCGACGTAAACGGCCACAAGTTCAGCGTGTCCGGCGAGGGCGAGGGCGATGCCACCTAC
YFP             ggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgccacctac
XFP             ggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgccacctac
mTurquoise2     ggcgacgtaaacggccacaagttcagcgtgtccggcgagggcgagggcgatgccacctac
                ************************************************************

GFP             GGCAAGCTGACCCTGAAGTTCATCTGCACCACCGGCAAGCTGCCCGTGCCCTGGCCCACC
YFP             ggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccacc
XFP             ggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccacc
mTurquoise2     ggcaagctgaccctgaagttcatctgcaccaccggcaagctgcccgtgccctggcccacc
                ************************************************************

GFP             CTCGTGACCACCCTGACCTACGGCGTGCAGTGCTTCAGCCGCTACCCCGACCACATGAAG
YFP             ctcgtgaccaccttcggctacggcctgcagtgcttcgcccgctaccccgaccacatgaag
XFP             ctcgtgaccaccttcggctacggcctgcagtgcttcgcccgctaccccgaccacatgaag
mTurquoise2     ctcgtgaccaccctgtcctggggcgtgcagtgcttcgcccgctaccccgaccacatgaag
                ************ *   **  *** ***********  **********************

GFP             CAGCACGACTTCTTCAAGTCCGCCATGCCCGAAGGCTACGTCCAGGAGCGCACCATCTTC
YFP             cagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttc
XFP             cagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttc
mTurquoise2     cagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttc
                ************************************************************

GFP             TTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGGGCGACACCCTG
YFP             ttcaaggacgacggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctg
XFP             ttcaaggacgacggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctg
mTurquoise2     ttcaaggacgacggcaactacaagacccgcgccgaggtgaagttcgagggcgacaccctg
                ************************************************************

GFP             GTGAACCGCATCGAGCTGAAGGGCATCGACTTCAAGGAGGACGGCAACATCCTGGGGCAC
YFP             gtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcac
XFP             gtgaaccgcatcgagctgaagggcatcgatttcaaggaggacggcaacatcctggggcac
mTurquoise2     gtgaaccgcatcgagctgaagggcatcgacttcaaggaggacggcaacatcctggggcac
                ************************** ******************************

GFP             AAGCTGGAGTACAACTACAACAGCCACAACGTCTATATCATGGCCGACAAGCAGAAGAAC
YFP             aagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcagaagaac
XFP             aagctggagtacaactactttagcgacaacgtctatatcaccgccgacaagcagaagaac
mTurquoise2     aagctggagtacaactactttagcgacaacgtctatatcaccgccgacaagcagaagaac
                ******************   *** ***************  ******************

GFP             GGCATCAAGGTGAACTTCAAGATCCGCCACAACATCGAGGACGGCAGCGTGCAGCTCGCC
YFP             ggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgcc
XFP             ggcatcaaggccaacttcaagatccgccacaacatcgaggacggcggcgtgcagctcgcc
mTurquoise2     ggcatcaaggccaacttcaagatccgccacaacatcgaggacggcggcgtgcagctcgcc
                **********  ********************************** **************

GFP             GACCACTACCAGCAGAACACCCCCATCGGCGACGGCCCCGTGCTGCTGCCCGACAACCAC
YFP             gaccactaccagcagaacaccccatcggcgacggccccgtgctgctgcccgacaaccac
XFP             gaccactaccagcagaacaccccatcggcgacggccccgtgctgctgcccgacaaccac
mTurquoise2     gaccactaccagcagaacaccccatcggcgacggccccgtgctgctgcccgacaaccac
                ************************************************************

GFP             TACCTGAGCACCCAGTCCGCCCTGAGCAAAGACCCCAACGAGAAGCGCGATCACATGGTC
YFP             tacctgagctaccagtccaagctgagcaaagaccccaacgagaagcgcgatcacatggtc
XFP             tacctgagcacccagtccaagctgagcaaagaccccaacgagaagcgcgatcacatggtc
mTurquoise2     tacctgagcacccagtccaagctgagcaaagaccccaacgagaagcgcgatcacatggtc
                *********  *******   ****************************************

GFP             CTGCTGGAGTTCGTGACCGCCGCCGGGATCACTCTCGGCATGGACGAGCTGTACAAG
YFP             ctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaag
XFP             ctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaag
mTurquoise2     ctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtacaag
                ************************************************************
```

B

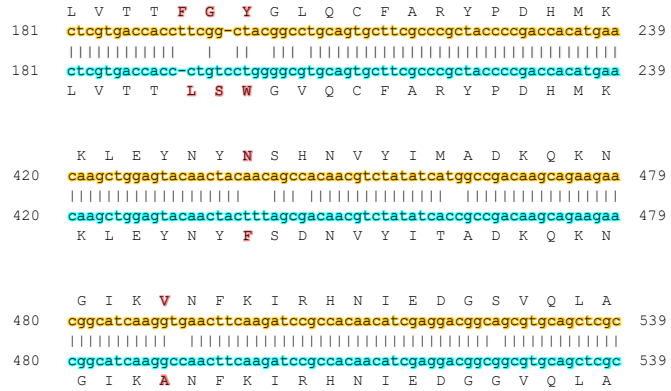
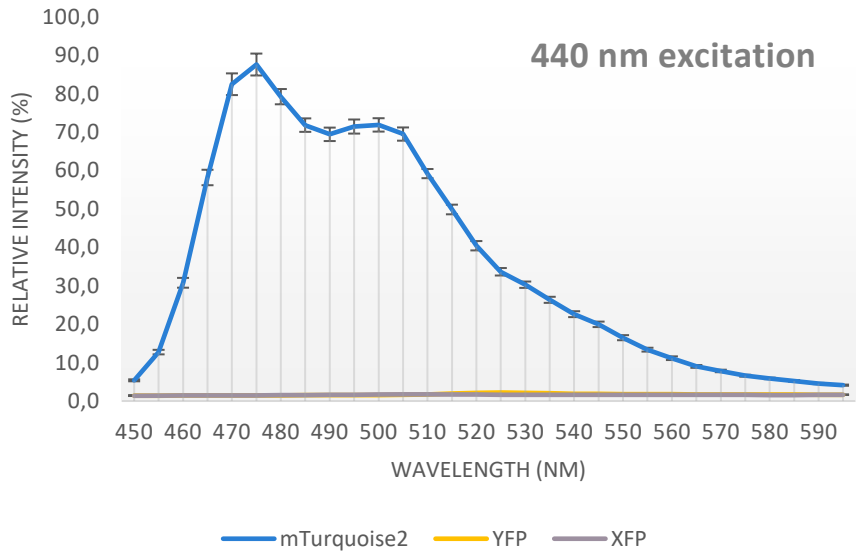


Figure 41: ***Sequence alignment of the fluorescent proteins.*** *(A) Multiple sequence alignment between GFP (in green), YFP (in yellow), XFP (in yellow at the beginning and in blue at the end) and mTurquoise2 (in blue). Lower lines represent nucleotide matching among aligned sequences with a positive match represented by a star. Alignment was performed using CLUSTAL O (1.2.4) multiple sequence alignment on https://www.ebi.ac.uk (Tools > Multiple Sequence Alignment > Clustal Omega). (B) Sequence alignment between YFP (in yellow) and mTurquoise2 (in blue) with matching amino acid chains. Differential amino acids between both sequences are highlighted in red. Vertical bars represent positive matching between nucleotides. Sequence alignments were generated by SerialCloner 2.6.1 (Align > Local Align Sequences). Raw sequences used for alignment are depicted in Appendix 3*

4.1.2 Fluorescence properties

Fluorescence properties were analyzed using epifluorescence on a spectral confocal microscope to characterize emission spectra (Figure 42). At 440nm excitation, mTurquoise2 is the only compound to emit fluorescence: mTurquoise2 is characterized by a large emission spectrum which peaks at 475nm. At 488nm, YFP and mTurquoise2 both emit fluorescence, but the signal is higher for YFP. Interestingly, a lower fluorescence intensity is emitted by XFP.

A



B

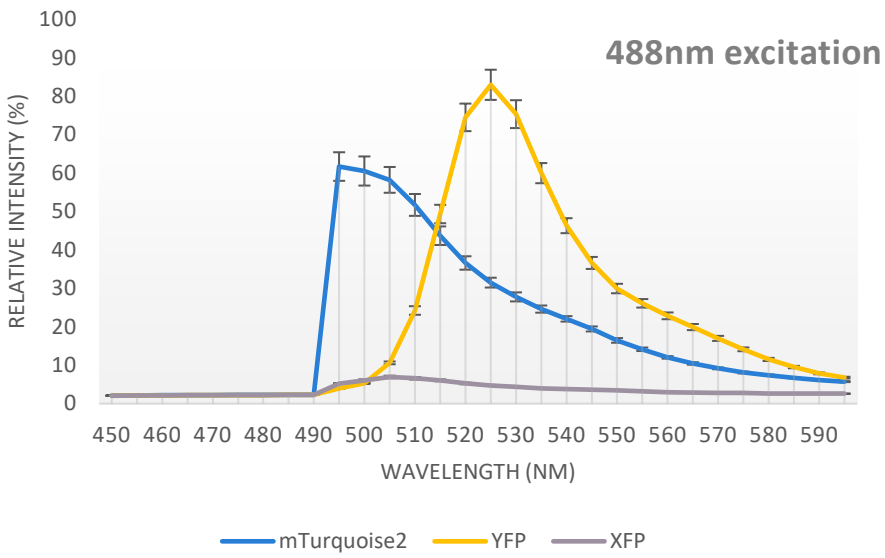


Figure 42: ***Emission spectra of mTurquoise2, YFP and XFP.*** *Emission spectra at (A) 440 nm and (B) 488 nm were acquired on a spectral confocal microscope. Mean fluorescence intensity (MFI) was quantified from regions of interest manually selected on fluorescent cells. Each point on the curves corresponds to the MFI averaged over six cells. Data are expressed as means ± SEM.*

These observations were validated by flow cytometry, demonstrating very distinct profiles with higher fluorescence intensity for mTurquoise2 (Figure 43). This data suggests that the XFP signal will be hidden by mTurquoise2 when TS occurs, so we decided to focus only on mTurquoise2 detection, suitable for detection of rare TS events.

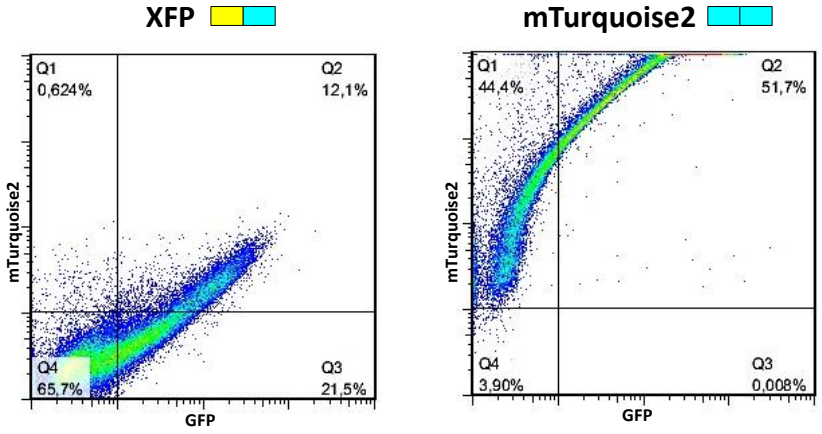


Figure 43: ***Cytometry profiles of XFP and mTurquoise2.*** *The mTurquoise2 channel is characterized by an excitation at 405 nm and 455/50 nm emission filter and GFP channel by an excitation at 488 nm and 536/40 nm excitation filter. Flow cytometry data were generated using CyFlow Space cytometer from Sysmex Partec company with Flomax 3.0 software. Dotplots were designed with FlowJo 7.6.*

4.2 Rationale for selection of *HTT* target introns

We adapted the TS fluorescence reporter system to *HTT* introns. These introns were selected based on multiple criteria (Table 6):

- **Intron location with regard to exon 1**: the cDNA replacement cassette of the PTMs should encode for all *HTT* exons located upstream of the *HTT* target intron. Consequently, when the target intron is far away from exon 1, the replacement cassette will be larger (see Table 6), thus impacting the PTM's capacity to be introduced in an AAV (adeno-associated virus) viral vector. AAVs are the most common viral vector for brain therapy applications because they exhibit a neuronal cell specificity with wide and stable expression, in particular in the striatum. However, their cloning capacity is limited to 4kb. PTM sequences are around 600bp without a replacement cassette and binding domain, so accordingly the replacement cassette must not be over 3kb. So we cannot target *HTT* introns located beyond the 20th.
- **Intron length**: hybrid mini-genes containing target *HTT* introns could have been introduced in lentiviral expression vectors allowing generation of stable cell lines. Lentiviral vectors have a limited cloning capacity, thus reducing target *HTT* intron length to less than 7kb. Cell lines were not generated in this project but they are still a possibility for assessing TS activity.
- **Splicing activity**: screening is mainly focused on binding domain impact on TS, but not only. Since introns exhibit differential splicing activity, we used parameters known to reveal differences in splicing signaling including:
    - the MaxEnt score which is an estimation of 5'SS strength (for more information refer to MaxEnt scoring page 84)
    - GC content which is likely to be involved in exon/intron recognition
    - intron/exon definition which is supposed to drive spliceosome sub-unit stabilization patterns
    - AGEZ (for AG Exclusion Zone) which is a characteristic of the 3' part of the intron in which "AG" motifs are excluded. These "AG" motifs are indeed involved in splicing signaling, and are systematically found in 3'SS. AGEZ corresponds to an extended region where no AG is found (See Figure 21 in Intron specifications page 57).

Taken together, these parameters led us to select *HTT* introns 3, 9, 17 and 20, whose characteristics are summarized in Table 6.

Table 6: ***Features of HTT target introns***

	Length (bp)	PTM replacement cDNA (bp)	GC %	5'SS in *HTT*	MaxEnt $_{HTT}$	5'SS in XFP	MaxEnt $_{XFP}$	Δ MaxEnt	Agez
Intron 3	4'429	784	47.6	aaaGTAGA	6.91	gatGTAAGA	5.46	1.45	43
Intron 9	1'456	2'196	43.3	tagGCAAGT	1.91	gatGCAAGT	1.36	0.55	36
Intron 17	93	2'711	35.5	cagGTAACG	10.52	gatGTAACG	4.99	5.53	33
Intron 20	258	3'013	29.8	gggGTAAGC	7.79	gatGTAAGC	6.16	1.63	34

HTT intron length, replacement cDNA and GC content were extracted using the AceView NCBI dataset (Homo sapiens complex locus HTT, encoding huntingtin > Introns and exons sequence) available at https://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/ For 5'SS in HTT and XFP, lowercase letters correspond to the 3' exon end and uppercase letters correspond to the 5' intron start. Maxent scores for HTT and XFP were generated using MaxEntScan::score5ss for human 5' splice sites online tool available at http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html with associated 9-mer sequences.

4.3 Splicing activity profiles

4.3.1 Cryptic Splice Site identification

The selected *HTT* introns were introduced in the hybrid mini-gene using the cloning strategy described in the Materials and methods section (Intron cloning page 84). Preliminary experiments revealed unexpected spliced-products deriving from the hybrid mini-gene constructs when amplified by PCR (Figure 44). This led us to identify a cryptic splicing site (CSS) in the 5' YFP sequence of the hybrid mini-gene, generating alternative splicing. CSSs are deleterious splicing sites, which are usually inactivated or used at a variable level in the cell when activated by mutations in the vicinity of the real splicing site as depicted in Introduction in Alternative splicing and exon skipping part, page 64. In our case, CSSs are activated by the introduction of an intron nearby, generating aberrant splicing. Most likely in XFP, CSS is activated, and it leads to two *cis*-splicing variants:

- a full XFP mRNA
- a 53nt-deleted XFP mRNA

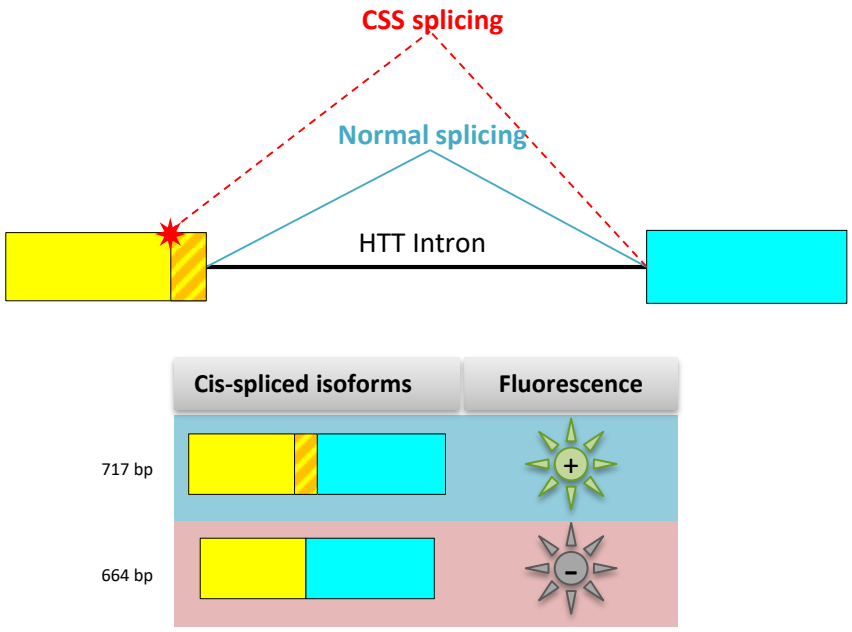
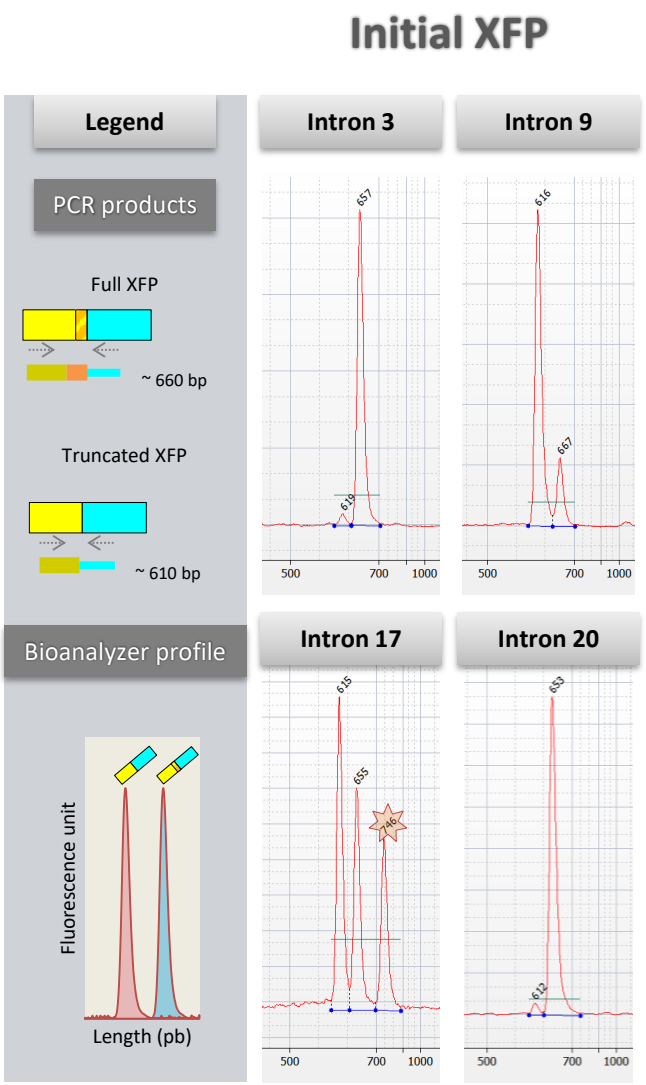


*Figure 44: **Schematic representation of CSS-mediated aberrant splicing**. (Upper) The hybrid mini-gene is pictured with a blue box as the mTurquoise2 exon and a yellow box as the YFP exon. A red star represents the CSS location on the YFP exon. The orange hatching represents the portion of YFP deleted when CSS induces splicing. (Lower) Both spliced products are represented with their fluorescence capacity (right) and their respective lengths (left).*

Located at position 334 from the beginning of the ORF, the CSS promotes alternative splicing by recruitment of the spliceosome machinery upstream from the conventional donor site. This leads to a mis-spliced mRNA, generating a non-fluorescent truncated XFP identified at the mRNA level by PCR (Figure 45). In introns 3, 9 and 20, two peaks were observed, validating mis-splicing's presence and thus CSS activation. Interestingly, the percentage of truncated XFP is intron specific, ranging from 3.6% for introns 3 and 20 to 84.3% for intron 9.

Intron 17 showed a distinct profile from the others, with 3 peaks. The first two were the same as the other introns, (truncated XFP and full XFP) whose respective proportions are depicted in Figure 45B. The last peak has the size of full XFP containing intron 17, suggesting a process of intron retention.

A



B

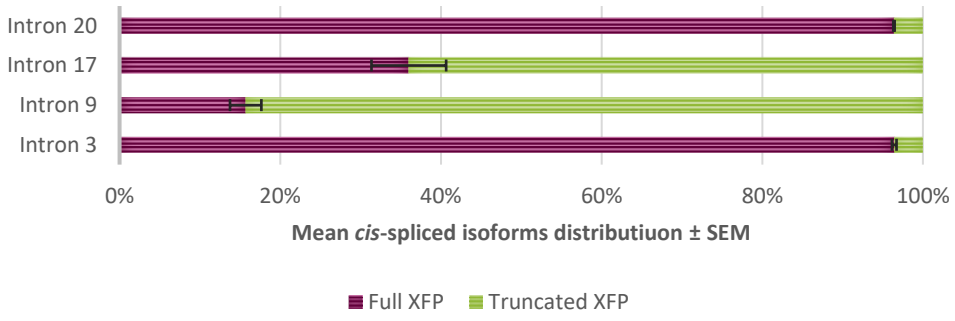


Figure 45: ***Identification of aberrant spliced products mediated by CSS activation in YFP exons.*** *(A) Bioanalyzer profile of PCR products representative of hybrid mini-gene mRNA level. Each peak represents mRNA variants resulting from different splicing events of the hybrid mini-gene. PCR product lengths are indicated on the top of each peak and on the horizontal axis in bp. The peak highlighted by a star in intron 17 corresponds to unspliced XFP resulting from an intron retention process. Around 10 ng per sample was used on a DNA 1000 Agilent Chip. Electropherograms were generated by 2100 Expert software. (B) Quantification of CSS activation rate from Bioanalyzer data. Graphical representation of full and truncated XFP distribution in %. Area under the curve, indicating DNA amount was averaged from 3 independent experiments. For intron 17, the 746 bp peak was excluded from percentage calculations.*

4.3.2 CSS correction

We have demonstrated above that CSS activation modifies normal splicing signaling and generates aberrant splicing events. In this context, TS activity could be impacted, and TS results will not be reliable. Before going further on the PTM screening project, it was necessary to remove the CSS site from our screening tools, i.e. the hybrid mini-genes.

For this purpose, we performed site-directed mutagenesis designed according to MaxEnt scores. We simulated different silent mutations targeting the CSS itself and selected the 9 mer motif with the lowest MaxEnt score (Figure 46A). The chosen motif is composed of the gaaGTCAAG sequence, with two nucleotide substitutions colored in red.

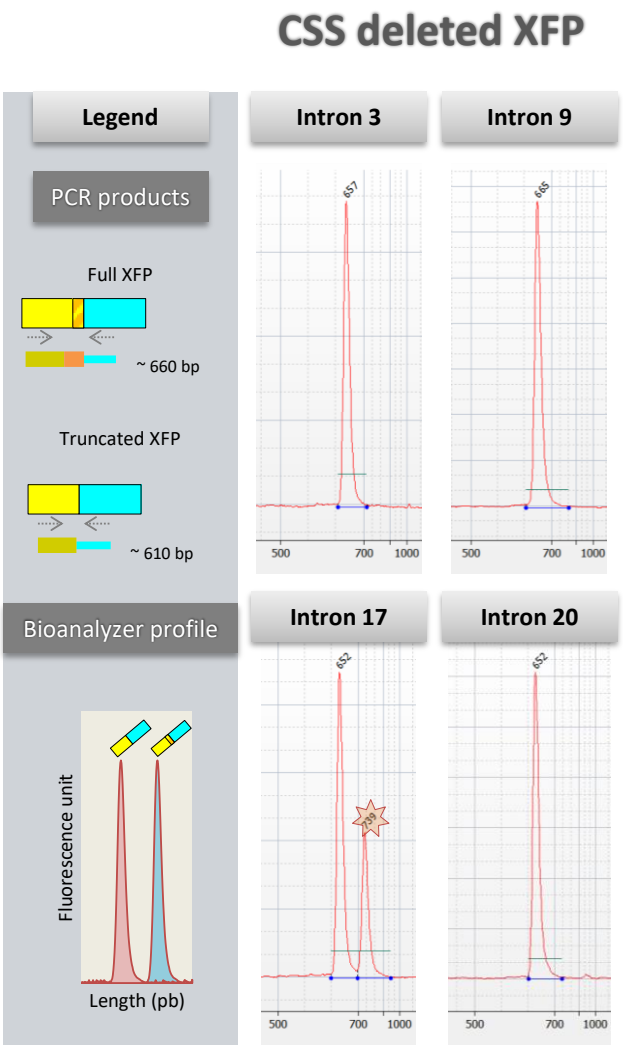
To validate the correction, we performed PCR on mutated hybrid mini-gene constructs for all *HTT* introns, and we observed a complete elimination of the 615bp mRNA variant (Figure 46B). However, bioanalyzer profiles of *HTT* intron 17 revealed an additional peak around 740bp, corresponding to unspliced mRNA, and representing about 17.5% of the total pool of XFP mRNAs. With intron 17 being very short compared to the others, spliced and unspliced products are very close in terms of length. In this condition, no PCR protocol optimization is possible to eliminate this unspliced product without deteriorating the spliced product signal.

Interestingly, this retention rate for intron 17 is similar in XFP with or without CSS correction, suggesting that intron retention is independent of CSS activation (Figure 46C).

A

CSS sequence	Maxent
gagGTGAAG	4,41
Simulation	
gaaGTGAAG	-3,83
gagGAAAAG	0,08
gagGTCAAG	0,13
gagGTAAAG	8,27
gaaGAAAAG	-7,36
gaaGTCAAG	-12,73
gaaGTAAAG	0,82

B



C

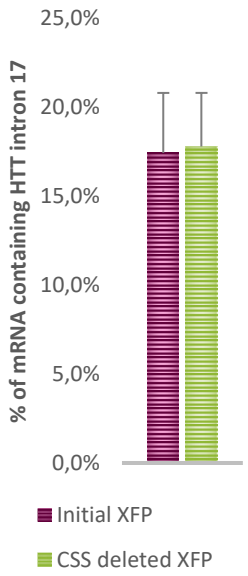
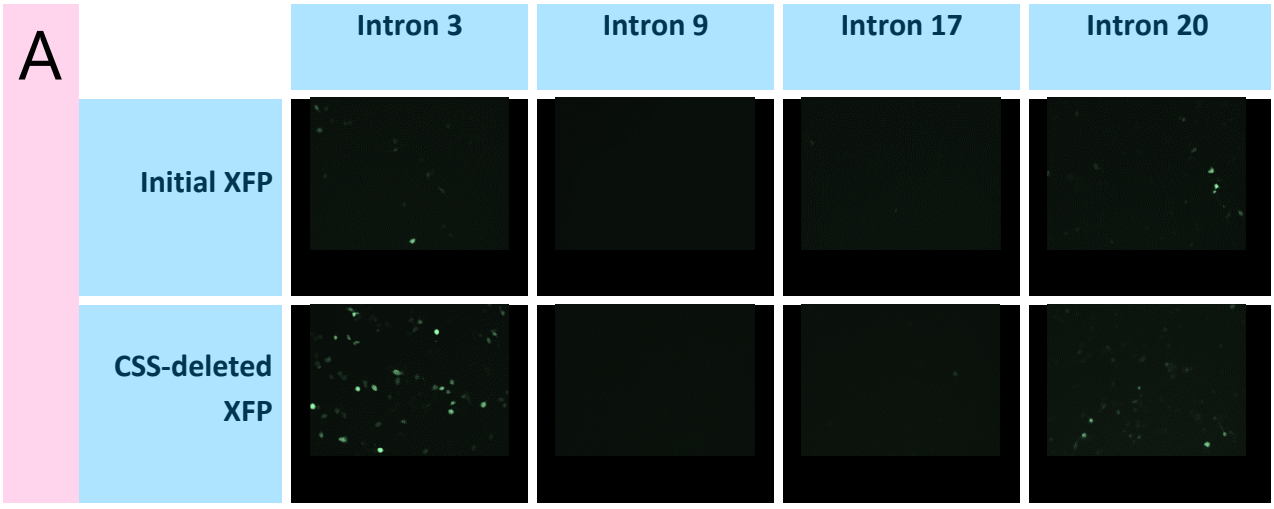


*Figure 46: **Validation of CSS correction.** (A) MaxEnt analysis on actual CSS motifs in initial XFP and on simulated motifs. 9 mer motifs were generated following genetic code (codon) in order to induce silent mutations. We selected the 9 mer motif colored in gray exhibiting the lowest MaxEnt score. Maxent scores for each simulated motif were generated using MaxEntScan::score5ss for human 5' splice sites online tool available at http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html (B) Bioanalyzer profile of PCR products from cDNA of corrected hybrid mini-gene mRNA extraction. PCR product lengths are indicated on the top of each peak and on the horizontal axis in bp. The peak highlighted by the star on intron 17 corresponds to unspliced XFP resulting from intron retention. Around 10 ng per sample was used on a DNA 1000 Agilent Chip and electropherograms were generated by 2100 Expert software. (C) Quantification of intron retention rate from bioanalyzer data. Area under the curve of the peak marked with a star, indicating DNA amount, was averaged from 3 independent experiments. Data are expressed as means ± SEM.*

Hybrid mini-gene constructs with and without CSS correction were also assessed at the protein level using their fluorescence properties (Figure 47A). We transfected HEK293T cells with both constructs of hybrid mini-genes, and 48h after transfection, fluorescence emitted from living transfected cells was observed using a 10x objective on an inverted epifluorescence microscope. We selected regions of interest using a threshold to focus on fluorescent cells and we observed important differences in terms of MFI (Figure 47B) and integrated density (Figure 47C) across introns. However, CSS correction seems to have no impact on fluorescence level except for:

- Intron 17 (p=0.036 t-Test) in mean fluorescence analysis, with higher signal when CSS is deleted, which is consistent with higher normal XFP expression. It is surprising that intron 9 with high CSS activation did not recover fluorescence intensity after correction.
- Intron 20 (NS) in integrated density with a reduction trend of fluorescence spread when CSS is deleted. We hypothesized that when CSS is activated, it increases splicing occurrence in both CSS and normal 5'SS leading to a better spreading of normal XFP expression. This synergistic action on overall XFP expression might be related to the MaxEnt score which is highest in intron 20.

We expected a low effect of CSS correction in introns 3 and 20 since truncated XFP was small compared to the full version in these conditions, and fluorescence recovery in introns 9 and 17 since higher rates of truncated XFP were observed. In intron 17, significant fluorescence recovery was observed but in other introns CSS deletion seems to have minimal impact, demonstrating that full XFP expression varies across introns. This variation can be mediated by intron dependent alternative splicing or other mechanisms such as 5'SS strength, regulating the standard XFP formation, thus impacting fluorescence level.



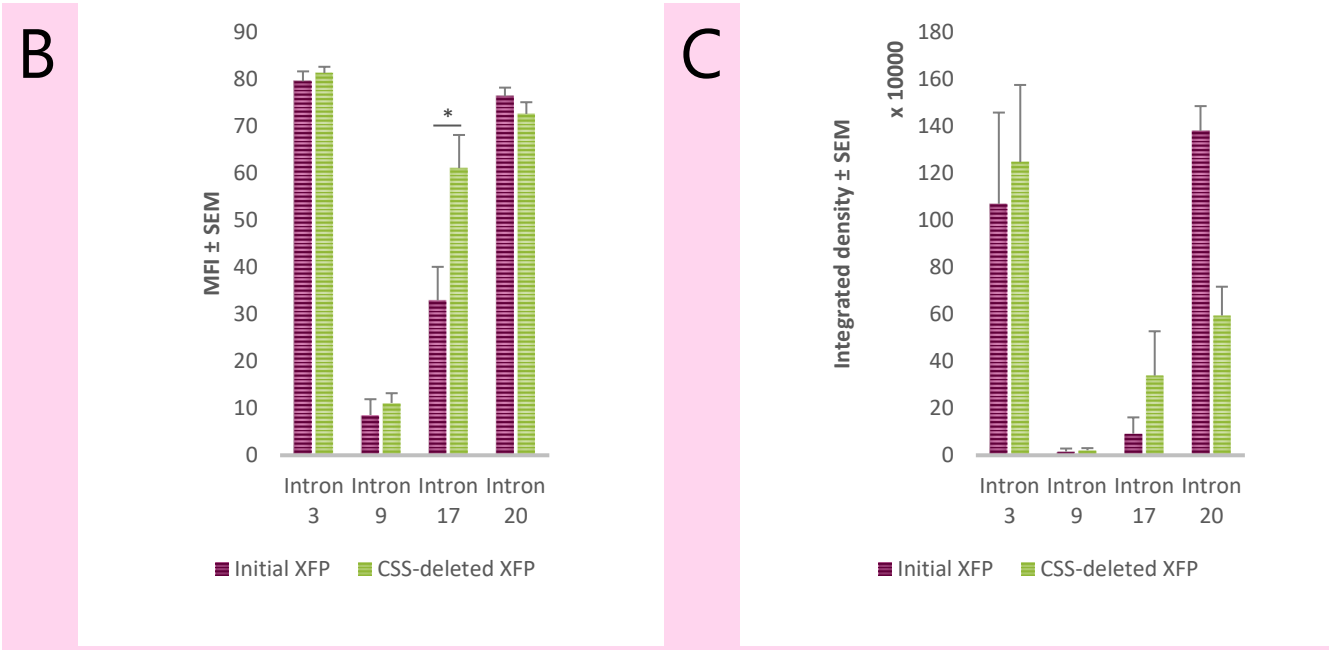


*Figure 47: **Evaluation of fluorescence intensity in initial and CSS-deleted XFP.** (A) Pictures of epifluorescence microscope acquisitions using the CFP channel (using a 480/40nm emission filter and a 436/20 nm excitation filter). Acquisition parameters were set at 800 ms exposure with a gain of 200 for all views. (B)Raw MFI and (C) raw integrated density quantification results from 3 independent experiments. Data are expressed as mean ± SEM. Student's t-test was performed (p-value: * < 0.05)*

4.3.3 Intron retention

Intron retention is a common and frequent process in mammals, with around 53% of potentially retained introns in human mRNAs[196]. Intron retention is not a disadvantageous process in the case of this TS study since it naturally occurs in cells. However, it is known to be highly regulated, so we wanted to evaluate if introns are also retained in our reporter system as suggested by the above PCR data of intron 17. Regarding fluorescence variation across introns, we wanted to evaluate the level of intron retention in the context of XFP. So, we performed qPCR to amplify the exon/intron boundaries at 5' and 3' as well, on mRNA extracted from transfected cells (as described in Figure 39 p92).

We observed that, in all conditions, 5' and 3' boundaries are expressed at similar levels, suggesting that introns are retained entirely in mRNAs as unspliced products (Figure 48A). Intron retention appears to occur with no significant differences across introns. CSS correction tends to induce higher intron retention in intron 9 as well as in intron 17 with higher differences. An inverse trend is observed for intron 20 but this result is due to a bias of amplification shown in Figure 48B. Indeed, when qPCR is analyzed on a DNA 1000 bioanalyzer chip, it is expected to have one single amplification product to validate quantification, such as observed for introns 3, 9 and 17. At the 5' junction PCR of intron 20, we observed multiple PCR products, and at the 3' junction, non-specific amplification of primer dimers was also detected (Figure 39, Figure 48B).

