Faculdade de Engenharia da Universidade do Porto

Identification of genetic modifiers of somatic CAG instability in Huntington's Disease by *in vivo* CRISPR--Cas9 genome editing

António Gil Cabral de Azevedo

Integrated Masters in Bioengineering
Major in Biomedical Engineering

Supervisor: Ricardo Mouro Pinto, PhD, Center for Genomic Medicine, Massachusetts General Hospital & Harvard Medical School, USA
Co-Supervisor: Isabel Alonso, PhD, Abel Salazar's Institute for Biomedical Sciences, University of Porto; Institute for Molecular and Cell Biology (IBMC/I3S), Portugal

March 2017

This page is intentionally left blank.

ABSTRACT

Huntington's Disease (HD) is a devastating, dominantly inherited, neurodegenerative disorder caused by the expansion of a CAG repeat within the huntingtin gene (*HTT*), with longer repeats being associated with earlier disease onset and more severe HD symptoms and phenotypes (Gusella et al. 2014). Despite this being a single gene disorder, no cure or disease-modifying therapy has yet been achieved, indicating that novel approaches are critical (Gusella et al. 2014).

The CAG repeat is highly unstable, both intergenerationally and in somatic tissues, where the repeat expands progressively over time in a cell-/tissue-specific manner (Kennedy et al. 2003). Notably, medium-spiny neurons of the striatum, which succumb most severely to the effects of the *HTT* mutation, exhibit the most dramatic CAG expansions (Kennedy et al. 2003).

A progressive CAG length increase in somatic tissues could therefore contribute to the HD pathogenic process, an hypothesis that is supported by findings in human studies indicating that longer somatic expansions in HD postmortem brains are associated with an earlier disease onset (Swami et al. 2009).

The current findings could imply that factors that modify instability might also modify disease progression. Conversely, modifiers of disease progression may act via a mechanism that alters repeat instability. Thus, understanding the role of disease modifiers in somatic CAG expansion may provide novel targets for therapeutic intervention directed at the mutation itself.

To this end, a study was developed, focusing on genes recently identified in a Genome-Wide Association Study (GWAS) as candidate modifiers of age at motor symptoms onset in HD patients (Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium 2015). These candidate modifiers include a large number of DNA repair pathway genes, whose effect on somatic CAG instability will be assessed in future studies, aiming at a greater understanding of the role of this pathway in HD pathology and of its potential as a target of new disease modifying therapies.

For this purpose, the development of a platform for efficient *in vivo* editing of candidate HD modifier genes in an HD mouse model was be optimized and validated, first *in vitro* and then *in vivo* using the MMR genes known to affect somatic CAG stability from previous experiments (Pinto et al. 2013; Wheeler et al. 2003; Dragileva et al. 2009) as test targets. A recent technological advance in the field of genome editing, the CRISPR-Cas9 system, was used to enable a more agile and precise analysis (Xue et al. 2014).

The present thesis focused on an experimental system assessing candidate gene effect on liver CAG instability, which could be a potential correlate of striatal CAG instability. This system could mean a more efficient and robust platform for screening novel modifiers of somatic CAG instability in HD.

The study thus used recent technological advances to further dissect HD disease mechanisms, investigate novel genetic risk factors and identify modifiers of disease-relevant pathways such as somatic CAG instability.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

HD – Huntington's Disease

CAG – Cytosine, Adenosine and Guanine nucleotide triplet

bp – Nucleotide base pair

HTT – Human huntingtin gene

HTT - Huntingtin protein

Htt - mouse huntingtin gene

GWAS – Genome-Wide Association Study

MSN – Medium spiny neurons

MMR – DNA Mismatch Repair mechanisms

CRISPR - Clustered regularly interspaced short palindromic repeats

EMC – Enzyme mismatch cleavage assay

AAV – Adeno-associated virus

Cas9 – CRISPR Associated Protein 9

FACS - Fluorescence-activated cell sorting

MSI – Microsatellite Instability

Q111 – knock in mouse model B6.Hdh^{Q111}

NGS – Next Generation Sequencing

PBS – Phosphate Buffer Saline

1. INTRODUCTION

1.1 MOTIVATION

Huntington's disease (HD) is an hereditary incurable neurodegenerative disorder caused by an autosomal dominant CAG repeat expansion mutation in the *HTT* gene (Gusella et al. 2014). The mutation promotes dysfunction and neuron loss in specific vulnerable brain structures, particularly affecting the medium spiny neurons of the striatum. This pathological process correlates with the onset, typically in midlife, of chorea (unintentional movements), bradykinesia, cognitive decline and psychiatric symptoms (Ross et al. 2014).

Current medical care can only help patients suffering from this prolonged, debilitating, highly stigmatized and ultimately fatal disease through palliative symptomatic treatment, which is not without secondary effects (Ross et al. 2014; Zielonka et al. 2015). Affecting 1 in 10,000 people in western society, HD is a rare disease (Roos 2010). However, due to the high level of assistance and several different types of specialized care the condition requires, HD represents a heavy burden not only for patients but also for caretakers, health systems and society as a whole (GBI Research 2012; Simpson & Rae 2012).

At the present moment, promising potentially disease modifying emerging therapies, such as lowering mutant protein dose by gene suppression, are still in early stages of development, not yet having direct evidence of efficacy in HD patients. Moreover, these approaches face technical and ethical challenges (Zielonka et al. 2015).

Under these circumstances, patient access to effective treatment in the near future is uncertain. There is thus a great need for alternative disease modifying therapeutic strategies and targets.

New in-human validated targets based on genetic modifiers from genome-wide association studies (Huntington's & Consortium 2015), together with technological breakthroughs such as CRISPR-CAS9 (Platt et al. 2014), hold the hope of accelerating insight into disease pathology to generate clues for new much-needed alternative rational therapies. The treatment design and effect would also be potentially amenable to optimization by traditional drug development leading to faster and easier validation.

1.2 CONTEXT AND PREVIOUS STUDIES

A promising strategy for meeting the pressing need of validated therapeutic targets for effective treatment (capable of delaying or preventing HD clinical onset or progression) has been to look for new genetic modifiers of the disease. Variation of key phenotypes such as age at symptom onset is not entirely predicted by the length of the mutation, the primary determinant of disease onset and severity (Gusella et al. 2014). This main determinant only accounts for 72% of age at motor symptom onset variation (Project & Wexler 2004), leaving room for deviations of up to 25 years of life without disease unexplained (Gusella et al. 2014). The 40% inheritable component of the remaining variability is therefore in-human validated biological proof of genetic modifiers existing and having potential as targets for future therapies (Project & Wexler 2004). Being genome bound these modifiers are also easier to study systematically than the infinite possible environmental modifiers.

Recently, a genome-wide study powered by DNA samples of over 4000 phenotypically well-documented HD patients of European ancestry mapped genome regions associated with motor symptom onset variation. The study yielded loci with genome-wide statistically significant association and global insight into the pathways and biological functions involved in HD pathology. Potential gene candidates yet to be validated were found, along with strongly implicated pathways, in particular DNA handling and repair (Huntington's & Consortium 2015).

MLH1, a gene involved in DNA repair more precisely mismatch repair (MMR), which had been previously implicated in another genome-wide unbiased study in an HD mouse model (Pinto et al. 2013), stood out as almost genome-wide statistically significant (Huntington's & Consortium 2015). The gene was associated in knock out mice studies with a more pronounced somatic CAG repeat expansion in the striatum as well as reduced early HD pathology phenotypes (Pinto et al. 2013). *Mlh1* could therefore plausibly modify HD disease onset through modification of CAG repeat length.

Somatic repeat instability through MMR or more broadly DNA handling, might be a mechanism shared by some of the newly implicated genes, representing a disease modifying mechanism possibly common not only to HD pathology but also to other trinucleotide repeat disorders such as DM1 (myotonic dystrophy type 1) and cerebellar ataxias (Bettencourt et al. 2016).

CRISPR-CAS9, a new genetic engineering tool, holds the promise of accelerated validation of this hypothesis. CRISPR technology based models could enable faster and more versatile dissection of somatic repeat instability biology, thus generating insight for new disease modifying therapies.

1.3 OBJECTIVES

This work aims to further the understanding of genetic modifiers of Huntington's Disease and their biology with the ultimate goal of accelerating the discovery of new disease modifying therapies for Huntington's disease.

The role of somatic CAG repeat instability in the molecular mechanisms of genetic HD modifiers will be investigated *in vivo* through CRISPR-CAS9 technology.

A platform will be optimized for this purpose, aiming at a more agile assessment of modification, by candidate genes, of striatum somatic repeat instability in an HD model.

Another platform for faster screening of modification, by both candidate genes and candidate drugs, will be optimized for liver somatic repeat instability. This is a correlate of striatum somatic CAG instability of easier access, which will be the focus of the current thesis.

2. HUNTINGTON'S DISEASE

Huntington's disease is a high burden hereditary neurodegenerative disorder with a complex and not fully understood aetiology and pathophysiology, for which no disease modifying therapy has yet been achieved.

The present chapter overviews some basic concepts of HD biology and examines the potential of genetic modifiers for new disease modifying therapies.

2.1 HISTORY AND SOCIAL IMPACT

Huntington's chorea was described in 1872 by George Huntington, as a neurodegenerative disorder of middle age onset characterized by involuntary choreatic movements, behavioural and psychiatric disturbances and dementia, which was passed within families from generation to generation. It was later renamed Huntington's disease (HD) in the 1980's, due to the awareness of its extensive non-motor symptoms (Roos 2010; Vale & Cardoso 2015).

In 1983, the genetic defect causing HD was mapped by genetic linkage to the short arm of chromosome 4 (Gusella et al. 1983). The HD-causing mutation was identified as an expanded CAG trinucleotide repeat at the N-terminal of the protein coding huntingtin gene (*HTT*) (MacDonald et al. 1993; Gusella et al. 2014).

Since the discovery of the mutated gene, extensive research has been done aiming at a better understanding of huntingtin (HTT) function and of its role in HD pathology. However, no effective disease modifying therapy has

been achieved and many aspects of HTT function and HD disease mechanisms remain unsolved (Gusella et al. 2014; Wexler 2012).

Historically, this incurable severe hereditary disease is one of the most stigmatised disorders, currently still being associated with discrimination within families, in social settings and by mortgage and insurance companies. This translates to reluctance of at risk individuals to have their status registered in medical records and to their reduced access to support (Bombard et al. 2009).

Considered to be largely underestimated, HD directly affects about 1 in 10,000 individuals in western societies (Roos 2010; Gusella et al. 2014) (approximately 25,000–30,000 individuals with manifest HD and a further 150,000–250,000 individuals at risk in the USA (Harper 2002)), making it a rare orphan disease. However, it has a global trend of increase in prevalence (Rawlins et al. 2016) with estimates more than doubling for some populations in recent studies, namely in the United Kingdom (Evans et al. 2013) and Italy(Squitieri et al. 2015) which is partly attributed to the improvement of diagnosis and a greater awareness of the disease (Rawlins et al. 2016).

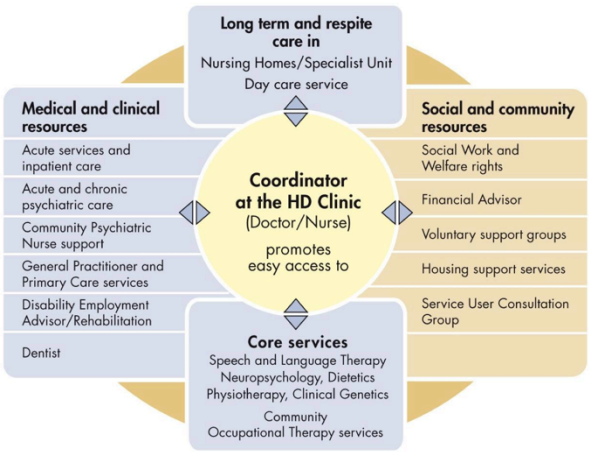

Figure 2.1 – Model of HD medical care. Source: (Rae et al. n.d.).

Representing a high burden for patients, caretakers and health systems (figure 2.1) HD had a global therapeutics market size of approximately $126.7 million in 2010 which is expected to grow to $786.5 million by 2017 with a compound annual growth rate of 29.8% (GBI Research 2012). As the economic

burden of HD increases with disease progression, new therapies for stabilizing or delaying progression would have a substantial net impact (Divino et al. 2013).

Even though there is a pressing need for new disease modifying therapies, current medical care only allows for symptomatic treatment (Kumar et al. 2015).

2.2 DIAGNOSIS AND CLINICAL PRESENTATION

HD can affect individuals of any age between infancy and senescence, although it primarily has adult onset of symptoms in middle age (Walker 2007). It can also present a variety of unspecific associated symptoms that overlap with other disorders (Craufurd et al. 2014). Being a rare disorder, Huntington's disease may therefore be sometimes difficult to diagnose for less experienced physicians, particularly in the absence of proven family history (Walker 2007; Craufurd et al. 2014).

Paradoxically however, there is a growing awareness of the disease, which is also becoming ever better described. Recent and ongoing natural history studies have enabled a finer understanding of HD's endophenotypes, signs and biomarkers that correlate with the disease and its progression, including during preclinical phases owing to predictive genetic testing (Ross et al. 2014).

2.2.1 Classification and diagnosis

Huntington's disease, classically characterized by motor, cognitive and psychiatric signs and symptoms is classified by the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) as a neurocognitive disorder (acquired cognitive decline and functional impairment as main features) (American Psychiatric Association 2013). The condition features early changes to executive function such as processing speed, organization and planning (as opposed to learning and memory) that along with change in behavior can precede its typical motor abnormalities: emergence of involuntary jerking movements (chorea) and slowing of voluntary movements (bradykinesia).

This disorder is diagnosed based on a proven family history and clinical symptoms. Although commonly accompanied by cognitive and psychiatric symptoms, motor changes together with family history are sufficient to meet clinical criteria, with unequivocal presence of an otherwise unexplained extrapyramidal movement disorder (e.g. chorea, dystonia, bradykinesia or

rigidity) defining the onset of the disease (American Psychiatric Association 2013; Ross et al. 2014; Roos 2010).

Mounting evidence from natural history studies however indicates cognitive impairment as a possible additional criterion. The same cannot be said of emotional and behavior changes which being implicated are not universally associated with the disorder in a steady and progressive way (Ross et al. 2014).

Genetic testing of the CAG expansion mutation in the *HTT* gene on chromosome, which requires consent, provides an alternative to proven family history. It is also an important tool for differential diagnosis as there are other conditions with similar symptoms that may be misleading as well as some rare cases phenocopies (presentation of identical symptoms without the underlying genetic cause) (American Psychiatric Association 2013; Ross et al. 2014). The most probable HD phenocopies are Huntington like 2 and 4 but in 97% of cases no genetic cause can be identified (Craufurd et al. 2014).

Juvenile Huntington's disease in which individuals develop manifest HD before the age of 20 is a variant of HD associated with a considerably longer CAG repeat length (usually >60 repeats), amounting to 5% of HD cases. It is associated to faster progression and more frequently to dystonia (sustained contraction of agonist and antagonist muscles leading to twisting movements) but not necessarily chorea (Ross et al. 2014; Walker 2007).

The most common tool of clinical and research assessment of HD is the Unified HD Rating Scale (UHDRS) that includes motor, cognitive, behavioural, emotional and functional components, allowing for a division of the manifest period in 5 descriptive stages (Ross et al. 2014).

2.2.2 Natural History and clinical presentation

Huntington's disease is a monogenic neurodegenerative disorder amenable to predictive genetic testing and within some limitations to estimation of age at symptom onset prediction. It has therefore been possible to follow HD disease progression since before the onset of clinical symptoms in individuals positive to the mutation. This has enabled a very detailed characterization of HD natural history unlike that of other neurodegenerative and late-onset diseases (Ross et al. 2014).

The course of Huntington's disease can be divided into two periods: a pre-manifest period before motor and neurological onset of symptoms and a manifest period after this formal onset of the disease. The pre-manifest period encompasses a presymptomatic (in which mutation carriers are clinically undistinguishable from controls) and a prodromal subdivion characterised by subtle motor, cognitive and behavioural changes. The prodromal period that can start up to 10 or 15 years before the age of motor onset of the disease, may feature signs such as irritability, restlessness, anxiety, dishinibition, difficulty with multitasking and forgetfulness. This period slowly merges with the manifest period with the onset of chorea, incoordination, motor impersistence (inability to maintain voluntary muscle contraction at a constant level, for instance applying steady pressure in a handshake or protruding the tongue), and slowed saccadic eye movements. Chorea is not present in all cases and can subside with the progress of the disease, as there is a growing dystonia, rigidity and bradykinesia.

Cognitive impairment and decline usually starts years before the motor age at onset and follows a profile closer to that of Parkinson's disease rather than of Alzheimer's, with decreased attention, mental flexibility, planning and organizational skills, visuospatial functions and emotion recognition preceding memory impairment.