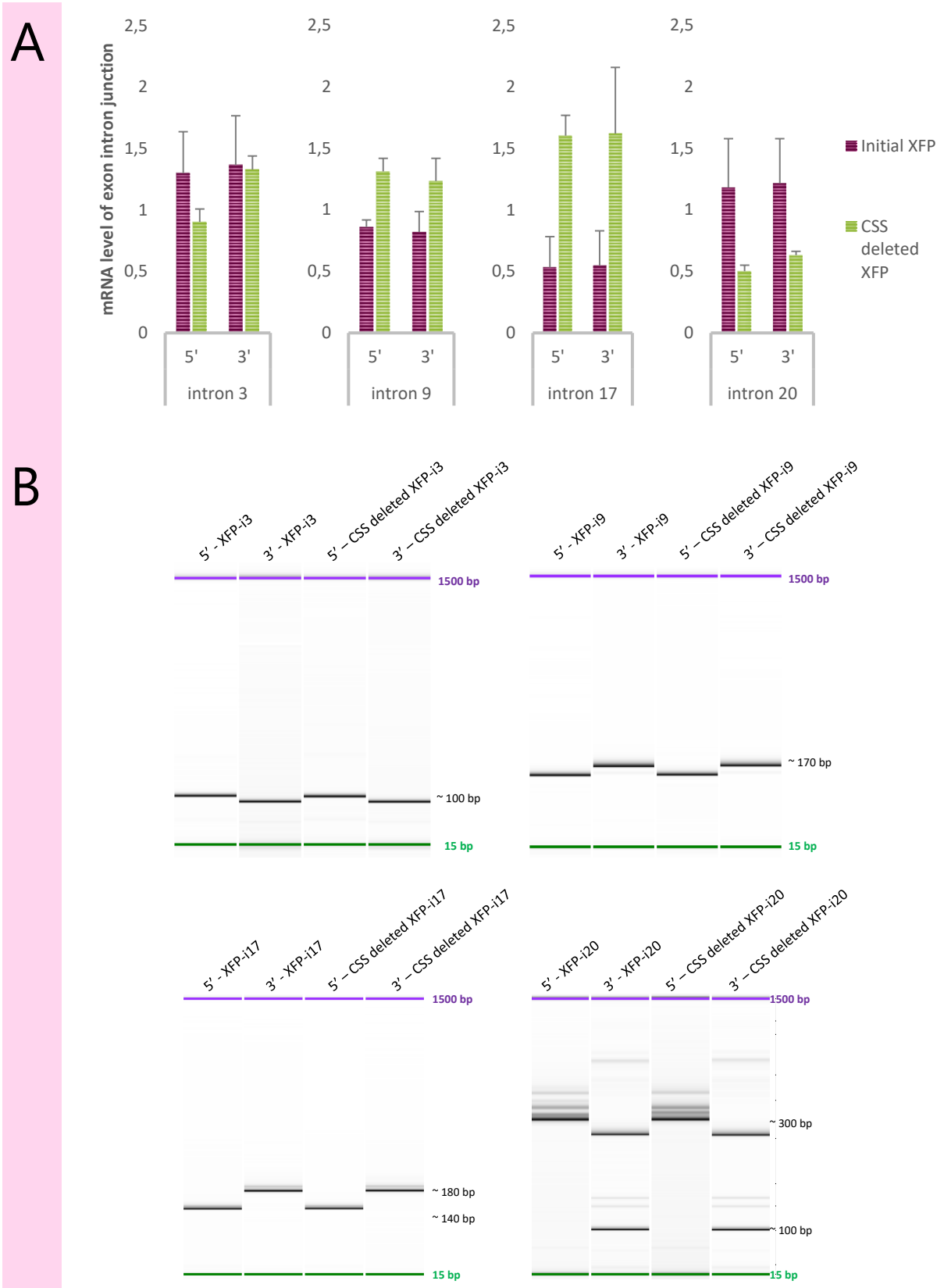


Figure 48: ***Intron retention in hybrid mini-genes*** *(A) Relative expression (normalized to PPIA mRNA levels) of exon/intron boundaries in 5' and in 3'. These results from 3 independent experiments. Data are expressed as mean ± SEM. (B) Gel representation of PCR products' profiles with 15bp and 1500bp migration markers. Around 10 ng per sample was used on a DNA 1000 Agilent Chip and electropherograms were generated by 2100 Expert software (Bioanalyzer).*

To finalize the splicing activity analysis of the hybrid mini-gene, we established a correlation matrix based on Pearson r scores to evaluate among the different splicing characteristics the possible relationship between XFP expression variability and other parameters related to:

- mRNA data (% of CSS activation, depicted as % of full XFP before correction)
- fluorescence (MFI, Integrated density)
- intron characteristics (GC content, MaxEnt score).

This correlation matrix (Figure 49A) reveals that $MFI_{initial\ XFP}$ and $integrated\ density_{initial\ XFP}$ are both positively correlated to CSS activation rate, suggesting that the use of fluorescence is an adequate indicator of XFP expression when CSS is activated. When CSS is corrected in XFP, the correlation between these items is weaker. This suggests that when CSS is activated, it is the major splicing mechanism getting the upper hand on the full XFP expression level.

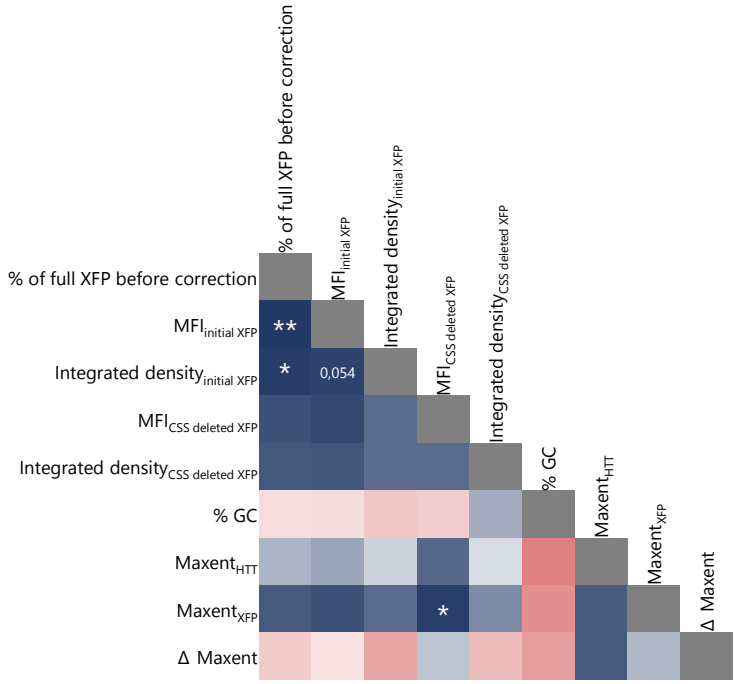
Moreover, we did not observe any fluorescence differences between XFP with or without the CSS correction (Figure 47). However, the correlation with MFI and Integrated density in both contexts is weak, indicating that those parameters show different aspects of XFP expression such as other minor splicing regulation pathways.

Basically, this correlation matrix highlights interesting aspects of the reporter system in a splicing context but also shows great variability of correlation strength.

However, we decided to exclude intron 17 from screening, since intron 17 exhibits a differential MaxEnt score between *HTT* and XFP, which could bias our estimation of intron 17 availability in an endogenous context.

So, we reproduced this correlation matrix without intron 17 data to evaluate the reliability of our parameters (Figure 49B). We observed more consistent results with higher overall correlation scores with slight changes, suggesting a more robust reliability of integrated density and MaxEnt scores to validate our reporter system.

A



B

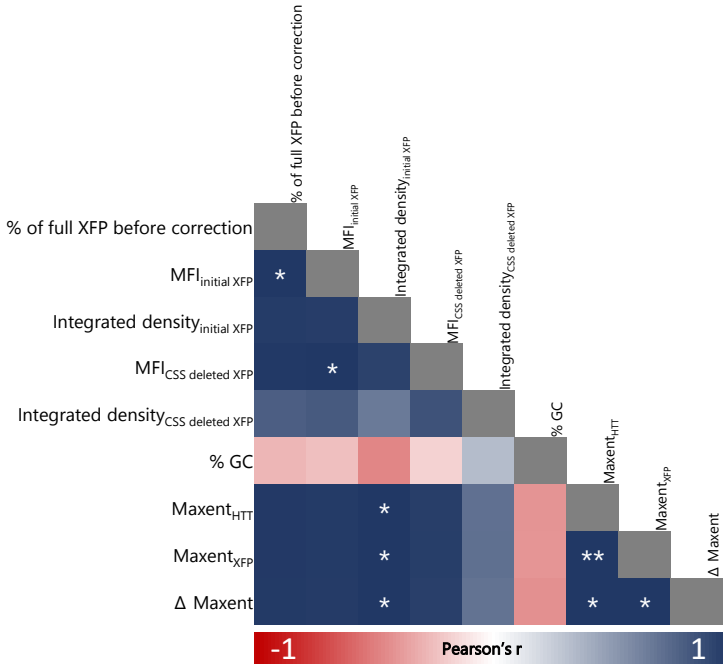


*Figure 49: **Correlation matrix with introns 3, 9, 17 and 20 data** in (A) and with intron 3, 9 and 20 data in (B). Scale from -1 to 1 with related colors from red to blue indicating respectively a trend of negative to positive correlation between datasets. Stars indicate significance level based on a p-value using Student's t-distribution (*< 0,05; **<0,01).*

4.4 PTM libraries for introns 3 and 9

PTM library construction was only performed for introns 3 and 9, being large enough to produce a high diversity of BDs. Intron 20 is 256bp long which is too close to the fragmentation target; no library was generated but a single PTM was assessed (See TS activity of intron 20 page 116).

4.4.1 PTMs contain random BDs with defined length

PTM libraries were generated using a universal PTM backbone containing random BDs from the different *HTT* target introns. These BDs were generated using focused-ultrasonication to obtain DNA fragments ranging from 150 bp to 200 bp long. We evaluated raw sonicated fragment length with a DNA 1000 Bioanalyzer chip. Electropherogram results reveal a gaussian-shaped distribution centered on 200bp (Figure 50). Fragments were then introduced in PTM backbones by TA cloning. Optimization of this ligation step allowed us to obtain 100% positive clones. The protocol for PTM library construction is detailed in Materials and methods (PTM libraries, page 85). Several hundred clones were generated with a mean of one hundred clones per µL of ligation.

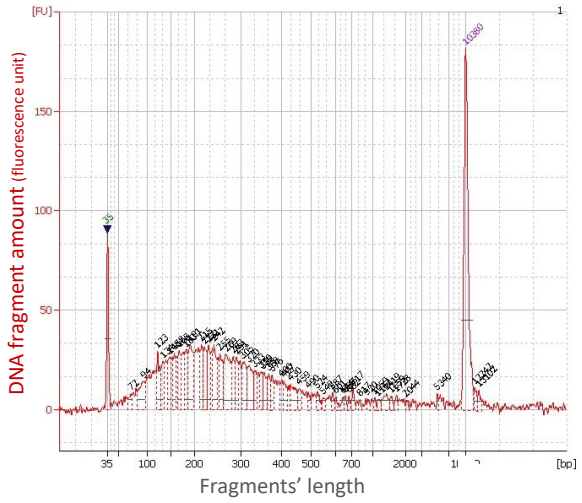


Figure 50: ***Electropherogram profile of raw sonicated intron (intron 3).*** *1µL of freshly sonicated intron was analyzed on a High Sensitivity DNA Bioanalyzer chip and an electropherogram was generated with 2100 Expert software with both peaks at 35 bp and 10,380 bp being respectively lower and upper markers.*

4.4.1 PTM libraries entirely cover *HTT* target introns

Once BDs were sub-cloned in the PTM backbone, 915 and 479 clones were sequenced for intron 3 and intron 9 respectively to assess different parameters (Table 7, Figure 51):

- insertion of the BD with respect to PTM backbone orientation, allowing (or not) PTM hybridization to pre-mRNA
- BD length
- location of the PTM hybridization site on the target intron
- coverage of the whole *HTT* associated intron by the different libraries.

Table 7: ***BD characteristics for intron 3 and intron 9***

	Orientation	# PTM (% of total)	Mean length (bp) ± SEM	Minimum length (bp)	Maximum length (bp)
Intron 3	*Sense*	*433 (47.3%)*	*150.3 ± 3,9*	*30*	*803*
	Anti-sense	**482 (52.7%)**	**139.3 ± 3,4**	**30**	**542**
Intron 9	*Sense*	*281 (58.7%)*	*140.1 ± 4,0*	*28*	*385*
	Anti-sense	**198 (41.3%)**	**139.6 ± 4,7**	**32**	**369**

Sequencing results (Table 7) revealed that binding domains integrated in the correct orientation to generate actual PTMs represent 52.7% (intron 3) and 41.3% (intron 9) of total sequenced PTM clones. Since intron 3 is 3-fold longer than intron 9, we sequenced more PTMs, obtaining a final intron 3 library 2.5-fold larger than that of intron 9. They have a mean BD length of 139 bp ranging from 30bp to 542 bp for intron 3 and 32 bp to 369 bp for intron 9. Moreover, libraries represent entirely corresponding introns, and so all intron sequences have matching PTMs (Figure 51A-B).

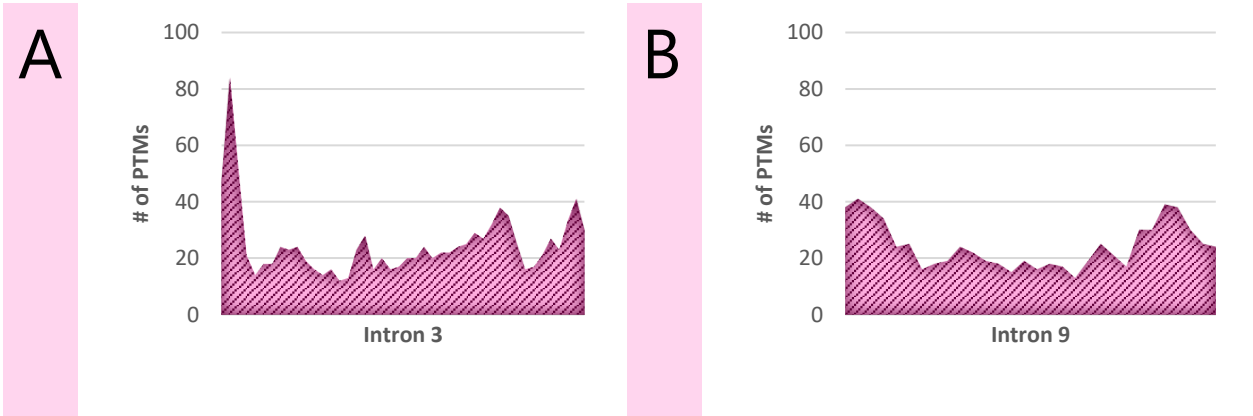


Figure 51: ***Sequencing results of PTM libraries.*** *(A) Sequencing results of BD of intron 3 (B) and intron 9 (C) representation of sequence targeting level along the HTT intron. Raw data were generated by Eurofins. Datasets were summarized using a custom made Galaxy pipeline, and NCBI alignment tools from fasta file.*

4.5 *HTT* subdomains exhibit differential TS potency in introns 3 and 9

To assess TS activity along *HTT* target introns, a selection of PTMs from each library was co-transfected in HEK293T cells together with the corresponding hybrid mini-gene. We used:

- as a negative control, co-transfection of PTM_0, corresponding to a PTM construct without BD, with the hybrid mini-gene.
- as a positive control, a plasmid encoding the mTurquoise2 cDNA without an intron, co-transfected with PTM_0.

Since hybrid mini-genes carrying *HTT* introns 3 and 9 exhibit differential fluorescence levels, screening was performed in two different conditions. In intron 3, co-transfection was performed with hybrid mini-gene and PTM in a 1:10 ratio, to reduce background noise emerging from basal levels of XFP fluorescence, whereas for intron 9, co-transfection was performed in a 1:1 ratio. 48h after transfection, mTurquoise2 fluorescence was observed using an inverted microscope equipped for epifluorescence detection. Compared to the positive control, the fluorescence signal corresponding to actual TS was overall very low (Figure 52). MFI and integrated density were quantified using an ImageJ macro made specifically for the project (See Figure 37 in Epifluorescence data analysis page 89). Data were transformed to a log scale suitable for fluorescence-based screening to have a normal distribution and then Z-scored, to be able to compare PTM scores derived from different experiments.

To identify TS hotspot(s) on target introns, we established subgroups of PTMs located in the same subdomains of each target intron, and averaged data of fluorescence for each subgroup. To that aim, target introns were divided into subdomains of equal length. We were thus able to deduce a "mean TS activity" for each subdomain (Figure 53 and Figure 55 for introns 3 and 9, respectively).

In intron 3, TS activity seemed to be detected with a similar pattern across subdomains between MFI and integrated density. Some subdomains showed significant (Student's t-test) TS activity compared to PTM_0. Several hot spots are likely to emerge including

- #18, #19 clusters, the two best subdomains in terms of integrated density
- #37, #38 clusters, the two best subdomains in terms of MFI
- #45, #46 clusters with lower scores than the other clusters in both integrated density and MFI analyses.

These subdomains exhibit significant TS activity and are all in the top 10 for MFI and integrated density analysis.

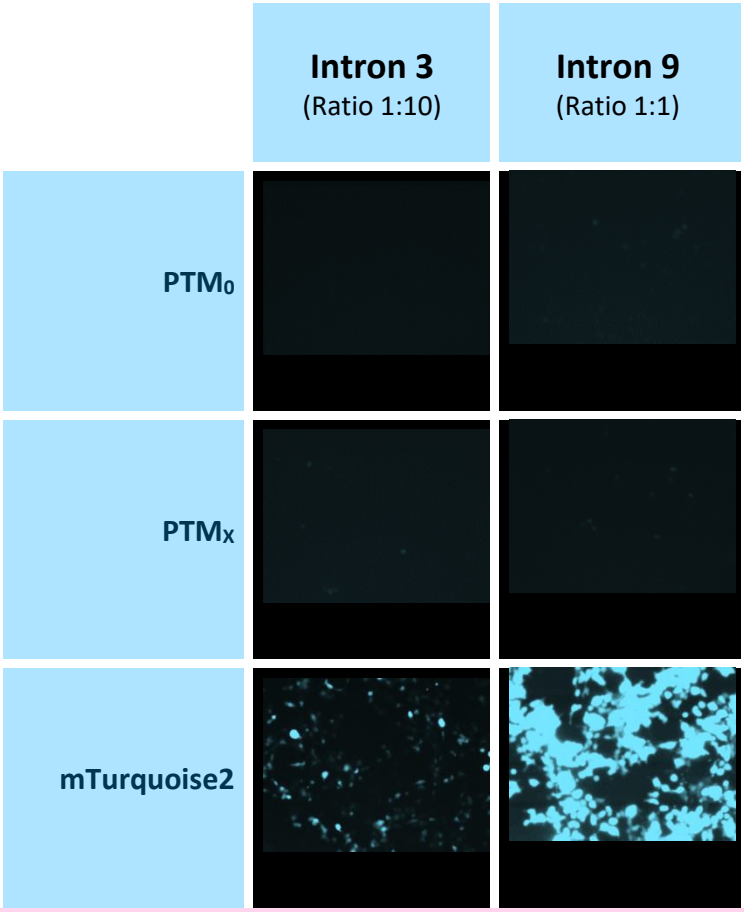


Figure 52: ***Pictures of transfected cells in a TS screening for HTT introns 3 and 9****. Acquisitions were performed on living transfected cells exhibiting fluorescence using a 10x objective on an inverted epifluorescence microscope. Auto-focus was performed on the bright field channel and mTurquoise2 fluorescence was observed on the CFP channel of the microscope mounted with an excitation filter of 436/20nm and an emission filter of 480/40 nm.*

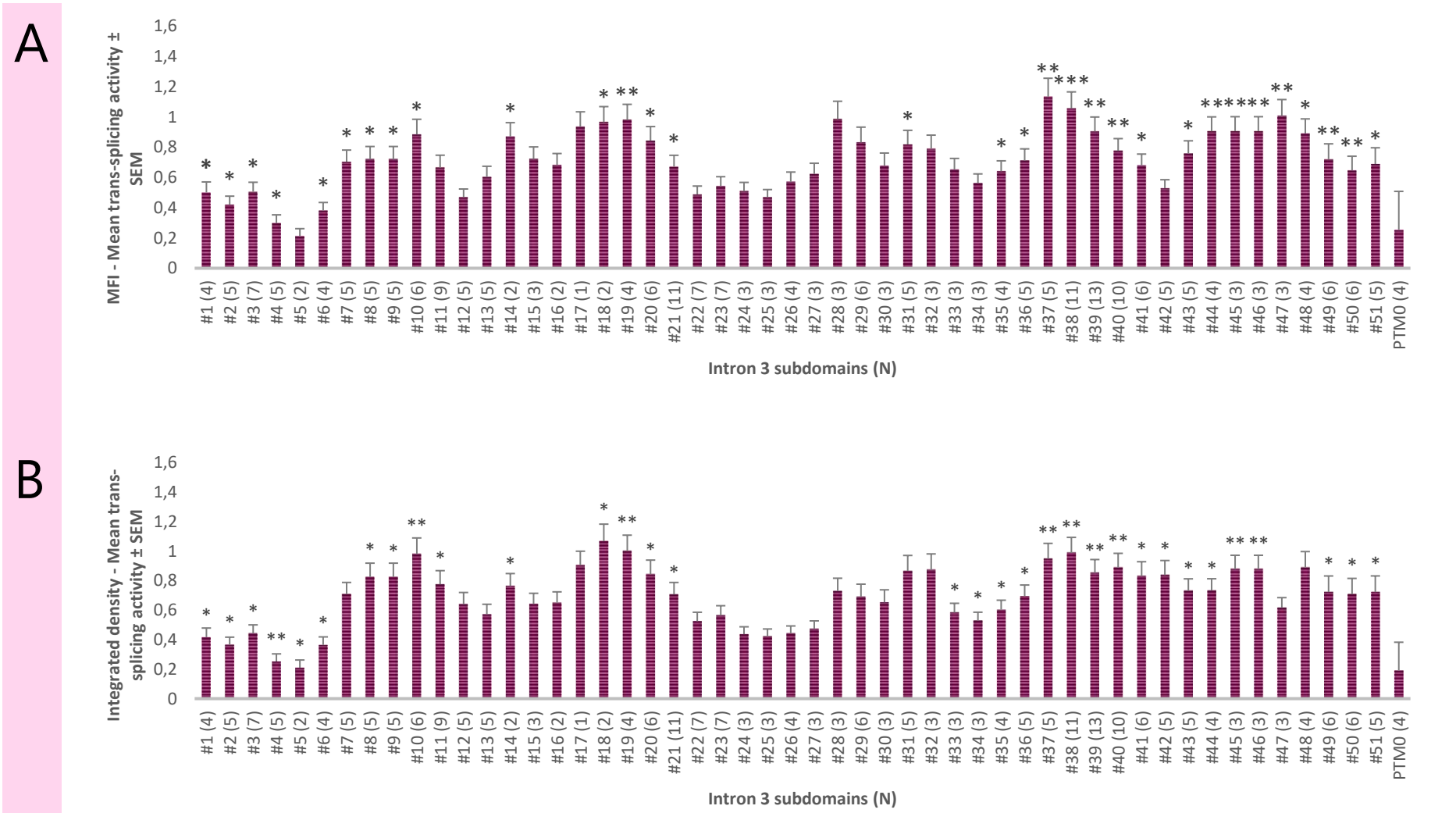


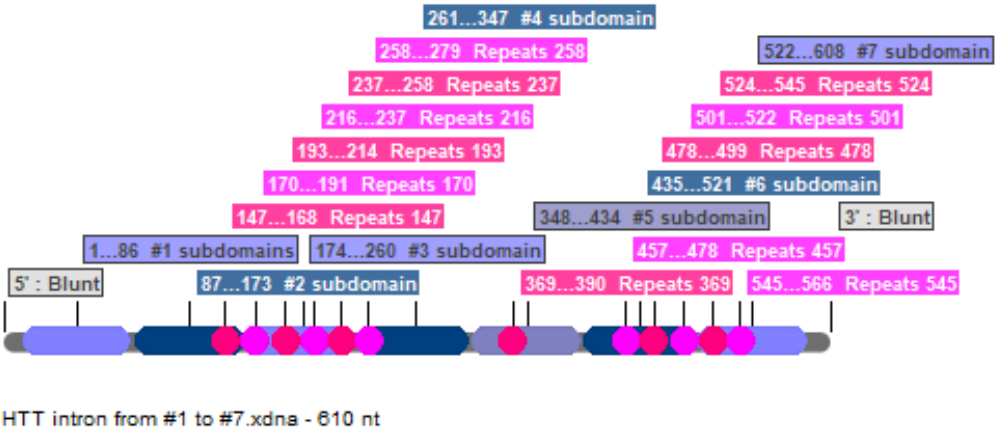
*Figure 53: **Screening results of HTT intron 3.** (A) Normalized MFI and (B) normalized integrated density of mTurquoise2 signal were expressed as mean ± SEM of PTM TS activity located in the same subdomains. In brackets, number of tested PTMs per subdomain which are all 86 bp long. Student's t-test was performed (p-value: * < 0.05; ** < 0.01; ***<0.001)*

Interestingly, intron 3 exhibits a fall-off of TS activity spanning subdomains #1 to #6 with subdomain #5 exhibiting the lowest TS activity. The 5' region of *HTT* intron 3 is characterized by twelve repeat regions which are 22bp long defined as "ttgtatggtttggaggtgctct". As depicted in Figure 54A, we found:

- **6 repeat motifs** ranging from the end of subdomain #2 to the beginning of #4, thus covering the whole of subdomain #3.
- **1 repeat motif** in subdomain #5
- **5 repeat motifs** overlapping subdomains #6 and #7.

Moreover, all these motifs are clustered in a small portion of the 5' part of *HTT* intron 3 as shown in Figure 54B. Taken together, this suggests a complex spatial conformation of the pre-mRNA, preventing an easy access of the PTMs to this portion of *HTT* intron 3.

A



B

Position	Size	Strand
147	22	+
170	22	+
193	22	+
216	22	+
237	22	+
258	22	+
369	22	+
457	22	+
478	22	+
501	22	+
524	22	+
545	22	+

*Figure 54: **Repeat motifs in HTT intron 3.** (A) Representation of the 5' end HTT intron with the seven first subdomains depicted in blue shades with a corresponding blue box indicative of relative position in HTT intron 3. Each purple circle corresponds to a repeat region characterized by 22 bp long motifs: "ttgtatggtttggaggtgctct". Purple boxes are indicative of each motif location on HTT intron 3. This representation was generated by Serial Cloner v2.6 (B) Exhaustive list of repeat region locations on HTT intron 3 in the first column, their length in second column and their strand in the third column.*

In intron 9, TS activity seems to be relatively stable along the whole intron. Indeed, when considering MFI quantification, all subdomains in this intron are significantly different from PTM_0 except for the last subdomain (#20). However, intron 9 exhibits differential TS activity depending on the analysis method. For MFI (Figure 55A), signal seems to be steady whereas in integrated density we observe more variations with the highest subdomains being #1, #6 and #15.

*Figure 55: **Screening results of HTT intron 9.** (A) MFI and (B) integrated density of mTurquoise2 signal are expressed as mean ± SEM for PTM TS activity located in the same subdomains. In brackets, number of tested PTMs per subdomain which are all 72 bp long. Student's t-test was performed on data (p-value : * < 0.05; ** < 0.01; ***<0.001).*

4.6 TS activity of intron 20

Intron 20 being significantly shorter (256 bp) than introns 3 and 9, we decided not to undertake a full screen for this intron. Indeed, one BD is sufficient to cover most of the length of intron 20. We thus tested only one PTM, called PTMi20, to evaluate TS activity on intron 20. PTMi20 BD is composed of the whole intron 20, allowing in theory hybridization all along the target intron. PTMi20 was evaluated using the same methodology as for the screening of intron 3 and 9 libraries. PTMi20 demonstrated TS activity compared to PTM_0 controls (Figure 56) but this data is a relative quantification since the fluorescence baseline depends on intron splicing activity. So, no PTM comparison with other target introns is possible, though PTMi20 has 2.3- and 3.7-fold higher TS activity in MFI and integrated density respectively, which is hopeful.

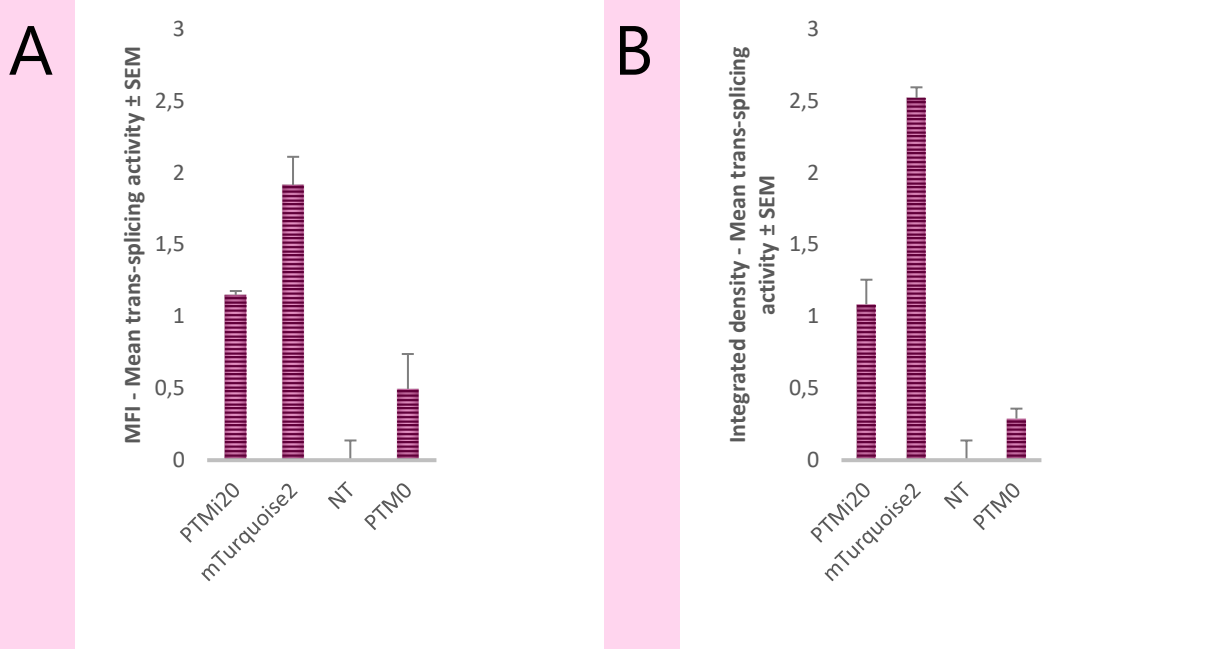


*Figure 56: **TS activity of PTMi20.** (A) MFI and (B) integrated density of mTurquoise2 signal were averaged from 2 independent experiments and are expressed as mean ± SEM.*

4.7 TS activity comparison with fluorescence

To compare the best PTMs coming from the screen of the different libraries, we selected two PTMs targeting intron 3 (PTMi3) and two PTMs targeting intron 9 (PTMi9) depicted in Figure 57. These PTMs were chosen because they are located in the hotspot subdomains mentioned above with:

- PTMi3-806 straddling subdomains #37 and #38
- PTMi9-296 straddling subdomains #1 and #2
- PTMi9-308 straddling subdomains #15 and #16

The exception is PTMi3-447, located over subdomains #49 and #50. Evaluation of this PTM was further extended because it came out as the best PTM in a preliminary screening and thus its characterization started before the end of the screening process.

Regarding mTurquoise2 signals in screening, these PTMs exhibited higher TS scores, depicted in purple diamond points in Figure 57, compared to other screened PTMs, whose mean scores are indicated in large purple boxes for each intron.

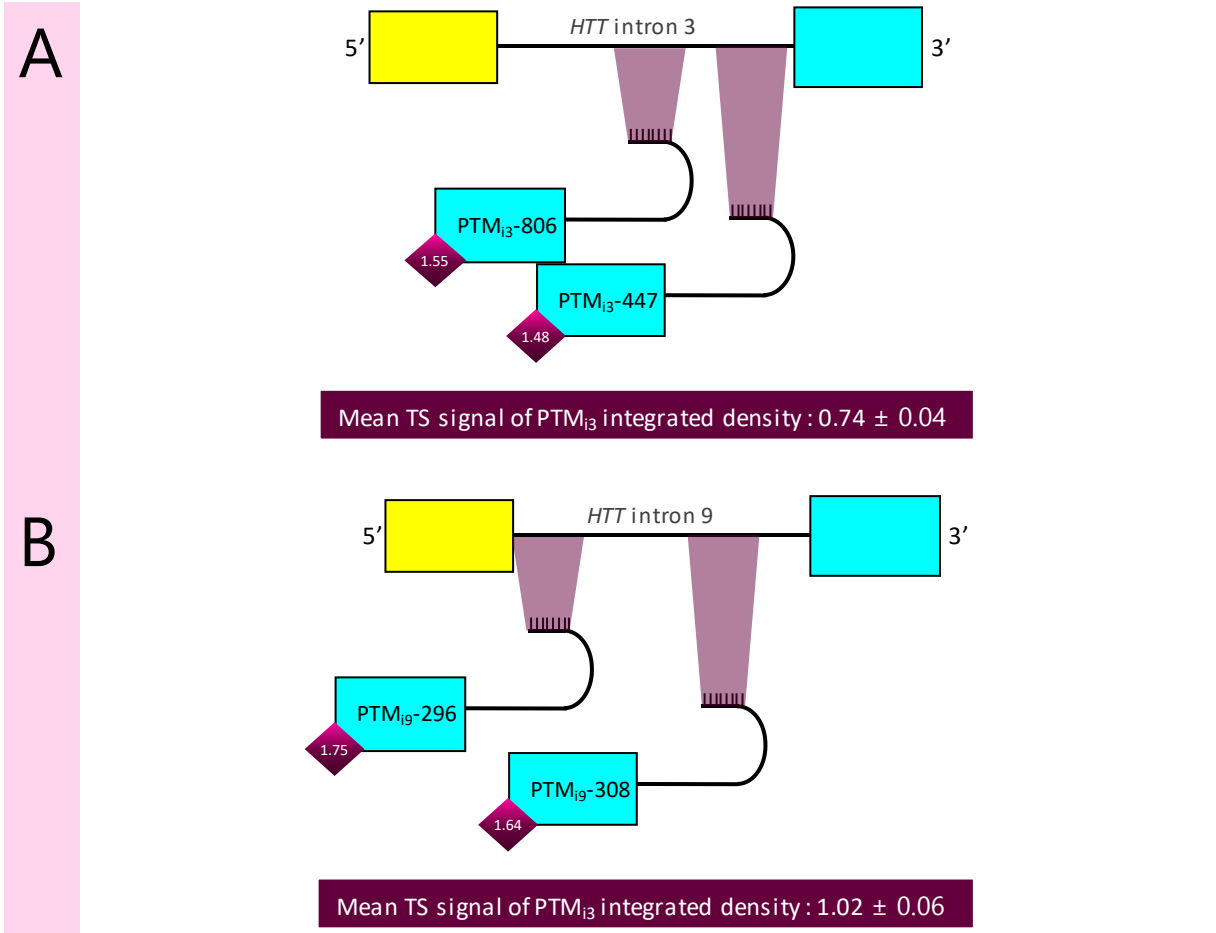


*Figure 57: **Schematic representation of selected PTMs.** (A) Selected PTMs targeting intron 3 are PTMi3-447 ranging from 4195 to 4313 nt corresponding to subdomains #49 and #50, and PTMi3-806 ranging from 3205 to 3288 nt corresponding to #37 and #38. (B) Selected PTMs targeting intron 9 are PTMi9-296 ranging from 1 to 102 nt, corresponding to subdomains #1 and #2 and PTMi9-308 ranging from 1091 nt to 1180 nt, corresponding to subdomains #15 and #16. Purple diamond points indicate corresponding integrated density level of each PTM. Large purple boxes indicate the mean integrated density level of all assessed PTMs for each library.*

Since Z-transformation is based on large population data, we cannot directly compare screening results with results obtained for intron 20, for which we assessed only a single PTM. Thus, we re-evaluated these PTMi3s and PTMi9s in an isolated manner. We validated their splicing activity at the fluorescence level (Figure 58) but no significant difference was observed between PTMs as already suggested by the results of the screening experiment.

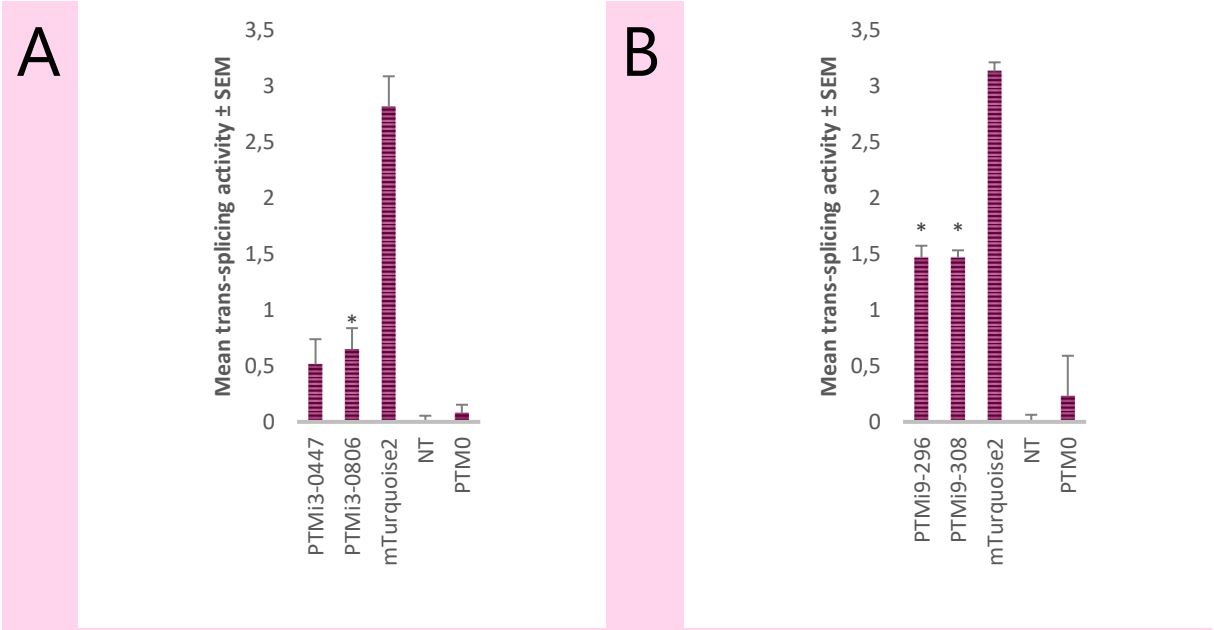


*Figure 58: **TS activity of selected PTMs.** (A) PTMi3 and (B) PTMi9 integrated density of mTurquoise2 signal was averaged from 3 independent experiments and is expressed as mean ± SEM. Student's t-test was performed on the data (p-value : * < 0.05; ** < 0.01; ***<0.001)*

To obtain a quantitative estimation of TS rate, we performed an exponential regression of the data obtained for PTM_0 and mTurquoise2 controls. This allowed us to deduce a specific regression equation for each target intron (Figure 59). Based on these equations, TS efficiency can be calculated. This revealed a low rate of TS, inferior to 1% for both PTMi3, and inferior to 2% for both PTMi9. PTMi20 seems to exhibit the highest rate of TS with 4.5% of estimated correction, consistent with the intron definition hypothesis in which normal *cis*-splicing is easier to disturb in short introns (See in Exon and intron definition page 63).

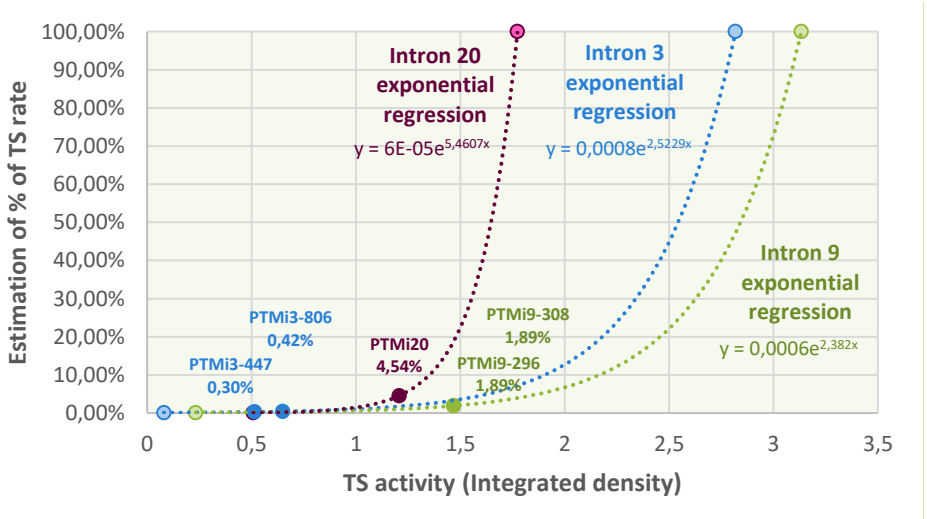


*Figure 59: **Exponential regression of TS rate**. In blue are data for intron 3, green for intron 9 and purple for intron 20. Exponential regressions are defined by each equation in which y corresponds to estimation (%) of TS rate and x to integrated density value.*

However, exponential regression directly correlated with the level of PTM0. Indeed, PTM0 is the negative control, and thus, the level of fluorescence measured for this condition corresponds to the background noise, and is partly dependent on the target intron. So, the comparison of PTM targeting different introns is biased by their intrinsic splicing dynamism. Thus, we needed to set up alternative methods to solve this issue. Consequently, we evaluated PTMi3s, PTMi9s and PTMi20 at the mRNA level with another method allowing a direct determination of the percentage of trans-spliced mRNA.

4.8 TS activity validation at the mRNA level

Previous experiments focused on mTurquoise2 detection at protein level with fluorescence. This method was adequate for "high content" PTM screening and allowed identification of efficient PTMs. However, this potent technique was not suitable for precise estimation of TS rate. To validate TS activity and to quantify their specific rate of correction, we investigated this at the mRNA level.

In this context we have developed an alternative method to quantify precisely TS rate. This method relies on several steps:

- cell-sorting by flow cytometry
- mRNA extraction
- PCR amplification of both *cis*- and *trans*-spliced products using a universal primer pair
- quantification of *cis*- and *trans*-spliced mRNA pools by restriction endonuclease

4.8.1 Fluorescent activated cell sorting (FACS)

HEK293T cells were transfected with the best PTM and the associated hybrid mini-gene. 48h after transfection, mTurquoise2 positive cells were isolated using flow cytometry to enrich the studied population in cells exhibiting TS activity. In all conditions, the proportion of positive cells was inferior to 1%, which is consistent with observed fluorescence intensity compared to positive control (Figure 60, Appendix 5).

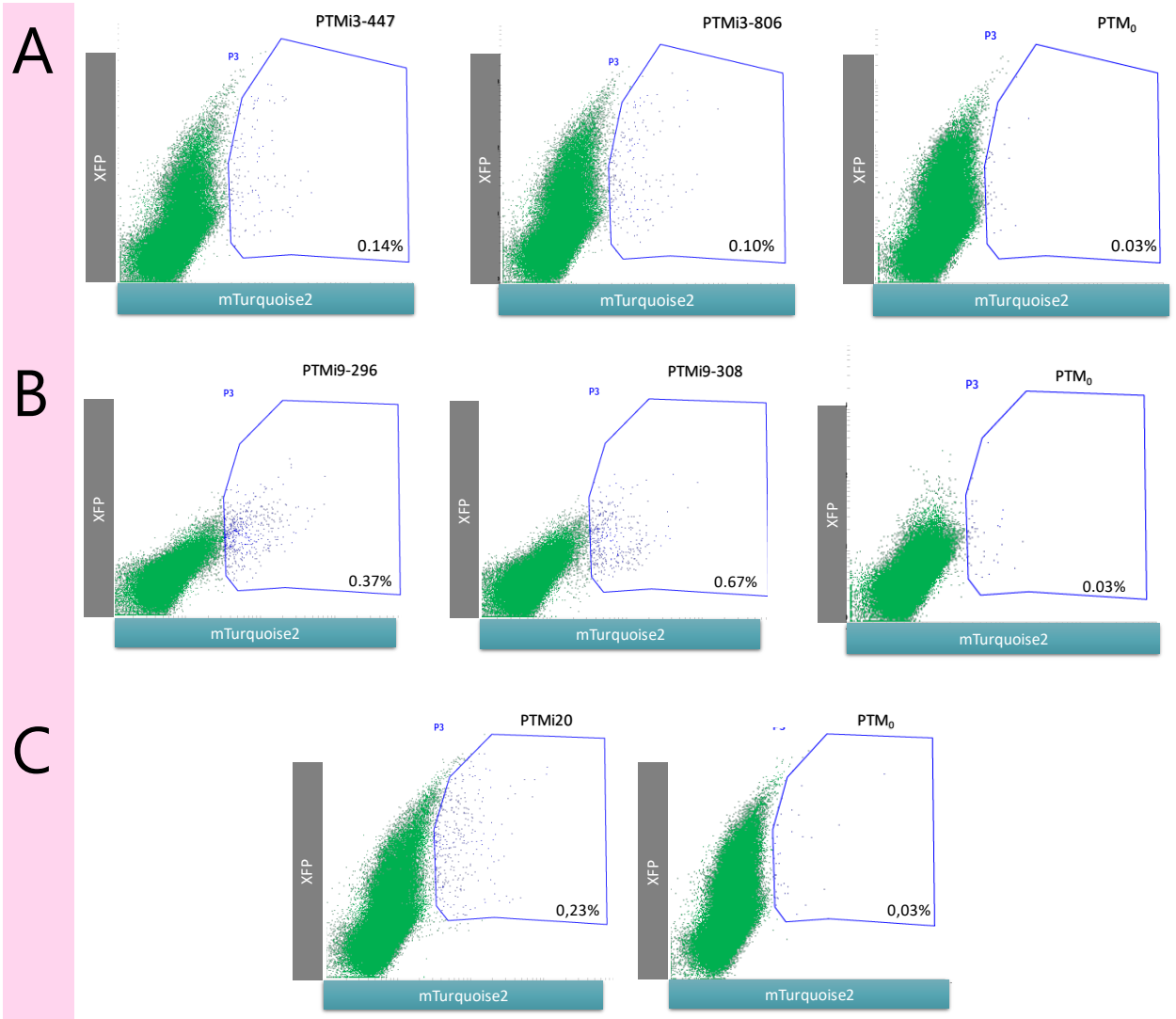


*Figure 60: **Cytometry profile** in intron 3 (A), intron 9 (B) and intron 20 (C) contexts. Annotated P3 gates correspond to mTurquoise2 positive cells that were collected for further analysis. Percentage in gate corresponds to % of singlet gate described in the Materials and Methods.*

4.8.2 mRNA quantification strategy

mRNAs were extracted from the mTurquoise2-positive cell population and retro-transcribed to amplify spliced products resulting from the fluorescent reporter system. In the same PCR reaction, both XFP *cis*-spliced products and mTurquoise2 *trans*-spliced products were amplified using a unique primer pair. PCR products were separated into two halves and cut with the restriction endonuclease specific to the XFP (PstI in cis-spliced products), or mTurquoise2 (DrdI in trans-spliced products), as depicted in Figure 61. Restriction products were analyzed using a DNA 1000 chip on a Bioanalyzer, allowing an easy detection and quantification of complete and partial digestions.

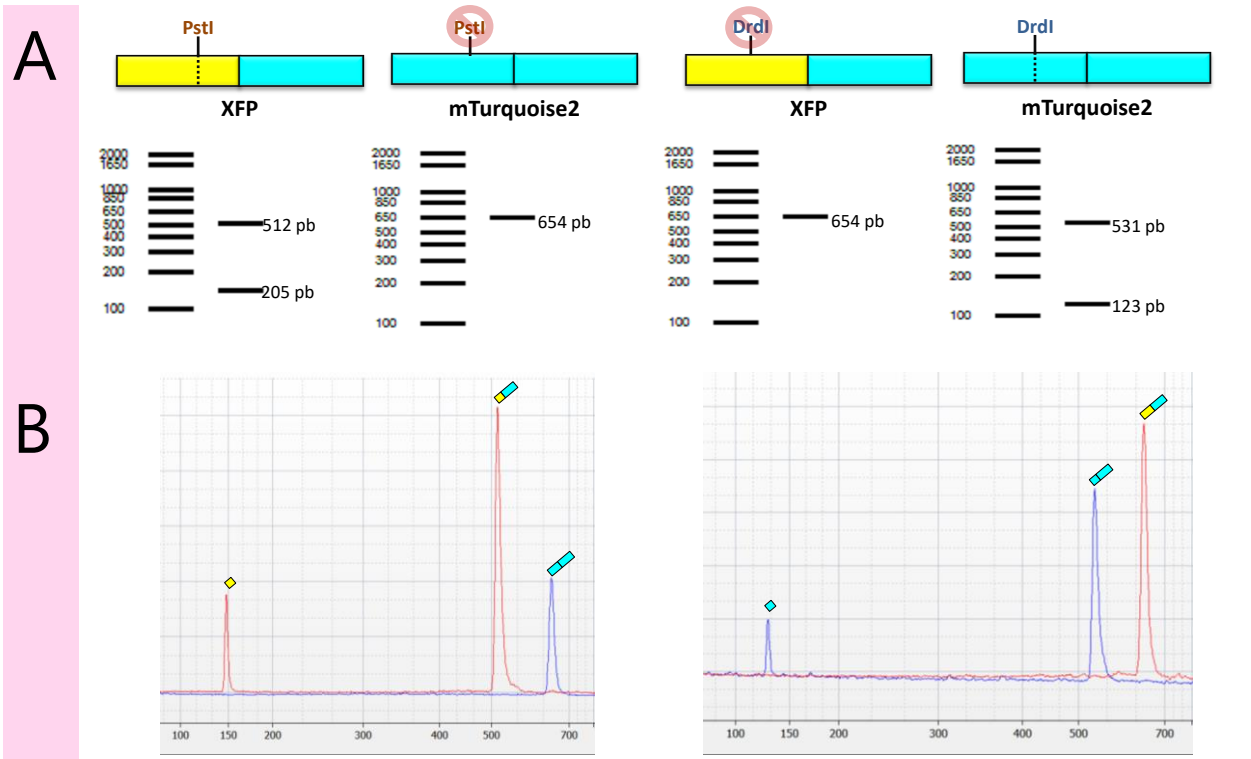


*Figure 61: **Quantification strategy** (A) Schematic representation of restriction sites used to identify spliced products at the mRNA level with PstI for cis-spliced products and DrdI for trans-spliced products. (B) DNA 1000 chip electropherogram of PTM_0 and mTurquoise2 controls digested by PstI (left) and DrdI (right). Red curves correspond to XFP mRNA profiles and blue to mTurquoise2. Around 10 ng per sample was used on a DNA 1000 Agilent Chip.*

The following experiments are preliminary data with only 2 independent experiments with a single PTM_0 control; however, some trends already emerge. In intron 3 (Figure 62), we observed a TS activity in PTMi3-447 ranging from 5.8% (DrdI) to 14.3% (PstI) as well as in PTMi3-806 ranging from 8.5% (DrdI) to 9.8% (PstI) compared to PTM_0 at 1.8% (DrdI) and 2.7% (PstI). For intron 9 (Figure 63), DrdI results revealed a TS rate of 2.3% for both PTMs (PTMi9-296 and -308) with only 0.7% for PTM_0. In contrast, PstI results revealed an inversed trend with a higher estimation of TS for PTM0 with 12.7% than for PTMi9 with 7.25% and 11.45% for respectively PTMi9-296 and PTMi9-308. For intron 20 (Figure 64), PstI also exhibits surprising results, with a high TS rate variability around 29% with a complete digestion of PTM_0 whereas DrdI revealed 4.0% TS with PTM_0 at 1.1%.

This data suggests that DrdI is more reliable for assessing TS since it induces minor variation. To use PstI for TS quantification, optimization is still needed. Nevertheless, these data validate that mTurquoise2 fluorescence results from TS process and both endonucleases are required to characterize TS, thus providing qualitative and quantitative information. However, for a firmer conclusion, higher numbers of replicates are also needed.

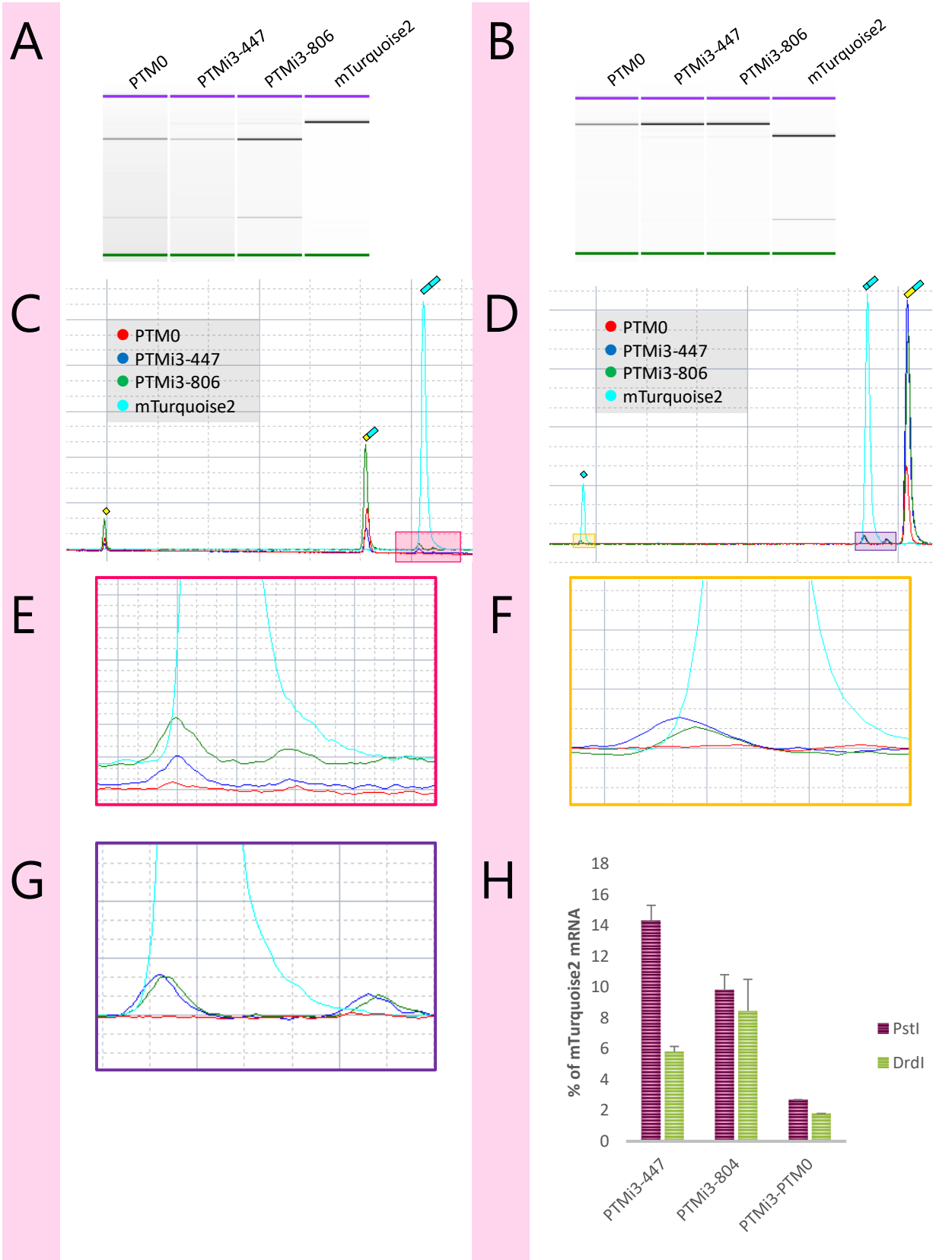


Figure 62: ***TS activity determined at the mRNA level of PTMs targeting intron 3****. (A,B) Gel reconstitution of PstI and DrdI restriction of PCR products (respectively). (C,D) Electropherogram of PstI and DrdI restricted PCR products (respectively). (E-G) Enlargement from electropherogram (D) at 654 bp in the pink box, 205 bp peaks in the orange box, and at 512 bp in the purple box, respectively (n=2). (H) quantification of PstI and DrdI restriction products obtained by calculating area under the electropherogram curves of mTurquoise2 and normalized to DNA length by 2100 Expert software. Around 5 ng per sample were used on a DNA 1000 Agilent Chip*

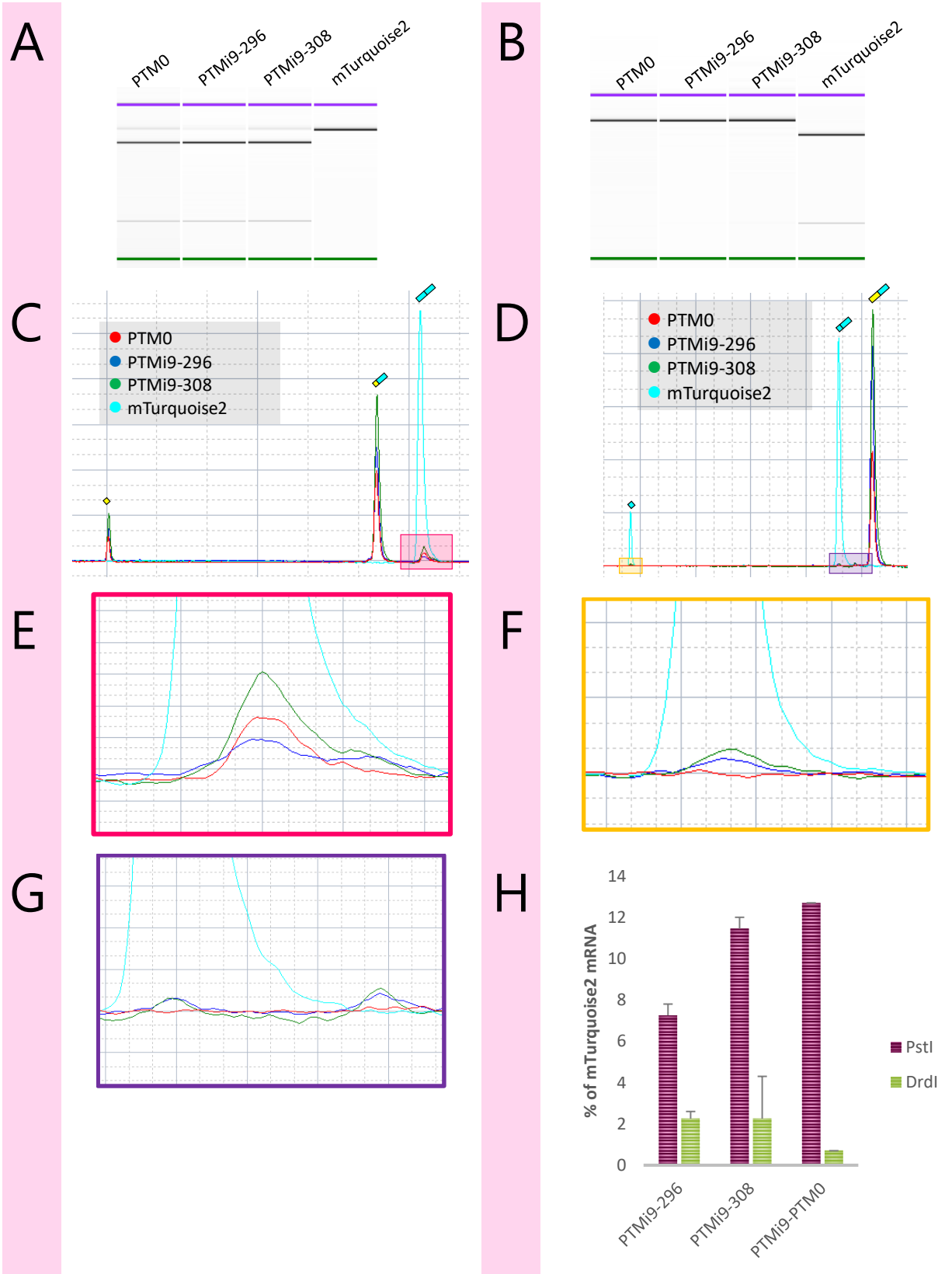


Figure 63: ***TS activity determined at the mRNA level of PTMs targeting intron 9.*** *(A,B) Gel reconstitution of PstI and DrdI restriction of PCR products (respectively). (C,D) Electropherogram of PstI and DrdI restricted PCR products (respectively). (E-G) Enlargement from electropherogram (D) at 654 bp in the pink box, at 205 bp in the orange box, and at 512 bp in purple box, respectively (n=2). (H) quantification of PstI and DrdI restriction products obtained by calculating area under the electropherogram curves of mTurquoise2 and normalized to DNA length by 2100 Expert software. Around 5 ng per sample were used on a DNA 1000 Agilent Chip.*

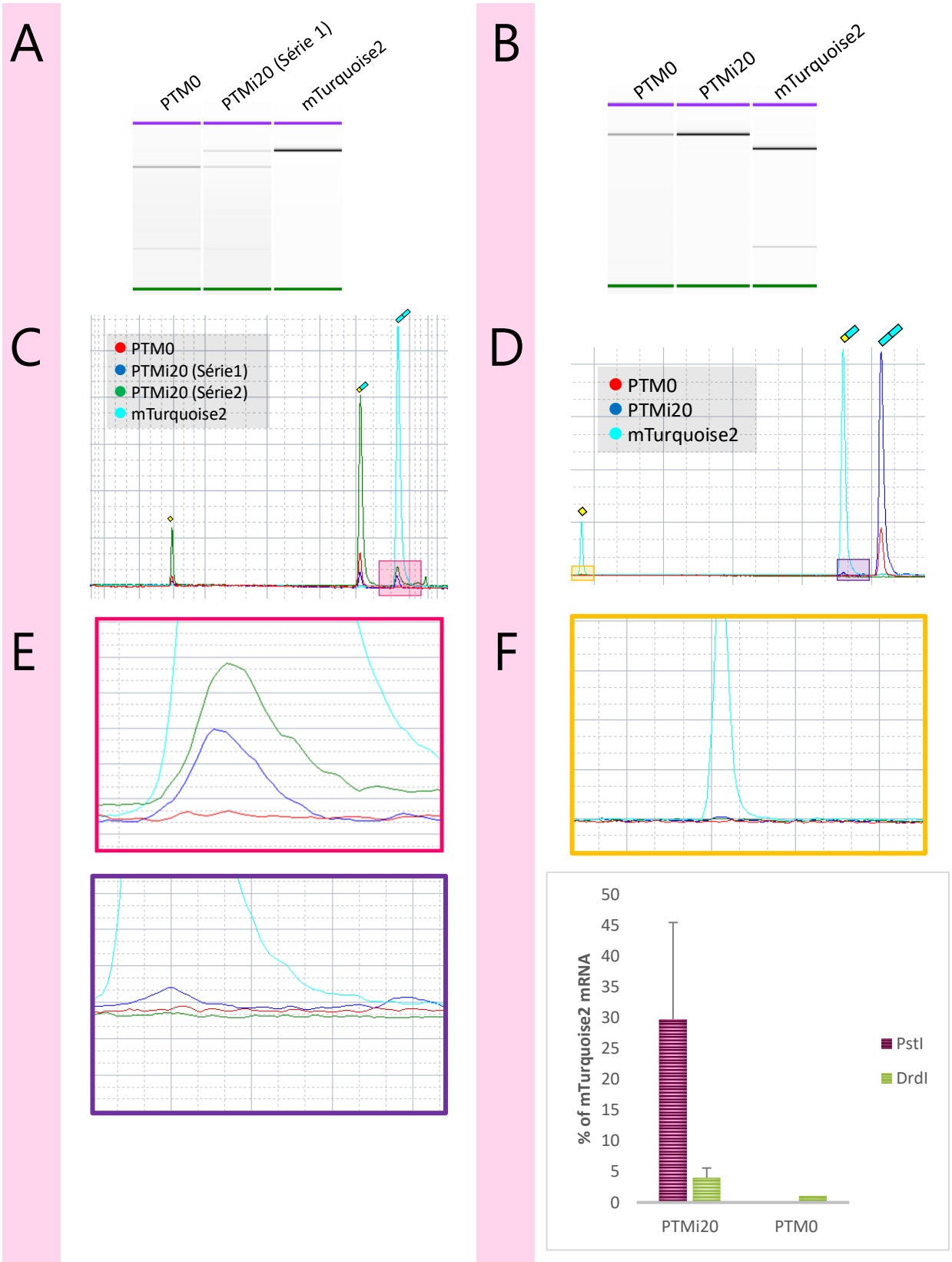


Figure 64: ***TS activity determined at the mRNA level of PTMs targeting intron 9.*** *(A,B) Gel reconstitution of PstI and DrdI restriction of PCR products (respectively). (C,D) Electropherogram of PstI and DrdI restricted PCR products (respectively). (E-G) Enlargement from electropherogram (D) at 654 bp in the pink box, at 205 bp in the orange box, and at 512 bp in the purple box, respectively (n=2). (H) quantification of PstI and DrdI restriction products obtained by calculating area under the electropherogram curves of mTurquoise2 and normalized to DNA length by 2100 Expert software. Around 5 ng per sample were used on a DNA 1000 Agilent Chip.*

5 Discussion

In this project, we demonstrated differential splicing activity of *HTT* introns 3, 9, 17 and 20 in an artificial exon context. We used these properties to assess TS activity for two different introns (*HTT* introns 3 and 9) by a screen of PTM libraries. We validated our fluorescent reporter system allowing easy detection of TS events and we showed that TS activity level is dependent on the PTM target along the intron. Additionally, TS in *HTT* intron 20 was also evaluated in this study with PTMi20, and we validated its TS capacity. However, precise quantification of TS rate was the main issue. Exponential regression did not allow a comparison between introns, and the mRNA level quantification must be considered as preliminary data that needs to be better characterized.

In this discussion, we will compare our screening strategy with other strategies that have been previously reported in the literature, then we will consider the main methodological issues of TS quantification and screening results, before giving some general considerations about the future of TS gene therapy.

5.1 The fluorescent reporter system

5.1.1 Hybrid mini-genes

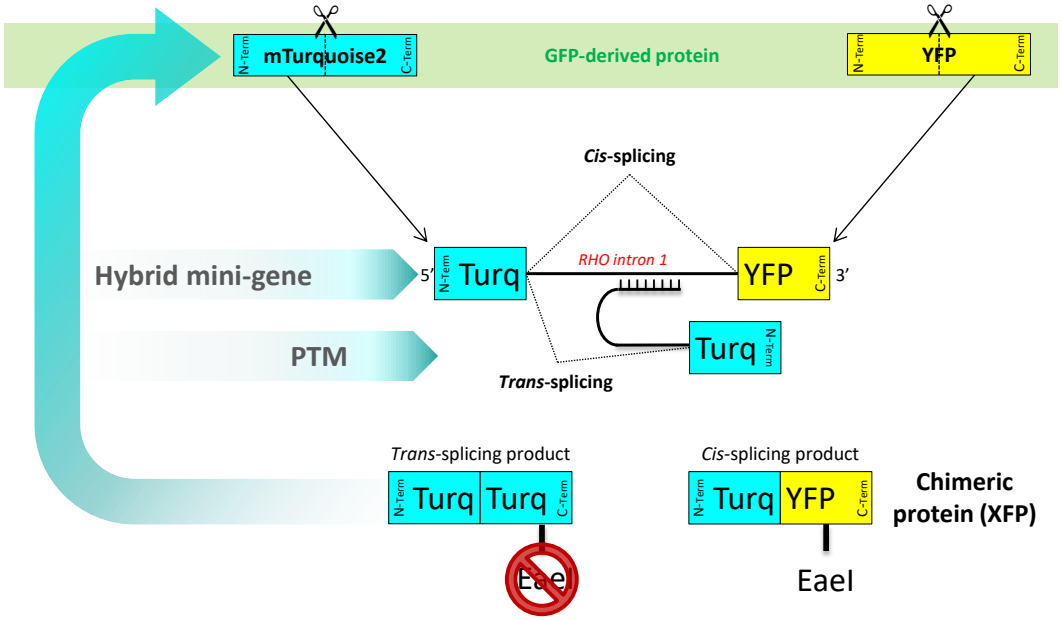
We have developed our own fluorescent reporter system based on two GFP-derived proteins: YFP and mTurquoise2. We took the 5' N-terminal part of YFP and the 3' N-terminal part of mTurquoise2 and introduced a target *HTT* intron in between. When a PTM encoding for the 5' half of mTurquoise2 is associated to this system, it enables a TS reaction, resulting in the formation of the full mTurquoise2. Thus, this reporter system is suitable for easy TS detection using the mTurquoise2 signal, and is appropriate for the screening strategy.

5.1.1.1 *Reliability of intron environment changes in an artificial context*

In this reporter system, we modified the intron environment at both 5' and 3' ends, by replacing endogenous *HTT* exons with artificial YFP and mTurquoise2 exons. These changes could impact the normal splicing pathway. Indeed, the only splicing elements potentially conserved are those located inside introns. Consequently, regulation signaling arising from exons (ESS/ESE) is excluded in our system. This drawback can be considered as the main limit of our screening methodology, because in theory, in superior eukaryotes exon/intron patterns are defined by short exons flanked by long introns which altogether are supposed to follow the rules of exon definition[197] (See Exon and intron definition page 63).

To investigate this issue, we assessed the reliability of our artificial system before the beginning of the thesis project. Previous work on TS in rhodopsin (*RHO*)[179] had already been implemented in the lab by Adeline Berger during her PhD. RHO is a gene whose mutations most of the time lead to autosomal dominant Retinitis pigmentosa. PTMs with a given TS activity targeting the first intron of *RHO* were generated and characterized by Adeline Berger. We thus decided to assess the TS activity of some PTMs from this work to compare with the TS rate in our reporter system. We thus adapted the BD of *RHO*-PTM1, 6 and 12 into the PTM backbone encoding the C-terminal part of mTurquoise2 (Figure 65 next page, Figure 67 page 132). We also generated the corresponding hybrid mini-gene composed of the 5' complementary N-terminal part of mTurquoise2, the 3' C-terminal part of YFP and in between the intron 1 of *RHO* as depicted in Figure 65. We used similar mRNA quantification methods based on restriction endonuclease EaeI, specific to the corresponding hybrid mini-gene and we observed a similar TS rate in both endogenous and artificial contexts. Thus, we validated our reporter system as a reliable tool to assess TS at the mRNA level in the 3' replacement.

A



B

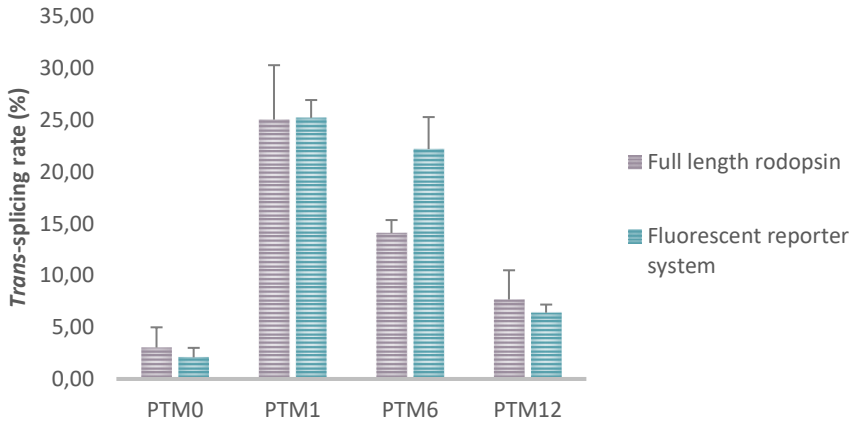


Figure 65: ***Assessment of a fluorescent reporter system for PTM screening.*** *(A) A fluorescent system was adapted to intron 1 of the rhodopsin gene (RHO), and the EaeI restriction site allowed the quantification of trans-splicing using the same methodology described in Berger et al. [179] To that aim, RHO intron 1 was subcloned between sequences coding respectively the N-terminal part of mTurquoise2 and the C-terminal part of YFP, while the PTM contained the C-terminal part of mTurquoise2. With this system, cis-splicing leads to an mRNA encoding a chimeric protein fluorescing minimally in green, and trans-splicing leads to full-length mTurquoise2, characterized by a bright blue fluorescence. (B) Three binding domains, previously characterized in the context of the full rhodopsin pre-mRNA (for details, see Berger et al.[179]), were tested with the fluorescent reporter system using end-point PCR amplification followed by enzymatic restriction, and led to the same trans-splicing rate, thus showing that trans-splicing efficiency is independent of the genomic context of the target intron. PTM0: negative control without binding domain. Location of binding domains of PTM1, 6, and 12 on RHO intron 1 have been described previously in Berger et al (2015)[179]. Bars represent mean ± SD; n = 6 for fluorescent reporter system; for full-length rhodopsin, n = 9 for PTM0 and 1, and n = 3 for PTM6 and 12.*

5.1.1.2 *Differential splicing activity in hybrid mini-genes*

The *HTT* introns selected for the screen were introduced in the hybrid mini-gene backbone, thus changing their exon environment. Evaluation of their splicing activities was needed before using them for PTM screening. Since fluorescence is emitted only if normal splicing occurs, we used an epi-fluorescence microscope to first assess the level of XFP expression. This data revealed high fluorescence variability among hybrid mini-genes (Figure 47), suggesting that introns exhibit differential dynamics in splicing activity. To understand this variability, we hypothesized that it is correlated with:

- CSS activation inducing alternative splicing
- Intron retention leading to unspliced mRNA which still contains introns, thus reducing the portion of normal and fluorescent XFP.

At the RNA level, we identified two different splice products: the expected full XFP and a truncated variant resulting from the activation of CSS located in XFP. Interestingly, we demonstrated that the occurrence of this CSS activation is intron dependent, since each intron induces different rates of mis-splicing, except for introns 3 and 20 (Figure 45B). Basically, TS events are likely to be in competition with *cis*-splicing events, but if additional aberrant processes occur in *cis*-splicing, it will result in three-way competition, which will bias the TS estimation rate.