In behavioural and psychiatric terms, there can be association with psychosis. Depression is commonly associated and there is a higher suicide rate in HD patients. However the more prevailing progressive persistent symptoms are irritability and apathy that can be present even before the motor onset.

The growing severity of motor and cognitive symptoms leads to mounting incapacity and loss of independence, and ultimately to death as a complication of falls or dysphagia (difficulties swallowing that may contribute to aspiration pneumonia). From diagnosis to death there is usually a latency of 20 years (Roos 2010; Walker 2007; Ross et al. 2014).

2.3 GENETICS

2.3.1 Mutation: Trinucleotide expansion

Huntington's disease is a hereditary autosomal dominant neurodegenerative disease caused by an abnormal expansion of a highly polymorphic CAG repeat in exon 1 of the *HTT* gene (Bean & Bayrak-Toydemir 2014). Trinucleotide expansion is common to a group of disorders known as trinucleotide repeat disorders to which HD belongs (repeats can be found in different regions of a gene locus in different diseases; larger nucleotide sequence repeat mutations are also associated with disorders) (Harper 2002).

Disease develops in individuals whose repeat length exceeds a given threshold, set at 40 repeats for HD (figure 2.2). While individuals with a repeat length below 36 repeats do not present symptoms, those with a repeat length between 36 and 40 repeats have a incomplete penetrance of the disease (Orr & Zoghbi 2007; Langbehn et al. 2004).

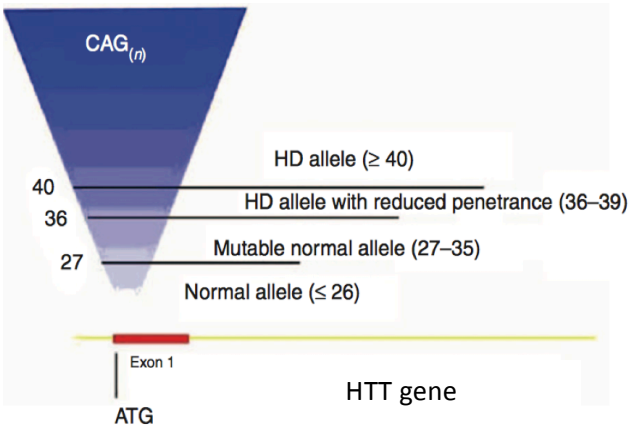

Figure 2.2 – Number of CAG repeats and HD penetrance. Source: (Bean & Bayrak-Toydemir 2014).

The threshold is thought to be associated with a balance between the formation of intermediate structures such as heteroduplex DNA (that can lead to repeat expansion upon resolution) and DNA repair machinery kinetics, when free DNA ends are generated, for instance in Okazaki fragments or single strand breaks (Lee & McMurray 2014).

It is also important to consider that trinucletide repeats and their expansion's effect on DNA structure can affect the interaction with non-genetic elements, namely of epigenetic nature (Orr & Zoghbi 2007; Nageshwaran & Festenstein 2015).

2.3.2 Mode of transmission: autosomal dominant

The HD mutation is inherited in an autosomal dominant manner with age-dependent penetrance (nearly full penetrance by the age of 65 years for a CAG repeat length equal to or above 40 repeats). This means both male and female descendants of affect individuals, have approximately a 50% chance of carrying the mutation (Ross & Tabrizi 2011).

2.3.3 Somatic and intergenerational instability

The length of repeats can expand, contract or remain stable when transmitted between generations leading to intergeneration variation (in about 80% of cases) (Duyao et al. 1993). There can also be repeat length instability somatically, leading to mosaicism in different tissues (with some vulnerable brain regions showing a age dependent increase in repeat length) (Telenius et al. 1994; Kennedy et al. 2003).

CAG repeat length accounts for up to 72% of age at onset variation (Project & Wexler 2004), explaining phenomena such as the anticipation of age at onset in some descendants (Duyao et al. 1993). Anticipation can be implicated in the juvenile form of HD (age of onset below 20 years; approximately 5% of cases) in which there is a substantially larger number of repeats relative to progenitors (Quarrell et al. 2012). The affected progenitor is usually male as testes and spermatozoids present a greater CAG repeat instability (Duyao et al. 1993). Moreover, repeat instability increases with repeat length making it easier for parents with a longer repeat to have descendants with an even longer repeat expansion (Pearson et al. 2005).

2.3.4 Polyglutaminic disorder

HD is a polyglutaminic (poly-Q) disorder given its repeats code for the glutamine amino acid (represented as a Q). The mutation thus leads to the expression of a mutant-expanded protein with an enlarged polyglutamine track, in this case within HTT's product, the huntingtin protein (Walker 2007; Orr & Zoghbi 2007).

2.4 PATHOPHYSIOLOGY

2.4.1 HTT and pathology: gain or loss of function?

The mutant huntingtin protein is thought to promote the pathologic hallmarks of HD, namely striatum cell loss and dysfunction, through loss of normal Htt function and simultaneously, to a greater extent, through Htt gain of toxic function (Bano et al. 2011; Gusella et al. 2014).

Huntingtin (Htt) is an essential protein expressed in all human cells whose absence can compromise normal embryogenesis and development. In some studies, Htt post-natal neuron-specific silencing was found to promote progressive apoptotic neuronal degeneration, which could indicate that a loss of functional Htt could lead to part of the deleterious processes associated with HD (Bano et al. 2011).

However, HD is thought to be mainly due to a toxic gain of function, since Htt loss of function models differ in phenotype from HD disease models. This hypothesis is supported by the existence of normal humans with only one copy of the HTT gene, which is not compatible with a simple loss of function effect. It is also supported by the known lethality of complete HTT deprivation, which is not compatible with a dominant negative interference of the mutant Htt with normal Htt function. The toxic gain of function does not appear to be protein dose dependent as having both HTT alleles mutated does not accelerate HD pathology relative to single allele mutation (only the repeat length of the allele with the longest repeat length is relevant for HD pathology) (Gusella et al. 2014).

2.4.2 HTT and mechanisms of pathology

The mutant protein is more prone to proteolysis and aggregate formation (figure 2.3), which is believed to lead to this toxic gain of function. The resulting aggregates generate intracellular inclusions and are associated with excitotoxicity, mitochondrial dysfunction, transcriptional dysregulation and apoptosis. This disease mechanism is thought to be shared with several

other trinucleotide repeat disorders. A longer repeat expansion is associated with faster aggregation kinetics which could contribute to the inverse correlation between repeat length and age at onset of the disease (Gusella et al. 2014; Walker 2007).

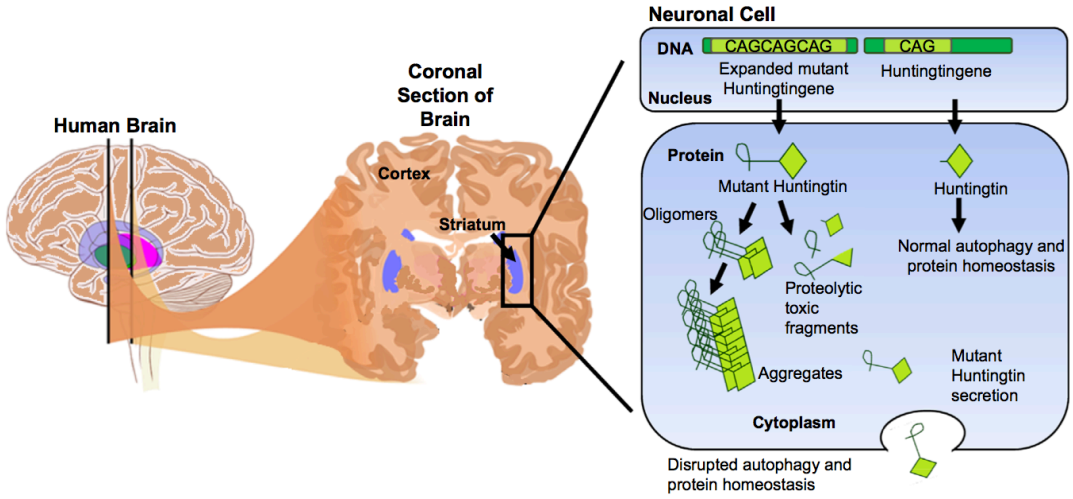

Figure 2.3 - Schematic drawing of Huntington's disease pathology. Source:.(Zhang et al. 2015)

Oligomers are considered by some authors to have a more important role in pathology than monomers or inclusions. Cells with inclusions were found in some studies to survive longer in affected tissues and inclusions are present in some disease models which do not develop symptoms (Bano et al. 2011; Hoffner & Djian 2014).

Nevertheless, inclusions can recruit, sequester and impair many functional proteins, namely those that naturally interact with native Htt, and those associated with proteostasis pathways such as proteasome and autophagy associated proteins. This effect could have, to a limited extent (see above), a negative effect on the normal function of the Htt expressed by the unaffected allele. The protein context of the poly-Q domain can lead to the preferential recruitment of different proteins which could explain the also preferential vulnerability of different brain regions in different trinucleotide repeat disorders (Bano et al. 2011; Hoffner & Djian 2014).

2.4.3 From neuropathology to clinical symptoms

The Huntingtin protein is present in higher concentrations in the brain (as well as in testes and in medium concentrations in the liver, heart and lungs) (Walker 2007). Medium spiny neurons in particular, that also present higher repeat length instability, are selectively affected in HD, especially those projecting to the external globus pallidum (GPe; figure 2.4) which are preferentially involved in the indirect pathway of the basal ganglia-thalamocortical circuitry (Walker 2007; Pinto et al. 2013; Kennedy et al. 2003; Wichmann & Delong 1996; Deng et al. 2004).

Degeneration of striatal neurons projecting to the GPe leads to disinhibition of the GPe, followed by increased inhibition of the subthalamic nucleus (STN). This leads to a reduced thalamo-cortical inhibition by the GPi (internal globus pallidum), causing facilitation of cortical motor areas and subsequent development of hyperkinetic symptoms such as involuntary movements (e.g. chorea) (Purves 2004; Wichmann & Delong 1996).

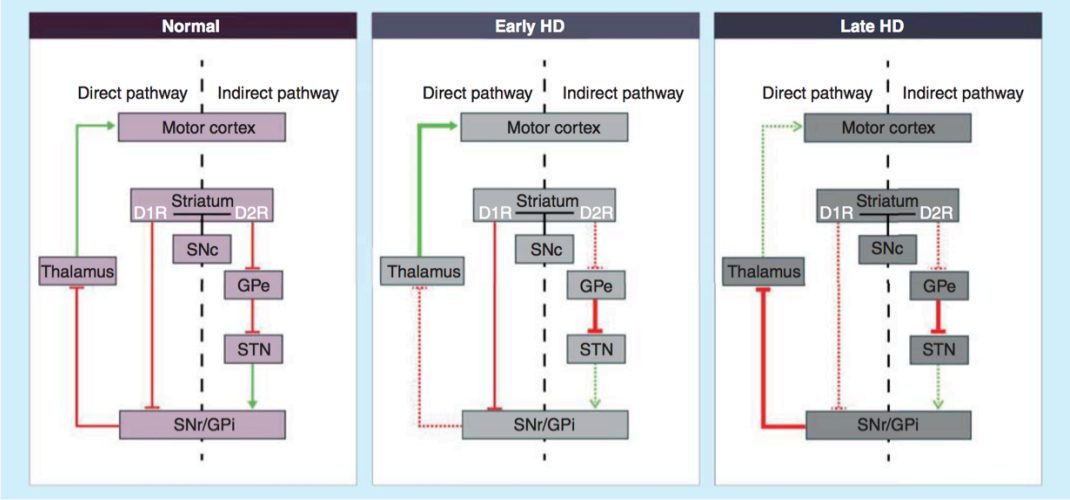

Figure 2.4- Schematic drawing of the direct and indirect pathways in early and late Huntington's disease. Source: (Schwab et al. 2015).

The intranuclear inclusions formed by Htt aggregates are hallmarks of HD pathology found in post-mortem samples of both prodromal (pre-manifest with early symptoms) and manifest HD brains. Other forms of inclusions such

as cytoplasmatic and axonal inclusions are also reported (Ross & Tabrizi 2011).

The progressive marked cell loss and atrophy of the striatum, which precedes the motor symptom onset of the disease, is accompanied by damage to a lesser degree to other brain structures such as the cerebral cortex, thalamus, hypothalamus and hippocampus (Kumar et al. 2015; Ross & Tabrizi 2011; Rangel-barajas et al. 2015).

Cell dysfunction even prior to cell loss in the affected brain regions is thought to be associated with the cognitive, behavioural and psychiatric symptoms that can start several years before the motor onset of the disease. The known role of the affected striatal cells in non-motor functions and complex behaviours could contribute to this phenotype (Paulsen & Long 2014).

2.4.4 Other mechanisms in HD pathology

Non cell autonomous mechanisms such as inflammation and excitotoxicity mediated by glial cells carrying the mutation are also reported to contribute to the HD pathology (Ross & Tabrizi 2011). Astrocytes in particular, contribute to HD pathology through astrogliosis and dysfunctional regulation of glutamate and potassium extra-cellular levels (Pekny et al. 2015).

Given the mutant protein is ubiquitously expressed, dysfunction of peripheral tissues should also be considered as they could contribute to symptoms such as weight loss and metabolic disturbance, which are characteristic of HD. Peripheral tissues may as well contribute to central nervous system pathology due to pro-inflammatory circulating cytokines (Carroll et al. 2015).

2.5 Disease Models

There is a great variety of HD disease models in different model cells (e.g. iPSC) or organisms such as mouse, Drosophila, C.elegans, pig, sheep and non-human primates among others (Harvey et al. 2011; Liu et al. 2015; Chang et al. 2015). Different models are better suited to different research questions.

2.5.1 Overview

Briefly, there are transgenic models in which only a portion (i.e. exon 1) or the full-length human mutant gene is expressed by exogenous promoters and there are knock in and conditional (e.g. only expressed in specific cell types) knock in models in which an homologue Htt gene is replaced by expanded CAG repeats or human mutant Htt exon 1 (Harvey et al. 2011; Chang et al. 2015).

In mice, while some transgenic models, for instance R6/2 or N171-82Q, produce strong early onset HD like phenotypes, they might not be the most faithful models, as they present phenotypes that differ from human pathology for instance comparatively reduced apoptotic neuron loss or weight gain (HD is associated with weight loss) and may have abnormally elevated expression levels. These concerns are also present for full-length mutant HTT expressing mice models such as YAC and BAC models (Chang et al. 2015).

Knock-in mice models, although having milder phenotypes, seem to more closely recapitulate HD pathology, presenting a progressive and late-onset phenotype. These models could therefore present a plausible option when studying underlying mechanisms of HD pathology such as preferential accumulation of mutant Htt and neuron loss in the striatum (Dragileva et al. 2009; Chang et al. 2015).

2.5.2 Hdh CAG knock-in mouse model

An important and accurate genetic HD disease model to gain insight on HD pathology in mammals is the huntingtin mouse homologue (Hdh) CAG knock in model (Dragileva et al. 2009). This model is particularly relevant for the investigation of HTT CAG instability in HD, for which it is extensively characterized in the literature (Lee et al. 2011; Wheeler et al. 2002). Several studies have shown the model's potential as a predictor of genetic background influence in human HD pathology, making it the most suitable model available for the current proposed work on genetic modifiers of somatic repeat instability in HD pathology (Pinto et al. 2013; Huntington's & Consortium 2015; Swami et al. 2009; Lloret et al. 2006).

2.5.3 HD phenotypes

The Hdh CAG knock-in mouse model presents several constitutive CAG repeat length dependent HD like phenotypes.

Some of the most relevant phenotypes are progressive diffuse nuclear accumulation of mutant HTT in striatal neurons and late-onset neurodegeneration and gait deficits (significantly shortened stride and imprecise hind–fore paw placement) (Pinto et al. 2013).

The model has another important phenotype in terms of somatic and intergenerational CAG repeat length instability (Lee et al. 2011; Pinto et al. 2013), presented in figure 2.5, which is associated with human HD pathology (Swami et al. 2009). In terms of somatic CAG repeat instability, the striatum presents a particularly strong instability and tissue specific age dependent repeat length expansion (Lee et al. 2011).

Another tissue that presents somatic repeat instability, in correlation with the striatum, although with differences in dynamics, is the liver. The liver being more easily accessible and easier to process and analyse, could be an interesting phenotype to screen for differences in striatum somatic instability (Pinto et al. 2013; Lee et al. 2011).

The influence of genetic background on HD phenotypes, for instance in the case of different mouse strains, has been studied extensively in this

model, having predicted a candidate gene that was later confirmed as an almost genome wide significant genetic modifier of disease, MLH1 (Pinto et al. 2013; Huntington's & Consortium 2015).