Surprisingly, for rhodopsin[179], only one mRNA was generated as a *cis*-splicing product and no alternative-spliced products were observed, suggesting that the CSS was not activated. Being positioned at the 334th nucleotide of both mTurquoise2 and YFP, we did not expect that it would be activated in *HTT*.

This differential CSS activation is still not completely understood, although we have demonstrated that it depends on intron splicing activity.

To use our screening system with confidence, we needed to eliminate CSS activation, and for this purpose, we did site-directed mutagenesis on CSS. We simulated all possible silent mutation motifs and calculated the corresponding MaxEnt score (Figure 46A). Two nucleotide substitutions in YFP reduced MaxEnt significantly, and resulted in a complete recovery of full-length XFP (Figure 46B). However, this CSS correction revealed no influence on XFP fluorescence, except for intron 17 exhibiting 85% of MFI increase and a more than 3-fold higher signal for integrated density after CSS deletion compared to initial XFP. These data suggest that in *HTT* intron 17, CSS correction allows better XFP expression.

In parallel, we assessed intron 17 retention. CSS is not likely to impact this aspect since retention rate is quite stable, around 17.5% (Figure 46C). This data is consistent with the impact of CSS on XFP expression.

In the case of *HTT* intron 9, CSS deletion results were more surprising. *HTT* intron 9 exhibited 84.3% of mis-spliced XFP before CSS correction which is the highest percentage among assessed *HTT* introns (Figure 45) and we observed a complete recovery of full-length XFP after CSS deletion; however, no fluorescence improvement was observed (Figure 47). We first hypothesized that this observation was related to the level of intron retention. However, this assumption was excluded since qPCR data showed that intron retention was nearly 50% higher when CSS was corrected (Figure 48). To sum up, low XFP expression is not related to CSS activation, nor to intron retention rate, suggesting that another basal mechanism is involved, such as 5'SS strength.

5.1.1.3 *5'SS MaxEnt score impact*

We then postulated that fluorescence variation could be related to 5'SS motifs as assessed by MaxEnt score, a probabilistic model to evaluate splicing site strength. Indeed, both intron 3 and 20 hybrid mini-genes, which showed the highest fluorescence intensities, also have the highest MaxEnt scores in an artificial context. This contrasts with intron 9 exhibiting the lowest MaxEnt of 1.91 and the lowest fluorescence intensity (Table 6, Figure 47). This data is consistent with our emerging hypothesis that 5'SS is the main regulator of XFP expression.

However, the changes of intron environment can impact the strength of splicing site in the fluorescent environment compared to the endogenous one. So, we decided to assess the change of 5'SS strength by calculating ΔMaxEnt which is the absolute difference between MaxEnt scores in the endogenous (*HTT* exon/intron junction) and the artificial (YFP/*HTT* intron junction) contexts. MaxEnt is calculated from 5'SS motifs, composed of the last three nucleotides of the exon and the six first of the intron. In our comparison, the intron portion is unchanged, and only exon nucleotides are modified between the two conditions. In all cases, we observed that only one nucleotide is conserved in the exon portion, but MaxEnt is modeled on a weighted matrix in which both relative position and composition of each nucleotide impact differently the final score.

Consistent evidence was provided with intron 17 as a counterexample. For *HTT*, intron 17 exhibits a very high score of 10.52, suggesting that this intron is very active in the endogenous gene. When intron 17 is introduced in an artificial context, MaxEnt is cut in half, reaching 4.99, corresponding to the highest difference (ΔMaxEnt=5.53), while much lower differences are found in other *HTT* introns, all with ΔMaxEnt inferior to 2 (Table 6). So, with intron 17 whose splicing signaling is largely reduced

compared to the endogenous *HTT* environment, it is likely to lead to an excessive sensitivity to CSS activation, thus impacting XFP expression. If 5'SS junction sequences are too different from endogenous motifs, it leads to changes in overall splicing signaling, suggesting that the basal splicing activity is altered, and thus estimation of TS could also be biased. We assumed in this case that changes in intron environment are likely to highly impact intron 17 splicing regulation and therefore we decided to exclude it from the screen. So ΔMaxEnt seems to be a good indicator of whether a given intron will be suitable for use in our reporter system.

The 5'SS strength hypothesis is also consistent with the correlation matrix revealing higher correlation without intron 17 (Figure 49B). The fact that integrated density data (before CSS deletion) correlates significantly with MaxEnt scores seems to indicate that 5'SS regulates CSS activation (Figure 49B). This suggests that the higher the Maxent score, the lower the CSS activation and so the higher fluorescence intensity.

To conclude, we have selected in this study *HTT* introns with differential splicing activity and we have developed efficient methods to characterize them. Additional experiments would be needed to investigate if these differential splicing levels are actually the same for the endogenous gene. A possible strategy to answer this question would be to generate hybrid mini-genes with *HTT* exons for each selected intron and compare their levels of expression by western-blot using an HA-tag located at the end of the hybrid mini-gene. This experiment would provide the final proof of a robust modeling of endogenous splicing activity by our reporter system.

5.1.1.1 Koller's fluorescent reporter system for TS

In this project, we developed molecular tools enabling TS detection using fluorescence. This strategy, depicted in Figure 66, was initially developed by Koller et al[198]. For 5' replacement (Figure 66b), they suggest cutting in half the sequence of a fluorescent protein such as GFP, and putting the 5' sequence on the PTM and the 3' sequence on the hybrid mini-gene. In case of TS, this results in the production of the full GFP.

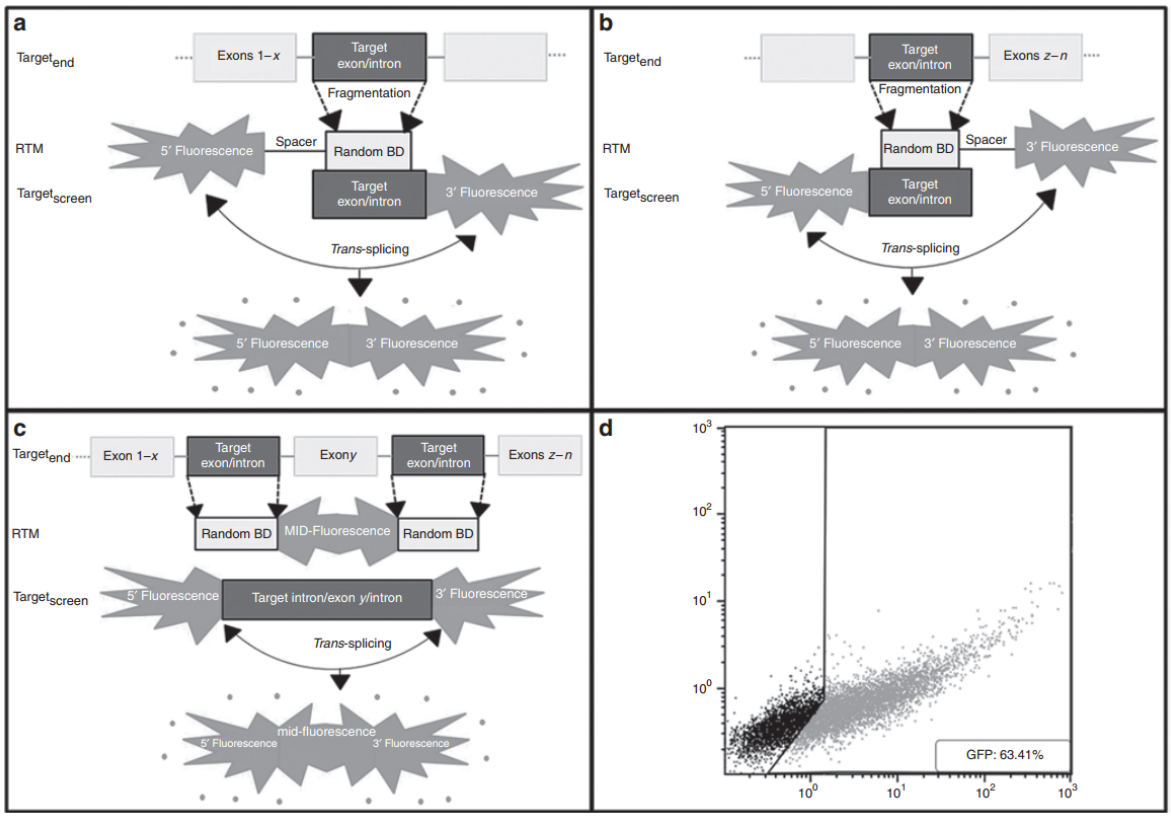


Figure 66: ***Schematic depiction of fluorescence-based RTM (=PTM) screening constructs****. Schematic representation of (a) 3′ trans-splicing, (b) 5′ trans-splicing, and (c) internal exon replacement between a GFP target and GFP–RTM constructs. (d) FACS analysis of cotransfected test cells expressing GFP after successful trans-splicing. BD, binding domain; FACS, fluorescence-activated cell sorting; GFP, green fluorescent protein; IER, internal exon replacement; RTM, RNA trans-splicing molecules (=PTM). Adapted from Koller et al (2014)* [198]

In this strategy, the gene portion to be replaced is conserved in the hybrid mini-gene, thus preserving one of the two endogenous exon/intron boundaries. This construct has the advantage of expanding possible targets to exon elements such as ESS or ESE, which are excluded in our strategy. Moreover, regarding exon/intron definition, Koller's strategy seems to be more suitable since half of the hybrid mini-gene remains in its endogenous context. So, it should provide a more robust estimation of final TS rate in the endogenous gene.

Nevertheless, fluorescence detection is only dedicated to TS detection and so characterization of intron behavior in this artificial environment is not easily conceivable with this setup. Indeed, to evaluate splicing dynamics, fluorescence is a good indicator to evaluate the level of normal *cis*-splicing since GFP-derived protein fluorescence is very sensitive to small changes in its amino acid chain. Even though splicing alteration should be reduced in this strategy, endogenous exons add complexity in this reporter system. The advantage of our reporter system is thus to compare splicing activity of different introns in a similar environment. This enables identification of whether intron elements are shared or

not across selected introns, and so it could bring out new evidence about the category of introns that could be more suitable to promote TS.

To conclude, we have validated the reliability of the hybrid mini-genes and demonstrated how to prevent bias inducing key changes in splicing signaling such as in intron 17. The other major part of this reporter system is PTM, which is an artificial pre-mRNA promoting TS event.

5.1.2 PTM libraries

5.1.2.1 The need for high BD diversity

Our PTMs are composed of three essential elements: the replacement cDNA encoding the 5' half of mTurquoise2, an artificial intron, and a binding domain (BD) an antisense sequence complementary to the hybrid mini-gene target intron.

BD is known to be critical for regulating TS efficiency, as described in previous work conducted in the lab; this also demonstrated that a small shift of BD highly impacts TS efficiency[179]. As illustrated in Figure 67 (adapted from Berger et al.[179]), PTM1 being the best PTM shares 100/150 nt with PTM8 and 50/150 nt with PTM7 and PTM9 with differential impact on TS activity, in particular for PTM9 in which the shared region does not seem to be involved in promoting TS. So, it appears that hotspots can be precisely defined thanks to the screen on multiple BDs overlapping on the same sequence to assess its real TS potential.

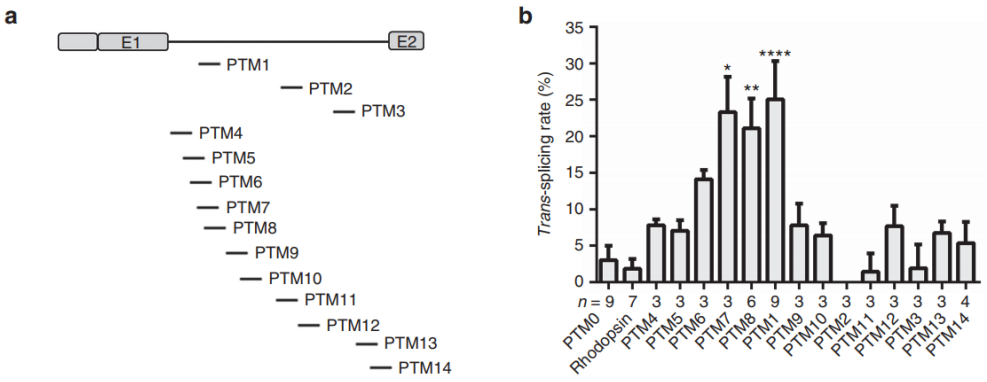


*Figure 67: **Effect of the binding domain and intron sequence on trans-splicing efficiency.** (a) Schematic diagram of the sites of binding domains in Rhodopsin intron 1. Each binding domain is represented by a dash, indicating its location, following a vertical line above the intron diagram. (b) Rate of trans-splicing of PTM0 to 14. PTM0 with no binding domain served as a negative control. Each experiment was performed with a minimum of three culture wells and was replicated several times. The number of replicates is indicated under the x-axis for each PTM. Statistical analysis was performed by comparing PTM0 with all other groups using Dunn's test. Trans-splicing rates were determined at the RNA level, after the transient cotransfection of HEK-293T cells with the PTM and RHO constructs. All values are presented as medians ± SD. Adapted from Berger et al (2015)[179]*

Consequently, we wanted to generate a high BD diversity in order to randomly target intron sequences. With high BD diversity, we expected to be able to detect PTMs with higher TS activity. Therefore, we decided to implement the construction of a random PTM library to target whole introns. Although targeting strategies have, for instance, given some clues as to target splicing site boundaries or spliceosome subunit recognition motifs[158,181,182,185–190], we needed to generate these libraries without prior assumptions.

We generated two PTM libraries targeting introns 3 and 9 respectively. Intron 20 being only 256 bp long, we estimated that it was not large enough to generate multiple PTMs, so only one PTM was assessed, whose BD covers *HTT* intron 20 entirely. However, for these three introns, we used the same "universal" PTM backbone, which is composed of the N-terminal mTurquoise2 sequence, followed by an artificial intron containing a 5'SS, a branch point and a polypyrimidine tract. At the end of the artificial intron, an AhdI restriction site was added to introduce by TA cloning the binding domains generated by random fragmentation of the target *HTT* intron.

5.1.2.2 Library characterization by sequencing

For introns 3 and 9, BDs were generated from sonicated *HTT* introns, and their lengths assessed:

- First, after sonication using a bioanalyzer to assess the quality of sonication
- Second, after transformation of the libraries in bacteria using colony PCR to select positive clones (Data not shown)
- At the end, using Sanger sequencing of individual clones for a precise BD analysis

The last but not least BD analysis was sequencing of a given number of clones.

The number of sequenced clones was defined based on the *HTT* intron length. Indeed, the number of PTMs included in the different libraries was proportional to target length since more PTMs are needed for long introns. However, among sequences in positive clones, only half of these constructs will in theory be included, since BDs inserted by TA cloning have a 50% chance of being in an antisense orientation and thus able to hybridize endogenous mRNA. So twice the number of PTMs was sequenced for intron 3 compared to intron 9. The selected antisense PTMs were characterized by a mean length of 139.3 bp and 139.6 bp, respectively, consistent with raw sonicated bioanalyzer profile and colony PCR data. Taken together, this indicates a high reproducibility of the sonication method used for generating random BDs.

With sequencing, we validated the representativeness of each library regarding respective target introns (Figure 51). Surprisingly, we observed an over-representation of small fragments in the 5' region of intron 3. This sequence is actually composed of multiple repeat motifs of 22-mers characterized by “ttgtatggtttggaggtgctct”, and PTMs containing this motif in its BD will target multiple regions in intron 3 depicted in Figure 54. So, this over-representation is not related to more PTMs with BDs targeting this region but to a pool of PTMs targeting multiple regions distributed in 5’ of intron 3.

To conclude, we have generated PTM libraries and validated their composition by sequencing. Associated with their corresponding hybrid mini-genes, PTM screening was implemented to identify potential hotspots in intron sequences and to select very efficient PTMs.

5.2 Screening analysis

PTM libraries were screened by transient transfection in HEK293T cells, and mTurquoise2 signal was observed using an epi-fluorescence microscope, with automatic acquisition in 16 fields per well. The corresponding images were analyzed with a custom-made macro in ImageJ.

5.2.1 Subdomain definition

To identify potential hot spots on target introns, we needed to assemble the TS activity of several PTMs dependent on target intron region, so we divided each intron sequence into subdomains and averaged the TS results from PTMs targeting each subdomain.

We aimed to have uniform subdomains with approximately the same length, but we faced the following issues:

- Whether we decide to choose small subdomains associated with a more precise identification of hotspot, though small subdomains would not be suitable with BDs exhibiting a mean length around 139 bp. In this case, a high level of overlap would be observed between adjacent subdomains thus reducing the precision of TS estimation for each subdomain.
- Or we decide to choose large subdomains, more relevant with large BD resolution but in this situation, identification of hotspots would be less clear.

In this context, we needed to have objective parameters to determine subdomain length. So, we decided to meet this issue halfway to have subdomain lengths around half the mean BD length. Moreover, as we wanted to compare the screening results obtained with the selected introns, i.e. 3 and 9, we needed to have subdomains with similar characteristics in term of splicing signaling such as BP and PPT, whose location can be more or less distant from the 3'SS. An adequate indicator to localize these essential splicing signals is AGEZ, which is a region in the 3' part of the intron where BPs are potentially found. So, we decided to choose AGEZ length as the intron specific criterion and to define subdomains as twice the length of AGEZ, which gives lengths of 86 bp and 72 bp respectively for intron 3 and 9 subdomains.

To conclude, we have based our TS analysis to define hotspots on introns 3 and 9. Target intron choice was motivated by their differential splicing activity, to investigate their impact on TS efficiency. However, we faced other limits to select target introns, described hereinafter.

5.2.2 Target introns

5.2.2.1 *Limitations of intron choice given AAV vector cloning capacity*

Targeted *HTT* introns were first chosen according to their location relative to exon 1, which is an overriding criterion. Since replacement cDNA of the final PTM will encode all exons located upstream of the target intron, the length of replacement cDNA is an essential parameter to consider. Indeed, PTM sequences will ultimately be introduced into an AAV vector, which, to date, is the vector of choice for sustained expression in the brain; however, they have a cloning capacity limited to 4.7kb[199]. Being close to exon 1 leads to PTMs encoding for a small replacement cassette and, as depicted in Table 8, if we target intron 20, the replacement cassette will be 3,013 bp long; but with promoter, 5' and 3' UTR (untranslated regions), maximum AAV cloning capacity is reached.

Table 8: ***List of the first twenty introns of HTT***

HTT intron	Intron length (bp)	Replacement cassette length (bp)	MaxEnt scores
Intron 1	11850	579	9.4
Intron 2	12251	663	8.55
Intron 3	**4429**	**784**	**6.91**
Intron 4	1473	844	7.93
Intron 5	1848	924	10.86
Intron 6	7880	1063	9.11
Intron 7	639	1205	9.31
Intron 8	4964	1384	9.79
Intron 9	**1456**	**1589**	**1.91**
Intron 10	2612	1637	5.88
Intron 11	1634	1718	6.3
Intron 12	2319	2059	8.54
Intron 13	257	2183	7.43
Intron 14	862	2302	9.16
Intron 15	240	2414	10.13
Intron 16	786	2552	8.14
Intron 17	**93**	**2711**	**10.52**
Intron 18	1489	2809	7.75
Intron 19	1363	2949	10.86
Intron 20	**258**	**3013**	**7.79**

Intron lengths were extracted from the NCBI Aceview data base. Replacement cDNA lengths were estimated by addition of all exon lengths upstream of corresponding introns.

5.2.2.2 *Introns excluded from the screen*

Among these first 20 *HTT* introns suitable for the design of a TS therapeutic product, we decided to quickly exclude some of them like *HTT* introns 1 and 2 which are too large (11.85kb and 12.25kb) to be introduced in a lentivirus whose maximum cloning capacity is 9kb[199]. Indeed, at the beginning of the project, we aimed to generate stable cell lines expressing the different hybrid mini-genes using lentiviral gene transfer which has the capacity to insert its transgene into the target cell genome.

HTT intron 6 was selected at the beginning of the project to be screened because being large (7.88kb) allowed a high diversity of PTM BDs. Cloning steps were challenging but we succeeded in generating the corresponding hybrid mini-gene and the associated PTM library. We realized that given the intron length, screening could not have been completed in the time course imposed by the PhD program. Indeed, it would involve a very high number of PTMs to cover all the intron. So, the study of *HTT* intron 6 was postponed, and we focused on introns 3 and 9.

In the next part, we will discuss the rationale for selection of each target intron and their screening results.

5.2.3 Intron 3

5.2.3.1 *The "standard" intron*

The first *HTT* intron included in the project was *HTT* intron 3 for the following reasons:

- It is very close to *HTT* exon 1, thereby leading to a PTM encoding for a small replacement cassette, which corresponds to *HTT* exons 1, 2 and 3 and is only 784 bp long.
- Being 4.4 kb long, it is long enough to allow a high diversity of BDs.

5.2.3.2 *Screening results*

The PTM_{i3} library showed high variability of TS activity across its target intron (Figure 53). We observed a drop in TS activity in the 5' region of the intron, and more specifically in subdomains containing repeat motifs (Figure 54). These repeat motifs seem to correspond to variable tandem repeats which are classified as minisatellite regions with in our case a motif length of 22bp defined as "ttgtatggtttggaggtgctct". Tandem repeats are often located in gene regulatory regions[200] such as exon/intron boundaries with a potential role in alternative splicing[201]. It is not clear by what mechanism they inhibit TS but we can speculate that repeat regions may not be available for PTM hybridization. Indeed, repeat sequences are known to exhibit complex tridimensional conformations[202–204], preventing PTM access.

To identify hotspot subdomains, we generated a top 10 list from MFI and integrated density scores (Table 9 and Table 10). Seven subdomains appear in both lists and they come with neighboring subdomains such as:

- 17, 18 and 19
- 37 and 38
- 45 and 46.

Table 9: ***Top 10 MFI subdomains***

Ranking	#Subdomains	MFI score
1	**18**	1.066
2	**19**	1.000
3	**38**	0.988
4	10	0.980
5	**37**	0.948
6	**17**	0.904
7	40	0.889
8	48	0.889
9	**45**	0.878
10	**46**	0.878

Table 10: ***Top 10 integrated density subdomains***

Ranking	#Subdomains	Integrated density score
1	**37**	1.133
2	**38**	1.055
3	47	1.007
4	28	0.986
5	**19**	0.981
6	**18**	0.966
7	**17**	0.935
8	**45**	0.906
9	**46**	0.906
10	44	0.905

However, we decided to not take into account subdomain 17 since it results in a single PTM evaluation.

So, these couples of subdomains are paired because they share PTMs whose BD is located on both subdomains. To identify more precisely hotspot areas in these regions, differential analysis on BD domains is needed. Indeed, we observed multiple overlaps of PTMs exhibiting different TS scores and it is necessary to align BD sequences to identify commonalities between efficient PTMs which are excluded from weaker PTMs. Using matching sequence alignment and TS, it would be possible to implement a TS probabilistic model in a weighted matrix in order to take full advantage of a high content screening approach.

This analysis is needed to further refine hotspot descriptions because as it stands, identified positive regions are too large and not defined enough to make conclusions. In parallel, more PTMs could also be added in a screening experiment to complete subdomain #17.

5.2.4 Intron 9

5.2.4.1 *The alternative intron*

The second *HTT* intron introduced in the study was intron 9, meeting the same criteria as intron 3, but with a more limited PTM diversity. However, the main reason for the choice of intron 9 was its unusual donor site motif, which is “GC” instead of “GT”, thus providing a particular splicing activity. This change leads to a very low MaxEnt score of 1.91 that is extremely different from the other first 20 introns of the *HTT* gene, the second lowest score being for intron 11 with a MaxEnt of 6.3. Taken together, this intron is considered as an alternative intron which is spliced in the most common *HTT* variant mRNA (aAug10) according to the Aceview NCBI dataset (Appendix 6)[205].

5.2.4.2 *Intron 9 and 5'SS competition*

Using MaxEnt analysis, we postulated that weak introns could be a preferential target to elaborate a TS therapeutic strategy. This hypothesis is based on the study of Philippi et al.[158], suggesting that introns with a low MaxEnt are more likely to perform exon skipping. TS indeed exhibits some similarity with exon skipping since it involves mis-matching between donor and acceptor sites which can be predicted by 5'SS strength[158]. This mis-matching is likely to result from less robust recognition or stabilization of splice elements at weak exon/intron boundaries. Based on this hypothesis, we postulated that if a PTM exhibits a higher MaxEnt than the target intron, it could favor its own splicing signaling and thus promote TS (Figure 68).

PTM MaxEnt is 4.36 (5'SS = ggaGTAAGA) and is universal across all PTMs tested in this study. Indeed, MaxEnt is determined from exon/intron boundary sequences corresponding to the junction between the mTurquoise2 N-terminal exon and the artificial intron, and the only parameter that changes between PTMs is the BD. In this context, we expected that intron 9 with MaxEnt 1.91 would exhibit higher TS activity compared to other introns (with a much higher score of 6.91 for intron 3 and 7.79 for intron 20).

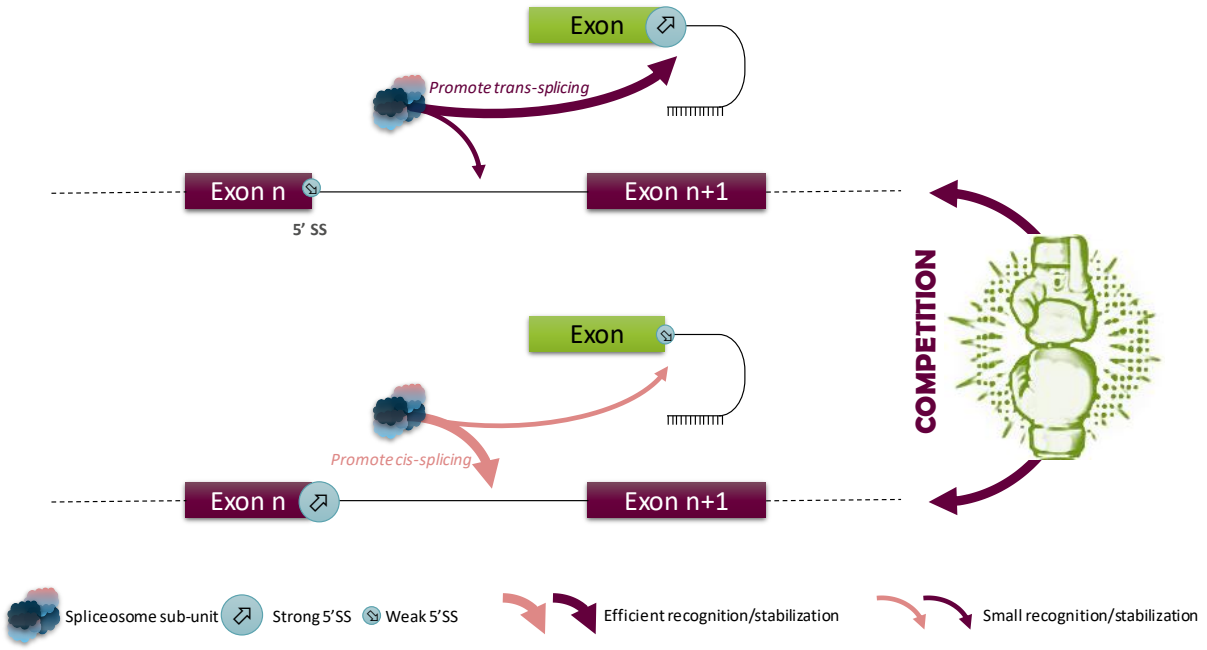


Figure 68: ***Representation of donor site competition between PTMs and endogenous 5'SS.*** *(Upper) Competition in favor of TS. (Lower) Competition in favor of cis-splicing. Cloud-like shapes represent spliceosome subunits recruited whether at PTM 5'SS or at endogenous 5'SS (depicted by orange circles, with size proportional to their strength). Arrows represent possible splicing processes based on 5'SS strength.*

5.2.4.3 *Screening results*

Regarding intron 9 results, no clear evidence demonstrated higher TS activity in a given subdomain even though integrated density showed more variation in TS score with trend of high TS activity around subdomains #1, #6 and #15 (Table 11). Analysis methods to improve hotspots precision described for intron 3 could also be implemented with intron 9 data.

Table 11: ***Top 5 subdomains in integrated density analysis***

Ranking	# Subdomains	Integrated density
1	15	1.308
2	6	1.235
3	16	1.203
4	14	1.196
5	1	1.175

Based on mRNA quantifications using DrdI, both PTMi9s exhibited a 2.3% TS rate, which is the lowest score. If further experiments confirmed this ratio, this would invalidate the MaxEnt competition hypothesis. This could be explained by the fact that in the study of Philipi et al., they worked on a 3' replacement strategy whereas in our study, we assessed a 5' replacement strategy. This would indicate that 5' and 3' replacement involve differential signaling and in case of 5' replacement, high recruitment of spliceosome subunits at endogenous 5'SS is required for an efficient TS.

5.2.5 Intron 20

5.2.5.1 *The small intron*

The last introns included in the project were the 17th and 20th, chosen because of their short length, 93 and 258 bp respectively. As mentioned before, our screening strategy does not consider the theory of exon/intron definition, which postulates that introns superior to 250 bp are regulated by splice elements located on exon sequences, while introns inferior to 250 bp are regulated by splice elements located on intron sequences. In our hybrid mini-genes, *HTT* introns are indeed taken out of their physiological genomic context, leading to potential deregulation of the exon/intron definition. As a consequence, we selected these two short introns to assess whether disturbing the intron elements is more efficient in short introns than in longer ones. According to this hypothesis, introns 13 and 15 could have been selected as well.

5.2.5.2 *Screening results*

Following this hypothesis, only intron 20 was evaluated since intron 17 was excluded mainly because of its high shift of MaxEnt score between *HTT* and *XFP*, indicative of a possible important change in splicing activity. We designed a single PTM targeting the whole of intron 20 and showed that it exhibits TS activity using fluorescence. mRNA data indicated a TS rate of 4.0% according to the DrdI restriction site, which was still lower than for intron 3. An alternative experiment (currently under way) could be to design a shorter BD targeting just portions of intron 20. Moreover, since this strategy is easy to implement, assessment of introns 13 and 15 by comparing their splicing activities to each other in a common artificial context, could be interesting as well.

To conclude, we have validated TS activity in all chosen introns, but we observed relatively low TS efficiency regarding the extensive number of PTMs assessed. However, the absolute level of TS also remains to be clarified. Indeed, our mRNA quantification methods based on restriction sites seem to be very accurate, as demonstrated in the study of Berger et al., but in our particular case, protocol optimization is needed.

Essentially, the main issue in comparing TS studies is TS quantification methodology. This differs greatly in the literature, thus providing a wide range of TS efficiency, an issue that will be discussed below.

5.3 TS quantification methods

5.3.1 Precise rate of correction

To precisely characterize TS activity of PTMs, it is necessary to develop tools to determine the exact rate of correction. In the literature, two categories of methods appear, involving direct and indirect approaches[177].

Indirect approaches involve the use of reporter genes, encoding for a functional reporter protein in the case of TS such as LacZ/ß-galactosidase activity[206,207] or our strategy using hybrid mini-gene/mTurquoise2 fluorescence. TS activity can also be indirectly determined by measuring phenotype improvement, such as normal protein localization[179] or cell death[191,208]. These methods present the advantage of having an easy experimental readout, suitable for the evaluation of high numbers of samples. However, the precise rate of correction mediated by TS events cannot be extrapolated from these kinds of approaches, since additional processes can impact final results. For instance, with our fluorescence-based approach, we performed exponential regression to estimate TS efficiency from negative (PTM_0) and positive (full mTurquoise2) fluorescent controls used in screening (Figure 59). However, background noise was dependent on introns and induced a small shift of the exponential equation definition resulting in potential bias in TS estimation. Thus, indirect approaches

are fine for screening and identification of potent PTMs, but do not allow precise determination of TS rate, which needs a direct quantification strategy.

Direct approaches consist of quantification of *cis*- and/or *trans*-spliced products at the mRNA or protein level. This estimation is commonly performed using end-point PCR based on band intensity or qPCR with primer pairs specific to *cis*- or *trans*-splicing products even though end-point PCR could potentially induce amplification bias [207,209,192,206,210,188,187]. At the protein level, western-blots are also used to identify corrected protein and to quantify TS rate. Normal protein expression is measured alone or compared with mutated proteins with size discrimination parameters, or using specific antibodies [188,207,211,212]. In our case, some of these approaches were impossible or very difficult to implement such as qPCR and western-blotting, because of high sequence homology between *cis*- and *trans*-spliced products (Figure 41). Both indeed are derived from GFP, and thus are the same size and recognized by the same antibodies, making western-blot quantification too challenging to be performed. Along the same lines, small differences in 5' sequences were not enough to design specific primers for each kind of spliced product, resulting in persistent background amplification in negative controls.

However, these quantification methods can also be categorized as relative and absolute methods. To date, most common methods use relative parameters such as % of GFP positive cells using flow (indirect) cytometry or level of fold change using qPCR (direct). These methods provide, respectively, qualitative and quantitative levels of TS but exclude an essential parameter, the correction rate in cells, i.e. the percentage of mRNA/protein pool which is actually *trans*-spliced. Indeed, % of positive cells corresponds to a binary signal about whether TS occurs or not in cells, so providing information about the frequency of splicing in cell populations, whereas qPCR compares the level of expression with controls, thus providing a relative TS efficiency.

With our screening strategy, direct TS quantification can be accurately performed at the mRNA level as demonstrated for *RHO*. Even if some protocol optimizations are still needed for *HTT*, we are likely to be closer to a *bona fide* TS rate.

We assessed fluorescence intensity by flow cytometry and sorted positive cells to analyze TS-mediated correction at the mRNA level. In this context, the high sequence homology between *cis*-spliced and *trans*-spliced raises the advantage of reducing bias in terms of amplification efficiency for PCR. Then we used a couple of restriction endonucleases with PstI for *cis*-spliced products and DrdI for *trans*-spliced products. It is essential to have both to quantify TS for several reasons. TS rate depends on restriction endonuclease efficiency, in particular on their capacity to recognize the restriction site and

to cleave DNA. However, each restriction endonuclease has specific requirements for a complete cleavage reaction, including a restriction time which has to be long enough to cut all target DNAs, but not so long as to produce star activity, resulting in non-specific cuts. An additional issue specific to our setup is that we have a high amount of small DNA fragments (650 bp), which corresponds to a more challenging case compared to plasmids which are bigger. The difference between the two is that for a similar level of DNA, small fragments require more endonuclease, since there will be more targets. The concentration of endonuclease is limited to 10% of total reaction in order to prevent star activity, which can also be induced by a high glycerol concentration. Moreover, the use of restriction endonucleases specific to each spliced product enables the comparison of the two TS rates, for a more accurate range of TS activity. So, many different parameters have to be taken into account for a successful restriction cleavage and using two different restriction sites will reduce the variability related to technical issues and enhance the quantification precision.

To our knowledge, this quantification method has never been developed and implemented in other TS studies with a reporter system, and it seems to be the most accurate method to compare TS efficiency across targeted introns. The main limit of this approach is that the quantification precision relies on restriction endonuclease type, and the cleavage protocol is critical for reliable results.

5.3.2 Complementary indirect quantification using our fluorescent reporter system

The reporter system we developed allows the detection of *cis*-splicing events as well. Indeed, before establishing emission spectra of each fluorescent spliced product, we aimed to evaluate TS events by the detection of an increased mTurquoise2 signal and simultaneously a decreased XFP signal in the same experiment. However, once we established the emission spectra of XFP (Figure 42), we realized that it had a very low peak at 488nm excitation, and although mTurquoise2 is preferentially excited at 440nm with wide and high emission spectra, it is also excited at 488nm, covering the XFP signal, therefore we decided to focus on mTurquoise2 signal only.

However, in theory, it is still possible to estimate TS rate by measuring the reduction of XFP fluorescence; if the PTM does not encode for the mTurquoise2 N-terminal part but for another replacement cDNA, the only fluorescence detected would be XFP's. This strategy could be implemented at a later stage, with PTMs encoding for *HTT* exon replacement cDNA, i.e. the final PTM construct to be assessed for therapeutic efficiency in cellular and/or animal models of HD.

More simply, in our screening strategy we have changed exon/intron definition of our artificial hybrid genes, compared to the endogenous genomic context of the *HTT* gene, but our approach brings essential advantages regarding multiple TS evaluation methods available like:

- Indirect quantification using fluorescence methods:
    - Level of fluorescence intensity (epi-fluorescence)
    - % of positive cells (Flow cytometry)
- Direct quantification at the mRNA level:
    - Endonuclease restriction specific to *cis*- and *trans*-spliced products (Bioanalyzer)

This strategy takes advantage of introducing various target introns in the same environment, allowing the possibility of comparing their splicing activity in a neutral context. Finally, this screening strategy could be used to improve PTMs. Indeed, it has been demonstrated that TS activity can be promoted by the addition of specific *cis*-acting element(s) in PTM sequences. For instance, introduction of artificial or endogenous intron sequences in replacement cDNA in between exons increases the TS rate significantly[179].

6 Conclusion

To conclude, we have successfully characterized and validated a new fluorescent reporter system to detect TS events. TS screening was performed in two *HTT* introns including 3 and 9, where hotspot regions were also identified. Intron 20 also demonstrated TS capacity when targeted by a PTM. Further preliminary experiments were performed at the mRNA level for a precise TS quantification with TS rate ranging from 2.5% for intron 9, to 8.5% for intron 3. These data are consistent with a previous TS study in HD led by Rindt et al., which reached a 5.4% TS rate in their first study. Recently, they optimized their PTM design by targeting branch point and obtained 10 to 15 % TS rates with some positive effect on HD cellular phenotypes in iPSCs. Indeed, they assessed ATP levels in HD neuronal progenitor cells which exhibit energy deficiency as found in HD patients. They observed that ATP levels were significantly improved in HD progenitors when PTMs are expressed. They also evaluated the survival of HD iPSCs which is known to be reduced when BDNF is withdrawn. Similarly, PTMs seem to mediate positive effects on iPSC vulnerability to death when BDNF and other factors promoting its expression are removed from culture media[182].

Although their quantification method is based on end-point PCR by measuring band intensity, which can result in overestimation, their data suggests that a low TS rate is enough to observe beneficial effects in relevant HD models.

So, we can assume that our quantification method is more accurate for low yields such as those observed in our study. Thus, our best PTMs could also exhibit some positive effects on cellular impairments related to energy and survival deficits, which seems to be more sensitive to the recovery of HTT function, as demonstrated in Rindt et al. We overall obtained low TS efficiency whereas by an extensive screening we expected more variability in favor of higher TS rate. We first hypothesized that 5’ replacement is more challenging than 3’ replacement since 5’ replacement efficiency would depend on the first pre-mRNA cleavage during splicing, which has to be inhibited in endogenous 5’SS.

The splicing process, being highly conserved across eukaryotic cells, triggering the 5’ splicing mechanism is likely to involve other complex processes through multiple factors and spliceosome subunit recruitment which are supposed to more demanding than in 3’. The challenge is to inhibit this natural process in endogenous pre-mRNA while promoting it on PTMs. This mechanism remains to be explored since no study ever evaluated this aspect of TS. What has been published is, however, interesting since the 3’ replacement strategy has been studied more compared to 5’ (23 versus 9 studies, respectively). So, a comparison of 3’ replacement efficiency versus 5’ would be hard to clarify because differences in TS activity would be related to the high number of studies in 3’ replacement, allowing a higher chance to observe high TS activity.

Nevertheless, a recent study evaluated TS efficiency in 5' and 3' with PTMs carrying the same BD. They observed the same rate of about 1% TS[194]. Moreover, their quantification method was the same in both strategies and was based on two end-point PCRs with similar amplicon length including:

- one specific for *trans*-spliced products
- the other amplifying both *cis*- and *trans*-spliced products

They then calculated the ratio between values of band intensity, thus providing a direct and absolute quantification. Band intensity quantification is less precise than bioanalyzer output, and even if the two PCR amplification products are close in terms of location and length, primer differences can also induce a bias in amplification efficiency. Despite these limitations, reducing slightly the precision of quantification, these results appeared to provide consistent TS quantification, suggesting that 5' and 3' replacement occur at the same level for a given binding domain. Regarding this above-mentioned study, it is becoming increasingly clear that TS efficiency variability as seen in the literature is partly related to quantification methods.

If this hypothesis is confirmed, then characterization of TS hotspots could be the next point to evaluate, which reinforces the relevance of the screening strategy that we developed during this work. Indeed, we have depicted TS activity patterns across two different introns. Using bioinformatics tools[213,214], it is possible to match a given intron region with key splicing elements. Open data are now more and more available online, allowing us to go further in these introns analyses.

In parallel, more PTMs could be screened. Indeed, only 18.7% and 20.7% of the clones isolated from intron 3 and 9 libraries respectively were assessed during this work. Moreover, from the identification of best hits presented here, other PTMs have been screened with higher mTurquoise2 signal, indicating very promising TS activity. Unfortunately, we were not able to show any further experiments from these hopeful PTMs due to time constraints regarding the PhD defense deadline, but PTM screening is still on going.

7 Perspectives

Once the most efficient PTM has been identified with the highest TS rate, it will be necessary to validate the possibility of using TS as a curative treatment for HD. For this purpose, the next step will be to investigate the relationship between the percentage of *HTT* correction and actual improvement of HD phenotype. Indeed, although the dominant nature of the mutation suggests that a maximum correction will be needed, it is still unknown how much TS must be reached to expect therapeutic benefit in HD patients. However, some speculation can be made from silencing strategies, which revealed promising results.

The use of ASO leads to a reduced level of mHTT mRNA decreasing by 72% in BACHD mice right after two weeks of ASO infusion[215]. A similar approach with shRNA also exhibits a long lasting effect with 65% to 75% reduction after 9 months of treatment, associated with better survival of striatal neurons, a reduction of mHTT inclusions and a recovery of vesicular transport in an HD rat model overexpressing the N-terminal fragment of mHTT [216,217]. In these studies, silencing demonstrated its capacity to reduce significantly, but not totally, mHTT expression with improved HD phenotype.

With TS, even if we have a lower silencing rate, we provide at the same time a corrected version. It has been demonstrated that depletion of normal HTT does not lead to any side effects in the adult brain, and would be safe in a silencing strategy[218]. However, controversial data indicates that it could lead to multiple deficits including behavior changes, motor impairments, mice survival decrease, and other neuropathological deficits[219]. The relevance of HTT function and HTT need in the normal adult brain is still under debate, although nonspecific allele silencing approaches gave the same results as a specific allele strategy, suggesting that in a pathological context, normal HTT could be negligible.

However, the added value of the combination of silencing and correction is complicated to estimate in terms of a therapeutic effect. We can hypothesize that in the normal adult brain, HTT can be removed quite safely, but in the HD brain expressing mHTT continuously from birth, it induces an environment which triggers neurodegeneration. This could be partly related to continuous BDNF withdrawal all lifelong. In this context, the use of TS, which silences and corrects mHTT at the same time, could significantly slow down disease progression. Indeed, lowering the ratio of mutant versus normal HTT could be used as a booster of the silencing effect, thus reducing synergy in HD pathology.

Taken together, a TS strategy seems to be optimum with HD pathology. With a highly efficient TS reaction, we could get closer to results obtained using silencing approaches, and even better since normal HTT expression will be restored to a physiological level. The most promising hypothesis is that future HD treatment could associate a PTM to correct *mHTT* pre-mRNAs and ASOs to eliminate the remaining mHTT mature mRNA. Forthcoming results from the Ionis-HTTRx clinical trial where a non-

allele specific ASO is evaluated in 48 early manifest HD patients (details in Appendix 7) will provide more evidence for impacts of a silencing approach in HD patients and thus for a potential bi-therapy. In this clinical trial, the main concerns are towards safety and tolerability of the active agent, but other outcomes such as ventricular volume and cognitive tests are also evaluated.

To validate the possibility of using TS alone or combined with silencing strategy, a lot of work remains to be accomplished. Once this TS project in the artificial context of hybrid mini-genes is completed, it will be necessary to evaluate TS capacity in *HTT* by removing the replacement cDNA encoding for the mTurquoise2 artificial exon, and replacing it with the corrected *HTT* exons.

In this context, the first goal will be to compare TS rate in artificial and *HTT* genes and then associate this correction rate with a given level of HD phenotype improvement.

To assess TS rate, direct and absolute quantification can be developed. To sort TS positive cells, we can add a sequence encoding for DsRed upstream to corrected HTT exons, in PTM, so that when PTM will be expressed such as in final *trans*-spliced mRNA, it will lead to the expression of DsRed. To discriminate mHTT from corrected HTT, it is also possible to introduce a silent mutation generating a restriction site into the replacement cDNA encoding for corrected *HTT* exons, thus allowing precise assessment of the percentage of trans-spliced HTT mRNA.

Then, to evaluate TS positive effects on HD pathology, it is essential to bear in mind key issues in gene therapy, including delivery system. In the context of HD, recombinant AAV, a non-integrative vector, is likely to be the best candidate since it is known to transduce neurons very efficiently and to stay as an extra-chromosomal episome in the cell nucleus. This allows a long-term expression in non-dividing cells such as neurons, leading to a continuous expression of PTMs in the brain. Moreover, various AAV pseudotypes lead to enhanced specificity of transduction in terms of cellular type. Some of them have demonstrated their capacity to be more specific to the brain with serotype 2, which is the most widely used serotype. For our concerns, AAV2/9 could be a good option with a large expression in striatal cells[199] The major limit of AAV use in therapy is its limited transgene capacity, but with a TS approach, this limitation might not limiting since only a portion of the target gene sequence is present in the PTMs.

In parallel to vector choice, it is also essential to consider each HD model is the most suitable to evaluate TS therapeutic benefit. For instance, the most studied HD model is the R6/2 mouse strain[220], which is characterized by a strong phenotype associated with a 150 CAG expansion. Motor and cognitive symptoms appear early in life from 5-6 weeks and mice are not likely to survive after 13

weeks. As described in Li et al in 2005[220], these mice with a severe phenotype are particularly suitable for disease modifier types of treatment such as those targeting :

- Aggregates and aggregate formation[221–223]
- Gene transcription[224–226]
- Protease inhibitors[227,228]
- Apoptosis inhibitors[227,229,230]
- Exotoxicity[231]
- Energy metabolism and diet[227,232,233]
- Anti-inflammatory agents[234,235]
- Trophic factors[236]

However, this kind of model is not suitable with our strategy since it expresses *HTT* exon 1 only. As we target intron sequences of the pre-mRNA, no TS reaction can occur with these models.

The actual model to evaluate TS effects must carry the full human *HTT* gene, including exon and intron sequences. The only matching animal models that exist to date are BACHD rodents, and we will focus on BACHD rats, because they have a more complex brain and thus are closer to humans. This rat model carries the full human *HTT* gene, with 97 CAG/CAA repeats under the control of the human HTT promoter. The CAG/CAA repeats also encode an expanded polyglutamine tract, but this motif is known to reduce the instability of the expansion in tissues.

This rat exhibits several aspects of HD neuropathology such as mHTT aggregates, a lower level of cortical BDNF, anxiety behavior[237], emotional[238] and cognitive[239] deficits and motor impairments which are particularly marked[237,240]. In parallel, this model is also characterized by an early altered transcription level of HTT from 3 months of age. It appears that the level of mHTT is correlated with the number of altered genes[241]. Indeed, two BACHD rat lines have been generated by Yu-Taeger et al expressing human mHTT differentially, amounting to a 4.5- (TG5) and 2.7- (TG9) fold increase compared to endogenous[57]. As depicted in Figure 69, they observed more down and up-regulated genes in TG5 than in TG9, so using this parameter, we could have a TS rate-dependent level of altered genes, adequate for a gene therapy evaluation.

However, alternative to animal models are also available such as iPSC derived from HD patients. In these models, phenotype depends on HD patient cell lines and has to be compared with age- and sex-matched healthy control iPSC lines. So, consideration of HD phenotype is critical. In the best characterized HD iPSC lines, common impaired processes can be observed like mHTT aggregate

formation and cell death after BDNF withdrawal[242–244]. Other HD mechanisms have also been detected such as:

- Gene and protein expression changes[244,245]
- neuronal differentiation and neurite length decrease[245]
- energy metabolism deficits[243]
- reduced capacity to fire spontaneous and evoked action potentials[243]
- altered calcium homeostasis [243,244]

To conclude, cellular and animal models are essential complementary tools to validate TS.

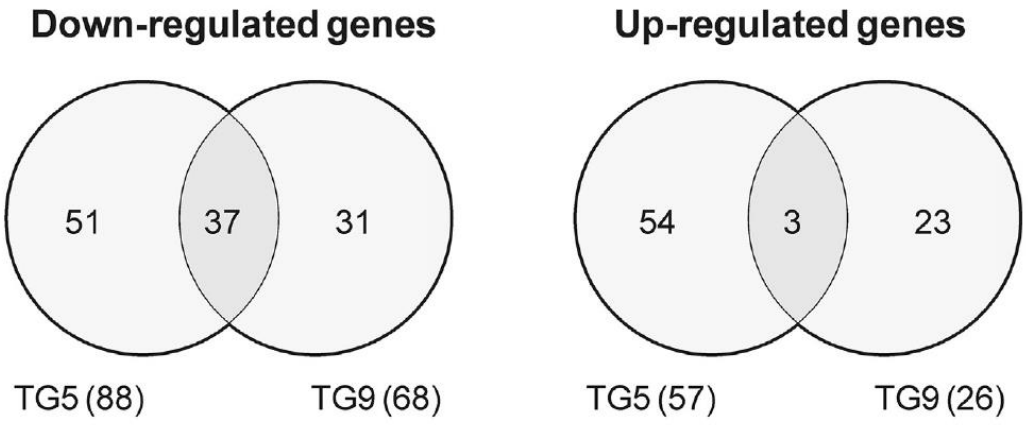


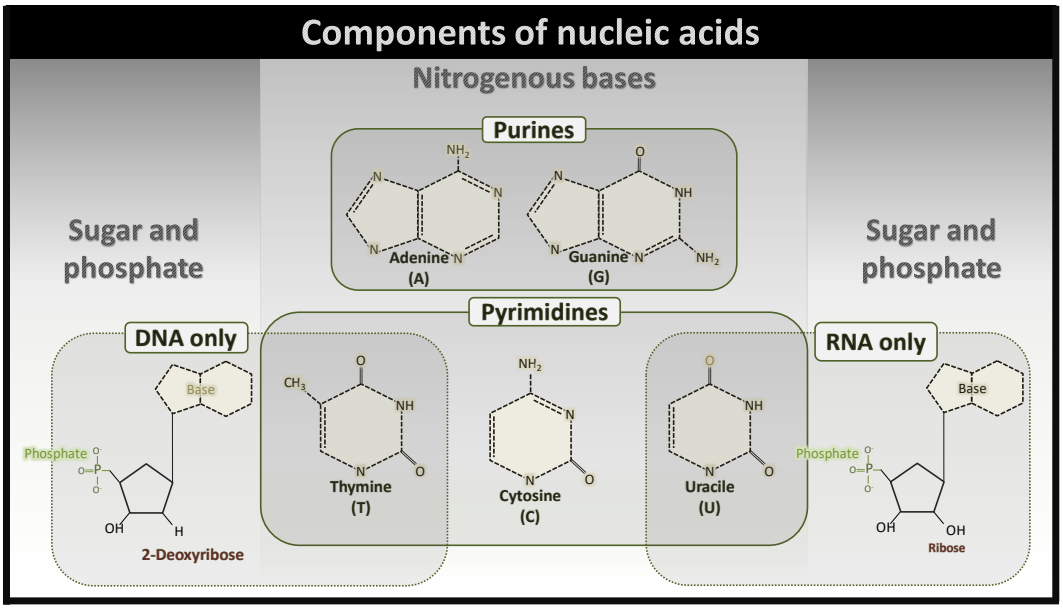
Figure 69: ***Venn diagram showing the number of up- and down-regulated striatal genes in BACHD TG5 and TG9 rats in comparison to wild type littermates at 12 months of age.*** *Genome-wide gene expression was analyzed in the striatum of BACHD transgenic rats and wild type controls. The numbers of up- and down-regulated genes are indicated in the brackets for each line. The intersections between lines indicate the numbers of shared genes, whereas the numbers of unique genes are shown in line-specific areas. Adapted from Yu-Taeger et al (2017)[241]*

To go further in the use of TS for HD, it should be emphasized that the TS strategy has already had multiple successes. One of the leading teams, led by Dr Murauer, focuses on a severe skin disease, dystrophic epidermolysis bullosa, caused by various mutations in the *COL7A1* gene. In the case of this pathology, the relevance of TS is that a single PTM can correct approximately 40% of mutations leading to the disease[246]. In their latest published study, they demonstrated the feasibility of using TS in ex vivo gene therapy. They have developed human keratinocyte cell lines expressing endogenously PTMs after lentiviral integration. They focused on design, packaging and delivery of their PTM, and after xenograft in the mouse model, they have shown that this strategy could be safely assessed in a clinical setting[246]. In the same category of disease caused by different mutations in the same gene, we can mention retinitis pigmentosa, which is a pathology leading to different degrees of blindness. In the lab, we have previously targeted mutations in the rhodopsin gene causing the disease using TS with in vivo experiments using the AAV viral vector. This study demonstrated the feasibility of TS and was a very encouraging prospect for recessive forms[179].

TS has also been used in tau pathology to act on the splicing process which mediates tauopathy. Tau is a microtubule-associated protein which is known to be expressed in six different isoforms in neurons. Two of these isoforms differ partly by the number of microtubule binding repeats with three or four repeats. In pathological cases, the ratio between these two variants can be unbalanced leading to cellular impairments that can trigger tauopathies, such as Alzheimer's disease. TS was used to model different ratios between these two isoforms. By modulating their relative level of expression, Lacovich et al. suggest that TS could be used as a therapeutic approach[247]. In this study, the low TS activity was an advantage to finely tune the splicing process, which is a novel strategy.

Other uses of TS have been reported in the literature including a TS reaction between two different but compatible AAV vectors. The rationale is to express a large gene by two different viral vectors separately and to recombine them by a TS reaction to obtain a large transcript whose size would otherwise not be compatible with the AAV packaging limit. This method was first described in retinal disease[248]. Later, this strategy was adapted to Duchenne muscular dystrophy (DMD)[249] with the use of three AAVs allowing the full length expression of the DMD gene. This suggests that this approach could be extended to larger and larger genes.

Although TS is still limited by low efficiency, the overall conclusion is quite promising regarding all the possibility provided by TS as a potent gene therapy tool. Although the first TS studies were dedicated to recessive genetic disease because therapeutic benefit is easier to observe, TS has demonstrated some evidence for dominant disease application as well. Moreover, TS can go beyond the raw expected achievement of silencing and correction performance and also be used to finely regulate the splicing process, which can lead to multiple diseases or to express large genes not suited to current vectors. If these aspects of TS develop, we can speculate that one day, TS could go one step further and be used to generate cellular and animal models by a precise control of gene expression. Regarding all the opportunities raised by TS, this project has provided essential information to improve this technology.



The genetic code

Codon language linking nucleotide triplets and amino acids

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gln	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

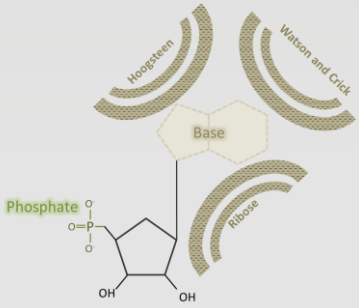
Base Pairing

Canonical pairing

- Through two (=) or three (≡) hydrogen bonds
- Follows Watson-Crick base pair rules:
  - G ≡ C
  - A = T (in DNA)
  - A = U (in RNA)
- In between any kind of nucleotide molecule
  - DNA ⇆ DNA
  - DNA ⇆ RNA
  - RNA ⇆ RNA

Non-canonical pairing

- Specific RNA capacity
- Includes:
  - Wobble pairing (G = U)
  - Hoogsteen pairing
  - Ribose face paring



Other direct comparisons

	DNA	RNA
Shape	• Mostly double-stranded molecules ▪ with complete complementary strand ▪ along its entire length • Regular and uniform shape	• Mostly single-stranded molecules ▪ with transient pairing with itself or other target • Irregular and multiple shapes
Size	• Very large molecule • Composed of millions of nucleotides	• Smaller molecules • from 20 to thousand nucleotides

Appendix 2: Factors regulating the splicing of exon 10 of MAPT pre-mRNA encoding for tau. Adapted from Park et al[160]

CIS-ELEMENTS	SEQUENCE	EFFECT ON E10
SC35-like enhancer	E10, TGCAGATA	Inclusion
Polypurine enhancer (PPE)	E10, AAGAAGCTG	Inclusion
A/C-rich enhancer (ACE)	E10, AGCAACGTCCAGTCC	Inclusion
Exonic splicing silencer (ESS)	E10, TCAAAGGATAATATCAAA	Exclusion
Exonic splicing enhancer (ESE)	E10, CACGTCCCGGGAGGCGGC	Inclusion
Intronic splicing silencer (ISS)	I10, tcacacgt	Exclusion
Intronic splicing modulator (ISM)	I10, cccatgcg	

SR PROTEINS	EFFECT ON E10
SRSF1: ASF, SF2, SRp30a	Inclusion
SRSF2: SC35, PR264, SRp30b	Inclusion
SRSF3: SRp20	Exclusion
SRSF4: SRp75	Exclusion
SRSF6: SRp55, B52	Exclusion
SRSF7: SRSF3 9G8	Exclusion
SRSF9: SRp30c	Inclusion
SRSF11: P54, SRp54	Exclusion

OTHER RNA BINDING PROTEINS	EFFECT ON E10
RBM4	Inclusion
Tra2β	Inclusion
DDX5 (RNA helicase p68)	Inclusion
hnRNPE2 and hnRNPE3	Inclusion
hnRNPG	Exclusion
SWAP	Exclusion
CELF2	Exclusion
CELF3, TNRC4	Inclusion
PTBP2	Inclusion
PSF	Exclusion

MIRNAS	EFFECT ON E10
miR-9	Exclusion
miR-124	Exclusion
miR-132	Exclusion

KINASES	EFFECT ON E10
CLK2	Exclusion
PKA	Inclusion
PKA	Inclusion
DYRK1	Exclusion
DYRK1	Inclusion
DYRK1	Exclusion
DYRK1	Exclusion
GSK-3β	Exclusion

Fluorescent protein	Sequences
GFP	atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccaca agttcagcgtgtccggcgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccaccggcaa gctgcccgtgccctggcccaccctcgtgaccaccctgacctacggcgtgcagtgcttcagccgctaccccgaccacatgaa gcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaact acaagacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaagga ggacggcaacatcctggggcacaagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcagaag aacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccgaccactaccagcaga acacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccgccctgagcaaagaccc caacgagaagcgcgatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtaca ag
mTurquoise2	atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccaca agttcagcgtgtccggcgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccaccggcaa gctgcccgtgccctggcccaccctcgtgaccaccctgtcctggggcgtgcagtgcttcgcccgctaccccgaccacatgaa gcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaact acaagacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaagga ggacggcaacatcctggggcacaagctggagtacaactactttagcgacaacgtctatatcaccgccgacaagcagaag aacggcatcaaggccaacttcaagatccgccacaacatcgaggacggcggcgtgcagctcgccgaccactaccagcaga acacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccaagctgagcaaagaccc caacgagaagcgcgatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtaca ag
YFP	atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccaca agttcagcgtgtccggcgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccaccggcaa gctgcccgtgccctggcccaccctcgtgaccaccttcggctacggcctgcagtgcttcgcccgctaccccgaccacatgaa gcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaact acaagacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgacttcaagga ggacggcaacatcctggggcacaagctggagtacaactacaacagccacaacgtctatatcatggccgacaagcagaag aacggcatcaaggtgaacttcaagatccgccacaacatcgaggacggcagcgtgcagctcgccgaccactaccagcaga acacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagctaccagtccaagctgagcaaagaccc caacgagaagcgcgatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtaca ag
XFP	atggtgagcaagggcgaggagctgttcaccggggtggtgcccatcctggtcgagctggacggcgacgtaaacggccaca agttcagcgtgtccggcgagggcgagggcgatgccacctacggcaagctgaccctgaagttcatctgcaccaccggcaa gctgcccgtgccctggcccaccctcgtgaccaccttcggctacggcctgcagtgcttcgcccgctaccccgaccacatgaa gcagcacgacttcttcaagtccgccatgcccgaaggctacgtccaggagcgcaccatcttcttcaaggacgacggcaact acaagacccgcgccgaggtgaagttcgagggcgacaccctggtgaaccgcatcgagctgaagggcatcgatttcaagga ggacggcaacatcctggggcacaagctggagtacaactactttagcgacaacgtctatatcaccgccgacaagcagaag aacggcatcaaggccaacttcaagatccgccacaacatcgaggacggcggcgtgcagctcgccgaccactaccagcaga acacccccatcggcgacggccccgtgctgctgcccgacaaccactacctgagcacccagtccaagctgagcaaagaccc caacgagaagcgcgatcacatggtcctgctggagttcgtgaccgccgccgggatcactctcggcatggacgagctgtaca ag

Appendix 4: Primer list.

Primer ID		Sequences
#1	Reverse **ClaI insertion** in YFP/mTurquoise2	GCTGAAGGGCATCGATTTCAAGGAGGACGG
	Forward **ClaI insertion** in YFP/mTurquoise2	CCGTCCTCCTTGAAATCGATGCCCTTCAGC
#2 Intron 3	Forward **intron cloning with ClaI site**	CTCTATCGATGTAAGAACCGTGTGGATG
	Reverse **intron cloning with ClaI site**	CTCTATCGATCTACAACAAATCACAAGAGA
#3 Intron 9	Forward **intron cloning with ClaI site**	CACCTCTATCGATGCAAGTTATTAGCAAGGT
	Reverse **intron cloning with ClaI site**	CTCTATCGATCTGCAAATCACAGAAAGTG
#4 Intron 17	Forward **intron cloning with ClaI site**	CACCTCTATCGATGTAACGGCCAGTTTTTCAGC
	Reverse **intron cloning with ClaI site**	CTCTATCGATCTAACCAGAAGAACAAGCAACA
#5 Intron 20	Forward **intron cloning with ClaI site**	CACCTCTATCGATGTAAGCGGTTTATTTTTGTGAG
	Reverse **intron cloning with ClaI site**	CTCTATCGAT CTTTAGAGAGAAGGATAATAAAAAT
#6 Intron 3	Reverse **ClaI deletion** in hybrid mini-gene	CTCTTGTGATTTGTTGTAGTTCAAGGAGGACGGCAAC
	Forward **ClaI deletion** in hybrid mini-gene	GTTGCCGTCCTCCTTGAACTACAACAAATCACAAGAG
#7 Intron 9	Reverse **ClaI deletion** in hybrid mini-gene	CACTTTCTGTGATTTGCAGTTCAAGGAGGACGGCAACATCC
	Forward **ClaI deletion** in hybrid mini-gene	GGATGTTGCCGTCCTCCTTGAACTGCAAATCACAGAAAGTG
#8 Intron 17	Reverse **ClaI deletion** in hybrid mini-gene	TGTTGCTTGTTCTTCTGGTTAGTTCAAGGAGGACGGCAAC
	Forward **ClaI deletion** in hybrid mini-gene	GTTGCCGTCCTCCTTGAACTAACCAGAAGAACAAGCAACA
#9 Intron 20	Reverse **ClaI deletion** in hybrid mini-gene	ATTTTTATTATCCTTCTCTCTAAAGTTCAAGGAGGACGGCAAC
	Forward **ClaI deletion** in hybrid mini-gene	GTTGCCGTCCTCCTTGAACTTTAGAGAGAAGGATAATAAAAAT
#10	Forward HTT i20 BD	GTAAGCGGTTTATTTTTGTGAG
	Reverse HTT i20 BD	CTTTAGAGAGAAGGATAATAAAAAT
#11	Forward **spliced product** XFP/mTurquoise2	GACGTAAACGGCCACAAGTT
	Reverse **spliced product** XFP/mTurquoise2	CTTGTACAGCTCGTCCATGC
#12 Intron 3 (5')	Forward **qPCR exon/intron junction**	CATCATCCACACGGTTCTTACATC
	Reverse **qPCR exon/intron junction**	CAAGGACGACGGCAACTACAA
#12bis Intron 3 (3')	Forward **qPCR exon/intron junction**	AAGTAGTTGTACTCCAGCTTGTGC
	Reverse **qPCR exon/intron junction**	TGATGGGATGTGTCTTCCAT
#13 Intron 9 (5')	Forward **qPCR exon/intron junction**	CAGGCCCATAATCAATAGAGTGT
	Reverse **qPCR exon/intron junction**	AAGGACGACGGCAACTACAA
#13bis Intron 9 (3')	Forward **qPCR exon/intron junction**	AAGTAGTTGTACTCCAGCTTGTGC
	Reverse **qPCR exon/intron junction**	CTGCGATGTTAAGTGTTTCCTGT

#14 Intron 17 (5')	Forward **qPCR exon/intron junction**	CTAACCAGAAGAACAAGCAACA
	Reverse **qPCR exon/intron junction**	CAAGGACGACGGCAACTACAA
#14bis Intron 17 (3')	Forward **qPCR exon/intron junction**	GTAACGGCCAGTTTTTCAGC
	Reverse **qPCR exon/intron junction**	AAGTAGTTGTACTCCAGCTTGTGC
#15 Intron 20 (5')	Forward **qPCR exon/intron junction**	CTTTAGAGAGAAGGATAATAAAAAT
	Reverse **qPCR exon/intron junction**	CAAGGACGACGGCAACTACAA
#15bis Intron 20 (3')	Forward **qPCR exon/intron junction**	GTAAGCGGTTTATTTTTGTGAG
	Reverse **qPCR exon/intron junction**	AAGTAGTTGTACTCCAGCTTGTGC

Appendix 5: Precise results of FACS gate

	NT		**Intron 3 - PTM0**		**Intron 3 - PTM1**		**Intron 3 - PTM2**	
	% of total events	% of total cells	% of total events	% of total cells	% of total events	% of total cells	% of total events	% of total cells
TOTAL CELLS FSC/SSC	94.80%	100%	82.90%	100%	85.89%	100%	76.49%	100%
SINGULET	85.87%	90.55%	71.75%	86.55%	70.49%	82.07%	63.96%	83.62%
POSITIVE CELLS	0	0	0.01%	0.03%	0.10%	0.14%	0.06%	0.10%

	NT		**Intron 9 - PTM0**		**Intron 9 - PTM1**		**Intron 9 - PTM2**	
	% of total events	% of total cells	% of total events	% of total cells	% of total events	% of total cells	% of total events	% of total cells
TOTAL CELLS FSC/SSC	94.80%	100%	85.20%	100%	59.69%	100%	51.81%	100%
SINGULET	85.87%	90.55%	66.56%	78.11%	52.36%	87.73%	39.08%	75.43%
POSITIVE CELLS	0	0	0.02%	0.03%	0.19%	0.37%	0.26%	0.67%

	NT		Intron 20 - PTM0		Intron 20 - PTM1	
	% of total events	**% of total cells**	**% of total events**	**% of total cells**	**% of total events**	**% of total cells**
TOTAL CELLS FSC/SSC	**94.80%**	**100%**	**72.02%**	**100%**	**84.00%**	**100%**
SINGULET	**85.87%**	**90.55%**	**64.60%**	**87.28%**	**78.53%**	**93.50%**
POSITIVE CELLS	**0**	**0**	**0.02%**	**0.03%**	**0.18%**	**0.23%**

mRNA variant	mRNA matching the genome	Best predicted protein	Transcription unit pre-mRNA	# of exon	From tissue (no strict specificity is implied)
aAug10	13646 bp	3247 aa	169451 bp	67	brain (29), eye (13), uterus (11), cerebellum (10) and 103 other tissues
bAug10	11506 bp	1147/1105/296 aa	112670 bp	53	brain (3), head neck (1)
cAug10	1122 bp	373 aa	19581 bp	8	embryonic stem cells, embryoid bodies derived from H1, H7 and H9 cells (1), lymph (1)
dAug10	1235 bp	243 aa	6340 bp	6	blood (1), eye (1), retinoblastoma (1)
eAug10	963 bp	189 aa	7612 bp	6	adenocarcinoma cell line (1), colon (1), lymph (1)
fAug10	763 bp	187 aa	6366 bp	3	denis drash (1), endometrium, adenocarcinoma cell line (1), uterus (1)
gAug10	921 bp	147 aa	12282 bp	5	brain (1), hippocampus (1)
hAug10	380 bp	126 aa	2448 bp	2	head neck (3)
iAug10	980 bp	123 aa	1401 bp	2	brain (2), embryonic stem cells (1)
jAug10$_{unspliced}$	446 bp	119 aa	446 bp	1	brain (1)
kAug10	344 bp	114 aa	3631 bp	3	lung tumor (2)
lAug10	429 bp	111 aa	12866 bp	3	pooled germ cell tumors (6), testis (1)
mAug10$_{unspliced}$	1569 bp	102 aa	1569 bp	1	lung normal (15), lung (7), prostate tumor (7), brain (4) and 44 other tissues
nAug10	386 bp	81 aa	756 bp	2	brain (1)
oAug10	177 bp	58 aa	1828 bp	3	placenta normal (1)
pAug10$_{unspliced}$	1733 bp	66 aa	1733 bp	1	brain (2), colon (1), colon tumor, RER+ (1), glioblastoma (1) and 5 other tissues
qAug10$_{unspliced}$	811 bp	72 aa	811 bp	1	
rAug10$_{unspliced}$	756 bp	26 aa	756 bp	1	eye (1), fetal eyes, lens, eye anterior segment,optic nerve, retina, retina foveal and macular, RPE andchoroid (1), germinal center B cell (1)
sAug10	656 bp	57 aa	8560 bp	5	colon normal (1), testis (1)
tAug10$_{unspliced}$	631 bp	37 aa	631 bp	1	cerebellum (1), pineal gland (1)
uAug10$_{unspliced}$	607 bp	18 aa	607 bp	1	liver and spleen (3), brain (2)
vaAug10$_{unspliced}$	226 bp	17 aa	226 bp	1	stomach (1)

This table documents the physical properties of each representative mRNA. In brackets is represented the number of observations.

Appendix 7: ***Details of the IONIS HTTRx clinical trial.*** *This study will test the safety, tolerability, pharmacokinetics and pharmacodynamics of multiple ascending doses of IONIS-HTTRx administered intrathecally to adult patients with early manifest Huntington's Disease.*

Study design

Study Type:	Interventional (Clinical Trial)
Actual Enrollment :	46 participants
Allocation:	Randomized
Intervention model:	Parallel Assignment
Masking:	Double (Participant, Investigator)
Primary purpose:	Treatment
Official title:	A Randomized, Double-blind, Placebo-controlled Study to Evaluate the Safety, Tolerability, Pharmacokinetics and Pharmacodynamics of Multiple Ascending Doses of Intrathecally Administered ISIS 443139 in Patients With Early Manifest Huntington's Disease
Study start date :	August 31, 2015
Estimated primary completion Date:	July 31, 2017
Estimated study completion date:	November 30, 2017
Current primary outcome measures:	Safety and Tolerability: The number of participants with adverse events Time Frame: Participants will be followed for the duration of the study; an expected 29 weeks

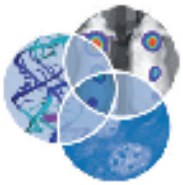
Study arms

Experimental	IONIS HTTRx is administered intrathecally at 4 week intervals over the course of a 13 week treatment period for different dose levels
Placebo comparator	A placebo is administered intrathecally at 4 week intervals over the course of 13 weeks.

Location

Canada, British Columbia	University of British Columbia Vancouver, British Columbia, Canada, V6T 2B5
Germany	eCharite University Berlin Berlin, Germany, 10117 Ruhr-University of Bochum Bochum, Germany, 44791 Ulm University Hospital Ulm, Germany, 89081
United Kingdom	University Hospitals Birmingham Birmingham, United Kingdom, B15 2TH Cambridge University Hospital Cambridge, United Kingdom, CB2 0PY University Hospital of Wales Cardiff, United Kingdom, CF14 4XN University College London London, United Kingdom, WC1N 3BG University of Manchester, St. Mary's Hospital Manchester, United Kingdom, M13 9WL

Advanced Review



mRNA *trans*-splicing in gene therapy for genetic diseases

Adeline Berger,[1] Séverine Maire,[2,3] Marie-Claude Gaillard,[2,3] José-Alain Sahel,[1,4,5,6] Philippe Hantraye[2,3] and Alexis-Pierre Bemelmans[2,3]*

Spliceosome-mediated RNA *trans*-splicing, or SMaRT, is a promising strategy to design innovative gene therapy solutions for currently intractable genetic diseases. SMaRT relies on the correction of mutations at the post-transcriptional level by modifying the mRNA sequence. To achieve this, an exogenous RNA is introduced into the target cell, usually by means of gene transfer, to induce a splice event in *trans* between the exogenous RNA and the target endogenous pre-mRNA. This produces a chimeric mRNA composed partly of exons of the latter, and partly of exons of the former, encoding a sequence free of mutations. The principal challenge of SMaRT technology is to achieve a reaction as complete as possible, i.e., resulting in 100% repairing of the endogenous mRNA target. The proof of concept of SMaRT feasibility has already been established in several models of genetic diseases caused by recessive mutations. In such cases, in fact, the repair of only a portion of the mutant mRNA pool may be sufficient to obtain a significant therapeutic effect. However in the case of dominant mutations, the target cell must be freed from the majority of mutant mRNA copies, requiring a highly efficient *trans*-splicing reaction. This likely explains why only a few examples of SMaRT approaches targeting dominant mutations are reported in the literature. In this review, we explain in details the mechanism of *trans*-splicing, review the different strategies that are under evaluation to lead to efficient *trans*-splicing, and discuss the advantages and limitations of SMaRT.