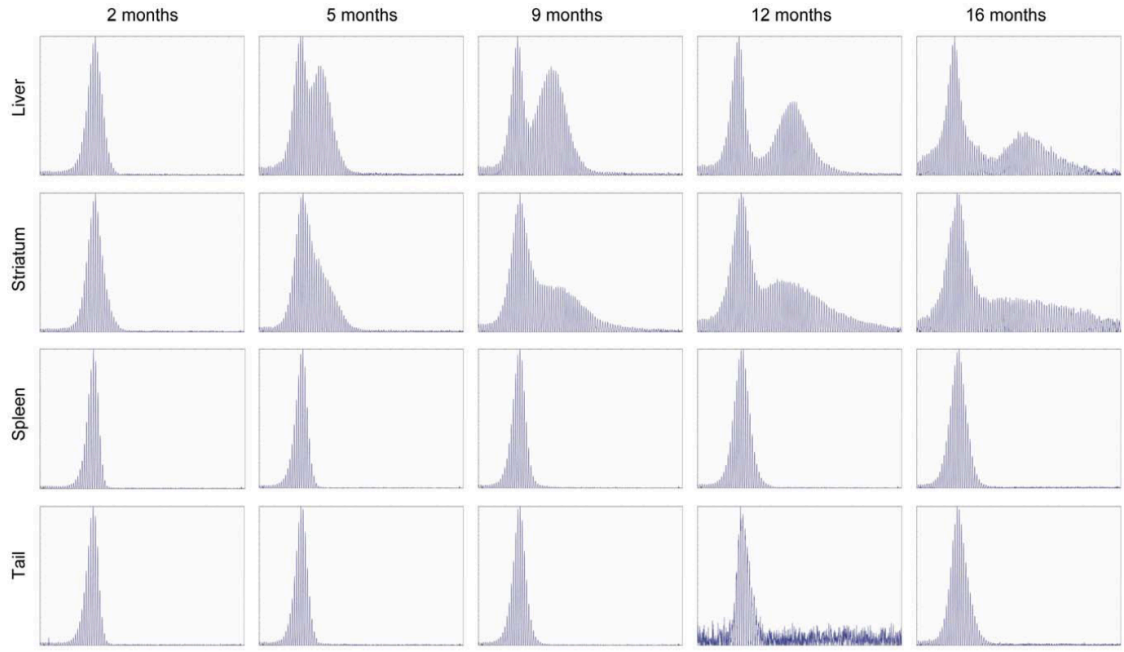

Figure 2.5 – Tissue specific age dependent CAG repeat length instability in the Hdh mouse model. Source (Lee et al. 2011)

2.5.4 CRISPR-CAS9 based changes to the model

Gene editing technology enables agile editing or disruption of (multiple) genes *in vivo* (Boettcher & McManus 2015), presenting an opportunity to expand the potential of the Hdh CAG KI model to dissect HD pathology mechanisms.

In vivo gene editing of candidate genetic disease modifiers will be performed in this mouse model to study their impact on the model's HD associated phenotypes, namely on striatum somatic repeat instability.

It is shown in the literature, that somatic instability in the liver, an organ more easily accessible and examined, correlates in this model with somatic instability in the striatum (Lee et al. 2011). *In vivo* gene editing of genetic modifiers in the liver might therefore enable an accurate screening of potential modifiers of somatic instability in the striatum.

2.6 Genetic Disease Modifiers

Currently there are no disease modifying therapies capable of delaying the onset or progression of HD, with medical care being limited to treatments that only help to alleviate some of the movement and psychiatric symptoms associated with the pathology (Kumar et al. 2015).

There is however, evidence that factors other than CAG repeat length can significantly affect disease age at onset, since for the typical 40 to 55 CAG repeat range associated with adult onset, the number of CAG repeats can only account for 56% of HD age at onset variability for motor symptoms (Gusella et al. 2014). The remaining variability leaves room for deviations of up to 20 or even 25 years from the mean age of onset predicted by repeat length (Gusella et al. 2014) and is 38% determined by genetic modifiers (Project & Wexler 2004), which can potentially be harnessed as new disease modifying therapeutic targets (figure 2.6).

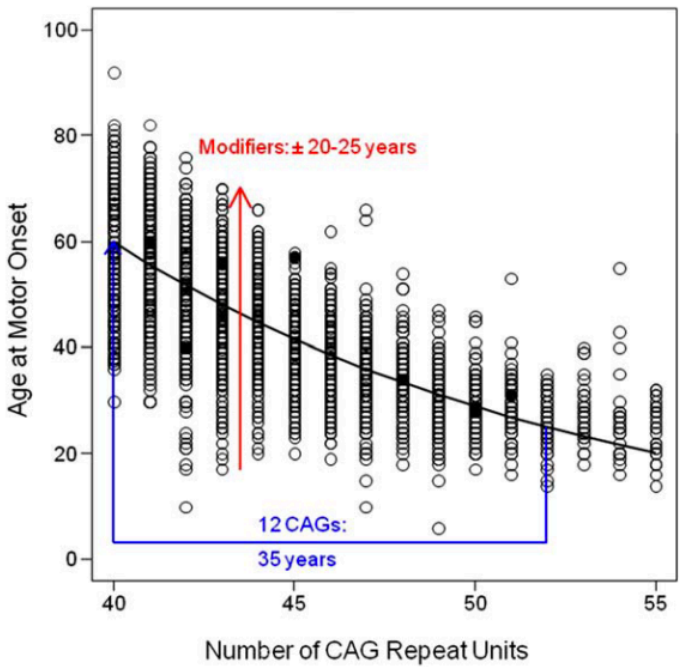

Figure 2.6 – CAG repeat length association with age at HD motor onset. Source (Kolodner 2015)

In the same way, new genetic modifiers might also influence other HD phenotypes, with a lower correlation to CAG repeat length indicating a

potentially greater influence by other factors. While cognitive age at onset correlates to CAG repeat length in a similar way to motor onset, the correlation is weaker for psychiatric symptoms (diagnosis subjectivity could partly affect the comparison). There is as well a difference in strength of correlation for more measurable phenotypes such as in pathophysiology, with striatum neuropathology presenting a stronger correlation than cortical neuropathology (Gusella et al. 2014).

Genetic modifiers have therefore been the object of intense research by the scientific community as they offer the hope of disease modification, meaning an intervention capable of producing long term slowing of accumulative disability, which nature and evolution have already shown to be biologically possible, rather than the transient palliative relief of symptoms currently available (Kieburtz & Olanow 2015; Gusella et al. 2014).

These modifiers are also thought to be technically easier to search than environmental modifiers since they are limited to a finite genome that can be studied in a more uniform way, much as was the case with the original linkage studies and mapping of the CAG expansion mutation. Additionally, they could lead to therapeutic targets amenable to rational therapies generated by traditional drug development, as opposed to other promising emerging therapies based of HTT gene suppression strategies, which still face several technical difficulties (Gusella et al. 2014).

The pursuit of new genetic disease modifiers has been made through both candidate and unbiased genome wide approaches.

In candidate gene studies, despite the study of genes chosen based of pathology mechanism assumptions yielding a great number of potential candidates, these did not meet statistical significance when tested in bigger and more controlled datasets. The confounding effect of population stratification of polymorphisms was also found to be difficult to overcome in this type of studies. These studies did however indicate an absence of common significant new modifiers in the HTT locus itself (Gusella et al. 2014).

Unbiased genome wide approaches enabled by technological advances hold the promise of a faster and thorough screening of all potentially disease modifying genes while also overcoming candidate studies' limitations to

provide data-driven findings. Not presupposing a specific hypothesis for HD pathogenesis, this type of approach has also the potential to generate new insight on HD biology.

Two important genome wide approaches used in HD literature are genetic linkage and genome-wide association studies (GWAS).

Genetic linkage studies are based on the principle that fragments of the genome that are found closer to each other in a parental chromosome are less likely to be separated by genetic recombination during crossing over. Neighbouring fragments are therefore more likely to segregate together during meiosis, having a greater chance to be found together in the offspring than predicted by chance (Ngeow & Eng 2015).

GWAS examine a great number of polymorphisms (common genetic variants), uniformly distributed in the genome, in many different individuals, to search for polymorphisms associated with a given trait (for instance deviation from CAG repeat length predicted age at motor onset in HD). Typically GWAS focus on common SNPs (single nucleotide polymorphisms). If the study is adequately powered and statistically corrected, a strong association of a trait with a SNP indicates that a genetic modifier of the trait could be in genetic linkage with the SNP (or that the SNP itself could be a genetic modifier). Gene loci genetically close to the SNP would therefore be potentially associated with the trait. Results can be further analysed by pathway analysis in which preferential clustering of trait correlated SNPs in certain gene pathways and biological functions can give insight on mechanisms potentially associated with the trait (Ngeow & Eng 2015; Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium 2015).

Unbiased genome wide approaches have implicated promising new HD genetic modifiers not previously studied in HD literature and confirmed some genes and pathways already implicated in HD and trinucleotide repeat literature, supporting their potential as therapeutic targets.

Recently, the Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium published the results of a genome-wide association assay (GWAS) identifying loci harbouring genetic variations capable of modifying the age at motor onset of HD.

The study, powered by DNA samples of 4082 HD patients of european ancestry (selected based on polymorphic profiles, from a total of 7410 subjects, to reduce confounding effects due to differences in SNP frequency between different populations), yielded 2 statistically significant modifier loci, one in chromosome 15 presenting 2 independent effects (6.1 years of onset acceleration and 1.4 years of onset delay) and another in chromosome 8 that accelerates onset by 1.6 years. Pathway analysis indicated clustering of SNPs associated with motor age at onset modification in specific pathways and biological functions, in particular DNA handling and repair.

Involvement of the DNA repair pathways in HD was further supported by the presence of near genome-wide significance centered at *MLH1*, which participates in DNA mismatch repair (MMR) and had already been implicated in a genome wide linkage study in a HD mouse model (Pinto et al. 2013; Huntington's & Consortium 2015). Repeatedly implicated in unbiased screenings of HD modifiers, *Mlh1* was show to modify somatic CAG repeat instability, which is associated with HD pathology and age at onset in human studies (Swami et al. 2009).

Mlh1 knockout in a B6.Hdh^{Q111} mice model of HD was shown to reduce somatic CAG repeat instability, namely in the striatum, and to strongly lower nuclear Htt immunostaining, a CAG length dependent sensitive marker of the pathologic process in these mice. *Mlh1* was thus implicated as a genetic enhancer of both somatic CAG repeat instability and early CAG length dependent pathology, indicating it could be exerting its effect in HD pathology through modification of somatic CAG repeat instability.

MLH1's effect on HD pathology through MMR or DNA repair mediated somatic repeat instability is consistent with data from other MMR participating genes such as *Msh2, Msh3 and Mlh3* (Wheeler et al. 2003; Dragileva et al. 2009; Pinto et al. 2013), and might be shared by some of the new in-human validated disease modifiers from the GWAS study. The implication of MMR through somatic repeat instability in other trinucleotide repeat expansion disorders (e.g. myotonic dystrophy and cerebellar ataxias (Foiry et al. 2006; Ezzatizadeh et al. 2014; Schmidt & Pearson 2015)) also points to a potentially

greater scope of mechanisms associated with the biology of HD genetic modifiers.

A greater insight into genetic modifiers of HD phenotypes and their mechanisms of action might thus contribute to the development of new rational disease modifying therapies.

2.6.1 MMR AND REPEAT INSTABILITY

DNA mismatch repair is pathway involved in different biological functions being implicated in maintenance of genomic integrity and stability (figure 2.7), as well as in class switch recombination, immunoglobulin somatic hypermutation and disease-associated trinucleotide repeat expansions. Disruption of some MMR genes (*MSH2*, *MSH6*, *MLH1* and *PMS2*) is also associated with cancer and the Lynch syndrome (hereditary non-polyposis colorectal carcinoma) (Sleana et al. 2008; Walsh 2015).

Its canonical activity is the recognition and correction of incorrectly paired (that is mismatched) nucleotides in DNA. During this process, fragments of the mismatched DNA strand are excised and resynthesized, in guidance with the homologous strand (Fishel et al. 2007).

Mismatched bases can arise during DNA replication (not expected post-mitotic in neurons), to spontaneous deamination of nucleotides (G-T and G-U transitions) or to other types of damage. Mismatchs can also occur during repair synthesis (Fishel et al. 2007).

In eukaryotes, DNA mismatches are recognized by heterodimeric protein complexes MutS-alpha (MSH2-MSH6) and MutS-beta (MSH2-MSH3). MutS-alpha is the most important complex, being required for base-base mismatches, however MSH2-MSH3 involved in repairing small insertions and deletions also has a role in repairing short CAG or CTG slip-outs. MSH2 implicated in both complexes is therefore necessary for MMR activity. (Fishel et al. 2007; Kolodner 2015) MutL-alpha (MLH1-PMS2) and MutL-gama (MLH1-MLH3) act downstream of MutS complexes promoting mismatch resolution, with MLH1 being a required component of MutL complexes (Schmidt & Pearson 2015).

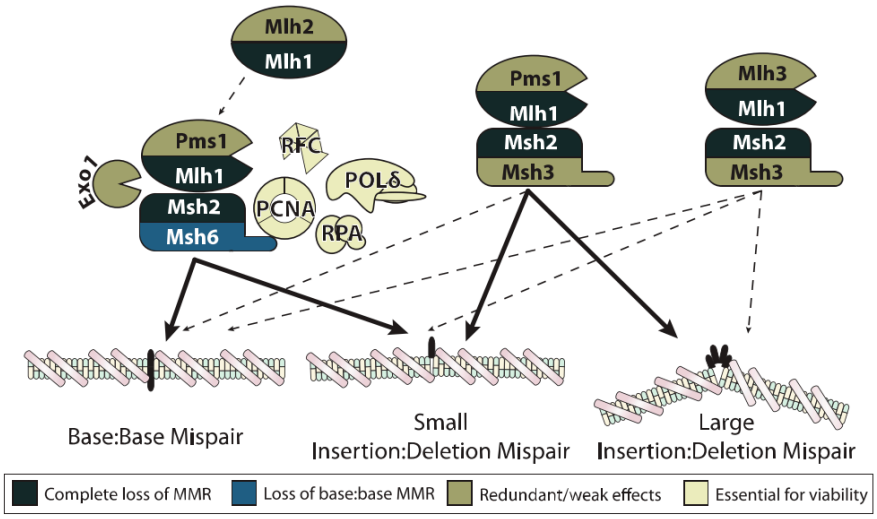

Figure 2.7 - Depiction of a model of MMR machinery. Source (Kolodner 2015)

MMR complexes are present in the brain and implicated in somatic trinucleotide repeat instability, having expression levels associated with the difference in CAG instability between the striatum and cortex of HD mice (Tomé et al. 2013; Mason et al. 2014; Schmidt & Pearson 2015).

Studies with HD mouse models knocked out for MMR genes strongly implicate them in somatic trinucleotide instability, with ablation of *Msh2*, *Msh3*, *Mlh1* or *Mlh3* preventing striatum somatic CAG instability. Some of these studies, namely for *Msh2, Msh3 and Mlh1* have also shown a reduction of HD early phenotypes in response to MMR gene ablation (Dragileva et al. 2009; Pinto et al. 2013; Schmidt & Pearson 2015).

It should be noted that there is ongoing discussion of non-canonical MMR pathways in the literature and that other DNA repair mechanisms are also implicated in trinucleotide disorders (Pinto et al. 2013; Schmidt & Pearson 2015; Usdin et al. 2015).

2.6.2 CANDIDATE GENETIC MODIFIERS OF CAG REPEAT INSTABILITY

The HD age of onset genetic modifier GWAS study suggested several genome wide and almost genome wide candidate modifiers of HD age of onset (Huntington's & Consortium 2015) that could be acting through somatic CAG repeat instability. From these, some of the most promising DNA repair relevant SNPs were again studied in a independent cohort of 1,462 subjects, comprising both HD patients and patients suffering from spinocerebellar ataxias (SCA) 1, 2, 3, 6, 7 and 17 which are other polyglutaminic disorders in which CAG somatic instability could also have an impact on age of onset variation.(Bettencourt et al. 2016) This study showed that a set of 22 DNA repair SNPs correlated with age of onset in this group of polyglutaminic disorders and that in particular 2 individual SNPs in *Fan1* and 1 in *PMS2* had a significant association with age of onset in the 8 disorders.

Modifiers of other forms of repetitive sequence instability, namely micro satellite instability (MSI), pertaining to repetitive sequences of 1–10 base pairs, could also be a source of potential candidate modifiers of CAG instability. FancJ, shown to prevent MSI in knock out mouse studies is one such example (Matsuzaki et al. 2015).