How to cite this article:
WIREs RNA 2016, 7:487–498. doi: 10.1002/wrna.1347

*Correspondence to: alexis.bemelmans@cea.fr

[1]Centre de recherche Institut de la Vision, Sorbonne Universités, Université Pierre et Marie Curie UM80, Paris, France

[2]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Département des Sciences du Vivant (DSV), Institut d'Imagerie Biomédicale (I2BM), Molecular Imaging Research Center (MIRCen), Fontenay-aux-Roses, France

[3]Centre National de la Recherche Scientifique (CNRS), Université Paris-Sud, Université Paris-Saclay, UMR 9199, Neurodegenerative Diseases Laboratory, Fontenay-aux-Roses, France

[4]Centre Hospitalier National d'Ophtalmologie des Quinze-Vingts, INSERM-DHOS, Paris, France

[5]Fondation Ophtalmologique Adolphe de Rothschild, Paris, France

[6]Institute of Ophthalmology, University College of London, London, UK

Conflict of interest: The authors have declared no conflicts of interest for this article.

INTRODUCTION

Following a fluctuating development during the 1990's and the first decade of the 21st century, gene therapy is now enjoying a renewed interest both in the pharmaceutical industry and academic laboratories engaged in translational research. We are now witnessing the emergence of new techniques in medicine that are a direct consequence of the progress in biotechnology and genetic engineering.[1] A large number of gene therapy modalities have been devised to address specific different pathologies. The techniques range from the simplest, such as gene supplementation, to the more complex, such as genome editing using the CRISPR/Cas9 technology.[2] However, there remain cases where existing tools are either not applicable, or not effective enough, to expect a significant

therapeutic effect. This is true for the example of recessive genetic diseases, in which the size of the cDNA that would correct the phenotype is not compatible with the currently available gene transfer vectors. In this case, one would need a technology allowing to introduce into the vector only the part of the mutant transcript that need to be corrected. We may also evoke dominant genetic diseases where a tight regulation of healthy gene expression is essential, disqualifying the only use of mRNA silencing, and causing a suppression of the expression of the mutant allele or both alleles. It would be interesting to repair the mutant transcripts to make clean copies without altering expression level. In addition, in case of multiple dominant mutations of the same gene, development of a specific therapeutic vector for each of them can be hardly feasible from a logistical perspective. In other scenarios, the appropriate strategy would be to express an exogenous factor, by taking advantage of the regulation of the expression of an endogenous gene. For example, it would be useful to specifically express a toxin in tumor cells, or in the case of neurodegenerative diseases, to express on demand a trophic factor during inflammatory episodes when up-regulation of inflammatory genes occurs in glial cells.

One sees that in all these cases, intervention by genetic engineering to modify a given transcript in the target cell population could be beneficial. This is what proposes the technology of spliceosome-mediated RNA *trans*-splicing by diverting the spliceosome machinery to modify part of the exons of an endogenous transcript, using an artificial RNA brought by gene transfer into the cells to be corrected. In this review, we will return to the theory and the principles of operation of *trans*-splicing, we will present the latest techniques that allow its implementation using selected examples in the literature, and finally, we will discuss its main limitations.

PRINCIPLE, ADVANTAGES, AND APPLICATIONS OF SPLICEOSOME-MEDIATED RNA-*TRANS*-SPLICING

Splicing, a reaction occurring in the nucleus of eukaryotic cells, is catalyzed by the spliceosome, one of the largest ribonucleoproteic complex of the cell, and results in the elimination of intronic sequences from pre-mRNA. Splicing requires several *cis*-acting elements on the pre-mRNA: (1) 5′ donor and 3′ acceptor splice site consensus sequences that constitute the exon–intron boundaries; (2) a branch point, consisting of an adenosine, located in a consensus sequence of the intron, 18–40 nucleotides upstream of the 3′ acceptor splice site, and (3) a polypyrimidine-rich sequence (PPT for polypyrimidine tract), just before the 3′ splice site, often composed of an uracil repetition. Progressively as it is transcribed in the nucleus, the pre-mRNA is handled by the spliceosome machinery and spliced.[3] To summarize the process, presplicing complexes composed of ribonucleoproteins are sequentially assembled on the pre-mRNA to form the active final spliceosome, a complex of about 12 MDa. Then, a sequence of two *trans*-esterifications occurs. The first one, involving the 5′ donor site and the branch point, leads to the formation of a lariat and the release of a 3′-OH free extremity at the end of the upstream exon. The second one involves this 3′-OH free extremity and the 3′ acceptor site, leading to the joining of the two exons and the release of the lariat. This reaction, universally present in eukaryotes, is called splicing, or more precisely *cis*-splicing as it involves a donor site, a branch point, and an acceptor site located on the same RNA molecule (Figure 1(a)).

The spliceosome can also catalyze a *trans*-splicing reaction. Unlike the above-defined *cis*-splicing reaction, *trans*-splicing occurs between two different RNA molecules. The molecular process is exactly the same, except that the final mRNA is composed of the first exon(s) from one of the pre-mRNA and the following(s) exon(s) of the other one, creating a chimeric molecule. Although anecdotic when compared to *cis*-splicing, the occurrence of *trans*-splicing was initially described in the mid 80's using recombinant mRNAs in *in vitro* cell-free systems,[4,5] and it was quickly shown that this reaction was of physiological relevance among different lower eukaryotes, in particular the trypanosome.[6,7] Thereafter, evidence of *trans*-splicing has also been shown in mammals, with the demonstration that carnitine octanoyltransferase mRNAs are subject to *trans*-splicing in rat hepatocytes,[8] as are transcripts from the human estrogen receptor gene.[9] Moreover, recent refinements of RNAseq analysis helped to unambiguously identify physiologic *trans*-splicing events in human embryonic stem cells and demonstrated the role of a *trans*-spliced transcript in the maintenance of pluripotency.[10] It is thus established that *trans*-splicing exists naturally in a wide variety of organisms, but its physiological value remains to be elucidated. One possibility, suggested by Dixon et al.,[11] is that *trans*-splicing is at the origin of the exon-repetition phenomenon, which results in the tandem repeat of exons on a mRNA, in the absence of such repetition at the genomic level. In a 2007 study, these authors identified the existence of complementary

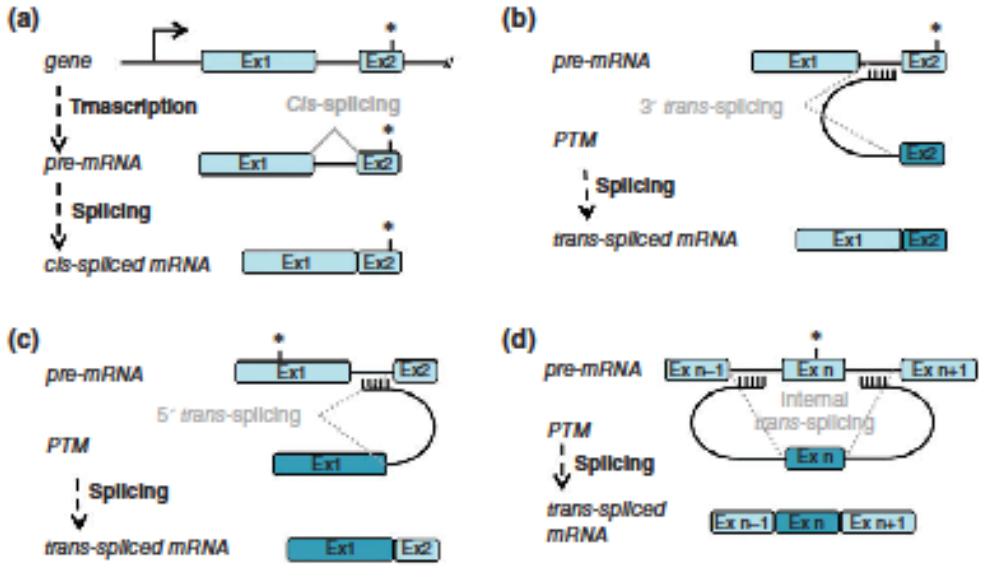


FIGURE 1 | Different modes of action of *trans*-splicing. (a) The classical splicing mechanism, or *cis*-splicing occurs within a single pre-mRNA molecule and leads to the elimination of the intronic sequences. (b, d) On the contrary, *trans*-splicing occurs between two splice sites located on two different pre-mRNAs. In SMaRT technology, an exogenous RNA, called pre-mRNA *trans*-splicing molecule (PTM) is used to replace one or several exons of an endogenous pre-mRNA. Depending of the orientation of the PTM, it is possible to replace 3′-exon(s) (b), 5′ exon(s) (c), or even internal exon(s) (d). Ex n: exon n; *: point mutation.

sequences in introns flanking the exons subjected to repetition, suggesting that exon-repetition occurs after *trans*-splicing between two pre-mRNA of the same gene, whose proximity is favored during transcription by the complementarity of intron sequences. This hypothesis is reinforced by the fact that exon-repetition operates strictly between transcripts encoded by the same allele,[12] and suggests that to promote *trans*-splicing between two pre-mRNA, it may be necessary to promote their closeness with complementary sequences allowing them to hybridize.

The discovery of these physiological mechanisms has prompted the use of *trans*-splicing for bioengineering purposes. Accordingly in 1999, using cell culture experiments, Puttaraju et al. demonstrated the feasibility of diverting *trans*-splicing for inducing the repair of an endogenous mRNA using exon exchange mediated by an artificial RNA capable of inducing *trans*-splicing.[13] Subsequently, they also showed the feasibility of using this strategy *in vivo*, leading to functional restoration of mutant cystic fibrosis trans-membrane conductance regulator (CFTR) in a human bronchial xenograft model system.[14] These studies paved the way for the use of spliceosome-mediated RNA *trans*-splicing, or SMaRT, as a gene therapy strategy. In this case, an artificial RNA, called pre-mRNA *trans*-splicing molecule (PTM), is engineered to specifically target an endogenous pre-mRNA expressed in the target cell (Figure 1(b)–(d)). To be successful, the PTM must bind the target pre-mRNA and induce the *trans*-splicing reaction more efficiently than the *cis*-splicing one. For this purpose, a typical PTM is composed of: (1) a binding domain able to recognize the target intron on the endogenous pre-mRNA by base pairing, (2) an artificial intron to catalyze the splicing reaction, and (3) the cDNA containing the coding sequence of substitution (Figure 2). The precise order of these three elements in the PTM will depend on the location of the exon(s) to be replaced in the target mRNA, either in 5′ position or 3′ position (Figure 1 (b) and (c)). Replacement of an internal exon is also theoritically possible, although it relies on the occurrence of two *trans*-splicing events within the same pre-mRNA (Figure 1(d)). From its mode of action, the SMaRT technology presents several valuable benefits when considering gene therapy application:

- When targeting the first intron of the pre-mRNA and hence replacing all the subsequent sequence (depending of course of the cDNA size), only one PTM is necessary to repair numerous mutations. For allele-specific knockdown or silencing of dominant mutations, this unique PTM has a large benefit over CRISPR technology, ASO or shRNA molecules, which can target only one specific mutation at a time.

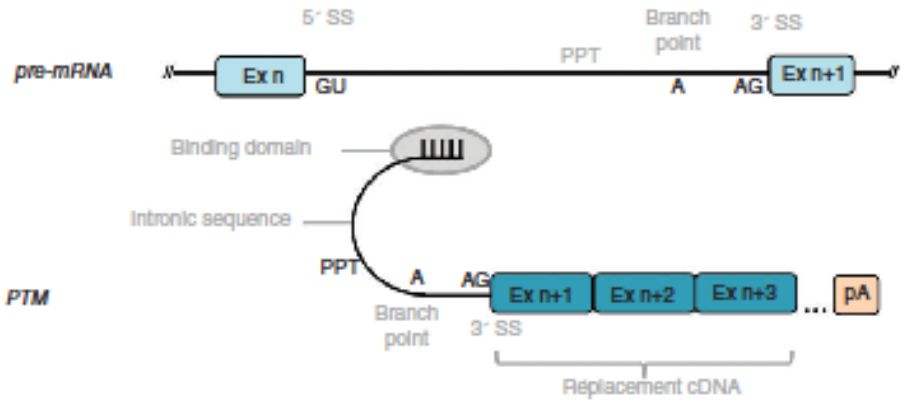


FIGURE 2 | Design of a PTM for 3′ *trans*-splicing. To achieve a 3′ replacement by *trans*-splicing, the PTM must be composed of, from 5′ to 3′: a binding domain, an artificial intron containing a polypyrimidine tract (PPT), a branch point, a 3′ acceptor site (3′ SS), and the replacement cDNA which contains the coding sequence of the exon(s) to be corrected. The necessary 5′ donor site (5′ SS) for the splice reaction is given by the endogenous pre-mRNA.

- The preservation in time, space and quantity of the endogenous regulation of the target mRNA expression. As the PTM is in theory inert except for the *trans*-splicing reaction, the regulation of expression of the chimeric RNA is thus totally dependent on the natural regulation of the targeted pre-mRNA. This information is crucial in gene therapy, where the expression of a gene has to be constrained to a strict spatiotemporal pattern, and kept at a physiologic level in the case of dominant diseases.
- The dual capacity of the PTM to reduce mutated protein synthesis and to simultaneously promote the normal one in a single reaction. In contrast to most of other approaches that require the delivery of two molecules, *trans*-splicing only requires the delivery of the engineered PTM, the two other components of the reaction (targeted pre-mRNA and spliceosome) being naturally present in the target cell.
- As the PTM includes only a part of the cDNA sequence needed for repair, SMaRT technology relies on a small-sized restorative molecule while other gene therapy systems require full-length cDNA. This facilitates the choice of a viral vector for PTM delivery in targeted cells. The size of transgenic cassette is indeed often a major limitation in the choice of a viral vector leading for example to the inability of using AAV vectors. SMaRT technology could therefore enlarge viral vector choice for diseases involving genes that are usually considered too large to be included in an AAV vector.

After an initial study leading to the proof of concept that SMaRT can efficiently promote a functional protein *in vitro*,[15] *trans*-splicing has been developed using different strategies, which can be roughly classified in two groups of applications, the first one being the repair of mutations at the mRNA level, and the second one, the expression of a therapeutic factor under the regulation of an endogenous gene. Although the latter case has not been extensively explored, we should mention here the study by Wang et al. that demonstrates the possibility to use partial *trans*-splicing of an abundant mRNA to express a therapeutic factor in the target cell.[16] However, the most common use of SMaRT concerns the RNA repair strategy for genetic diseases and is summarized in Table 1. In the case of a recessive disease, it is likely that correcting only part of the mRNA will result in therapeutic benefit, because one ends up in the scenario of a heterozygous individual who presents no phenotype. SMaRT technology has thus been evaluated for a significant number of recessive pathologies, such as cystic fibrosis,[14,19,30] hemophilia A,[26] dysferlinopathies and titinopathies,[27] X-linked immunodeficiency with hyper-IgM,[25] spinal muscular atrophy,[17,18] severe combined immune deficiency,[21] Duchenne muscular dystrophy,[22,31] and epidermolysis bullosa.[20,24,32] All these studies have shown that it is possible to partially correct the cellular pool of mutated mRNA. However, when compared with each other, they also show that there is no consensus on how to orientate PTM design to reach a significant *trans*-splicing efficiency, nor to lead to a therapeutic effect, or even the way to test the PTM efficiency. Still, a characteristic that does emerge is

TABLE 1 | Summary of the Main Publications Reporting the Implementation of a *Trans*-Splicing Strategy for Gene Therapy of Genetic Diseases

Disease	Target Gene	Mutation	Authors	Models Vitro/Vivo	Efficiency
Cystic fibrosis	CFTR	Recessive	Mansfield et al.[15], Liu et al.[11,14]	Vitro	12%[1,5,7] 16%[1,1,2] 14.2%[1,4,2]
Duchenne muscular dystrophy	MDX	Recessive	Lorain et al.[17,18]	Vitro + vivo	29.6%[1,8,7]
Dysferlinopathies/Titinopathies	DYSF/TTN	Dominant/ recessive	Monjaret et al.[19]	Vitro + vivo	NA
Dystrophia myotonica type 1	DMPK	Dominant	Chen et al.[20]	Vitro	7.4%[1]
Epidermolysis bullosa simplex	COL7A1, K14	Recessive	Wally et al.[21], Murauer et al.[22],	Vitro	15.5%[2,1,1]
Frontotemporal dementia with parkinsonism linked to chromosome 17	MAPT	Recessive	Rodriguez-Martin et al.[23]	Vitro	34%[1]
Hemophilia A	FVIII	Recessive	Chao et al.[16]	Vitro	13%[2] 6.3%[1]
Huntington's disease	HTT	Dominant	Rindt et al.[24]	Vitro	5.4%[1]
Retinitis pigmentosa	RHO	Dominant	Berger et al.[24]	Vitro + vivo	40%
Severe combined immune deficiency	DNA-PKcs	Recessive	Zayed et al.[25]	Vitro	15%[2]
Spinal muscular atrophy	SMN2	Recessive	Coady et al.[26–28], Shababi et al.[29]	Vitro + vivo	NA
X-linked immunodeficiency with hyper-IgM	CD40L	Recessive	Tahara et al.[25]	Vivo	10.4%[1]

NA, data not available.
[1] *trans*-Splicing efficiency determined by quantification of corrected mRNA or protein and expressed as percentage of normal level (direct quantification approach).
[2] *trans*-Splicing efficiency determined by quantification of the activity of the corrected protein and expressed as percentage of normal level (indirect quantification approach).

the partial nature of *trans*-splicing. In the best case, the reported efficacy is approximately 20–40%, meaning that only 20–40% of the mutant mRNA pool of the target cell is converted to wild-type mRNA. This is probably why there are far fewer studies reporting application of *trans*-splicing for dominant mutations. In this context, it is necessary to reach a significant percentage of efficiency to expect a therapeutic effect. To our knowledge, there are only four studies that reported the implementation of *trans*-splicing in the case of a dominant mutation (Table 1). In 2009, Chen et al. reported the correction of *DMPK* pre-mRNA, the gene responsible for *dystrophia myotonica* type 1, with an efficiency ranging from approximately 1.5 to 8%.[33] In 2012, Rindt et al. demonstrated in cellular models of Huntington's disease that the mutated exon can be eliminated by 5′ replacement from the Huntingtin encoding mRNA.[23] Recently, Monjaret et al. described *trans*-splicing of the Titin pre-mRNA, whose mutations lead to muscular dystrophies.[27] In this study, the *trans*-spliced product was not accurately quantified and appeared clearly minor. Thus, the *trans*-splicing rate achieved in these studies, by the order of only a few percent, will likely not lead to a therapeutic threshold in patients affected by these dominant conditions to be reached. Our team has recently reported a rate of *trans*-splicing of about 40% in a cell-line expressing a dominant mutation of rhodopsin, the gene responsible for retinitis pigmentosa.[34] Although encouraging, this rate is probably still insufficient for the point mutations that cause the most aggressive cases of retinitis pigmentosa. This highlights the fact that, to achieve successful SMaRT applications for dominant mutations, it is necessary to implement rational means for optimizing design and characterization of an effective PTM.

DESIGN OF HIGHLY EFFICIENT PTM

The first step in designing powerful PTMs is the implementation of methods for accurate and reproducible quantification of *trans*-splicing efficiency. These methods can be divided into two main categories, i.e., direct and indirect approaches. The former consist of direct quantification of *cis* and/or *trans*-splicing products detected either at the mRNA or the protein level. Spliced mRNAs are usually quantified using end-point or quantitative RT-PCR

with specific primers and probes to discriminate *cis*- and *trans*-spliced products. This kind of method estimates the relative amount of repaired mRNAs establishing proof that *trans*-splicing occurs while assessing PTM activity.[14,24,26,30,32,35–39] An alternative approach of quantification at the mRNA level consists in the addition of a silent mutation in the replacement cDNA of the PTM in order to produce different restriction profiles between *cis* and *trans*-spliced mRNAs. Thus, amplification of *cis*- and *trans*-spliced products is achieved in a single RT-PCR reaction using the same primer set, and precise quantification of PTM activity is accomplished by restriction analysis of the amplicon.[23,34] The essential advantage of this approach is to avoid the bias of comparing two different PCR amplifications, one for the *cis*-spliced product, and one for the *trans*-spliced, reflecting respective level of *cis*- and *trans*-spliced events. For *trans*-splicing assessment at the protein level, western-blot analysis is the most frequently used technique, and quantification is performed by measuring level of normal protein expression alone or compared with mutated proteins with size discrimination parameters,[24,30] or using specific antibodies.[14,37] Another direct way to quantify *trans*-splicing is to use an artificial reporter system. In this case, the target intron is introduced into the open reading frame of a reporter gene, which can encode a functional reporter protein only after *trans*-splicing. The first systems described were based on the use of the LacZ gene and measuring the ß-galactosidase activity generated by *trans*-splicing.[28,35] More recently, the team of J. Bauer described a reporter system based on fluorescent proteins, which allows the efficient pre-screening of the PTM target on the intron to be *trans*-spliced.[29] *Trans*-splicing efficiency may also be measured indirectly. For instance, as *trans*-splicing gene therapy aims to provide phenotype improvements, PTM activity may be assessed by measuring therapeutic indexes in *in vitro* or *in vivo* models. Several methods have been used for that purpose, such as immunofluorescence assay enabling for example the assessment of normal protein localization[34] or expression[24], cell death measurement in the case of anti-tumor *trans*-splicing,[39,40] and functional assessment of ion channel by electrophysiology for CFTR repair.[14,19]

Effective *trans*-splicing requires to favor *trans*-splicing over *cis*-splicing. That can be achieved by a fine-tuning of the PTM design. In their initial study, Puttaraju et al. demonstrated that *trans*-splicing requires the presence of several motifs in the PTM molecule, such as a functional branch point, a PPT and the AG consensus motif as splice acceptor site at the 3′ end for a 3′ replacement strategy (Figure 2). Other factors are potentially important for triggering an efficient *trans*-splicing reaction, their influence on the *trans*-splicing rate was therefore investigated by different teams. Firstly, the localization of the PTM binding domain on the target intron plays a significant role in this challenge and could be determinant when targeting sequences involved in splicing. For instance, Murauer et al. tried to block *cis*-splicing by designing a binding domain complementary to the acceptor site located just downstream the target intron.[29] They described that blocking the endogenous 3′ splice site led to an increase of *trans*-splicing efficiency, although this is not a general rule.[34] Secondly, the choice of the target intron is likely to modulate the rate of *trans*-splicing event and in particular the potency of 5′ and 3′ splice site. Exon and intron definition models, which depict the recognition, binding and stability pattern of spliceosome first step assembly, might be of importance.[41,42] It is known that eukaryotic genes are characterized by short exons and long introns, preventing recognition across introns in favor of the exons. Accordingly, Philippi et al. tested whether the strength of 3′ splice sites is a parameter to take into account to favor the *trans*-splicing reaction.[38] This strength can be estimated by calculation of the MaxEnt (maximum entropy) score, which assigns a score to the splicing sites based on their homology to the canonical sequence.[43] 5′ and 3′ splice sites have consensus motifs, which in theory modulate splicing efficiencies based on their capacity to bind splicing factors. Thus, the splice site strength is highly relevant in the choice of the target intron, as it represents the splicing dynamic in terms of splicing factor interaction. Strong splice sites associated with high MaxEnt scores should promote splicing more efficiently than weak splice sites associated with low MaxEnt scores. Applied to *trans*-splicing, this assumption underlies that weak 3′ splice site in the target intron tends to enhance *trans*-splicing events since this splice site is in competition with a stronger PTM splice site.[38] Third, Coady et al.[44] and Shababi et al.[45] have demonstrated that the combination of PTM with antisense oligonucleotides (ASO) capable of blocking the splice-site targeted by the PTM increases the *trans*-splicing rate, most likely by inhibition of *cis*-splicing by the ASO. Coady et al. developed a single vector for ASO and PTM delivery and showed its efficiency *in vivo* in a mouse model of spinal muscular atrophy.[17] Similarly, Wang et al. suggested that splicing efficiency is correlated to the time of intron synthesis.[46] This mechanism is certainly due to the competition between *cis*-splicing, which requires the synthesis of both endogenous

splice sites, and *trans*-splicing which only needs the endogenous 5′ splice site. Fourthly, another way to increase *trans*-splicing efficiency is to insert an intron in the replacement cDNA of the PTM.[22,28,34] In this case, the addition of intronic splicing enhancer (ISE) in the PTM sequence may promote *trans*-splicing events. It is currently difficult to predict the effect of an ISE according to its sequence, its location related to which exon to include, and its potential to recruit splicing enhancer factors. However, new methods are currently being developed to better locate ISEs as well as splicing repressor sequences and should be useful to increase PTM efficiency.[47,48] Splicing is highly dependent on RNA/protein interactions which promote or prevent pre-mRNA availability to the spliceosome.[49] This is true also for PTMs which have to reach the target intron despite complex pre-mRNA conformations resulting from RNA/protein interactions. These interactions are a clue for PTM optimization and better knowledge of this aspect is needed to understand its real impact on *trans*-splicing. Finally, to maximize *trans*-splicing efficiency, the delivery mode of the PTM has to be taken into account. We have shown that this efficiency is directly correlated to the amount of PTM molecules available in the target cell.[34] For *in vivo* applications, it will thus be important to use the most efficient gene transfer vector to transduce target cells.

To summarize, several aspects of the PTM's design, target, environment and delivery can be modulated and are constantly improved leading to better understanding. However, an obvious clue in PTM design is that a great part of the success remains in the binding domain. Most of the groups working on *trans*-splicing have tried to modify the size and the location of this sequence with direct consequences on *trans*-splicing efficiency and specificity.[15,20,28,34,50] The presence of *trans*-splicing hotspots in the target intron has been demonstrated. In these hotspots, a shift on the target intron of about 30 base pairs (bp) applied to a 150-bp long binding domain can lead to significant variability in PTM efficiency.[34] Nevertheless, predicting an effective binding domain is hard to achieve. Even if bioinformatics analyses based on secondary structures of PTMs and their targets are used to perform a splice site availability prediction, the process still needs to be improved, as some aspects of splicing sequence remain unclear.

Therefore, screening strategies are needed to select effective binding domains. To date, very few binding domains have been tested for each PTM reported in the literature, because the techniques of PTM synthesis and efficiency assessment are cumbersome, thus limiting the discovery of highly efficient PTMs. For this reason, a larger PTM screening and efficient reporter system suitable for automated analysis needs to be implemented. The team of J. Bauer has described a PTM screening system in which *trans*-splicing of an artificial minigene generates a fluorescent protein, thus enabling an easy and fast detection of *trans*-splicing events.[29,51] To achieve this, the target introns are introduced into a minigene, suitable for testing every kind of *trans*-splicing orientation. This system consists of two elements. The first is a reporter minigene including the target intron flanked by two exons, one of which is the partial ORF of a fluorescent protein, placed upstream or downstream of the intron depending on whether one wishes to promote 5′ or 3′ *trans*-splicing. The second one is a library of PTMs that differ only in the binding domain to the target intron and whose cDNA can reconstruct the complete ORF of the fluorescent protein, allowing the detection of *trans*-splicing. For generating a high diversity of binding domain, it is possible to proceed by fragmentation of the target intron using one or more high-frequency cutting restriction enzymes. In this case, however, the fragments are not generated in a totally random manner, and parts of the intron, such as GC-rich or repetitions containing regions can escape fragmentation. Alternatively, the use of physical methods may allow higher fragment diversity. Thanks to the rise of next generation sequencing, sonication-based DNA fragmentation methods that can generate reproducibly fragments of specific lengths are now available. For generating the library of PTMs once the target intron has been fragmented, it is cloned into a vector containing the unvarying part of the PTM. This library can then be screened by cotransfection with the minigene that generates a fluorescent protein following *trans*-splicing. Although attractive, this screening strategy requires changing the genomic environment: the target intron is no longer surrounded by endogenous exons but by artificial ones and this change may affect the PTM activity. Thus, in a recent experiment we aimed to assess the reliability of this fluorescent reporter system to isolate efficient binding domains regardless of the genomic context of the target intron. We constructed a mini-gene encoding fluorescent proteins and containing the first intron of the human rhodopsin gene. Using this system, we showed that binding domains previously characterized on the full-length human rhodopsin gene induce the same rate of *trans*-splicing on the mini-gene (Figure 3), demonstrating that the genomic environment of the target intron has no major influence on the binding domain efficiency.

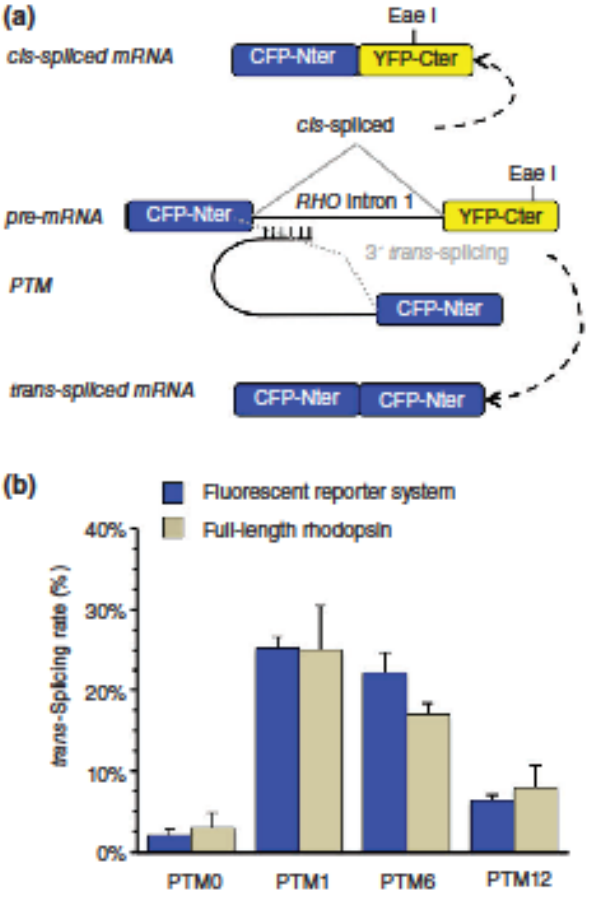


FIGURE 3 | Assessment of a fluorescent reporter system for PTM screening. (a) A fluorescent system was adapted to intron 1 of the rhodopsin gene (RHO), and the EaeI restriction site allowed quantifying *trans*-splicing using the same methodology described in Berger et al.[34] To that aim, RHO intron 1 was subcloned between sequences coding respectively the N-terminal part of CFP (CFP-Nter) and the C-terminal part of YFP (YFP-Cter), while the PTM contained the C-terminal part of CFP (CFP-Cter). With this system, *cis*-splicing leads to an mRNA encoding a chimeric protein fluorescing minimally in green, and *trans*-splicing leads to full-length CFP, characterized by a bright blue fluorescence. (b) Three binding domains, previously characterized in the context of the full rhodopsin pre-mRNA (for details, see Berger et al.[34]), were tested with the fluorescent reporter system using the end-point PCR amplification followed by enzymatic restriction method, and led to the same *trans*-splicing rate, thus showing that *trans*-splicing efficiency is independent of the genomic context of the binding domain. PTM0: negative control without binding domain. Location of binding domains of PTM1, 6, and 12 on RHO intron 1 have been described previously.[32] Bars represent mean ± SD; $n = 6$ for fluorescent reporter system; for full-length rhodopsin, $n = 9$ for PTM0 and 1, and $n = 3$ for PTM6 and 12.

DRAWBACKS AND LIMITATIONS OF THE SMART TECHNOLOGY

One of the main limitations in the development of SMaRT technology is linked to the lack of information regarding the mechanism involved in *trans*-splicing. While the *cis*-splicing mechanism is now largely understood and documented in the literature, *trans*-splicing as a natural event in human has raised only a few publications and most of the iceberg remains subsurface.[9,10,52,53] This lack of data is undoubtedly due to the difficulty to isolate *trans*-splicing products compared to those of *cis*-splicing, chromosomal rearrangements or technical artifacts. Because the *trans*-splicing reaction mechanism still presents some black holes, it is currently difficult or impossible to predict the efficiency of a given PTM. Notably the importance of the 3D-structure of both target and PTM RNA are to date insufficiently taken into consideration. Despite several tools developed to predict the secondary structures of pre-mRNA (ViennaRNA Package[a]; RNAfold[b]; mfold[c]...), it remains quite difficult to estimate the final shape of RNA and thus to design binding sequences based on this information. Interestingly, a recent report suggests that it is also important to consider the sequence of the target intron, and particularly the homology of the splice sites to the consensus sequences.[38] Based on this characteristic, the authors showed that it is possible to rank the different introns of a target pre-mRNA to determine which are the most likely to lead to an efficient *trans*-splicing reaction. To progress on this purpose, we can count on progress of bioinformatic tools in terms of 3D structure prediction, and, as discussed above, PTM screening using randomly-generated libraries. In addition to this specific limitation, like all molecular mechanisms used to modify gene expression, SMaRT technology still presents some demonstrated or theoretical drawbacks, which are covered below.

The main drawback of SMaRT technology is common to most nucleic acids-targeting engineering tools: the specificity of the molecule. One cannot imagine the PTM, generated to replace a mutated sequence, as a source of production of aberrant proteins due to off-target *trans*-splicing with random pre-mRNAs. Although these potential hybrid mRNAs should be processed by nonsense mediated decay or nonstop decay,[54] it is essential to validate the specificity of the PTM for the target sequence and limit nonspecific events. Several approaches should be considered for this purpose. (1) One must

determine the right size of the binding site that might limit off-target binding. Puttaraju et al. showed that increasing the length of the binding domain sequence up to 153 bases significantly enhanced the efficiency of the PTM.[15] A binding sequence of approximately 150 bases dramatically decreases the probability of finding the entire and exact corresponding sequence in a human genome, thus ensuring in theory a highly specific homology of the designed sequence. However, long binding sequences may only partially bind to nonspecific targets, decreasing the specificity of the molecule. The design of the PTM has to take in consideration these parameters and find a balance to obtain a good *trans*-splicing efficiency with a high specificity. (2) Take into account the importance of the 3′ splice site, which can induce more or less off-target *trans*-splicing depending on its potency.[50] (3) Increase the specificity of the PTM delivery to the targeted cells. By using an AAV vector for example, the adjustment in the capsid choice is crucial to introduce the PTM in the right cells, thus limiting potential nonspecific *trans*-splicing in cells that do not express the target pre-mRNA. (4) Similarly, take into consideration the expression level of the targeted mRNA, and adapt the PTM expression appropriately. We have shown that this efficiency is directly correlated to the amount of PTM molecules available in the target cell,[34] but an overproduction of PTM would be useless, and could potentially increase the off-target effect, or even be potentially toxic for the cell as can be overexpression of exogenous small RNA.[55,56]

The issue of binding domain specificity also raises the following question: what happens if the PTM does not bind to the specific sequence? Unfortunately, the production of truncated proteins due to the translation of PTM designed for 3′-replacement has already been observed in the absence of *trans*-splicing events.[15,27] Although 3′-replacement PTMs do not possess the first exon(s) of the endogenous mRNA, they are transcribed in the cell by a RNA polymerase II, and thus will be naturally capped and stabilized.[57] They are therefore likely to possess all the necessary elements for promoting translation, i.e., a 5′-cap, an alternative AUG start codons and a poly-adenylation sequence. Moreover, while the great majority of transcripts are translated from the first AUG of the KOZAK sequence, a secondary AUG can also be used to initiate translation. Although to our knowledge, there are no data available in the literature, this problem may be avoided in 5′-replacement PTMs for which the poly-adenylation signal is missing. In this case, this signal is given by the target pre-mRNA and is therefore excluded from the PTM design.