While an extensive literature of DNA repair genes with suggested potential for trinucleotide repeat instability modification exists from other in vivo and *in vitro* studies, the current thesis focused as a starting point on:

- Known modifiers of striatum somatic CAG instability from knock out studies in HD mouse models: *Mlh1* (Pinto et al. 2013), *Mlh3* (Pinto et al. 2013), *Msh2* (Wheeler et al. 2003) and *Msh3* (Dragileva et al. 2009)
- Candidate genes associated with DNA repair most significantly implicated in the GWAS study and shown to associate with age of onset in a broader set of polyglutamic disorders: *Fan1, Rrm2b, Ubr5, Mlh1, Msh3, Mlh3, Msh6, Ercc3, Pms1, Pms2 and Lig1* (Huntington's & Consortium 2015; Bettencourt et al. 2016)
- Candidate genes from MSI literature: *FancJ* (Matsuzaki et al. 2015)
- Other MMR genes: *Exo1* (Goellner et al. 2015; Iyer et al. 2015)

3. GENE EDITING TOOLS

3.1 CRISPR-CAS9

Gene editing, silencing and disruption has been key to further the understanding of gene function and biology.

The recent addition of a new gene editing technology CRISPR-CAS9 (clustered regularly interspaced short palindromic repeat; figure 3.1) to the biologist's toolbox has however brought significant breakthroughs as it simple (uses an RNA guided nuclease not requiring nuclease engineering as was needed using zinc finger and TALENs), fast (compared to the time consuming method to produce knock in and knock out homologous recombination based mouse models) and leads to strong, precise and lasting interventions (unlike RNA interference approaches that can have unspecific effects and only produce transient incomplete effects).

The CRISPR-CAS9 system was engineered from a bacterial immune response to virus. It uses RNA guided nucleases to target specific genome sequences, recognised by the guide RNA, where it produces a double stranded break.

The DNA repair machinery can repair double strand breaks through non-homologous end joining (NHEJ). This mechanism can introduce mutations, namely small insertions and deletions of a number of nucleotides non-multiple of 3, causing a shift in the reading frame (frameshift), that due to the triplet nature of the genetic code leads downstream triplets to code for different aminoacids and to the appearance of early stop codons. Frameshift inducing

mutations near the start of the coding sequence can therefore induce protein loss of function and functional gene disruption. This enduring change in the genome can thus lead to loss of function phenotypes that can give insight on gene function.

In the presence of a homologous template DNA sequence, for instance introduced along with nuclease, the DNA repair machinery of cells, though less frequently, can repair the double strand break by homologous recombination, using the exogenous nucleotide sequence as a template. This mechanism results in a change to the genome that now replicates the template sequence that may differ in some bases from the original genomic sequence. Thus, RNA guided cleavage by CRISPR-CAS9 allows editing of the genome sequence by homologous recombination with exogenous template sequences.

The CRISPR-CAS9 system can therefore be used to dissect gene function and biological phenomena by either gene disruption or editing. Further improvements to the system also enable epigenetic changes and transient gene inhibition through derived systems (Moore 2015; Teimourian & Abdollahzadeh 2014; Barrangou et al. 2015; Singh et al. 2014; Dow 2015).

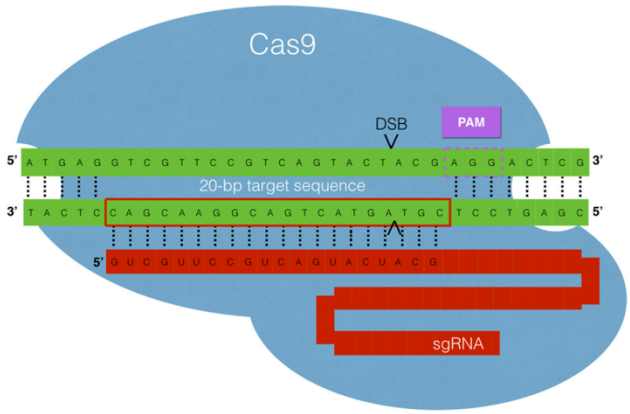

Figure 3.1 – CRISPR-CAS9 system. Source (LaFountaine et al. 2015)

In order to disrupt a specific gene using CRISPR-CAS9, a single guide RNA (sgRNA) must be designed to generate site-specific DNA breaks in its locus. Guide RNA targets are about 20 nucleotides long and must be preceded by a NGG PAM sequence required by the nuclease. As the NGG PAM sequence is relatively frequent, targets can be found in most parts of the genome.

Libraries of sgRNA have been created for most exons in the human genome. Specific applications however may require optimization as the design of the sgRNA and targeted site in the locus affect editing efficiency (Doench et al. 2014). An sgRNA designed for high specificity is considered to have very limited chance of producing non-specific breaks, making it safe for most applications (O'Geen et al. 2015).

3.2 *IN VIVO* DELIVERY

The efficiency of CRISPR-CAS9 gene editing is dependent on nuclease and sgRNA access the target DNA, for most applications within the nucleus of cells. Strategies are thus required to effectively deliver these elements to cells.

Ex-vivo or *in vitro* systems are compatible with the use of many technics such as lipofection, electroporation or heat shock for competent cells to help nuclease and sgRNA permeation of cells. Delivery efficiency however will depend on several factors and in most cases require optimization.

For efficient specific *in vivo* delivery, the challenge increases substantially with less viable options to overcome more complex obstacles in many domains (immune reaction, enzymatic degradation, barriers such as the brain blood barrier, diffusion through considerable volumes, ethical and safety issues, to name a few).

Nevertheless, gene-editing tools have a great potential to unlock new insight of physiological pathological processes. Their potential *in vivo* applications also extend to medical therapy although it is only now reaching clinical trials (Crowley & Rice 2015; Ran et al. 2015; LaFountaine et al. 2015; Eguchi et al. 2015; Schmidt & Grimm 2015; Bryant et al. 2013).

3.2.1 Viral and non viral delivery strategies

Strategies for CRISPR-CAS9 *in vivo* delivery fall in two categories: viral and non viral.

In either type of strategy, the ultimate aim is to deliver a RNA guided nuclease (Cas9) as well as a sgRNA to guide it. This implies direct delivery of these elements or delivery of genetic material that codes for these elements. Another option for animal models is to use a CAS9 constitutive or conditional knock in animal to which only the sgRNA or genetic material coding for it must be added (Ran et al. 2015; Swiech et al. 2015; Schmidt & Grimm 2015).

Among the non-viral strategies there are many options according to the application. One possibly promising strategy is to use lipid formulations developed for *in vivo* delivery of siRNA to directly deliver sgRNAs through tail vein injection. Invivofectamine 3.0, a commercially available product of this nature, is particularly interesting for the current project aims as it is designed for high efficacy accumulation in the liver and has shown successful results for siRNA in peer reviewed literature (Eguchi et al. 2015).

Another type of strategy is to use viral vectors. In this case the considerable size of the most commonly used nuclease usually requires using different vectors for the nuclease and the sgRNA or using a shorter form of nuclease, or delivering the vector to as nuclease expressing model (Ran et al. 2015; Schmidt & Grimm 2015; Swiech et al. 2015).

3.2.2 AAV-mediated delivery

Among viral vector systems, one particularly promising option for CRISPR delivery is the adeno-associated virus (AAV) mediated gene delivery system.

The AAV is a small, non-enveloped virus that packages a single-stranded linear DNA genome (approximately 5,000 bp). The AAVs were adapted through modification to be used as gene transfer vectors.

Approved for clinical applications by the European Commission in November 2012 (Bryant et al. 2013), the AAV gene delivery system is characterized by its safety profile and efficacy in transducing both dividing and non dividing cells in different tissues and species (namely in rodents and humans) (Samulski & Muzyczka 2014; Calcedo et al. 2009).

There are different AAV serotypes with different tropisms and efficiencies in transducing different tissues. AAV's tropism to different tissues according to serotype can therefore be harnessed to increase delivery efficiency to specific targets, namely for the purpose of this monograph the liver and the brain (Asokan et al. 2012; Samulski & Muzyczka 2014).

For liver targeted strategies, AAV8 shows the most promise having specific tropism for the liver when delivered by tail vein injection (Asokan et al. 2012).

For brain targeted strategies, AAV9 presents a promising option as it is reported to have tropism and to be able, to some extent, to cross the blood brain barrier. This makes it possible to deliver AAV9 to the brain and striatum not only by sterotaxy (direct intracranial injection to specific tridimensional coordinates corresponding to the brain structures of interest in a mouse brain atlas (Cetin et al. 2006)) but possibly, in a less specific way (also has tropism to the heart and muscles), by tail vein injection. It is additionally reported to present a good brain distribution when administered to cerebrospinal fluid (Asokan et al. 2012).

Other alternatives for brain delivery are AAVrh.10, reported to be at least as efficient as AAV9, and AAV5 and AAV1 that presented a more efficient transduction of striatum neurons than AAV8 (Asokan et al. 2012).

Another very promising alternative for brain delivery more recently developed, AAV-PHP.B, is a recombinant AAV variant with 40x better transduction efficiency relative to AAV9, that can more easily cross the blood brain barrier when delivered by tail vein injection (Deverman et al. 2016).

Though not associated with disease or tissue toxicity, AAV can induce a relatively mild innate and adaptive immune response. This can limit effective gene transfer due to induced immunity when there are multiple exposures to a given AAV serotype. Therefore serotype may be an important feature of AAVs to consider when there are multiple target genes with sgRNAs in different vectors administered at different times.

More than implying constraints to multiple treatments with AAV, this situation has important consequences for delivery efficiency and choice of AAV serotype. Immunoprivileged tissues, namely the brain are more ameanable, to some extent, to repeated AAV transduction. Studies indicate however that transient use of immunosuppressors may reduce the impact of immune response on AAV transduction efficiency which could be important namely in liver targeted multiple CRISPR delivery.(Samulski & Muzyczka 2014; Calcedo et al. 2009)

3.3 MUTATION HIT-RATE QUANTIFICATION

Given genome editing does not target all loci with similar efficiencies, mutation hit-rate should be assessed both for efficacy and efficiency when delivering engineered nucleases.(Vouillot et al. 2015)

3.3.1 DNA sequencing

Genome editing induced mutations in transfected cells can be detected by sequencing (through Sanger or next-generation technologies) of representative polymerase chain reaction (PCR) products.

While sanger sequencing is the historic golden standard to sequence and detect specific mutations, it has limitations when sequencing DNA from heterogeneously mutated cell populations (Arsenic et al. 2015), as is the case of cells mutated through CRISPR induced non-homologous end joining (Doench et al. 2014). For this type of samples, next generation sequencing (NGS) of a PCR enriched amplicon can be more sensitive and informative enabling the quantification of the relative proportion of each type of mutation (Arsenic et al. 2015; Doench et al. 2014).

3.3.2 Enzyme mismatch cleavage: T7 assay

DNA sequencing is expensive and time consuming, making it possibly unsuitable for the preliminary screening of genome editing constructs. There is thus a need for other specific methods of mutation detection, capable of detecting a mutated allele in a background of wild-type (WT) alleles. While there are other alternatives (such as high-resolution melting curve analysis, denaturing high-performance liquid chromatography (DHPLC), capillary electrophoresis-based single strand conformation polymorphism (CE-SSCP), denaturing gradient gel electrophoresis (DGGE) and chemical cleavage), the present thesis will focus on enzyme mismatch cleavage, which stands out as a popular, simple and sensitive option.(Vouillot et al. 2015; Huang et al. 2012)

Enzyme mismatch cleavage (EMC) is the one of the commonly used mutation hit-rate quantification methods (figure 3.2). This method consists in analysing DNA heteroduplexes, formed by denaturated WT and potentially mutant DNA PCR products when mixed and slowly cooled down (the mutation leads to a mismatch between hybridizing DNA chains from the two PCR

products). Heteroduplexes are then digested by specific enzymes that cleave heteroduplex DNA at mismatches and extrahelical loops formed by single or multiple nucleotides, where DNA structure differs.(Gohlke et al. 1994; Vouillot et al. 2015) The digestion products are separated by standard gel electrophoresis and bands of DNA separated by molecular weight detected in the gel through the presence of a DNA intercalating dye. Non-mutated DNA samples will only lead to one specific band (no cleavage due to heteroduplex mismatch) corresponding to the size of the uncleaved heteroduplex. Mutated DNA samples though, will lead to cleavage of part of the set of heteroduplexes that combine both mutated DNA and non-mutated DNA. Under these circumstances the band corresponding to uncleaved heteroduplexes will be followed by two other bands of smaller molecular size, corresponding to cleaved DNA heteroduplexes. (Vouillot et al. 2015)

Bacteriophage resolvases, namely T7E1, are the most commonly used mismatch cleavage enzymes in this assay. (Vouillot et al. 2015; Freeman et al. 2013) These enzymes recognise and cleave polymorphic structures in dsDNA having a preferential activity on mismatched base pairs (although they can also, to a lesser extend, act on Watson-Crick base pairs, cleaving homoduplex DNA). The sequence, the number of mismatched nucleotides and the flanking sequences affect heteroduplex sctructure therefore influencing T7E1 cleavage efficiency, which is higher for deletions than for single base mutations. (Vouillot et al. 2015)

The T7 assay, an enzyme mismatch cleavage assay using the T7E1 resolvase, can have a high sensitivity being for instance able to detect a 20-bp deletion mutation in a 5% mutant/WT DNA ratio.(Vouillot et al. 2015) For the purpose of mutation screening, namely within the context of deletion mutations, the more versatile T7E1 outperforms its main enzyme mismatch cleavage assay alternative, Surveyor, being an arguably better option.(Vouillot et al. 2015; Huang et al. 2012)

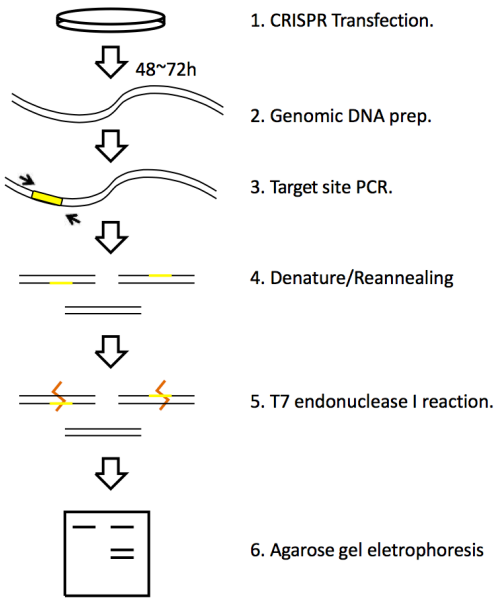

Figure 3.2– Enzyme mismatch cleavage Source:(Anon n.d.)

4. METHODS

4.1 BASIC TECHNIQUES

4.1.1 DNA extraction

4.1.1.1 Direct lysis with detergent and protease k (van der Burg et al 2011)

This method was used for the quick extraction of DNA from samples with low cell number, in the context of this thesis, from antibiotic or FACS selected NIH/3T3 cell samples. Cells were collected either from a supernatant with dead cells by pelleting or from attached cells by detaching and pelleting them. Pellets were washed in PBS (resuspended and repelleted) and treated with 100µL of lysis buffer (10mM TrisHCl, 50mM NaCl, 6.25mM MgCl2, 0.045% NP4O, 0.45% Tween 20; as described in van der Burg et al 2011) and 5µL of protease K (bioline; 1mg/mL final concentration). Samples were transferred to strips of PCR tubes and incubated for 1h30 at 56°C. Protease K was then heat inactivated (15minutes incubation at 90°C).

4.1.1.2 DNA extraction from mouse tissues

DNA was extracted from mouse tissues, in the scope of this thesis, the brain (striatum and cerebellum; left hemisphere), tail and liver (3 to 4mm samples), using a spin-column based nucleic acid purification kit (DNeasy Blood and Tissue kit; QIAGEN). Extracted DNA was eluted in 100µL of EB buffer.

4.1.2 DNA quantification

4.1.2.1 Quantification by UV-Vis Absorbance: nanodrop

This method was used for routine quantification of DNA from tissue extractions and PCR products. DNA was quantified based on sample absorbance at 260nm using a nanodrop equipment. The DNA purity was determined by the 260nm/280nm absorbance ratio (≥ 1.8). Contaminated samples were re quantified by a fluorimetric method (qubit).

4.1.2.2 Quantification by Fluorimetric assay: Qubit

This method was used to quantify DNA extracted by 'direct lysis with detergent and protease k (van der Burg et al 2011)'. A qubit high sensitivity dsDNA detection kit (ThermoFisher Scientific) was used following manufacturer instructions.

Fluorimetric DNA quantification assays are based on the detection of fluorescence from dyes that fluoresce when binding specifically to a type of biomolecule of interest, for the purpose of this thesis, double stranded DNA. This method is more sensitive, being a good option for samples with low DNA concentration and is not as affected by common contaminants, such as salts, free nucleotides, solvents, detergents or proteins, making it very useful for quantification of non-purified samples. Fluorimetric quantification however requires the preparation of sample dilutions and their mixing with fluorescent dye solutions, which makes it less practical for routine DNA quantification applications, particularly for purified samples that can be directly quantified by their absorbance at 260nm, using a nanodrop.

4.1.3 Agarose Gel Eletrophoresis

Agarose gel electrophoresis was used to separate DNA products of different molecular weight. Gels of 0.8%-1% agarose in 1xTBE were prepared, by mixing agarose with TBE, microwaving the mix in a standard microwave until melted without lingering grains of agarose, letting the mix cool to about 60°C, adding a fluorescent DNA intercalating agent (gel red; Biotium), mixing until visually homogeneous and pouring to set in an appropriate cast. Small gels, of approximately 50mL, were run in mini-gel tanks and bigger gels of 100mL in midi-gel tanks, in both cases being immersed in enough 1xTBE to

cover the tank's electrodes. Samples were pre-mixed with 6x Orange G loading dye before being loaded on gels, upon which an electric field (50 to 120V) was induced forcing negatively charged DNA molecules to migrate through the gel towards the tank's cathode. Molecules of low molecular weight migrate faster enabling a separation of molecules by molecular weight that becomes more apparent as the run progresses. Orange G loading dye migrates at approximately 50 bp giving a visual indication of the extent of the gel already run, that can be used to estimate if the desired resolution between the expected products has been achieved, at which point the gel electrophoresis can be stopped. DNA products bound by gel red can be seen as bands under UV light and their molecular weight estimated by comparison with products of know molecular weight run on the same gel. Gels were visualised under UV light and documented.

4.1.4 Gel purification

This method was used to purify PCR products by molecular weight from contaminated samples. Samples were loaded in a 0.8% agarose gel and separated by electrophoresis. The gel was periodically checked under UV to check if the band at the desired molecular weight was already adequately separated from contaminant bands of other molecular weights. Once good resolution was achieved, the desired bands were visualized under high wavelength UV light (362nm), excised from the gel using a sterile scalpel blade and transferred to 1.5mL tubes. A kit was used for DNA gel extraction (QIAquick gel extraction kit) and purification (QIAquick PCR purification kit). The resulting purified DNA was stored at -20°C.

For purification of sample volumes up to 75µL (3x a 25µL PCR reaction), loading was performed on wider wells.

4.1.5 Polymerase Chain Reaction (PCR)

DNA was amplified by polymerase chain reaction using Phusion High-Fidelity DNA Polymerase (Thermo Scientific). A high fidelity polymerase was chosen to avoid the accumulation of mutations during amplification, as this would bias both the T7 and sequencing assays. A negative control without DNA

template was used to check for DNA contaminants in all PCR reactions performed.

Table 4.1 General PCR reaction conditions. PCR components were combined according to this table and incubated for 1minute at 98°C, they were then submitted to 35 cycles of denaturation for 10 seconds at 98°C, annealing at a assay specific melting temperature (Tm) for 30 seconds and elongation at 72°C for 30seconds. The PCR reaction was finally incubated for an extra 10minutes at 72°C and cooled to 4°C.

Component	Amount
dH20	14.5 µL
Phusion HF Buffer (5x)	5 µL
dNTPs (10mM)	0.5 µL
DMSO	0.75 µL
Forward primer (5µM)	1.5 µL
Reverse primer (5µM)	1.5 µL
Phusion polymerase (2U/µL)	0.25 µL
DNA	1 µL
Total	25 µL

In the context of this thesis, PCR was mainly used to amplify specific genomic regions targeted by sgRNA guides. These amplified regions were further analysed for detection, quantification and characterization of induced mutations.

4.1.6 Plasmid preparation

Plasmid was prepared by growth of plasmid carrying bacteria, plasmid DNA extraction and DNA purification.

4.1.6.1 Small scale plasmid preparation: Minipreparation

This method was used to extract low quantities of plasmid DNA for validation of correctly transformed bacterial clones during molecular cloning. It was also used to produce small quantities of plasmid DNA for *in vitro* transfection experiments.

A volume of 5mL of antibiotic supplemented liquid LB (100 µg/mL of ampycilin) was inoculated with bacteria from glycerol-stocks or single colonies (grown post-transformation in agar plates). Cultures were grown overnight at 37°C under agitation, after which cells were pelleted by centrifugation (6800g for 10minutes at room temperature). The supernatant was discarded and plasmid DNA was extracted using a QIAprep Spin Miniprep kit (Qiagen) following the manufacturer instructions.

4.1.6.2 Larger scale plasmid preparation: Maxipreparation

This method was used to produce large quantities of plasmid DNA to submit for AAV production at the Gene Transfer Vector Core facility (Schepens Eye Research Institute and Massachusetts Eye and Ear Infirmary). It was also used to produce large quantities of constructs commonly used as controls in *in vitro* transfection experiments.

A volume of 5mL of antibiotic supplemented liquid LB was inoculated with bacteria from glycerol-stocks and grown overnight at 37°C under agitation. A volume of 200µL of this starting culture was then used to inoculate 200mL of antibiotic supplemented liquid LB (100µg/mL ampicillin). The new culture was grown overnight under agitation. Cells were pelleted by centrifugation (6800g for 30minutes at room temperature) and the supernatant discarded. DNA was extracted from the pellet using a Qiagen endofree plasmid maxi kit (cat12362), producing yields of approximately 500µg of plasmid DNA.

4.1.7 Cell Culture

In vitro validation of sgRNA guides designed for *in vivo* mouse gene editing was performed using a mouse derived cell line, the murine fibroblast NIH/3T3 cell line (ATCC, CRL-1658TM).

NIH/3T3 cells were cultured in Dulbecco's modified Eagle's Medium (DMEM, Sigma) supplemented with 10% fetal bovine serum (FBS, Thermo Fisher Scientific), and 1% penicillin/streptomycin (Gibco, Thermo Fisher Scientific). Cells were maintained at 37°C and 5% CO2 in a humidified incubator and sub-cultured upon reaching 80 to 90% confluence. Sub-culture was performed by washing plated cells with DPBS, detaching them at 37°C with TrypLE (Gibco, Thermo Fisher Scientific), quenching the enzymatic reaction with complete medium, collecting and pelleting cells (5min at 800RPM) and finally resuspending them in fresh medium and plating them at the desired cell density. Cell density of cell suspensions was assessed using a Scepter 2.0 Cell Counter (60µm tips, Millipore Sigma).

For long term storage of modified NIH/3T3 cells, cryovials were prepared by washing cells with DPBS, detaching them with TrypLE (as in the subculture technique), quenching enzymes with complete medium, pelleting cells and then resuspending them in 10% DMSO, 20% FBS, DMEM medium. Cryovials were then frozen in a -80°C freezer at a cooling rate close to -1°C/minute using a Mr. Frosty™ Freezing Container (Thermo Scientific). Frozen vials were later transferred to liquid nitrogen for permanent storage.

Long term stored cells were revived when required by thawing cryovials for 2minutes in a water bath at 37°C, diluting cells in 5mL pre-warmed full DMEM medium, pelleting cells, discarding the DMSO containing supernatant, resuspending cells in medium and platting them in the desired cell culture format.

4.1.8 Protein extraction

Mouse tissue (brain and spleen) and cell pellets were homogenized in 50 to 200µL of cold RIPA Buffer (Cat#BP-115, Boston Bioproducts) supplemented with EDTA and proteinase inhibitors (Halt Protease Inhibitor cocktail 100x ; #78429; Thermo Scientific). Tissue samples were homogenized

using a tissue grinder and cell pellets by repeatedly pipetting up and down with a P200 micropipette. Samples were then sonicated on ice for 10 seconds (sonicated twice for tissue samples) and centrifuged at 15,000RPM for 30min at 4°C. The protein lysate supernatant was collected.

4.1.9 Protein quantification: BCA

Protein in lysates was quantified by Pierce BCA Protein Assay (Thermo scientific)

4.1.10 Western blot

Equal amounts of protein from each lysate sample were diluted with RIPA buffer to the same volume and denatured in 25% loading buffer (NP0007, NuPAGE LDS sample buffer) and 10% reducing agent (#NP0009) at 70°C. Samples were loaded in precast NuPAGE SDS-PAGE gels and run together with a molecular weight marker (Precision Plus Dual color Standard; Biorad) for 20 minutes at 80V and then 100V at 4°C until the desired molecular weight resolution was achieved. Choice of gel percentage and running buffer was adjusted according to the protein being assayed so as to have more resolution in its molecular weight range. Proteins were wet transferred at 4°C, soaked in transfer buffer (10% methanol, 0.025% SDS, 10% transfer buffer 10x(#BP-190) in cold ultrapure water) for 70 minutes at 100volts, using a transfer system, to 0.45um pore size nitrocellulose membranes (Biorad).

Membranes were blocked for 2hours with 5%milk (#M0841; Lab scientific), 0.1% tween TBS (pH7.4), incubated with primary antibody in 5% milk, 0.1% tween TBS overnight at 4°C or for 1hour at room temperature, washed 3 times with 0.1% tween TBS for 5minutes and incubated for 1 hour with secondary antibody (1:10,000 dilution, in 1% milk 0.1% tween TBS, of either anti-mouse-HRP, NA931VS, or anti-rabbit-HRP ,NA934VS). Membranes were washed 3 times with 0.1% tween TBS for 5minutes and incubated for 4minutes with ECL western blotting substrate (#32106 Pierce Thermo scientific) or PICO ECL (#34087, thermo scientific; for samples with a faint signal). Membranes were transferred to a sealed cassette and in a dark room, film (Amersham hyperfilm ECL high performance chemiluminescence, GE healthcare) was exposed and revealed for different exposure times.

Alpha-tubulin was used as a loading control to estimate the relative

protein amount loaded in different samples. The expression of this protein was used to correct the measured relative expression of the proteins being studied.

4.1.11 Western blot quantification

Blots were quantified using ImageJ (version 1.48; NIH). For an exposure in which the signal of bands was not saturated, band intensity was measured as the mean intensity of a rectangle of fixed area containing the band. Background was subtracted to these measurements by subtraction of the mean intensity of an equivalent rectangle of background. Band intensity measurements were further corrected for loading differences between wells based on tubulin expression, by dividing each band intensity by the band intensity of the corresponding tubulin band. Relative protein expression could then be more confidently compared between samples through their corrected band intensity levels.

4.2 CRISPR sgRNA DESIGN

Guide RNA's targeting known and candidate somatic CAG repeat instability modifiers were picked from published libraries (O. Shalem et al. 2014; Sanjana et al. 2014) or designed using validated algorithms (Doench et al. 2014; Doench et al. 2016).

In this thesis, guides were intended not to introduce a particular mutation in a well defined site within a gene, but rather to induce loss of function mutations, shifting the focus to maximum on-target efficiency and low off-target efficiency rather than distance to a desired target site.

Initially guides were selected from a published genome-scale CRISPR-Cas9 knockout (GeCKO) library validated for *in vitro* screening assays (Ophir Shalem et al. 2014), whose updated version 2 (Sanjana et al. 2014), included 6 sgRNA's per gene. The GeCKO library was subdivided into libraries A and B each with 3 sgRNA's per gene.

Guides were later designed using new published guide design algorithms, incorporated in the Broad Institute's sgRNA designer tool (Doench et al. 2014). This tool predicts on-target efficiency (ability to produce null alleles of the target gene) based on data from *in vitro* testing in cell lines and quantitative assessment by antibody staining and flow cytometry of 1,841 sgRNA's targeting 6 mouse and 3 human endogenous genes. (Doench et al. 2014).

Guide design was again adapted following the publication of updated algorithms (Doench et al. 2016), that incorporated more data, doubling the size of the sgRNA on-target activity dataset , and using a more effective modelling approach (Doench et al. 2016). The new version also ranks guides in terms of off-target efficiency using a model derived from a study in mammalian cells in which the impact of using guides with different types of mismatches and mutations was assessed on their ability to cut their original target sites. In this study a total of 27,897 sgRNAs were tested on the coding sequence of the human CD33. (Doench et al. 2016)

The guide nomenclature used in this thesis reflects the method used when designing the guide in the following manner:

[name of target gene]-[name of library/algorithm used]-[identifying number/rank]

Guides picked from the GeCKO library were named either "A" or "B" according to their sub-library and numbered as they were ordered. For example "Mlh1-A1"

. Guides designed using version 1 of the sgRNA designer tool were named "JD" (short for John Doench, the first author of the paper with the algorithm). Guides with high predicted on-target efficiency within the 5 to 65% extent of the gene's coding sequence (where indels causing frame shift mutations are more prone to result in loss of function inducing mutations(Doench et al. 2016)) were numbered according to the relative position of their target in the coding sequence of the gene of interest (guides designed for *Mlh1* are presented in the results as an example). For instance "Mlh1-JD4".

Finally guides designed using version 2 of the sgRNA designer tool were named "JD2" and separated from their identifying number by a dot. For example "Msh3-JD2.1". The identifying number in this case, is the picking order suggested by the sgRNA designer tool, which is based on the predicted combined ranking, considering on-target and off-target efficiency. The tool picks from the top ranking guides those that target within the 5 to 65% of the extension of the coding sequence. It also excludes overlapping guides, giving preference to guides targeting different regions of the gene for better chance of finding a highly efficient guide in terms of producing loss of function mutations.

Additional sgRNA's independently designed by a collaborator, Jacob Loup, from the MacDonald group, were named JL (short for Jacob Loup).

Guide specificity for its target site was confirmed by BLAT analysis against the GRCm38/mm10 assembly of the mouse genome (https://genome.ucsc.edu/cgi-bin/hgBlat).

4.3 Generation of CRISPR Constructs

4.3.1 Molecular cloning

Molecular cloning is a set of techniques to insert recombinant DNA, either synthesized or derived from different sources, into a replicating vehicle such as plasmids or viral vectors.

In order to clone DNA elements into pre-existing plasmids, both the element to be inserted and the accepting plasmid must be prepared for ligation.

4.3.2 Generating sgRNA oligo duplexes

DNA elements of interest to be inserted into the accepting plasmid should be double stranded (with or without overhanging sticky ends) and 5' prime phosphorylated.

Inserts with the mentioned properties can be derived from synthesized single stranded forward and reverse DNA sequences. 5' prime phosphorylation can be achieved by treatment with a polynucleotide kinase and oligo duplexes can be generated by annealing forward and reverse complementary oligos (Table 4.2).

In the scope of this thesis, sgRNA oligo duplexes were prepared from synthesized single stranded forward and reverse DNA sequences corresponding to the sgRNA sequence, with the addition of nucletides at their 5' and 3' ends (followed a *5'-CACCG N(20)* and a *5'-AAAC N(20) C* template for the forward and reverse oligos respectively) to produce stiky ends compatible with the digested vector (Cong et al. 2013).

Table 4.2 5’ prime phosphorylation and oligo duplex annealing conditions. Forward and reverse oligos were treated with T4 polynucleotide kinase (30minutes at 37°C) so as to phosphorylate 5'-hydroxyl terminus and enable subsequent ligation reactions. The 5’ phosphorylated oligos were then denatured (5minutes at 95°C) and slowly annealed (lowering of temperature by 6°C/min, down to 25°C) forming duplexes. The resulting 5’ prime phosphorylated oligo duplex is thus ready to be used in a ligation reaction with a previously linearized vector.

Component	Amount
dH2O	6.5 µL
10x T4 ligase buffer	1 µL
oligoForward [100uM]	1 µL
oligoReverse [100uM]	1 µL
T4 PNK	0.5 µL
total	10 µL

4.3.3 Plasmid digestion

A plasmid intended to receive an insert with a new DNA element can be digested and linearized using restriction enzymes specific for the insertion site.

Table 4.3 Plasmid Digestion conditions. pAAV and plentiCRISPRv2 plasmids were digested with BsmBI (NEB), according to the formulation above, for 1h15min at 55°C and heat treated for 20min at 80°C for enzyme inactivation; pX458 plasmids were digested with BbsI (NEB), for 1h15min at 37°C and heat treated for 20min at 65°C for enzyme inactivation. In both cases, water volume was adjusted so as to reach the total reaction volume. Digestion with these restriction enzymes generated stiky ends compatible with the oligo duplexes to be inserted.

Component	Amount
Restriction enzyme	10 U
DNA	1 µg
10x Buffer 2.1/3.1	5 µL
H20	made up to 50 µL
Total	50 µL

Digestion efficiency can be assessed by detection of digested products of different molecular weight compared with the undigested plasmid in a 1% agarose gel.

For better cloning efficiency, the 5' prime ends of linearized plasmids can be dephosphorylated by treatment with a phosphatase (treatment for 20minutes at 37°C with 0.6U FastAP per µg of digested plasmid). This treatment prevents self-ligation (or ligation with digestion released spacers) and subsequent re-circularization of digested vectors before the digestion reaction with the insert, increasing the efficiency of the desired ligation.