The last main drawback of SMaRT technology is inevitably linked to the efficiency of the reaction. In the context of an aggressive dominant disease, the level of mutated protein synthetized is determinant for the evolution of the disease. Even a low level of expression of a protein which leads to a dominant-negative effect will induce the disease. However, to date, nobody has been able to reach a 100% *trans*-splicing efficiency. Besides playing with the design of the PTM, the reaction efficiency could be ameliorated by increasing the likelihood of the target pre-mRNA coming into contact with the PTM, promoting more *trans*- than *cis*-splicing events. One major obstacle to achieve this goal is the coupling of mRNA transcription and splicing. Although this coupling is not required for the splice event to take place, the recruitment of elements making up the splicing machinery by the carboxy-terminal domain of the RNA polymerase II often makes the two processes simultaneous. Because it appears difficult to play on the transcription mechanism, the only way to intervene before *cis*-splicing is by accelerating the speed of action of the PTM, playing on its availability and its proximity with the target pre-mRNA. This availability will be defined by the quantity of PTM. As previously discussed, it is necessary to find a balance between providing enough PTM molecules, and avoiding a potential toxicity. Regarding the proximity of the two molecules, the reaction involves a closed localization of the PTM to the transcription site of the target, to bind and *trans*-splice the pre-mRNA before *cis*-splicing occurs. For now the PTM is simply delivered by conventional gene transfer means, either transfection for *in vitro* studies, or viral-mediated gene transfer *in vivo*, and lost in the huge pool of nucleic acids of the nucleus of the target cell. Increasing the likelihood of PTM and target pre-mRNA meeting could thus dramatically change the reaction efficiency. To that end, we could imagine the future development of a viral vector that allows targeted integration in the chromatin of the target cell (recently reviewed by Ott de Bruin et al.[58]). This development would be useful to introduce the PTM sequence near its target gene, or even in an intron of the target gene, ensuring proximity between the two RNA during and/or after transcription.

CONCLUSION

In the vast arena of human genetic diseases, advances in screening and sequencing technologies has allowed the identification of numerous genes which, when mutated, induce a specific condition. In both

recessive and dominant diseases, gene therapy is one of the most promising treatments for now, by cDNA supplementation in the first case and selective inhibition of the mutated form in the second one. In both cases, *trans*-splicing can bring a valuable alternative when technological barriers prohibit the use of more conventional approaches. In the context of cDNA supplementation for recessive conditions, the main limitation is the size of the expression cassette that needs to be introduced in the appropriate gene transfer vector. *Trans*-splicing overcomes this limit by allowing only a portion of the cDNA to be included in the vector: the PTM can target any of the intronic sequences of a given gene, thereby permitting the adaptation of the cDNA size to the capacity of the vector. In case of a very long pre-mRNA with a mutation in the middle, the internal *trans*-splicing approach can even be considered. Finally, this advantage facilitates the choice of the gene transfer vector, which can thus be dictated only by its tropism for the target cell. In the case of a dominant disease, despite the impressive progress made in the development of molecular tools in this field, including RNAi, nucleases and CRISPR systems, no technology is currently able to provide a universal solution, with broad application, high efficiency and no side-effect. The SMaRT technology could bring the answer to many if not all of these obstacles. Multiple advantages of SMaRT technology, that are unique to it and make it a major competitor for the other molecular tools, include: the requirement of only one exogenous RNA (the other components, i.e., the spliceosome, being provided by the cell), the possibility to use the same PTM for multiple mutations in the same gene, the reduction/inhibition of the mutated protein synthesis, the promotion of the normal protein synthesis at the same time, and finally the preservation of the endogenous gene expression regulation. Nevertheless, *trans*-splicing has not been widely developed so far, and still needs to be explored and better understood, allowing the development of more powerful PTMs. One of the main lacks concerns the prediction of the 3D structures of both target pre-mRNA and PTM, which are most likely determinant for *trans*-splicing efficiency. We and others have already shown that any change in the sequence of the PTM, not only in the binding domain but also in the cDNA replacement sequence, can dramatically modify this reaction efficiency. Despite this common observation, there are no clear and strict rules to define how to design the perfect PTM. All recommendations are based on each personal experience, and the best way to select an efficient PTM is for now to screen as many molecules as possible. In addition to complete all unexplored and unknown *trans*-splicing natural events that occur in all species including humans, a detailed and profound study on *trans*-splicing mechanism should enlighten us on the design of the PTM. This idea opens a door regarding the potential of *trans*-splicing for gene therapy. The knowledge to define the rules to achieve 100% reaction efficiency, while ensuring the absence of off-target effects, would make of the SMaRT technology an incredibly effective and powerful gene therapy tool.

NOTES

[a] http://www.tbi.univie.ac.at/RNA/

[b] http://rnhjy.hzau.edu.cn/kech/swxxx/jakj/dianzi/Bioinf4/miRNA/miRNA1.htm

[c] http://mfold.rna.albany.edu/?q=mfold/RNA-Folding-Form

ACKNOWLEDGMENTS

Adeline Berger received a PhD fellowship from Association Française contre les Myopathies (AFM) and Séverine Maire is currently supported by a PhD fellowship from DIM-Biothérapies. The authors sincerely thank Dr Susannah Williams and Dylan Murphy for the linguistic revision of the manuscript.

REFERENCES

1. Collins M, Thrasher A. Gene therapy: progress and predictions. *Proc Biol Sci* 2015, 282:20143003.
2. Riordan SM, Heruth DP, Zhang LQ, Ye SQ. Application of CRISPR/Cas9 for biomedical discoveries. *Cell Biosci* 2015, 5:33.
3. Matera AG, Wang Z. A day in the life of the spliceosome. *Nat Rev Mol Cell Biol* 2014, 15:108–121.
4. Konarska MM, Padgett RA, Sharp PA. Trans splicing of mRNA precursors in vitro. *Cell* 1985, 42:165–171.
5. Solnick D. Trans splicing of mRNA precursors. *Cell* 1985, 42:157–164.
6. Murphy WJ, Watkins KP, Agabian N. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing. *Cell* 1986, 47:517–525.

7. Sutton RE, Boothroyd JC. Evidence for trans splicing in trypanosomes. *Cell* 1986, 47:527–535.
8. Caudevilla C, Serra D, Miliar A, Codony C, Asins G, Bach M, Hegardt FG. Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc Natl Acad Sci USA* 1998, 95:12185–12190.
9. Flouriot G, Brand H, Seraphin B, Gannon F. Natural trans-spliced mRNAs are generated from the human estrogen receptor-alpha (hER alpha) gene. *J Biol Chem* 2002, 277:26244–26251.
10. Wu CS, Yu CY, Chuang CY, Hsiao M, Kao CF, Kuo HC, Chuang TJ. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res* 2014, 24:25–36.
11. Dixon RJ, Eperon IC, Samani NJ. Complementary intron sequence motifs associated with human exon repetition: a role for intragenic, inter-transcript interactions in gene expression. *Bioinformatics* 2007, 23:150–155.
12. Rigatti R, Jia JH, Samani NJ, Eperon IC. Exon repetition: a major pathway for processing mRNA of some genes is allele-specific. *Nucleic Acids Res* 2004, 32:441–446.
13. Puttaraju M, Jamison SF, Mansfield SG, Garcia-Blanco MA, Mitchell LG. Spliceosome-mediated RNA trans-splicing as a tool for gene therapy. *Nat Biotechnol* 1999, 17:246–252.
14. Liu X, Jiang Q, Mansfield SG, Puttaraju M, Zhang Y, Zhou W, Cohn JA, Garcia-Blanco MA, Mitchell LG, Engelhardt JF. Partial correction of endogenous DeltaF508 CFTR in human cystic fibrosis airway epithelia by spliceosome-mediated RNA trans-splicing. *Nat Biotechnol* 2002, 20:47–52.
15. Puttaraju M, DiPasquale J, Baker CC, Mitchell LG, Garcia-Blanco MA. Messenger RNA repair and restoration of protein function by spliceosome-mediated RNA trans-splicing. *Mol Ther* 2001, 4:105–114.
16. Wang J, Mansfield SG, Cote CA, Jiang PD, Weng K, Amar MJ, Brewer BH Jr, Remaley AT, McGarrity GJ, Garcia-Blanco MA, et al. Trans-splicing into highly abundant albumin transcripts for production of therapeutic proteins in vivo. *Mol Ther* 2009, 17:343–351.
17. Coady TH, Lorson CL. Trans-splicing-mediated improvement in a severe mouse model of spinal muscular atrophy. *J Neurosci* 2010, 30:126–130.
18. Coady TH, Shababi M, Tullis GE, Lorson CL. Restoration of SMN function: delivery of a trans-splicing RNA re-directs SMN2 pre-mRNA splicing. *Mol Ther* 2007, 15:1471–1478.
19. Liu X, Luo M, Zhang LN, Yan Z, Zak R, Ding W, Mansfield SG, Mitchell LG, Engelhardt JF. Spliceosome-mediated RNA trans-splicing with recombinant adeno-associated virus partially restores cystic fibrosis transmembrane conductance regulator function to polarized human cystic fibrosis airway epithelial cells. *Hum Gene Ther* 2005, 16:1116–1123.
20. Wally V, Brunner M, Lettner T, Wagner M, Koller U, Trost A, Murauer EM, Hainzl S, Hintner H, Bauer JW. K14 mRNA reprogramming for dominant epidermolysis bullosa simplex. *Hum Mol Genet* 2010, 19:4715–4725.
21. Zayed H, Xia L, Yerich A, Yant SR, Kay MA, Puttaraju M, McGarrity GJ, Wiest DL, McIvor RS, Tolar J, et al. Correction of DNA protein kinase deficiency by spliceosome-mediated RNA trans-splicing and sleeping beauty transposon delivery. *Mol Ther* 2007, 15:1273–1279.
22. Lorain S, Peccate C, Le Hir M, Griffith G, Philippi S, Precigout G, Mamchaoui K, Jollet A, Voit T, Garcia L. Dystrophin rescue by trans-splicing: a strategy for DMD genotypes not eligible for exon skipping approaches. *Nucleic Acids Res* 2013, 41:8391–8402.
23. Rindt H, Yen PF, Thebeau CN, Peterson TS, Weisman GA, Lorson CL. Replacement of huntingtin exon 1 by trans-splicing. *Cell Mol Life Sci* 2012, 69:4191–4204.
24. Wally V, Klausegger A, Koller U, Lochmuller H, Krause S, Wiche G, Mitchell LG, Hintner H, Bauer JW. 5' trans-splicing repair of the PLEC1 gene. *J Invest Dermatol* 2008, 128:568–574.
25. Tahara M, Pergolizzi RG, Kobayashi H, Krause A, Luettich K, Lesser ML, Crystal RG. Trans-splicing repair of CD40 ligand deficiency results in naturally regulated correction of a mouse model of hyper-IgM X-linked immunodeficiency. *Nat Med* 2004, 10:835–841.
26. Chao H, Mansfield SG, Bartel RC, Hiriyanna S, Mitchell LG, Garcia-Blanco MA, Walsh CE. Phenotype correction of hemophilia A mice by spliceosome-mediated RNA trans-splicing. *Nat Med* 2003, 9:1015–1019.
27. Monjaret F, Bourg N, Suel L, Roudaut C, Le Roy F, Richard I, Charton K. Cis-splicing and translation of the pre-trans-splicing molecule combine with efficiency in spliceosome-mediated RNA trans-splicing. *Mol Ther* 2014, 22:1176–1187.
28. Mansfield SG, Clark RH, Puttaraju M, Kole J, Cohn JA, Mitchell LG, Garcia-Blanco MA. 5' exon replacement and repair by spliceosome-mediated RNA trans-splicing. *RNA* 2003, 9:1290–1297.
29. Murauer EM, Koller U, Hainzl S, Wally V, Bauer JW. A reporter-based screen to identify potent 3' trans-splicing molecules for endogenous RNA repair. *Hum Gene Ther Methods* 2013, 24:19–27.
30. Mansfield SG, Kole J, Puttaraju M, Yang CC, Garcia-Blanco MA, Cohn JA, Mitchell LG. Repair of CFTR mRNA by spliceosome-mediated RNA trans-splicing. *Gene Ther* 2000, 7:1885–1895.
31. Lorain S, Peccate C, Le Hir M, Garcia L. Exon exchange approach to repair Duchenne dystrophin transcripts. *PLoS One* 2010, 5:e10894.

32. Murauer EM, Gache Y, Gratz IK, Klausegger A, Muss W, Gruber C, Meneguzzi G, Hintner H, Bauer JW. Functional correction of type VII collagen expression in dystrophic epidermolysis bullosa. *J Invest Dermatol* 2011, 131:74–83.
33. Chen HY, Kathirvel P, Yee WC, Lai PS. Correction of dystrophia myotonica type 1 pre-mRNA transcripts by artificial trans-splicing. *Gene Ther* 2009, 16:211–217.
34. Berger A, Lorain S, Josephine C, Desrosiers M, Peccate C, Voit T, Garcia L, Sahel JA, Bemelmans AP. Repair of rhodopsin mRNA by spliceosome-mediated RNA trans-splicing: a new approach for autosomal dominant retinitis pigmentosa. *Mol Ther* 2015, 23:918–930.
35. Dallinger G, Puttaraju M, Mitchell LG, Yancey KB, Yee C, Klausegger A, Hintner H, Bauer JW. Development of spliceosome-mediated RNA trans-splicing (SMaRT) for the correction of inherited skin diseases. *Exp Dermatol* 2003, 12:37–46.
36. Rodriguez-Martin T, Garcia-Blanco MA, Mansfield SG, Grover AC, Hutton M, Yu Q, Zhou J, Anderton BH, Gallo JM. Reprogramming of tau alternative splicing by spliceosome-mediated RNA trans-splicing: implications for tauopathies. *Proc Natl Acad Sci USA* 2005, 102:15659–15664.
37. Avale ME, Rodriguez-Martin T, Gallo JM. Trans-splicing correction of tau isoform imbalance in a mouse model of tau mis-splicing. *Hum Mol Genet* 2013, 22:2603–2611.
38. Philippi S, Lorain S, Beley C, Peccate C, Precigout G, Spuler S, Garcia L. Dysferlin rescue by spliceosome-mediated pre-mRNA trans-splicing targeting introns harbouring weakly defined 3' splice sites. *Hum Mol Genet* 2015, 24:4049–4060.
39. He X, Liu F, Yan J, Zhang Y, Yan J, Shang H, Dou Q, Zhao Q, Song Y. Trans-splicing repair of mutant p53 suppresses the growth of hepatocellular carcinoma cells in vitro and in vivo. *Sci Rep* 2015, 5:8705.
40. He X, Liao J, Liu F, Yan J, Yan J, Shang H, Dou Q, Chang Y, Lin J, Song Y. Functional repair of p53 mutation in colorectal cancer cells using trans-splicing. *Oncotarget* 2015, 6:2034–2045.
41. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip Rev RNA* 2013, 4:49–60.
42. Robberson BL, Cote GJ, Berget SM. Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol Cell Biol* 1990, 10:84–94.
43. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 2004, 11:377–394.
44. Coady TH, Baughan TD, Shababi M, Passini MA, Lorson CL. Development of a single vector system that enhances trans-splicing of SMN2 transcripts. *PLoS One* 2008, 3:e3468.
45. Shababi M, Lorson CL. Optimization of SMN trans-splicing through the analysis of SMN introns. *J Mol Neurosci* 2012, 46:459–469.
46. Takahara T, Tasic B, Maniatis T, Akanuma H, Yanagisawa S. Delay in synthesis of the 3' splice site promotes trans-splicing of the preceding 5' splice site. *Mol Cell* 2005, 18:245–251.
47. Wang Y, Ma M, Xiao X, Wang Z. Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. *Nat Struct Mol Biol* 2012, 19:1044–1052.
48. Wang Y, Xiao X, Zhang J, Choudhury R, Robertson A, Li K, Ma M, Burge CB, Wang Z. A complex network of factors with overlapping affinities represses splicing through intronic elements. *Nat Struct Mol Biol* 2013, 20:36–45.
49. Martinez-Contreras R, Fisette JF, Nasim FU, Madden R, Cordeau M, Chabot B. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biol* 2006, 4:e21.
50. Kikumori T, Cote GJ, Gagel RF. Promiscuity of pre-mRNA spliceosome-mediated trans splicing: a problem for gene therapy? *Hum Gene Ther* 2001, 12:1429–1441.
51. Koller U, Wally V, Bauer JW, Murauer EM. Considerations for a successful RNA trans-splicing repair of genetic disorders. *Mol Ther Nucleic Acids* 2014, 3:e157.
52. Finta C, Zaphiropoulos PG. Intergenic mRNA molecules resulting from trans-splicing. *J Biol Chem* 2002, 277:5882–5890.
53. Romani A, Guerra E, Trerotola M, Alberti S. Detection and analysis of spliced chimeric mRNAs in sequence databanks. *Nucleic Acids Res* 2003, 31:e17.
54. Wagner E, Lykke-Andersen J. mRNA surveillance: the perfect persist. *J Cell Sci* 2002, 115:3033–3038.
55. Grimm D, Streetz KL, Jopling CL, Storm TA, Pandey K, Davis CR, Marion P, Salazar F, Kay MA. Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* 2006, 441:537–541.
56. Boudreau RL, Martins I, Davidson BL. Artificial microRNAs as siRNA shuttles: improved safety as compared to shRNAs in vitro and in vivo. *Mol Ther* 2009, 17:169–175.
57. McCracken S, Fong N, Rosonina E, Yankulov K, Brothers G, Siderovski D, Hessel A, Foster S, Shuman S, Bentley DL. 5'-Capping enzymes are targeted to pre-mRNA by binding to the phosphorylated carboxy-terminal domain of RNA polymerase II. *Genes Dev* 1997, 11:3306–3318.
58. Ott de Bruin LM, Volpi S, Musunuru K. Novel genome-editing tools to model and correct primary immunodeficiencies. *Front Immunol* 2015, 6:250.

1. Bates, G. P. History of genetic disease: the molecular genetics of Huntington disease - a history. *Nat. Rev. Genet.* **6,** 766–773 (2005).
2. Pringsheim, T. *et al.* The incidence and prevalence of Huntington's disease: A systematic review and meta-analysis. *Mov. Disord.* **27,** 1083–1091 (2012).
3. Huntington (maladie de). Available at: https://www.inserm.fr/information-en-sante/dossiers-information/huntington-maladie. (Accessed: 3rd January 2018)
4. Almqvist, E. W., Elterman, D. S., MacLeod, P. M. & Hayden, M. R. High incidence rate and absent family histories in one quarter of patients newly diagnosed with Huntington disease in British Columbia. *Clin. Genet.* **60,** 198–205 (2001).
5. Siesling, S. *et al.* Family history and DNA analysis in patients with suspected Huntington's disease. *J. Neurol. Neurosurg. Psychiatry* **69,** 54–59 (2000).
6. Margolis, R. L., Rudnicki, D. D. & Holmes, S. E. Huntington's disease like-2: review and update. *Acta Neurol. Taiwanica* **14,** 1–8 (2005).
7. Paradisi, I., Hernández, A. & Arias, S. Huntington disease mutation in Venezuela: age of onset, haplotype analyses and geographic aggregation. *J. Hum. Genet.* **53,** 127–135 (2008).
8. Harper, P. S. The epidemiology of Huntington's disease. *Hum. Genet.* **89,** 365–376 (1992).
9. Conneally, P. M. Huntington disease: genetics and epidemiology. *Am. J. Hum. Genet.* **36,** 506–526 (1984).
10. Al-Jader, L. N., Harper, P. S., Krawczak, M. & Palmer, S. R. The frequency of inherited disorders database: prevalence of Huntington disease. *Community Genet.* **4,** 148–157 (2001).
11. Warby, S. C. *et al.* HTT haplotypes contribute to differences in Huntington disease prevalence between Europe and East Asia. *Eur. J. Hum. Genet.* **19,** 561–566 (2011).
12. DeLong, M. R. & Wichmann, T. Circuits and circuit disorders of the basal ganglia. *Arch. Neurol.* **64,** 20–24 (2007).
13. Nieuwenhuys, R., Voogd, J. & Huijzen, C. van. *The Human Central Nervous System: A Synopsis and Atlas*. (Springer Science & Business Media, 2013).

14. Vonsattel, J. P. & DiFiglia, M. Huntington disease. *J. Neuropathol. Exp. Neurol.* **57,** 369–384 (1998).

15. Rosas, H. D. *et al.* Striatal volume loss in HD as measured by MRI and the influence of CAG repeat. *Neurology* **57,** 1025–1028 (2001).

16. Tabrizi, S. J. *et al.* Biological and clinical manifestations of Huntington's disease in the longitudinal TRACK-HD study: cross-sectional analysis of baseline data. *Lancet Neurol.* **8,** 791–801 (2009).

17. MacDonald, M. E. *et al.* A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell* **72,** 971–983 (1993).

18. Clabough, E. B. D. Huntington's Disease: The Past, Present, and Future Search for Disease Modifiers. *Yale J. Biol. Med.* **86,** 217–233 (2013).

19. Szabo, L. *et al.* Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. *Genome Biol.* **16,** 126 (2015).

20. Fagerberg, L. *et al.* Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics MCP* **13,** 397–406 (2014).

21. Perutz, M. F., Johnson, T., Suzuki, M. & Finch, J. T. Glutamine repeats as polar zippers: their possible role in inherited neurodegenerative diseases. *Proc. Natl. Acad. Sci. U. S. A.* **91,** 5355–5358 (1994).

22. Steffan, J. S. *et al.* SUMO Modification of Huntingtin and Huntington's Disease Pathology. *Science* **304,** 100–104 (2004).

23. Andrade, M. A. & Bork, P. HEAT repeats in the Huntington's disease protein. *Nat. Genet.* **11,** 115–116 (1995).

24. Neuwald, A. F. & Hirano, T. HEAT repeats associated with condensins, cohesins, and other complexes involved in chromosome-related functions. *Genome Res.* **10,** 1445–1452 (2000).

25. Goldberg, Y. P. *et al.* Cleavage of huntingtin by apopain, a proapoptotic cysteine protease, is modulated by the polyglutamine tract. *Nat. Genet.* **13,** 442–449 (1996).

26. Kalchman, M. A. *et al.* Huntingtin is ubiquitinated and interacts with a specific ubiquitin-conjugating enzyme. *J. Biol. Chem.* **271,** 19385–19394 (1996).

27. Thompson, L. M. *et al.* IKK phosphorylates Huntingtin and targets it for degradation by the proteasome and lysosome. *J. Cell Biol.* **187,** 1083–1099 (2009).

28. Aiken, C. T. *et al.* Phosphorylation of threonine 3: implications for Huntingtin aggregation and neurotoxicity. *J. Biol. Chem.* **284,** 29427–29436 (2009).

29. Colin, E. *et al.* Huntingtin phosphorylation acts as a molecular switch for anterograde/retrograde transport in neurons. *EMBO J.* **27,** 2124–2134 (2008).

30. Zala, D. *et al.* Phosphorylation of mutant huntingtin at S421 restores anterograde and retrograde transport in neurons. *Hum. Mol. Genet.* **17,** 3837–3846 (2008).

31. Yanai, A. *et al.* Palmitoylation of huntingtin by HIP14 is essential for its trafficking and function. *Nat. Neurosci.* **9,** 824–831 (2006).

32. Jeong, H. *et al.* Acetylation targets mutant huntingtin to autophagosomes for degradation. *Cell* **137,** 60–72 (2009).

33. Glozak, M. A., Sengupta, N., Zhang, X. & Seto, E. Acetylation and deacetylation of non-histone proteins. *Gene* **363,** 15–23 (2005).

34. Zuccato, C., Valenza, M. & Cattaneo, E. Molecular mechanisms and potential therapeutical targets in Huntington's disease. *Physiol. Rev.* **90,** 905–981 (2010).

35. Bates, G., Tabrizi, S. & Jones, L. *Huntington's Disease*. (Oxford University Press, 2014).

36. Andrew, S. E. *et al.* The relationship between trinucleotide (CAG) repeat length and clinical features of Huntington's disease. *Nat. Genet.* **4,** 398–403 (1993).

37. Trottier, Y., Biancalana, V. & Mandel, J. L. Instability of CAG repeats in Huntington's disease: relation to parental transmission and age of onset. *J. Med. Genet.* **31,** 377–382 (1994).

38. Hannan, A. J. Trinucleotide-Repeat Expansions and Neurodegenerative Disease: A Mechanism of Pathogenesis*. *Clin. Exp. Pharmacol. Physiol.* **23,** 1015–1020 (1996).

39. Ratovitski, T. *et al.* N-terminal proteolysis of full-length mutant huntingtin in an inducible PC12 cell model of Huntington's disease. *Cell Cycle Georget. Tex* **6,** 2970–2981 (2007).

40. Pearson, C. E. Slipping while sleeping? Trinucleotide repeat expansions in germ cells. *Trends Mol. Med.* **9,** 490–495 (2003).

41. Panegyres, P. K., Beilby, J., Bulsara, M., Toufexis, K. & Wong, C. A study of potential interactive genetic factors in Huntington's disease. *Eur. Neurol.* **55,** 189–192 (2006).

42. Squitieri, F. *et al.* Family and molecular data for a fine analysis of age at onset in Huntington disease. *Am. J. Med. Genet.* **95,** 366–373 (2000).

43. Tobin, A. J. & Signer, E. R. Huntington's disease: the challenge for cell biologists. *Trends Cell Biol.* **10,** 531–536 (2000).

44. Shao, J. & Diamond, M. I. Polyglutamine diseases: emerging concepts in pathogenesis and therapy. *Hum. Mol. Genet.* **16 Spec No. 2,** R115-123 (2007).

45. Zuccato, C. *et al.* Progressive loss of BDNF in a mouse model of Huntington's disease and rescue by BDNF delivery. *Pharmacol. Res.* **52,** 133–139 (2005).

46. Godin, J. D. *et al.* Huntingtin Is Required for Mitotic Spindle Orientation and Mammalian Neurogenesis. *Neuron* **67,** 392–406 (2010).

47. Schulte, J. & Littleton, J. T. The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. *Curr. Trends Neurol.* **5,** 65–78 (2011).

48. Sathasivam, K. *et al.* Aberrant splicing of HTT generates the pathogenic exon 1 protein in Huntington disease. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 2366–2370 (2013).

49. Graham, R. K. *et al.* Cleavage at the caspase-6 site is required for neuronal dysfunction and degeneration due to mutant huntingtin. *Cell* **125,** 1179–1191 (2006).

50. Krench, M. & Littleton, J. T. Modeling Huntington disease in Drosophila. *Fly (Austin)* **7,** 229–236 (2013).

51. Parker, J. A. *et al.* Expanded polyglutamines in Caenorhabditis elegans cause axonal abnormalities and severe dysfunction of PLM mechanosensory neurons without cell death. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 13318–13323 (2001).

52. Parker, J. A. *et al.* Huntingtin-interacting protein 1 influences worm and mouse presynaptic function and protects Caenorhabditis elegans neurons against mutant polyglutamine toxicity. *J. Neurosci. Off. J. Soc. Neurosci.* **27,** 11056–11064 (2007).

53. Yang, S.-H. *et al.* Towards a transgenic model of Huntington's disease in a non-human primate. *Nature* **453,** 921–924 (2008).

54. Yang, D. *et al.* Expression of Huntington's disease protein results in apoptotic neurons in the brains of cloned transgenic pigs. *Hum. Mol. Genet.* **19,** 3983–3994 (2010).

55. Baxa, M. *et al.* A transgenic minipig model of Huntington's Disease. *J. Huntingt. Dis.* **2,** 47–68 (2013).

56. Jacobsen, J. C. *et al.* An ovine transgenic Huntington's disease model. *Hum. Mol. Genet.* **19,** 1873–1882 (2010).

57. Yu-Taeger, L. *et al.* A Novel BACHD Transgenic Rat Exhibits Characteristic Neuropathological Features of Huntington Disease. *J. Neurosci.* **32,** 15426–15438 (2012).

58. Abada, Y. K., Schreiber, R. & Ellenbroek, B. Motor, emotional and cognitive deficits in adult BACHD mice: A model for Huntington's disease. *Behav. Brain Res.* **238,** 243–251 (2013).

59. Beal, M. F. *et al.* Neurochemical and histologic characterization of striatal excitotoxic lesions produced by the mitochondrial toxin 3-nitropropionic acid. *J. Neurosci.* **13,** 4181–4192 (1993).

60. Brouillet, E. *et al.* Chronic mitochondrial energy impairment produces selective striatal degeneration and abnormal choreiform movements in primates. *Proc. Natl. Acad. Sci. U. S. A.* **92,** 7105–7109 (1995).

61. Coyle, J. T. & Schwarcz, R. Lesion of striatal neurones with kainic acid provides a model for Huntington's chorea. *Nature* **263,** 244–246 (1976).

62. Isacson, O., Brundin, P., Kelly, P. A., Gage, F. H. & Björklund, A. Functional neuronal replacement by grafted striatal neurones in the ibotenic acid-lesioned rat striatum. *Nature* **311,** 458–460 (1984).

63. Beal, M. F. *et al.* Replication of the neurochemical characteristics of Huntington's disease by quinolinic acid. *Nature* **321,** 168–171 (1986).

64. Régulier, E., Trottier, Y., Perrin, V., Aebischer, P. & Déglon, N. Early and reversible neuropathology induced by tetracycline-regulated lentiviral overexpression of mutant huntingtin in rat striatum. *Hum. Mol. Genet.* **12,** 2827–2836 (2003).

65. Almeida, L. P. de, Ross, C. A., Zala, D., Aebischer, P. & Déglon, N. Lentiviral-Mediated Delivery of Mutant Huntingtin in the Striatum of Rats Induces a Selective Neuropathology Modulated by Polyglutamine Repeat Size, Huntingtin Expression Levels, and Protein Length. *J. Neurosci.* **22,** 3473–3483 (2002).

66. Palfi, S. *et al.* Expression of Mutated Huntingtin Fragment in the Putamen Is Sufficient to Produce Abnormal Movement in Non-human Primates. *Mol. Ther.* **15,** 1444–1451 (2007).

67. Mangiarini, L. *et al.* Exon 1 of the HD Gene with an Expanded CAG Repeat Is Sufficient to Cause a Progressive Neurological Phenotype in Transgenic Mice. *Cell* **87,** 493–506 (1996).

68. Schilling, G. *et al.* Intranuclear inclusions and neuritic aggregates in transgenic mice expressing a mutant N-terminal fragment of huntingtin. *Hum. Mol. Genet.* **8,** 397–407 (1999).

69. Yu, Z.-X. *et al.* Mutant Huntingtin Causes Context-Dependent Neurodegeneration in Mice with Huntington's Disease. *J. Neurosci.* **23,** 2193–2202 (2003).

70. Gray, M. *et al.* Full Length Human Mutant Huntingtin with a Stable Polyglutamine Repeat Can Elicit Progressive and Selective Neuropathogenesis in BACHD Mice. *J. Neurosci. Off. J. Soc. Neurosci.* **28,** 6182–6195 (2008).

71. Wang, N. *et al.* Neuronal targets for reducing mutant huntingtin expression to ameliorate disease in a mouse model of Huntington's disease. *Nat. Med.* **20,** 536–541 (2014).

72. Hodgson, J. G. *et al.* A YAC Mouse Model for Huntington's Disease with Full-Length Mutant Huntingtin, Cytoplasmic Toxicity, and Selective Striatal Neurodegeneration. *Neuron* **23,** 181–192 (1999).

73. Slow, E. J. *et al.* Selective striatal neuronal loss in a YAC128 mouse model of Huntington disease. *Hum. Mol. Genet.* **12,** 1555–1567 (2003).

74. Bradford, J. *et al.* Expression of mutant huntingtin in mouse brain astrocytes causes age-dependent neurological symptoms. *Proc. Natl. Acad. Sci. U. S. A.* **106,** 22480–22485 (2009).

75. Southwell, A. L. *et al.* A novel humanized mouse model of Huntington disease for preclinical development of therapeutics targeting mutant huntingtin alleles. *Hum. Mol. Genet.* **26,** 1115–1132 (2017).

76. Southwell, A. L. *et al.* A fully humanized transgenic mouse model of Huntington disease. *Hum. Mol. Genet.* **22,** 18–34 (2013).

77. Wheeler, V. C. *et al.* Long glutamine tracts cause nuclear localization of a novel form of huntingtin in medium spiny striatal neurons in HdhQ92 and HdhQ111 knock-in mice. *Hum. Mol. Genet.* **9,** 503–513 (2000).

78. Menalled, L. B., Sison, J. D., Dragatsis, I., Zeitlin, S. & Chesselet, M.-F. Time course of early motor and neuropathological anomalies in a knock-in mouse model of Huntington's disease with 140 CAG repeats. *J. Comp. Neurol.* **465,** 11–26 (2003).

79. Lin, C. H. *et al.* Neurological abnormalities in a knock-in mouse model of Huntington's disease. *Hum. Mol. Genet.* **10,** 137–144 (2001).

80. Woodman, B. *et al.* The Hdh(Q150/Q150) knock-in mouse model of HD and the R6/2 exon 1 model develop comparable and widespread molecular phenotypes. *Brain Res. Bull.* **72,** 83–97 (2007).

81. Heng, M. Y., Tallaksen-Greene, S. J., Detloff, P. J. & Albin, R. L. Longitudinal evaluation of the Hdh(CAG)150 knock-in murine model of Huntington's disease. *J. Neurosci. Off. J. Soc. Neurosci.* **27,** 8989–8998 (2007).

82. Shaw, G., Morse, S., Ararat, M. & Graham, F. L. Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells. *FASEB J.* (2002). doi:10.1096/fj.01-0995fje

83. Cell atlas - HTT - The Human Protein Atlas. Available at: https://www.proteinatlas.org/ENSG00000197386-HTT/cell. (Accessed: 3rd January 2018)

84. Bowles, K. R., Brooks, S. P., Dunnett, S. B. & Jones, L. Huntingtin Subcellular Localisation Is Regulated by Kinase Signalling Activity in the StHdhQ111 Model of HD. *PLoS ONE* **10,** (2015).

85. Ritch, J. J. *et al.* Multiple Phenotypes in Huntington Disease Mouse Neural Stem Cells. *Mol. Cell. Neurosci.* **50,** 70–81 (2012).

86. Boussicault, L. *et al.* Impaired brain energy metabolism in the BACHD mouse model of Huntington's disease: critical role of astrocyte–neuron interactions. *J. Cereb. Blood Flow Metab.* **34,** 1500–1510 (2014).

87. Kwan, W. *et al.* Mutant huntingtin impairs immune cell migration in Huntington disease. *J. Clin. Invest.* **122,** 4737–4747 (2012).

88. Skotte, N. H. *et al.* Allele-Specific Suppression of Mutant Huntingtin Using Antisense Oligonucleotides: Providing a Therapeutic Option for All Huntington Disease Patients. *PLOS ONE* **9,** e107434 (2014).

89. Liu, L. *et al.* Induced Pluripotent Stem Cells in Huntington's Disease: Disease Modeling and the Potential for Cell-Based Therapy. *Mol. Neurobiol.* **53,** 6698–6708 (2016).

90. Valenza, M. *et al.* Dysfunction of the Cholesterol Biosynthetic Pathway in Huntington's Disease. *J. Neurosci.* **25,** 9932–9939 (2005).

91. Seo, H., Sonntag, K.-C. & Isacson, O. Generalized brain and skin proteasome inhibition in Huntington's disease. *Ann. Neurol.* **56,** 319–328 (2004).

92. Takahashi, K. & Yamanaka, S. Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell* **126,** 663–676 (2006).

93. Zhang, N., Bailus, B. J., Ring, K. L. & Ellerby, L. M. iPSC-based drug screening for Huntington's disease. *Brain Res.* **1638,** 42–56 (2016).

94. Coppen, E. M. & Roos, R. A. C. Current Pharmacological Approaches to Reduce Chorea in Huntington's Disease. *Drugs* **77,** 29–46 (2017).

95. Priller, J., Ecker, D., Landwehrmeyer, B. & Craufurd, D. A Europe-wide assessment of current medication choices in Huntington's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **23,** 1788 (2008).

96. Orth, M. *et al.* Observing Huntington's Disease: the European Huntington's Disease Network's REGISTRY. *PLoS Curr.* **2,** RRN1184 (2010).

97. Burgunder, J.-M. Recent advances in the management of choreas. *Ther. Adv. Neurol. Disord.* **6,** 117–127 (2013).

98. Brown, C. S., Markowitz, J. S., Moore, T. R. & Parker, N. G. Atypical antipsychotics: Part II: Adverse effects, drug interactions, and costs. *Ann. Pharmacother.* **33,** 210–217 (1999).

99. Videnovic, A. Treatment of huntington disease. *Curr. Treat. Options Neurol.* **15,** 424–438 (2013).

100. Braak, H. & Braak, E. Allocortical involvement in Huntington's disease. *Neuropathol. Appl. Neurobiol.* **18,** 539–547 (1992).

101. Vonsattel, J. P. *et al.* Neuropathological classification of Huntington's disease. *J. Neuropathol. Exp. Neurol.* **44,** 559–577 (1985).

102. Bachoud-Lévi, A. C. [Intracerebral graft in Huntington disease: debates and prospects]. *Rev. Neurol. (Paris)* **162,** 687–689 (2006).

103. Nicoleau, C., Viegas, P., Peschanski, M. & Perrier, A. L. Human pluripotent stem cell therapy for Huntington's disease: technical, immunological, and safety challenges human pluripotent stem cell therapy for Huntington's disease: technical, immunological, and safety challenges. *Neurother. J. Am. Soc. Exp. Neurother.* **8,** 562–576 (2011).

104. Cisbani, G. *et al.* Presence of tau pathology within foetal neural allografts in patients with Huntington's and Parkinson's disease. *Brain* **140,** 2982–2992 (2017).

105. Ramaswamy, S. & Kordower, J. H. Gene therapy for Huntington's disease. *Neurobiol. Dis.* **48,** 243–254 (2012).

106. Bemelmans, A. P. *et al.* Brain-derived neurotrophic factor-mediated protection of striatal neurons in an excitotoxic rat model of Huntington's disease, as demonstrated by adenoviral gene transfer. *Hum. Gene Ther.* **10,** 2987–2997 (1999).

107. Kells, A. P., Henry, R. A. & Connor, B. AAV–BDNF mediated attenuation of quinolinic acid-induced neuropathology and motor function impairment. *Gene Ther.* **15,** 966 (2008).

108. Gharami, K., Xie, Y., An, J. J., Tonegawa, S. & Xu, B. Brain-derived neurotrophic factor over-expression in the forebrain ameliorates Huntington's disease phenotypes in mice. *J. Neurochem.* **105,** 369–379 (2008).

109. Mittoux, V. *et al.* Restoration of cognitive and motor functions by ciliary neurotrophic factor in a primate model of Huntington's disease. *Hum. Gene Ther.* **11,** 1177–1187 (2000).

110. Bachoud-Lévi, A. C. *et al.* Neuroprotective gene therapy for Huntington's disease using a polymer encapsulated BHK cell line engineered to secrete human CNTF. *Hum. Gene Ther.* **11,** 1723–1729 (2000).

111. Bloch, J. *et al.* Neuroprotective gene therapy for Huntington's disease, using polymer-encapsulated cells engineered to secrete human ciliary neurotrophic factor: results of a phase I study. *Hum. Gene Ther.* **15,** 968–975 (2004).

112. Pérez-Navarro, E., Arenas, E., Marco, S. & Alberch, J. Intrastriatal grafting of a GDNF-producing cell line protects striatonigral neurons from quinolinic acid excitotoxicity in vivo. *Eur. J. Neurosci.* **11,** 241–249 (1999).

113. Hoffman, D., Breakefield, X. O., Short, M. P. & Aebischer, P. Transplantation of a polymer-encapsulated cell line genetically engineered to release NGF. *Exp. Neurol.* **122,** 100–106 (1993).

114. Boussicault, L. *et al.* CYP46A1, the rate-limiting enzyme for cholesterol degradation, is neuroprotective in Huntington's disease. *Brain* **139,** 953–970 (2016).

115. Wild, E. J. & Tabrizi, S. J. Therapies targeting DNA and RNA in Huntington's disease. *Lancet Neurol.* **16,** 837–847 (2017).

116. IONIS-Community-Statement-12082017-FINAL.pdf.

117. Garriga-Canut, M. *et al.* Synthetic zinc finger repressors reduce mutant huntingtin expression in the brain of R6/2 mice. *Proc. Natl. Acad. Sci. U. S. A.* **109,** E3136–E3145 (2012).

118. A Complete Suite of Reagents, From Cas9 Delivery Tools to Cell Function Assays. Available at: https://www.thermofisher.com/fr/fr/home/references/newsletters-and-journals/bioprobes-journal-of-cell-biology-applications/bioprobes-74/bioprobes-74-crispr-cas9-reagents.html. (Accessed: 4th January 2018)

119. Shin, J. W. *et al.* Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. *Hum. Mol. Genet.* **25,** 4566–4576 (2016).

120. Yang, S. *et al.* CRISPR/Cas9-mediated gene editing ameliorates neurotoxicity in mouse model of Huntington's disease. *J. Clin. Invest.* **127,** 2719–2724 (2017).

121. Kolli, N., Lu, M., Maiti, P., Rossignol, J. & Dunbar, G. L. CRISPR-Cas9 Mediated Gene-Silencing of the Mutant Huntingtin Gene in an In Vitro Model of Huntington's Disease. *Int. J. Mol. Sci.* **18,** (2017).

122. Merienne, N. *et al.* The Self-Inactivating KamiCas9 System for the Editing of CNS Disease Genes. *Cell Rep.* **20,** 2980–2991 (2017).

123. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).

124. Staahl, B. T. *et al.* Efficient genome editing in the mouse brain by local delivery of engineered Cas9 ribonucleoprotein complexes. *Nat. Biotechnol.* **35,** 431 (2017).

125. Crick, F. H. On protein synthesis. *Symp. Soc. Exp. Biol.* **12,** 138–163 (1958).

126. Crick, F. Central dogma of molecular biology. *Nature* **227,** 561–563 (1970).

127. Crick, F. Ideas on Protein Synthesis. *Wellcome Library for the History and Understanding of Medicine. Francis Harry Compton Crick Papers* (1956). Available at:

https://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/SCBBFT. (Accessed: 20th June 2017)

128. Crick, F. H. The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379 (1968).
129. Gesteland, R. F. *The Rna World*. (Cold Spring Harbor Laboratory Press, 1999).
130. Bernhardt, H. S. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a. *Biol. Direct* **7,** 23 (2012).
131. Crick, F. H. The origin of the genetic code. *J. Mol. Biol.* **38,** 367–379 (1968).
132. Alberts, B. *et al.* The RNA World and the Origins of Life. (2002).
133. Tafforeau, L. Qu'en est-il de la biogenèse des ribosomes chez l'homme ? *médecine/sciences* **31,** 622–628
134. Klemm, B. P. *et al.* The Diversity of Ribonuclease P: Protein and RNA Catalysts with Analogous Biological Functions. *Biomolecules* **6,** (2016).
135. Delihas, N. Regulation of gene expression by trans-encoded antisense RNAs. *Mol. Microbiol.* **15,** 411–414 (1995).
136. Gong, S., Wang, Y., Wang, Z. & Zhang, W. Computational Methods for Modeling Aptamers and Designing Riboswitches. *Int. J. Mol. Sci.* **18,** (2017).
137. Farrell Jr., R. E. Chapter 1 - RNA and the Cellular Biochemistry Revisited. in *RNA Methodologies (4th Edition)* 1–44 (Academic Press, 2010). doi:10.1016/B978-0-12-374727-3.00001-2
138. Koonin, E. V. Why the Central Dogma: on the nature of the great biological exclusion principle. *Biol. Direct* **10,** 52 (2015).
139. Berg, J. M., Tymoczko, J. L. & Stryer, L. Eukaryotic Transcription and Translation Are Separated in Space and Time. (2002).
140. Gooding, C. *et al.* A class of human exons with predicted distant branch points revealed by analysis of AG dinucleotide exclusion zones. *Genome Biol.* **7,** R1 (2006).
141. Corvelo, A., Hallegger, M., Smith, C. W. J. & Eyras, E. Genome-Wide Association between Branch Point Properties and Alternative Splicing. *PLoS Comput. Biol.* **6,** (2010).

142. Zhang, M. Q. Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.* **7,** 919–932 (1998).

143. Wahl, M. C., Will, C. L. & Lührmann, R. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell* **136,** 701–718 (2009).

144. Verma, B., Akinyi, M. V., Norppa, A. J. & Frilander, M. J. Minor spliceosome and disease. *Semin. Cell Dev. Biol.* (2017). doi:10.1016/j.semcdb.2017.09.036

145. Lin, C.-L., Taggart, A. J. & Fairbrother, W. G. RNA structure in splicing: An evolutionary perspective. *RNA Biol.* **13,** 766–771 (2016).

146. MacMillan, A. M. *et al.* Dynamic association of proteins with the pre-mRNA branch region. *Genes Dev.* **8,** 3008–3020 (1994).

147. Sun, J. S. & Manley, J. L. A novel U2-U6 snRNA structure is necessary for mammalian mRNA splicing. *Genes Dev.* **9,** 843–854 (1995).

148. Query, C. C., Moore, M. J. & Sharp, P. A. Branch nucleophile selection in pre-mRNA splicing: evidence for the bulged duplex model. *Genes Dev.* **8,** 587–597 (1994).

149. Suñé-Pou, M. *et al.* Targeting Splicing in the Treatment of Human Disease. *Genes* **8,** (2017).

150. Cech, T. R. The generality of self-splicing RNA: relationship to nuclear mRNA splicing. *Cell* **44,** 207–210 (1986).

151. Cech, T. R. Self-splicing of group I introns. *Annu. Rev. Biochem.* **59,** 543–568 (1990).

152. Kornblihtt, A. R. *et al.* Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nat. Rev. Mol. Cell Biol.* **14,** 153 (2013).

153. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA* **4,** 49–60 (2013).

154. Wang, Z. & Burge, C. B. Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA N. Y. N* **14,** 802–813 (2008).

155. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40,** 1413–1415 (2008).

156. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456,** 470–476 (2008).

157. Vuong, C. K., Black, D. L. & Zheng, S. The neurogenetics of alternative splicing. *Nat. Rev. Neurosci.* **17,** 265–281 (2016).

158. Philippi, S. *et al.* Dysferlin rescue by spliceosome-mediated pre-mRNA trans-splicing targeting introns harbouring weakly defined 3′ splice sites. *Hum. Mol. Genet.* **24,** 4049–4060 (2015).

159. Roca, X., Sachidanandam, R. & Krainer, A. R. Intrinsic differences between authentic and cryptic 5′ splice sites. *Nucleic Acids Res.* **31,** 6321–6333 (2003).

160. Park, S. A., Ahn, S. I. & Gallo, J.-M. Tau mis-splicing in the pathogenesis of neurodegenerative disorders. *BMB Rep.* **49,** 405–413 (2016).

161. Dandekar, T. & Sibbald, P. R. Trans-splicing of pre-mRNA is predicted to occur in a wide range of organisms including vertebrates. *Nucleic Acids Res.* **18,** 4719–4725 (1990).

162. Bruzik, J. P. & Maniatis, T. Spliced leader RNAs from lower eukaryotes are trans-spliced in mammalian cells. *Nature* **360,** 692–695 (1992).

163. Martínez-Abarca, F. & Toro, N. Group II introns in the bacterial world. *Mol. Microbiol.* **38,** 917–926 (2000).

164. Murphy, W. J., Watkins, K. P. & Agabian, N. Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing. *Cell* **47,** 517–525 (1986).

165. Sutton, R. E. & Boothroyd, J. C. Evidence for trans splicing in trypanosomes. *Cell* **47,** 527–535 (1986).

166. Chiara, M. D. & Reed, R. A two-step mechanism for 5′ and 3′ splice-site pairing. *Nature* **375,** 510–513 (1995).

167. Shimizu, A., Nussenzweig, M. C., Mizuta, T. R., Leder, P. & Honjo, T. Immunoglobulin double-isotype expression by trans-mRNA in a human immunoglobulin transgenic mouse. *Proc. Natl. Acad. Sci.* **86,** 8020–8023 (1989).

168. Caudevilla, C. *et al.* Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. *Proc. Natl. Acad. Sci. U. S. A.* **95,** 12185–12190 (1998).

169. Lei, Q. *et al.* Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biol. Evol.* **8,** 562–577 (2016).

170. Kowarz, E., Merkens, J., Karas, M., Dingermann, T. & Marschalek, R. Premature transcript termination, trans-splicing and DNA repair: a vicious path to cancer. *Am. J. Blood Res.* **1,** 1–12 (2011).

171. Zhou, J., Liao, J., Zheng, X. & Shen, H. Chimeric RNAs as potential biomarkers for tumor diagnosis. *BMB Rep.* **45,** 133–140 (2012).

172. Wu, C.-S. *et al.* Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Res.* **24,** 25–36 (2014).

173. Chuang, T.-J. *et al.* Integrative transcriptome sequencing reveals extensive alternative trans-splicing and cis-backsplicing in human cells. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky032

174. Benton, M. J. & Donoghue, P. C. J. Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24,** 26–53 (2007).

175. Lei, Q. *et al.* Evolutionary Insights into RNA trans-Splicing in Vertebrates. *Genome Biol. Evol.* **8,** 562–577 (2016).

176. Puttaraju, M., Jamison, S. F., Mansfield, S. G., Garcia-Blanco, M. A. & Mitchell, L. G. Spliceosome-mediated RNA *trans*-splicing as a tool for gene therapy. *Nat. Biotechnol.* **17,** 246 (1999).

177. Berger, A. *et al.* mRNA trans-splicing in gene therapy for genetic diseases. *Wiley Interdiscip. Rev. RNA* **7,** 487–498 (2016).

178. Peking, P. *et al.* An RNA-targeted therapy for dystrophic epidermolysis bullosa. *Nucleic Acids Res.* **45,** 10259–10269 (2017).

179. Berger, A. *et al.* Repair of rhodopsin mRNA by spliceosome-mediated RNA trans-splicing: a new approach for autosomal dominant retinitis pigmentosa. *Mol. Ther. J. Am. Soc. Gene Ther.* **23,** 918–930 (2015).

180. Goyenvalle, A., Leumann, C. & Garcia, L. Therapeutic Potential of Tricyclo-DNA antisense oligonucleotides. *J. Neuromuscul. Dis.* **3,** 157–167

181. Rindt, H. *et al.* Replacement of huntingtin exon 1 by trans-splicing. *Cell. Mol. Life Sci. CMLS* **69,** 4191–4204 (2012).

182. Rindt, H., Tom, C. M., Lorson, C. L. & Mattis, V. B. Optimization of trans-Splicing for Huntington's Disease RNA Therapy. *Front. Neurosci.* **11,** 544 (2017).

183. Koller, U. *et al.* A novel screening system improves genetic correction by internal exon replacement. *Nucleic Acids Res.* **39,** e108 (2011).

184. Wally, V., Koller, U. & Bauer, J. W. High-Throughput Screening for Highly Functional RNA-Trans-Splicing Molecules: Correction of Plectin in Epidermolysis Bullosa Simplex. (2011). doi:10.5772/24149

185. Chen, H. Y., Kathirvel, P., Yee, W. C. & Lai, P. S. Correction of dystrophia myotonica type 1 pre-mRNA transcripts by artificial *trans*-splicing. *Gene Ther.* **16,** 211 (2009).

186. Lorain, S., Peccate, C., Hir, M. L. & Garcia, L. Exon Exchange Approach to Repair Duchenne Dystrophin Transcripts. *PLOS ONE* **5,** e10894 (2010).

187. Avale, M. E., Rodríguez-Martín, T. & Gallo, J.-M. Trans-splicing correction of tau isoform imbalance in a mouse model of tau mis-splicing. *Hum. Mol. Genet.* **22,** 2603–2611 (2013).

188. Wally, V. *et al.* 5′ Trans-Splicing Repair of the PLEC1 Gene. *J. Invest. Dermatol.* **128,** 568–574 (2008).

189. Peking, P. *et al.* A Gene Gun-mediated Nonviral RNA trans-splicing Strategy for Col7a1 Repair. *Mol. Ther. Nucleic Acids* **5,** e287 (2016).

190. Uchida, N. *et al.* RNA Trans-Splicing Targeting Endogenous β-Globin Pre-Messenger RNA in Human Erythroid Cells. *Hum. Gene Ther. Methods* **28,** 91–99 (2017).

191. He, X. *et al.* Functional repair of p53 mutation in colorectal cancer cells using trans-splicing. *Oncotarget* **6,** 2034–2045 (2015).

192. Chao, H. *et al.* Phenotype correction of hemophilia A mice by spliceosome-mediated RNA *trans*-splicing. *Nat. Med.* **9,** 1015 (2003).

193. Mearini, G. *et al.* Repair of Mybpc3 mRNA by 5'-trans-splicing in a Mouse Model of Hypertrophic Cardiomyopathy. *Mol. Ther. Nucleic Acids* **2,** e102 (2013).

194. Prondzynski, M. *et al.* Evaluation of MYBPC3 trans-Splicing and Gene Replacement as Therapeutic Options in Human iPSC-Derived Cardiomyocytes. *Mol. Ther. Nucleic Acids* **7,** 475–486 (2017).

195. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **11,** 377–394 (2004).

196. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24,** 1774–1786 (2014).

197. De Conti, L., Baralle, M. & Buratti, E. Exon and intron definition in pre-mRNA splicing. *Wiley Interdiscip. Rev. RNA* **4,** 49–60 (2013).

198. Koller, U., Wally, V., Bauer, J. W. & Murauer, E. M. Considerations for a Successful RNA Trans-splicing Repair of Genetic Disorders. *Mol. Ther. Nucleic Acids* **3,** e157 (2014).

199. Serguera, C. & Bemelmans, A.-P. Gene therapy of the central nervous system: General considerations on viral vectors for gene transfer into the brain. *Rev. Neurol. (Paris)* **170,** 727–738 (2014).

200. Gemayel, R., Vinces, M. D., Legendre, M. & Verstrepen, K. J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.* **44,** 445–477 (2010).

201. Yang, F., Hanson, N. Q., Schwichtenberg, K. & Tsai, M. Y. Variable number tandem repeat in exon/intron border of the cystathionine beta-synthase gene: a single nucleotide substitution in the second repeat prevents multiple alternate splicing. *Am. J. Med. Genet.* **95,** 385–390 (2000).

202. Zumwalt, M., Ludwig, A., Hagerman, P. J. & Dieckmann, T. Secondary structure and dynamics of the r(CGG) repeat in the mRNA of the fragile X mental retardation 1 (FMR1) gene. *RNA Biol.* **4,** 93–100 (2007).

203. Dodd, D. W., Tomchick, D. R., Corey, D. R. & Gagnon, K. T. Pathogenic C9ORF72 Antisense Repeat RNA Forms a Double Helix with Tandem C:C Mismatches. *Biochemistry (Mosc.)* **55,** 1283–1286 (2016).

204. Zhang, Y., Roland, C. & Sagui, C. Structure and Dynamics of DNA and RNA Double Helices Obtained from the GGGGCC and CCCCGG Hexanucleotide Repeats That Are the Hallmark of C9FTD/ALS Diseases. *ACS Chem. Neurosci.* **8,** 578–591 (2017).

205. AceView: Gene:HTT, a comprehensive annotation of human, mouse and worm genes with mRNAs or ESTsAceView. Available at: https://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/av.cgi?db=human&c=Gene&l=HTT. (Accessed: 5th January 2018)

206. Dallinger, G. *et al.* Development of spliceosome-mediated RNA trans-splicing (SMaRT^{TM}) for the correction of inherited skin diseases. *Exp. Dermatol.* **12,** 37–46 (2003).

207. Mansfield, S. G. *et al.* Repair of CFTR mRNA by spliceosome-mediated RNA trans-splicing. *Gene Ther.* **7,** 1885–1895 (2000).

208. He, X. *et al.* Trans-splicing repair of mutant p53 suppresses the growth of hepatocellular carcinoma cells in vitro and in vivo. *Sci. Rep.* **5,** 8705 (2015).

209. Liu, X. *et al.* Partial correction of endogenous DeltaF508 CFTR in human cystic fibrosis airway epithelia by spliceosome-mediated RNA trans-splicing. *Nat. Biotechnol.* **20,** 47–52 (2002).

210. Rodriguez-Martin, T. *et al.* Reprogramming of tau alternative splicing by spliceosome-mediated RNA trans-splicing: Implications for tauopathies. *Proc. Natl. Acad. Sci. U. S. A.* **102,** 15659–15664 (2005).

211. Murauer, E. M. *et al.* Functional Correction of Type VII Collagen Expression in Dystrophic Epidermolysis Bullosa. *J. Invest. Dermatol.* **131,** 74–83 (2011).

212. Koller, U. *et al.* Trans-Splicing Improvement by the Combined Application of Antisense Strategies. *Int. J. Mol. Sci.* **16,** 1179–1191 (2015).

213. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37,** e67 (2009).

214. Tang, R., Prosser, D. O. & Love, D. R. Evaluation of Bioinformatic Programmes for the Analysis of Variants within Splice Site Consensus Regions. *Advances in Bioinformatics* (2016). doi:10.1155/2016/5614058

215. Kordasiewicz, H. B. *et al.* Sustained therapeutic reversal of Huntington's disease by transient repression of huntingtin synthesis. *Neuron* **74,** 1031–1044 (2012).

216. Drouet, V. *et al.* Sustained effects of nonallele-specific Huntingtin silencing. *Ann. Neurol.* **65,** 276–285 (2009).

217. Drouet, V. *et al.* Allele-Specific Silencing of Mutant Huntingtin in Rodent Brain and Human Stem Cells. *PLoS ONE* **9,** (2014).

218. Wang, G., Liu, X., Gaertig, M. A., Li, S. & Li, X.-J. Ablation of huntingtin in adult neurons is nondeleterious but its depletion in young mice causes acute pancreatitis. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3359–3364 (2016).

219. Dietrich, P., Johnson, I. M., Alli, S. & Dragatsis, I. Elimination of huntingtin in the adult mouse leads to progressive behavioral deficits, bilateral thalamic calcification, and altered brain iron homeostasis. *PLoS Genet.* **13,** (2017).

220. Li, J. Y., Popovic, N. & Brundin, P. The use of the R6 transgenic mouse models of Huntington's disease in attempts to develop novel therapeutic strategies. *NeuroRx J. Am. Soc. Exp. Neurother.* **2,** 447–464 (2005).

221. Schiefer, J. *et al.* Riluzole prolongs survival time and alters nuclear inclusion formation in a transgenic mouse model of Huntington's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **17,** 748–757 (2002).

222. Sánchez, I., Mahlke, C. & Yuan, J. Pivotal role of oligomerization in expanded polyglutamine neurodegenerative disorders. *Nature* **421,** 373 (2003).

223. Tanaka, M. *et al.* Trehalose alleviates polyglutamine-mediated pathology in a mouse model of Huntington disease. *Nat. Med.* **10,** 148–154 (2004).

224. Hockly, E. *et al.* Suberoylanilide hydroxamic acid, a histone deacetylase inhibitor, ameliorates motor deficits in a mouse model of Huntington's disease. *Proc. Natl. Acad. Sci. U. S. A.* **100,** 2041–2046 (2003).

225. Ferrante, R. J. *et al.* Histone deacetylase inhibition by sodium butyrate chemotherapy ameliorates the neurodegenerative phenotype in Huntington's disease mice. *J. Neurosci. Off. J. Soc. Neurosci.* **23,** 9418–9427 (2003).

226. Ferrante, R. J. *et al.* Chemotherapy for the brain: the antitumor antibiotic mithramycin prolongs survival in a mouse model of Huntington's disease. *J. Neurosci. Off. J. Soc. Neurosci.* **24,** 10335–10342 (2004).

227. Chen, M. *et al.* Minocycline inhibits caspase-1 and caspase-3 expression and delays mortality in a transgenic mouse model of Huntington disease. *Nat. Med.* **6,** 797–801 (2000).

228. Smith, D. L. *et al.* Minocycline and doxycycline are not beneficial in a model of Huntington's disease. *Ann. Neurol.* **54,** 186–196 (2003).

229. Ona, V. O. *et al.* Inhibition of caspase-1 slows disease progression in a mouse model of Huntington's disease. *Nature* **399,** 263–267 (1999).

230. Ferrante, R. J. *et al.* Therapeutic effects of coenzyme Q10 and remacemide in transgenic mouse models of Huntington's disease. *J. Neurosci. Off. J. Soc. Neurosci.* **22,** 1592–1599 (2002).

231. Schiefer, J. *et al.* The metabotropic glutamate receptor 5 antagonist MPEP and the mGluR2 agonist LY379268 modify disease progression in a transgenic mouse model of Huntington's disease. *Brain Res.* **1019,** 246–254 (2004).

232. Dedeoglu, A. *et al.* Creatine therapy provides neuroprotection after onset of clinical symptoms in Huntington's disease transgenic mice. *J. Neurochem.* **85,** 1359–1367 (2003).

233. Clifford, J. J. *et al.* Essential fatty acids given from conception prevent topographies of motor deficit in a transgenic model of Huntington's disease. *Neuroscience* **109,** 81–88 (2002).

234. Norflus, F. *et al.* Anti-inflammatory treatment with acetylsalicylate or rofecoxib is not neuroprotective in Huntington's disease transgenic mice. *Neurobiol. Dis.* **17,** 319–325 (2004).

235. Klivenyi, P. *et al.* Increased survival and neuroprotective effects of BN82451 in a transgenic mouse model of Huntington's disease. *J. Neurochem.* **86,** 267–272 (2003).

236. Popovic, N., Maingay, M., Kirik, D. & Brundin, P. Lentiviral gene delivery of GDNF into the striatum of R6/2 Huntington mice fails to attenuate behavioral and neuropathological changes. *Exp. Neurol.* **193,** 65–74 (2005).

237. Abada, Y.-S. K., Nguyen, H. P., Schreiber, R. & Ellenbroek, B. Assessment of motor function, sensory motor gating and recognition memory in a novel BACHD transgenic rat model for huntington disease. *PloS One* **8,** e68584 (2013).

238. Lamirault, C. *et al.* Altered reactivity of central amygdala to GABAAR antagonist in the BACHD rat model of Huntington disease. *Neuropharmacology* **123,** 136–147 (2017).

239. Clemensson, E. K. H., Clemensson, L. E., Riess, O. & Nguyen, H. P. The BACHD Rat Model of Huntington Disease Shows Signs of Fronto-Striatal Dysfunction in Two Operant Conditioning Tests of Short-Term Memory. *PloS One* **12,** e0169051 (2017).

240. Manfré, G. *et al.* The BACHD Rat Model of Huntington Disease Shows Specific Deficits in a Test Battery of Motor Function. *Front. Behav. Neurosci.* **11,** 218 (2017).

241. Yu-Taeger, L., Bonin, M., Stricker-Shaver, J., Riess, O. & Nguyen, H. H. P. Dysregulation of gene expression in the striatum of BACHD rats expressing full-length mutant huntingtin and associated abnormalities on molecular and protein levels. *Neuropharmacology* **117,** 260–272 (2017).

242. Zhang, N., An, M. C., Montoro, D. & Ellerby, L. M. Characterization of Human Huntington's Disease Cell Model from Induced Pluripotent Stem Cells. *PLoS Curr.* **2,** (2010).

243. HD iPSC Consortium. Induced pluripotent stem cells from patients with Huntington's disease show CAG-repeat-expansion-associated phenotypes. *Cell Stem Cell* **11,** 264–278 (2012).

244. HD iPSC Consortium. Developmental alterations in Huntington's disease neural cells and pharmacological rescue in cells and mice. *Nat. Neurosci.* **20,** 648–660 (2017).

245. Chae, J.-I. *et al.* Quantitative proteomic analysis of induced pluripotent stem cells derived from a human Huntington's disease patient. *Biochem. J.* **446,** 359–371 (2012).

246. Peking, P. *et al.* An RNA-targeted therapy for dystrophic epidermolysis bullosa. *Nucleic Acids Res.* **45,** 10259–10269 (2017).

247. Lacovich, V. *et al.* Tau Isoforms Imbalance Impairs the Axonal Transport of the Amyloid Precursor Protein in Human Neurons. *J. Neurosci.* **37,** 58–69 (2017).

248. Trapani, I. *et al.* Effective delivery of large genes to the retina by dual AAV vectors. *EMBO Mol. Med.* **6,** 194–211 (2014).

249. Koo, T., Popplewell, L., Athanasopoulos, T. & Dickson, G. Triple trans-splicing adeno-associated virus vectors capable of transferring the coding sequence for full-length dystrophin protein into dystrophic mice. *Hum. Gene Ther.* **25,** 98–108 (2014).

Résumé en français

La maladie de Huntington est une maladie neurodégénérative, génétique à transmission autosomale dominante. Une mutation présente dans l'exon 1 du gène codant pour la Huntingtine (HTT) est directement responsable de la maladie. Cette mutation correspond à l'expansion d'une répétition CAG qui, chez le sujet sain n'excède pas les 36 répétitions, alors qu'à partir de 40 répétitions on observe une pénétrance complète de la maladie. Lorsque ce gène est muté, il entraine la production d'une protéine toxique ayant à la fois un gain et une perte de fonction, qui va altérer de nombreux processus cellulaires menant à la mort de neurones épineux moyen GABAergiques du striatum. A l'heure actuelle aucun traitement curatif n'est disponible pour les patients, c'est pourquoi le développement de nouvelles stratégies thérapeutiques est essentiel, notamment les stratégies ciblant directement la mutation au niveau de l'ARN ou de l'ADN ; les autres stratégies n'ayant pas donné de résultats satisfaisants jusqu'à présent.

Dans ce projet de thèse, nous proposons de développer une stratégie de thérapie génique basée sur le trans-épissage (TE). Cette stratégie utilise un mécanisme naturel, bien que rare chez les mammifères, qui permet la production d'ARN messager (ARNm) mature après fusion d'exons provenant de précurseurs d'ARNm différents (pré-ARNm). Le développement de cette stratégie repose sur le design d'une molécule : le PTM pour « pre-trans-splicing molecule ». Cette molécule est un pré-ARNm artificiel composé de 3 éléments : (i) la version sans mutation des exons à remplacer, (ii) un intron artificiel contenant tous les éléments de signalisations requis pour être épissé, (iii) un domaine de liaison qui va permettre l'hybridation spécifique sur le pré-ARNm endogène contenant la mutation. Cette stratégie a déjà été adaptée pour différentes maladies génétiques avec des degrés d'efficacité fluctuant en fonction de différents paramètres, dont le domaine de liaison du PTM. Il a été démontré que ce domaine de liaison jouait un rôle très important dans le niveau d'activité de TE du PTM, or ce paramètre n'a jusqu'ici pas été extensivement étudié au regard de la faible diversité de PTMs testés dans la plupart des études publiées.

Dans ce contexte, l'objectif de ce projet est d'évaluer l'activité de TE d'une centaine de PTMs ciblant 3 introns HTT cibles en utilisant un système rapporteur fluorescent compatible avec une stratégie de criblage haut débit. Nous avons choisi de cibler les introns 3, 9 et 20 en suivant différentes contraintes impliquant leur taille (clonage dans des vecteurs viraux, et modèle de définition de l'exon et de l'intron) et leur dynamisme d'épissage (force intronique). Nous avons introduit ces introns HTT cibles dans un mini-gène hybride composé de deux introns artificiels avec en 5' une séquence codant la partie N-terminale de la YFP et en 3' la partie C-terminale de la mTurquoise2. Ces deux protéines fluorescentes sont dérivées de la GFP et présentent un fort taux d'homologie. Ainsi, en cas de cis-épissage du mini-

gène hybride, la fusion de ces deux exons artificiels mène à un ARNm codant pour une protéine fluorescente chimérique (XFP). Ce système rapporteur est complété par un PTM codant pour la partie N-terminale de mTurquoise2, et permet la production de la version entière de cette protéine après TE du pré-ARNm codé par le mini-gène hybride. La mTurquoise2 a une intensité de fluorescence plus importante que la XFP et est la seule à être détectée à 440nm. Ces premiers résultats nous ont permis de valider une détection spécifique des évènements de TE grâce à l'émission de fluorescence mTurquoise2.

Nous avons ensuite généré une banque de PTMs pour les introns HTT 3 et 9, par fragmentation aléatoires de ces derniers par sonication (en collaboration avec K. Labadie, C. Menguy et E. Mahieu du Génoscope), permettant de générer des PTMs avec des domaines de liaison de 139pb en moyenne. L'intron 20 étant beaucoup plus petit (256pb), un unique PTM a été généré avec un domaine de liaison couvrant la totalité de l'intron. L'activité de ces PTMs a ensuite été évaluée individuellement par co-transfection avec le mini-gène hybride correspondant et l'intensité de fluorescence mTurquoise2 a été mesurée après acquisition au microscope à épifluorescence. Pour les introns 3 et 9, nous avons pu mettre en évidence des variations de niveau de TE au sein des différents sous-domaines que nous avons définis le long de ces introns.

Pour chacun de ces introns, nous avons identifiés des PTMs avec une activité de TE significative et nous avons développé une méthode pour quantifier précisément leur taux de TE au niveau de l'ARN. Nous avons collecté les cellules positives mTurquoise2 (en collaboration avec J. Baijer et N. Deschamps, CEA) et nous avons extrait les ARNm de cette population. Après une PCR amplifiant à la fois la XFP et la mTurquoise2 avec des amorces universelles, nous avons effectué un traitement avec des enzymes de restrictions spécifiques soit de la XFP (produit cis-épissé), soit de la mTurquoise2 (produit trans-épissé). Bien que cette méthode demande encore certaines optimisations, les deux digestions du PTM #804 ciblant l'intron 3 HTT semblent indiquer un taux de TE de 8-10%. Ces résultats sont à considérer comme des données préliminaires mais cela nous confirme que cette méthode permet une quantification précise et absolue du taux de TE.

Pour conclure, nous avons mis en évidence la faisabilité du TE sur différents introns du gène HTT et identifié des régions introniques favorables pour le TE. Une méthode de quantification absolue du TE a également été développé au niveau de l'ARNm pour compléter la caractérisation des PTMs identifiés après le criblage. La prochaine étape du projet est de sortir du système rapporteur en adaptant les meilleurs PTMs au contexte endogène de la HTT, d'évaluer leur activité de TE dans un modèle de la maladie de Huntington (iPSC dérivées de patients, rat BACHD) et d'analyser leur potentiel thérapeutique en évaluant les améliorations des caractéristiques de la pathologie.

Titre : Vers la mise au point d'une thérapie génique par *trans*-épissage pour la maladie de Huntington : identification de cibles introniques dans le gène Huntingtine

Mots clés : Thérapie génique, Maladie de Huntington, Huntingtine, *trans*-épissage, ARN

Résumé : La maladie de Huntington (MH) est une maladie autosomale dominante causée par une expansion de la répétition CAG codant pour une expansion de la polyglutamine dans le premier exon du gène Huntingtine (HTT). Ce gène code pour une protéine ubiquitaire dont la mutation entraine de graves symptômes moteurs, psychiatriques et cognitifs, dus à la dégénérescence spécifique des neurones GABAergique épineux moyens du striatum. Nous proposons d'utiliser le trans-épissage pour développer un vecteur de thérapie génique qui réduira significativement voir éliminera l'expression de la protéine mutée tout en restaurant un niveau physiologique de HTT normale dans les cellules affectées par la mutation du gène Huntingtine. Cette technologie est basée sur le remplacement de l'exon muté par un exon sans mutation pendant l'étape de maturation de l'ARNm. Du fait du caractère dominant de la mutation,l'efficacité thérapeutique nécessitera une réaction de trans-épissage très efficace capable de convertir une portion significative de pre-ARNm HTT mutés en en ARNm HTT normaux. Nous avons donc développé un système rapporteur fluorescent permettant la détection des évènements de trans-épissage afin d'identifier les séquences les plus performantes parmi une centaine de molécules candidates. Nous avons validé notre stratégie de criblage basée sur la fluorescence et réalisé le criblage sur plusieurs introns HTT (3, 9 et 20) qui ont démontré des zones favorables au trans-épissage. Une méthode de quantification directe et absolue du taux de trans-épissage a également été validée pour déterminer très précisément le taux de correction. L'ensemble de ce travail a permis de contribuer à la mise en évidence de la faisabilité du trans-épissage dans le contexte de la MH

Title : Towards *trans*-splicing gene therapy for Huntington's disease : intronic target identification in the Huntingtin gene

Keywords : Gene therapy, Huntington disease, Huntingtin, *trans*-splicing, RNA

Abstract : Huntington's disease (HD) is an autosomal dominant genetic disorder caused by the expansion of a CAG repeat encoding a polyglutamine tract in the first exon of the Huntingtin gene (HTT). This gene encode a ubiquitous protein in which mutation lead to severe motor, psychiatric and cognitive deficits and causes degeneration of specific neuronal populations, in particular the GABAergic medium spiny neurons of the striatum. We propose to use trans-splicing to develop a gene therapy vector that will significantly reduce or eliminate the expression of the mutant protein while restoring a physiological level of normal HTT in cells affected by the HD mutation. This technology is based on replacement of the mutated exon by a normal version during the mRNA maturation process. HTT mutation being dominant, therapeutic benefits necessitates a highly efficient trans-splicing reaction that would convert a significant proportion of mutant-HTT pre-mRNA into normal HTT mRNA. For this purpose, we developed a fluorescent reporter system enabling the detection of trans-splicing events in high content screening in order to identify the most potent trans-splicing sequences among hundreds of molecules. We validated our fluorescent screening strategy and implement trans-splicing screening on 3 HTT introns (3, 9 and 20), in which we demonstrated the presence of hotspot promoting trans-splicing reactions. A direct and absolute quantification method was also validated to accurately assess the correction rate. Overall, this work generated additional evidences of trans-splicing feasibility in HD.