4.3.4 Ligation

The double stranded DNA elements of interest were ligated to the digested plasmids using T4 ligase.

Table 4.4 Ligation with T4 ligase reaction conditions. Linearized vectors and oligo duplexes to be inserted were incubated for 30 minutes at room temperature with T4 ligase The enzyme was subsequently heat inactivated for 10min at 65°C

Component	Amount
Digested vector [12.5ng/ µL]	4 µL (=50ng)
T4 ligase	1 µL
T4 ligase buffer (10x)	1 µL
sgRNA oligo duplex (1:100)	1 µL
dH2O	3 µL
total	10 µL

4.3.5 Transformation of competent cells

Transformation is the process by which foreign DNA is introduced into a cell. Re-circularised constructs can be transformed into host cells such as bacteria so as to be replicated and multiplied as cells divide and grow.

Stbl3 chemically competent E.coli were used for transformation in this thesis for having a high transformation efficiency and a lower homologous recombination frequency, contributing to a lower risk of undesired changes especially to unstable regions as is the case of long terminal repeats found in lentiviral plasmids, that are required for their efficient host genome integration.

These chemically competent cells were taken from the -80°C and thawed for 30 minutes on ice. They were then incubated with approximately 20ng of re-circularised construct for 30minutes on ice and heat-shocked by immersion in a 42°C water bath for 45seconds, followed by 2minutes of incubation on ice. The heat-shock treatment induces a transient state of cell membrane permeability that allows for the uptake of foreign DNA, namely the construct of interest. Cell are then incubated for 1hour at 37°C under gentle agitation, in a nutritive rich medium, S.O.C. (Super Optimal broth with Catabolite repression), so as for them to have time to recover and start expressing antibiotic resistance proteins encoded in the construct.

4.3.6 Selection of successfully transformed cells

Plasmids harboring antibiotic resistance inducing genetic material allow for positive selection of individual bacterial colonies carrying the resistance inducing constructs, by enabling them to survive and grow better relative to other colonies in the presence of antibiotic.

For this purpose, transformed cell suspension was spread (20μL per 90mm diameter plate) in pre-warmed LB-ampicillin agar plates and incubated overnight at 37°C, after which individual ampicillin resistant bacterial colonies could be picked and used to inoculate separate tubes of 5mL of LB with 100μg/mL of ampicillin (within the scope of this thesis all cloning vectors expressed ampicillin resistance).

As colonies grow from a single cell or a reduced number of cells, picking single colonies increases the odds of finding homogeneous bacterial

populations, presenting copies of the same transformed construct from their common ancestor.

4.3.7 Glycerol stocks for long term storage

The single colony derived, overnight expanded (at 37°C, under agitation), LB-ampicillin cultures yield cells that can be stored for long periods of time at -80°C in glycerol-stocks (750 µL of cell suspension in LB and 250 µL of glycerol).

4.3.8 Molecular cloning validation

Plasmid DNA was extracted using a QIAprepSpin miniprep kit, and tested for the presence of the desired modified construct by sanger sequencing, using a primer flanking the region where the new element is intended to be introduced (U6-forward for guide cloning).

4.4 *IN VITRO* VALIDATION OF CRISPR CONSTRUCTS

The CRISPR constructs developed in the previous step were tested *in vitro* in the NIH/3T3 mouse cell line. *In vitro* testing allowed for construct validation and selection of the most promising candidate sgRNA's to advance to *in vivo* validation.

The pipeline for *in vitro* validation of sgRNA guides was optimized so as to incorporate lessons learned from previously validated guides and to be able to move to a higher throughput when validating new candidate modifiers. This section of the thesis presents the *in vitro* validation pipeline followed and its adaptations.

4.4.1 Transfection

NIH/3T3 cells were seeded in a 24 well format at a density of 10k cells per cm^2 and cultured for 24 hours until 60% confluent. Transfection was performed with Lipofectamine 3000 according to manufacture's instructions (500ng of plasmid per well). A negative control, with non-transfected cells, and a positive control, transfected with a GFP-expressing plasmid (pmaxGFP) were used for each transfection. The positive control was used to assess transfection efficiency by microscopy at 48h. Samples transfected with fluorescent protein expressing constructs were also assessed by microscopy at 48h.

For Mlh1-A1, the first sgRNA to be tested, 2 different experiments with independent transfections were performed.

For some of the most recent experiments, transfection was performed in a 6 well format, yielding more transfected cells in less time.

4.4.2 Selection of transfected cells

Successfully transfect cells carrying the CRISPR construct were enriched by either FACS or puromycin selection. Changes induced by the construct are expected to only affect transfected cells. To be able to detect and quantify those changes when there is a low transfection rate, non-

transfected cells should be removed by selection as they will otherwise be quantified as non-modified and mask and dilute changes.

The first approach tested for the selection of transfected cells was FACS sorting as it had already been in lab validated in previous experiments.

4.4.2.1 FACS selection

For the selection of pX458 transfected cells 72 hours post transfection, cells were washed with DPBS, detached with TrypLE (as in the subculture technique), enzymes were quenched with complete medium and cells were pelleted and resuspended in complete medium. Cell suspensions were then filtered with a cell strainer to remove cell clumps. Cells were FACS sorted according to green fluorescence (pX458 derived constructs express GFP) and cells with a green fluorescence signal above background collected. FACS selected and non-FACS selected cells were seeded and cultured and sub-cultured until enough material could be collected for DNA and protein analysis.

4.4.2.2 From FACS to puromycin selection

At a later stage, when analyzing the *in vitro* validation pipeline, the selection of transfected cells step was identified as a key limiting factor due the logistics of scheduling shared equipment and the added costs of FACS sorting. In order to overcome the current limitations on throughput imposed by using FACS sorting, an alternative method for selection of transfected cells, puromycin selection, was tested.

Puromycin selection is an antibiotic selection method compatible with the pLentiCRISPRv2 vector described in the literature for *in vitro* CRISPR screening (Sanjana et al. 2014) . Cell selection is achieved through this kind of method by treating cells with a high enough dosage of antibiotic (in this case puromycin) such that cells successfully transfected with a resistance conferring plasmid survive and recover while non-transfected cells do not, rendering the surviving cell population greatly enriched in successfully transfected cells.

At this point, a protocol using puromycin selection in pLentiCRISPRv2 transfected NIH/3T3 cells was not yet established in the lab or described in the literature and the recommended minimum puromycin concentrations for this cell line ranged widely from 1 to 5µg/mL according to the vector used (Tamura et al. 1999; Whalen et al. 2005).

Preliminary tests were performed so as to find a suitable puromycin dose. The chosen approach was to treat non-transfected NIH/3T3 cells with different puromycin concentrations and to choose the lowest concentration capable of completely eliminating the cells.

So as to be able to compare the puromycin selection method with FACS sorting, a previously *in vitro* validated guide, Mlh1-A1 was cloned into pLentiCRISPRv2 and again *in vitro* validated using the new method to assess if the new pipeline could be a viable alternative.

4.4.2.3 Puromycin selection

For the selection of plentiCRISPRv2 transfected cells 72 hours post transfection, cells were incubated for 3 days with 4µg/mL of puromycin (Sigma), the antibiotic for which this vector confers resistance. Cells were medium changed every day, with antibiotic supplemented full DMEM during the 3 days of selection. After the antibiotic treatment, cells kept being gently medium changed daily until no more cells of unviable appearance remained. Surviving cells were cultured until confluent and subcultured until enough material could the collected for DNA and protein extraction. As mentioned, the chosen puromycin concentration was derived from a preliminary experiment in which it was the lowest concentration to eliminate non-transfected NIH/3T3 cells.

4.4.3 DNA validation

DNA was extracted using the 'direct lysis with detergent and protease k (van der Burg et al 2011)' method and quantified by Qubit. The extracted DNA was used as a template for PCR amplification of the targeted locus.

The detection and quantification of CRISPR induced mutations in the

target DNA locus was performed through T7 assay and next generation sequencing. Initially T7 assay, a more affordable though less informative validation method was used to screen candidate guides before performing NGS analysis.

4.4.3.1 T7 assay

Primers were designed using primer-blast (https://www.ncbi.nlm.nih.gov/tools/primer-blast) to amplify sgRNA targeted regions so as to meet criteria to perform T7 assays. Amplicons of approximately 1000bp containing the cut site were chosen such that it would be nearer to one of the primers than the other. This condition ensures that asymetric cleavage products generated can be resolved by molecular weight in a 1.5% agarose gel. The melting temperature was optimized for each assay as described.

DNA from targeted regions was amplified by PCR using primers for T7 (3x 25 µL reactions; see 'Polymerase Chain Reaction' section) and either gel purified or column purified and eluted in 20 µL EB.

A solution with 200ng of DNA in 19uL of 1x NEB2 was incubated for 5 minutes at 98°C for denaturation, and then slowly cooled, first to 85°C at 2°C/second then to 25°C at 0.1°C/second, so as to allow mutated DNA to form heteroduplexes with wild type DNA present in the sample.

The annealed DNA was incubated for 20minutes at 37°C with 10 units of T7 endonuclease (#MO302; NEB). The enzymatic reaction was stopped by addition of 1 µL of 0.5M EDTA.

The reaction product was loaded and resolved in a 1.5% agarose gel and analysed under UV lighting. A non T7 treated sample was used as a negative control.

4.4.3.2 From T7 assay to NGS assay

The T7 assay, though effective, proved not to be compatible with a higher throughput *in vitro* validation pipeline. This assay is time consuming, as it requires the design and optimization of primers with specific conditions, not compatible with sample submission for NGS, the more informative

analysis method intended for the *in vivo* stage of this project. Due to the nature of the assay, PCR products also tend to require gel purification prior to analysis, another time-consuming step, as any unspecific PCR product would result in mismatches during the heteroduplex annealing step leading to background bands and inconclusive results.

Based on this experience, more recent guides have been directly validated by NGS, so as to achieve a higher throughput.

4.4.3.3 NGS assay

Primers for the NGS assay were designed using primer-blast (https://www.ncbi.nlm.nih.gov/tools/primer-blast) so as to select 200-280bp amplicons for which the cut site is within the first 100 bp (where there is maximum resolution). PCR conditions were optimized for primer pair as described. However, for part of the PCR assays to prepare samples for NGS, a melting temperature of 66°C was chosen based on the optimized temperatures for previous assays with similar design, rather than new optimizations for each assay. That was the case of PCR assays for *Fan1* (JL1.8), *Exo1, Ercc3, Pms1, Rrm2b and Msh6*.

DNA from targeted regions was amplified by PCR using primers for NGS (3x 25 µL reactions; see 'Polymerase Chain Reaction' section). DNA from empty construct treated cells was used as a negative control for which no mutations were expected. When this control template DNA was not available, DNA from cells treated with guides targeting a different gene was used.

PCR products were submitted to the CHGR DNA core for Sanger sequencing in both directions as a first screen for CRISPR induced mutations and then submitted for more detailed characterisation by NGS sequencing at the Massachusetts General Hospital's Center for Computational and Integrative Biology (CCIB core).

In more recent experiments, so as to make the best possible use of each NGS sequencing submission, samples were submitted pooled for sequencing (2 PCR products for sgRNA treated samples at a time and up to 6 PCR products for empty vector treated samples at a time). Care was taken not to pool samples amplified by PCR using the same primers, so that contigs from each sample could be easily distinguishable during posterior analysis.

Initially, PCR products were gel purified to remove contaminant unspecific PCR products prior to submission for NGS so as to meet the NGS core's standards for sample purity.

As mentioned, gel purification steps were identified as a limiting factor when optimizing the *in vitro* validation pipeline. For this reason, the feasibility of just performing column purification before submission and filtering out unspecific product during output data analysis was considered.

A direct comparison experiment was performed, when validating Msh2-JD1 and Msh3-JD2.3, so as to evaluate if it would be possible to submit non-gel purified (only mini-elute purified) PCR products for NGS sequencing and remove unspecific products when analysing results.

Analyses of NGS results, included separating contigs according to the assay that generated them (different primers), and manual annotation of contigs as noise (contigs not mapping to the targeted genome sequence), unmodified, modified with frameshift mutations or modified with other types of mutations. Annotation was based on comparison with the reference sequence using clustering and blast software. NGS assays were also ran as a controls on samples from empty construct treated cells, or from cells treated with constructs not targeting the same genome region being assayed. High frequency polymorphisms (in this dataset only SNPs) also present in the empty-construct treated sample were annotated as unmodified contigs.

4.4.4 Protein validation

Protein was extracted quantified and analysed by western blot as described. Protein samples from control and knock out mice for the target protein were used as a control, as well as samples from non-transfected, empty vector transfected and empty vector transfected and selected NIH/3T3 cells.

As at the time there were no in lab validated MSH3 antibodies for western blot, 3 antibodies were validated on mouse and NIH/3T3 control samples. *Msh3* guide-treated cells were analysed using the most promising validated antibody. *Msh3* guide validation by western blot without performing cell selection was attempted in an independent experiment and was inconclusive.

Protein analysis was not possible for MLH3 for lack of specific

antibodies.

Table 4.5 Primary antibodies used in western blots

Protein	Antibody	Dilution used	Total Protein Loaded µg/lane	ECL used	Polyacrylamide gel % & Running Buffer	Protein size (a.a.)
MLH1	N20 (sc-581; J2108)	1:200	40µg/lane	Regular ECL	12% Bis-tris	760
MSH2	AB70270	1:1000	50µg/lane	Pico ECL	3-8% gel TA	935
MSH3	7H12	1:300	50µg/lane	Pico ECL	3-8% gel TA	1095
Alpha-Tubulin	DM1A (cell signalling; #38735; mouse)	1:3000	Detected in the same lanes for normalization	Regular ECL	Detected in the same lanes for normalization -	

4.5 GENERATING AAV FOR *IN VIVO* CRISPR

4.5.1 pAAV design and development

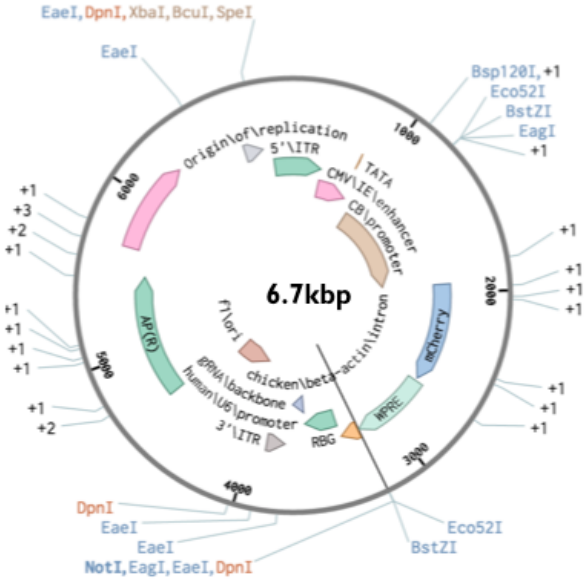

Figure 4.1 pAAV construct.

A new construct for sgRNA delivery *in vivo* by AAV was developed and assembled in collaboration with the Gene Transfer Vector Core facility (Schepens Eye Research Institute and Massachusetts Eye and Ear Infirmary).

Among its key elements are:

- The small polymerase III promoter U6, with receptor sites for BsmBI that can drive sgRNA expression upon its molecular cloning.
- The mCherry reporter gene under a strong constitutive CAG promoter for the detection of transfected cells within the GFP positive background of the Rosa26-Cas9 mice. A woodchuck hepatitis virus posttranscriptional regulatory element is also present to further potentiate expression levels
- Inverted Terminal Repeat (ITR) sequences to enable efficient packaging into AAV

4.5.2 pAAV construct validation

4.5.2.1 Restriction mapping

The plasmid was further validated by restriction mapping . Common enzymes from the Anza™ 10-Pack Starter Kit (Invitrogen™) capable of cutting the plasmid in only 1 or 2 restriction sites were chosen and the position of their restriction sites relative to different elements of the vector determined. Of these, 2 restriction enzymes were selected such that digestion at their restriction sites would produce products sufficiently different in molecular weight to be distinguishable by agarose gel electrophoresis. For validation of this particular construct, preference was given to pairs of restriction enzymes that, meeting the previous criteria, also flanked the newly inserted mCherry expressing element, so that the presence, absence or duplication of this element would be suggested by the size of the digestion product expected to contain it. The chosen pair of enzymes meeting both criteria consisted of Xba1, expected to cut the plasmid in 2 different restriction sites (in the multiple cloning site and the chicken beta actin intron sequence just before the mCherry sequence), and Eco321, expected to cut the plasmid in only 1 restriction site (immediately after the mCherry sequence and before the WPRE sequence).

The vector was digested with the chosen restriction enzymes (with each one separately and with both combined) and the molecular weight pattern of digestion products was compared with the pattern predicted for the designed vector. Undigested vector was used as a negative control of the digestions.

Table 4.6 Restriction enzyme digestion conditions for mCherry restriction mapping validation. Plasmid DNA was digested for 2h at 37°C with enzymes Xba1 and Eco321 from the Anza™ 10-Pack Starter Kit (Invitrogen™) .The enzymes were subsequently heat inactivated for 10min at 65°C

Component	Digestion with 1 restriction enzyme	Digestion with 2 restriction enzymes
H20	16 µL	15 µL
Anza Buffer (10x)	2 µL	2 µL
DNA	1 µL (=1.5µg)	1 µL (=1.5µg)
Restriction enzyme	1 µL	1 µL of each enzyme
Total	20 µL	20 µL

4.5.2.2 Whole plasmid sequencing

Plasmid DNA was extracted from a bacterial colony carrying the pAAV construct and submitted for NGS based whole plasmid sequencing at the CCIB core (35 µL of 50ng/µL in EB).

4.5.2.3 ITR integrity validation

ITR regions are important elements for AAV encapsulation that being repetitive elements are more prone to mutations during molecular cloning. Being repetitive elements, ITR's are also hard to sequence with confidence (low read count with NGS based whole plasmid sequencing methods)

These elements were therefore validated for more confidence through digestion with restriction enzymes. Plasmid was digested with 3 separate restriction enzymes (Xma-I, AHDL-I and MSC-I) that have restriction sites in the ITR regions, and the resulting molecular weight patterns of digested products compared with predictions. A restriction enzyme that successfully digests a given restriction site is informative of the integrity of its sequence in the plasmid as the enzyme would otherwise not recognize and digest it. Undigested plasmid was again used as a negative control of the digestions.

Tabel 4.7 Restriction enzyme digestion conditions for ITR restriction mapping validation. Plasmid DNA was digested for 2h at 37°C with restriction enzymes specific for the ITR region of the pAAV construct. Enzymes were subsequently heat inactivated by 20 minutes incubation at 80°C. A volume of 20uL corresponding to 600ng of digested DNA from each condition was run and analysed by agarose gel electrophoresis.

	Conditions			
Components	UnDigested	Xma1	AhdI1	MSC1
Restriction Enzyme	0 µL	1.5 µL (10kU/mL)	1.5 µL (10kU/mL)	3 µL (5kU/mL)
DNA	1uL (=1.5µg)	1uL (=1.5µg)	1uL (=1.5µg)	1uL (=1.5µg)
CutSmart Buffer (10x)	5 µL	5 µL	5 µL	5 µL
dH2O to reach Total Volume	44 µL	42.5 µL	42.5 µL	41 µL
Total Reaction Volume	50 µL	50 µL	50 µL	50 µL

4.5.2.4 *In vitro* validation of mCherry expression

In order to validate the pAAV vector in terms of mCherry expression, NIH/3T3 were transfected with pAAV construct using lipofectin 3000 . Fluorescence of mCherry was assessed 48hours post transfection by microscopy. Non-transfected cells were used as a negative control for this experiement and pMaxGFP transfected cells as a control for transfection efficiency.

4.5.3 Cloning in vitro validated guides into pAAV

Following pAAV construct validation, sgRNA duplexes of the previously *in vitro* validated guides Mlh1-A1 and Mlh1-JD4 were produced and cloned to pAAV as above described.

Plasmid DNA from clones of cells transformed with pAAV-guide construct was extracted and submitted for sanger sequencing, using a U6 forward primer, to check for the correct insertion of the desired guide into the plasmid backbone.

The bacterial clones with sanger validated pAAV-guide constructs and the clone with the validated empty pAAV backbone were then maxipreped

and revalidated by sanger and restriction mapping.

The validated maxipreped constructs were submitted for production of AAVs of AAV8 serotype at the Gene Transfer Vector Core facility (Schepens Eye Research Institute and Massachusetts Eye and Ear Infirmary) as 500µg of plasmid in 400 µL EB.

4.6 *IN VIVO* VALIDATION OF CRISPR CONSTRUCTS: *MLH1*

The potential of the previously developed guides and constructs to genetically modify CAS9 constitutively expressing mice by *in vivo* delivery using AAV8 was validated for the known CAG somatic instability genetic modifier *Mlh1*.

A colony of constitutively and ubiquitously CAS9 expressing mice described in (Platt et al. 2014) was obtained and established from the Jackson Laboratory (Gt(ROSA)26Sor$^{tm1.1(CAG-cas9*,-EGFP)}$Fezh/J ; Stock No: 024858).

These mice expressing a mammalian codon-optimized cas9 gene derived from Streptococcus pyogenes, can be genetically modified by *in vivo* sgRNA delivery. (Platt et al. 2014). The mice are reported to breed normally without morphological abnormalities or up regulation in DNA damage and apoptosis markers.

Adult CAS9 constitutively expressing mice were treated by hydrodynamic tail vein injection with 3x1011 copies in 200 µL of AAV8 for Mlh1-A1 or Mlh1-JD4. Treatment was performed in mice of 3 age groups of 2, 7 and 9 month old mice (table 4.8). PBS treated mice were used as negative controls (table 4.8). At this preliminary stage only male mice were studied. Mice were sacrificed 10 days post-injection and liver tissue samples collected and stored at -80°C.

Table 4.8 Mice used in preliminary in vivo CRISPR experiments. Mice from 3 age groups,2 7 and 9 months, heterozygous for constitutive CAS9 expression were treated with *Mlh1* sgRNA's A1 and/or JD4, by tail vein injection(TVI) or intraperitoneal injection (IP). PBS treated wild type mice were used as negative controls.

Mouse #	Sex	Age (months)	Cas9	Delivery Route	Treatment
1	M	7	HET	TVI	Mlh1-A1
2	M	7	HET	TVI	Mlh1-JD4
3	M	7	WT	TVI	PBS
4	M	9	HET	TVI	Mlh1-A1
5	M	9	HET	TVI	Mlh1-JD4
6	M	9	HET	TVI (failed)	Mlh1-JD4
7	M	2	HET	TVI	Mlh1-A1
8	M	2	HET	TVI	Mlh1-JD4
9	M	2	WT	TVI	PBS
10	M	2	HET	IP	Mlh1-A1 & Mlh1-JD4

DNA was extracted from a whole liver sample, as described, and the targeted sites for genetic modification amplified by PCR and analysed by sanger and NGS sequencing so as to assess the presence and efficiency of modification.

DNA samples from AAV treated mice were analysed by both Mlh1-A1 and Mlh1-JD4 NGS assays. Mice treated with Mlh1-A1 AAV were used as an additional negative control for Mlh1-JD4, since they are not expected to be modified at the DNA region Mlh1-JD4 targets. DNA samples from mice treated with Mlh1-JD4 AAV were similarly used as negative controls for the analysis of Mlh1-A1 treated mice DNA.

AAV delivery of CRISPR constructs *in vivo*, for the genetic modification of liver cells, was also attempted by intraperitoneal injection (IP) using a surplus mouse from the 2month aged mouse group. The mouse was injected with an ill-defined volume of both Mlh1-A1 and Mlh1-JD4 AAV making it

impossible to directly compare IP and tail vein injections in this very preliminary side experiment.

4.7 SOMATIC CAG INSTABILITY IN CONSTITUTIVE CAS9 EXPRESSING MICE

The ROSA26-CAS9 knockin mice were crossed with Q111 mice to produce offspring expected to reproduce the CAG somatic instability phenotype while simultaneously being susceptible to CRISPR genetic modification by *in vivo* delivery of sgRNA's.

Somatic CAG repeat instability was characterized in 7 and 9 month old Q111-CAS9 mice (table 4.9) in a preliminary study so as to validate the presence of this phenotype in Q111-CAS9 mice and investigate whether different CAS9 genotypes could affect it.

Table 4.9 Q111-CAS9 mice with different CAS9 genotypes used in preliminary somatic CAG repeat instability assays. Mice were genotyped for CAG mutation and CAS9 using DNA extracted from tail tissue at weaning. A total of 7 Q111 mice from 2 age groups (7 and 9 months) and different CAS9 genotypes were used.

Mouse #	Sex	Age (months)	CAG	CAS9
1	M	9	Q111	HOMO
2	M	9	Q111	HET
3	F	9	Q111	WT
4	F	9	Q111	WT
5	M	7	Q111	HET
6	M	7	Q111	HET
7	M	7	Q111	HOMO

DNA from both CAG repeat unstable (striatum and liver) and stable (cerebellum) tissues was extracted and quantified. The humanized genetic region in Htt exon1 containing the CAG repeat mutation was amplifyed by PCR with fluorescent primers (following conditions in table 4.10 and using primers in table 4.11) and submitted to the CHGR DNA core to be resolved by size through capillary electrophoresis and quantified by fluorescence intensity as in (Lee et al. 2011). Results were analyzed using the gene mapper software (Applied Biosystems, version 5).

Table 4.10 CAG instability assay conditions. A PCR mix was prepared according to the table. The mix was denatured for 5minutes at 95°C and then submitted to 30 cycles of 30seconds at 95°C, 30seconds at 65°C and 1minute and 30 seconds at 72°C.

Component	Amount (µL)
dH20	6.3
10xBuffer	2
5xQ	4
dNTPs[10uM]	0.4
CAG1-FAM [10uM]	1.6
Hu3 [10uM]	1.6
Taq [5U/ µL]	0.1
DNA [20ng/ µL]	4
Total	20

Table 4.11 CAG instability assay primers

Primer	Sequence 5'->3'
CAG1-FAM	ATGAAGGCCTTCGAGTCCCTCAAGTCCTTC
Hu3	GGCGGCTGAGGAAGCTGAGGA

5. RESULTS

5.1 CRISPR SGRNA DESIGN

Guides for known genetic modifiers and candidate modifiers of somatic CAG instability were designed for maximum on-target efficiency and low off-target efficiency, using the best library or design tool available at the time.(Doench et al. 2016)

Guides were successfully picked or designed for each intended target gene (table 5.1). From these it was possible to select only those targeting their specific cut site within the 5 to 65% range of distance to the N' terminal in the corresponding coded protein, where frame shift mutations have a greater expected chance of disrupting protein expression. (Doench et al. 2016) Picked guides meeting this criteria and simultaneously having a predicted on-target efficiency above 50% were found for each gene and chosen for further testing *in vitro*.

An exception was made for the very promising Msh2-JD1 guide, with a cut site at 4.28% of the corresponding protein extension, that had a JD version 1 predicted on-target efficiency of 96% while the next best guide had only 84% (data not shown).

For the 12 studied genes, there was a variable number and quality of top candidate guides which is reflected in the predicted on-target and off-target efficiency of the chosen guides for each gene.

Table 5.1 Design of sgRNA's to target known and candidate modifiers of CAG repeat instability. Guides selected for molecular cloning and *in vitro* testing. Guides were picked from the GeCKO library or designed with the Broad's sgRNA design tool (http://portals.broadinstitute.org/gpp/public/analysis-tools/sgrna-design) version 1 or 2, referred here as JD version 1 and 2. Guide features considered when choosing each guide (predicted on-target efficiency, off-target rank within the list of candidate guides and distance from the N' terminal of the cut site in the corresponding coded protein in terms of percentage of the total protein extension) are presented. The targeted exon is also presented for extra information.

Gene	Guide ID	Design Tool/Library	On-target Efficiency (JD version1)	On-target Efficiency (JD version2)	Off-target rank	Position in protein of cut (%)	Exon
Mlh1	A1	GeCKO A	51.96%	63.70%	286	12.7	3
	JD4	JD version 1	84.75%	68.53%	210	34.9	10
Msh2	JD1	JD version 1	95.65%	73.91%	3	4.28	1
Mlh3	JD2	JD version 1	77.69%	65.28%	288	18.2	2
Msh3	A1	GeCKO A	4.32%	55.84%	1	23.29	5
	A2	GeCKO A	16.38%	57.93%	180	22.56	5
	JD2.1	JD version 2	54.27%	72.10%	22	45.5	11
	JD2.3	JD version 2	27.59%	74.82%	62	57.4	14
Fan1	JD2.1	JD version 2	ND	77.86%	23	32.9	2
	JD2.2	JD version 2	ND	64.65%	2	6.6	2
	JL1.8	From JL	ND	48.67%	253	39.2	2
Rrm2b	JD2.1	JD version 2	ND	59.78%	4	16.5	2
	JD2.2	JD version 2	ND	59.77%	13	11.5	2
Pms2	JD2.1	JD version 2	ND	64.31%	8	23.8	6
	JD2.2	JD version 2	ND	67.63%	30	8.4	3
Pms1	JD2.1	JD version 2	ND	72.66%	8	33	8
	JD2.2	JD version 2	ND	68.14%	12	11.1	3
Ercc3	JD2.1	JD version 2	ND	64.45%	1	25.8	5
	JD2.2	JD version 2	ND	66.90%	21	13.7	3
Exo1	JD2.1	JD version 2	ND	72.26%	22	20.9	6
	JD2.2	JD version 2	ND	68.32%	13	56.6	11
Msh6	JD2.1	JD version 2	ND	72.83%	29	26.7	4
	JD2.2	JD version 2	ND	63.55%	9	58.1	4
FancJ	JD2.1	JD version 2	ND	70.28%	12	37.5	9
	JD2.6	JD version 2	ND	82.51%	37	20.4	7

The significant differences between the JD algorithm versions translated to substantial differences in the predicted efficiency of guides for the set of studied genes. For instance, Msh3-A1 and Msh3-A2, picked from the GeCKO library at an early stage, were found to have a low predicted on-target efficiency with JD version 1 (approximately 4 and 16%), not being the ideal candidates for further analysis. However, after re-evaluation with the new algorithm, their on-target efficiency above 55% was considered compelling enough to pursue further testing.

One feature that was remarkably improved by using the new JD algorithm to pick guides was the predicted off-target rank, not included in the previous version.

As an example of the guide picking methods used in this thesis, the guides picked for *Mlh1* from the GeCKO library and using JD version 1 are represented in terms of on-target efficiency according to each JD version (for version 1 in figure 5.1 and for version 2 in figure 5.2). A direct comparison between the predicted on-target efficiency is also presented (figure 5.3).

When analyzed with JD version 1, the GeCKO library picked Mlh1-A1 guide was not part of the predicted top candidate guides, having a lower predicted efficiency. Mlh1-JD4 was picked as the most promising according to JD version 1 for having a high predicted efficiency for a cut site near the N'terminal of the corresponding coded protein.

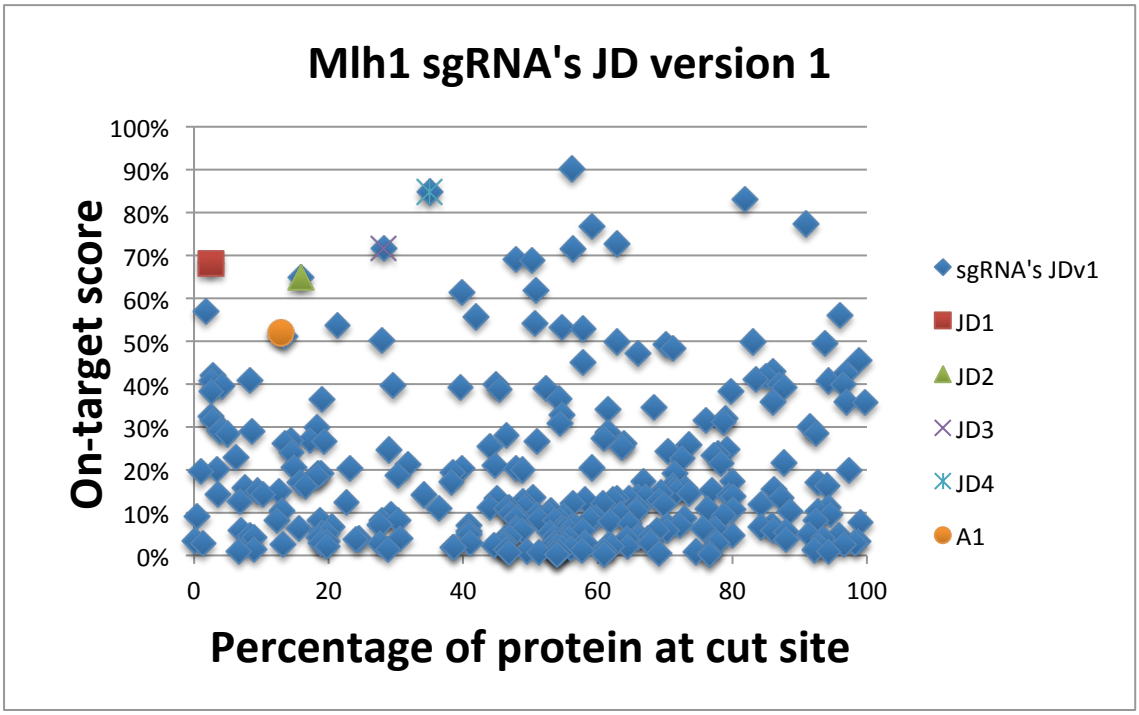

Figure 5.1Design of sgRNA's to target Mlh1 using JD version 1. Predicted on target efficiency is plotted against cut site position for all sgRNA's identified with this JD version. Top candidates in terms of on-target efficiency were identified and numbered according to their cut site position. Mlh1-A1 from the GeCKO library is also represented for comparison.

Overall, for *Mlh1*, top ranked sgRNA's had a lower predicted on-target efficiency using the second JD version. However, top candidates from version 1 remained within the most promising in terms of on-target efficiency in version 2. The same was true for Mlh1-A1, whose predicted efficiency even increased compared with the previous prediction, though remaining less promising than the JD picked Mlh1-JD4. New interesting candidates, not previously evidenced, could be identified in terms of on-target efficiency using JD version 2 for the 5 to 65% range of the protein.

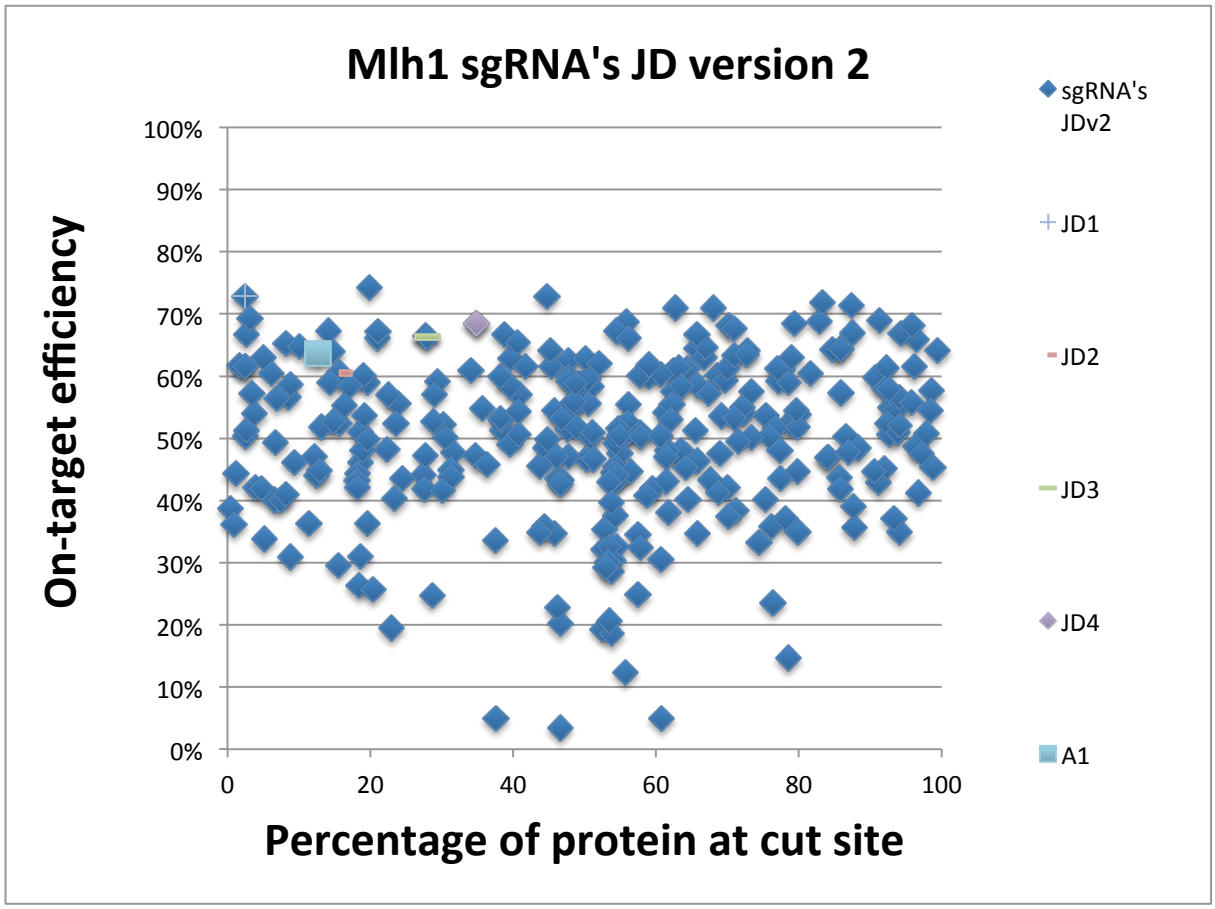

Figure 5.2 JD version 2 analysis on-target efficiency of sgRNA's designed with JD version 1 to target *Mlh1*. Predicted on target efficiency is plotted against cut site position for all sgRNA's identified with JD version 2.

Figure 5.3 Direct comparison of JD version 1 and JD version 2 predicted sgRNA on-target efficiency for *Mlh1*. All guides common to libraries designed using both algorithms for *Mlh1* were plotted in terms of their predicted on-target efficiency in each version. Mlh1-A1 and top guides from JD version were identified.

The same *Mlh1* guides are also represented here in terms of on and off target ranking according to JD version 2 (figure 5.4). Previously very similar guides in terms of predicted on-target efficiency ranked very differently in terms of predicted off-target efficiency. Mlh1-JD4 ranked better than Mlh1-A1, further pointing to its greater comparative potential. Both of the picked guides however, ranked poorly in terms of off-target efficiency, with more than 200 other guides being better ranked for this criterion. Among better off-target ranked guides (rank closer to one) are many guides with equally better or comparable on-target rank, namely some of the previously identified JD version 1 top candidates.

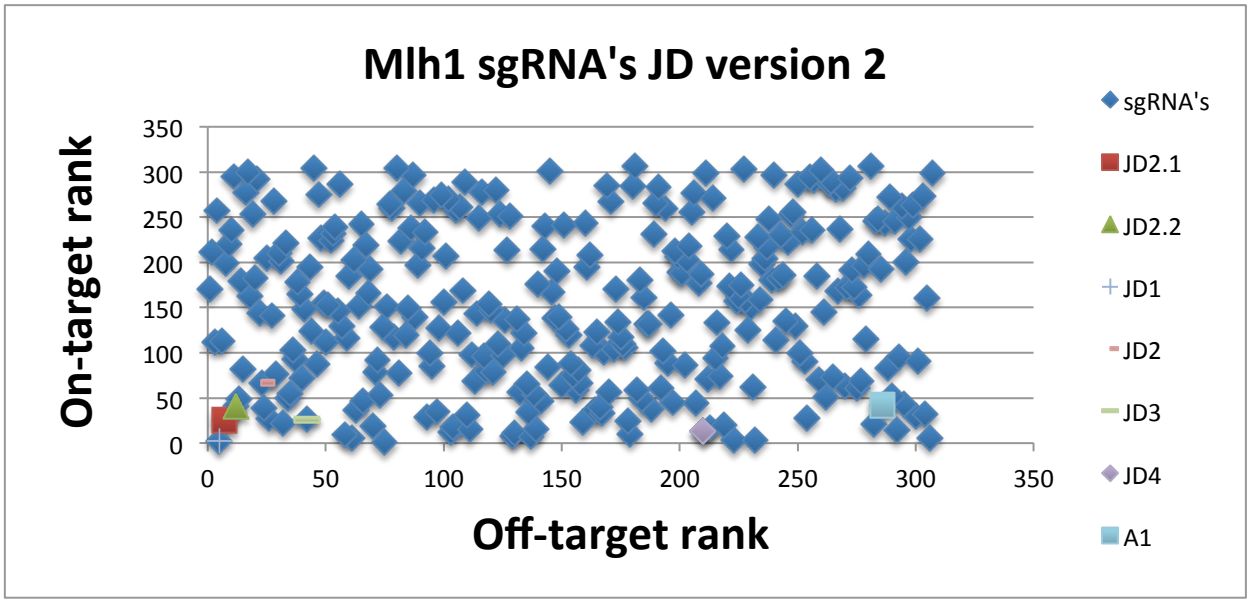

Figure 5.4 JD version 2 combined analysis of on- and off-target rank of sgRNA's designed to target *Mlh1*. Predicted on-target rank is plotted against off-target rank, with a sgRNA ranking closer to one being more efficient and specific respectively.

The 2 best guides considering both predicted on and off target ranking for *Mlh1* as assessed by JD version 2, that meet the 5-65% position criterion, are also presented as "JD2.1" and "JD2.2". Interestingly, Mlh1-JD1 would have an even better combined ranking but its cut site is at 2.5% of the associated coded protein.

These guides picked for maximized on-target efficiency and a low off-target rank as assessed by an algorithm based on a bigger dataset and better modeling would be the first choice if new guides were to be designed with the same goals for *Mlh1* in the future.

The guide Fan1-JL1.8, designed and cloned independently by a different group, was assessed using the algorithm JD version 2 (figure 5.5). When compared with Fan1-JD2.1 and Fan1-JD2.2, it was found to have a lower predicted on-target efficiency and to be ranked worse in terms of off-target efficiency.

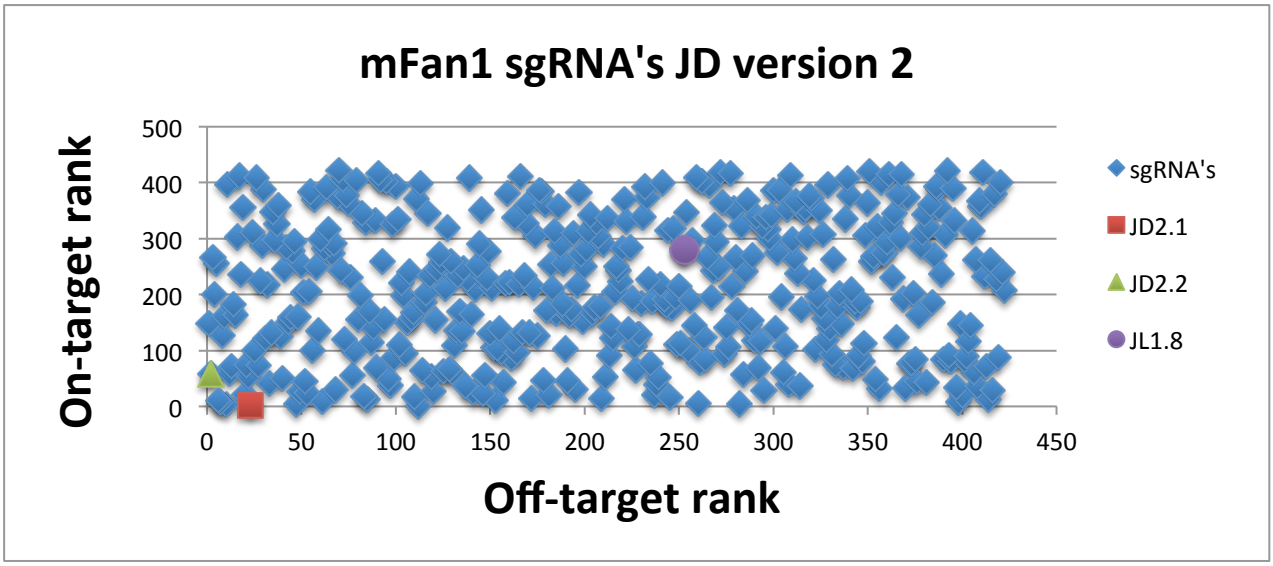

Figure 5.5 JD version 2 combined analysis of on- and off-target rank of sgRNA's designed to target *Fan1*. Predicted on-target rank is plotted against off-target rank, with a sgRNA ranking closer to one being more efficient and specific respectively.

5.2 In vitro validation of CRISPR Constructs

5.2.1 DNA validation: T7 assay

Primers were designed and optimised for T7 as described (table 5.2).

Table 5.2 Design of primers to amplify sgRNA targeted genome regions meeting criteria for the T7 assay.

GENE	GUIDE ID	T7 Forward Primer	T7 Reverse Primer	Tm (°C)	Product Length (bp)
Mlh1	A1	TGTTTGCCAATAAGGAAGTTGGT	ACCAGCTGACAGCCTATTATATCT	66	1000
Msh3	JD2.1	TCTTGCCCTTGGTGTCACAG	TGATTTGGAAGGGGCGTTGA	72	848
Msh3	JD2.3	GCTTCCCTTCCGCCTGTAAT	AAGCACTCCAGGGTGTCAAC	72	943

T7 assay was successfully performed with control samples from the surveyor kit (figure 5.6).

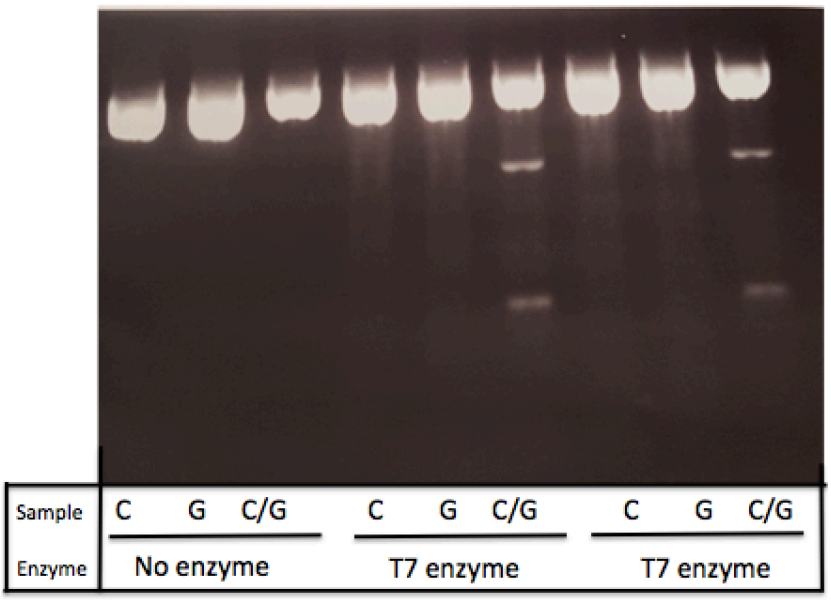

Figure 5.6 T7 assay performed with control samples from the Surveyor kit. DNA control samples C and G, which are 633bp sequences that differ by a single base pair mismatch, were amplified by PCR, gel purified, denaturated, reannealed into heteroduplexes and incubated with T7 enzyme. Two diferente batches of T7 enzyme were tested and samples that were not treated with T7 enzyme included as controls for the specificity of the T7 enzyme.

The T7 assay was then used to validate Mlh1-A1 (figure 5.7) and the *Msh3* guides JD2.1 and JD2.3 (figure 5.8). A high level of background noise is found in control samples in both experiments but in particular for the *Msh3*

guides which limits the interpretation of results. The clear presence of specific lower molecular weight bands, not present in the control, for Mlh1-A1 and Msh3-JD2.3 suggests their ability to induce mutations. It should be considered that these results could only be produced with this quality after several steps of assay optimization (Tm screening, gel purification, switching to a high fidelity enzyme, increasing gel loaded product).

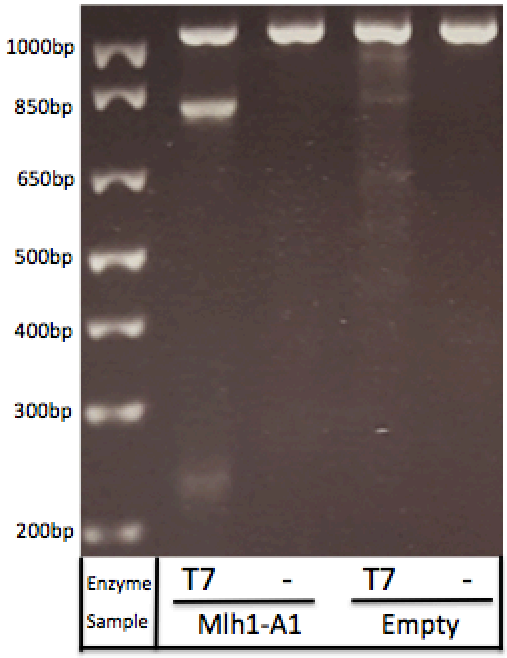

Figure 5.7. Mlh1-A1 sgRNA validation by T7 assay. DNA from NIH/3T3 cells transfected with pX458-guide constructs and FACS sorted for GFP was amplified by PCR, gel purified , denaturated, reannealed into heteroduplexes and incubated with T7 enzyme. The reaction product was resolved in a 1.5% agarose gel. As a control for sgRNA specificity, a sample from empty construct treated cells was included. Samples that were not treated with T7 enzyme were also included as controls for the specificity of the T7 enzyme.

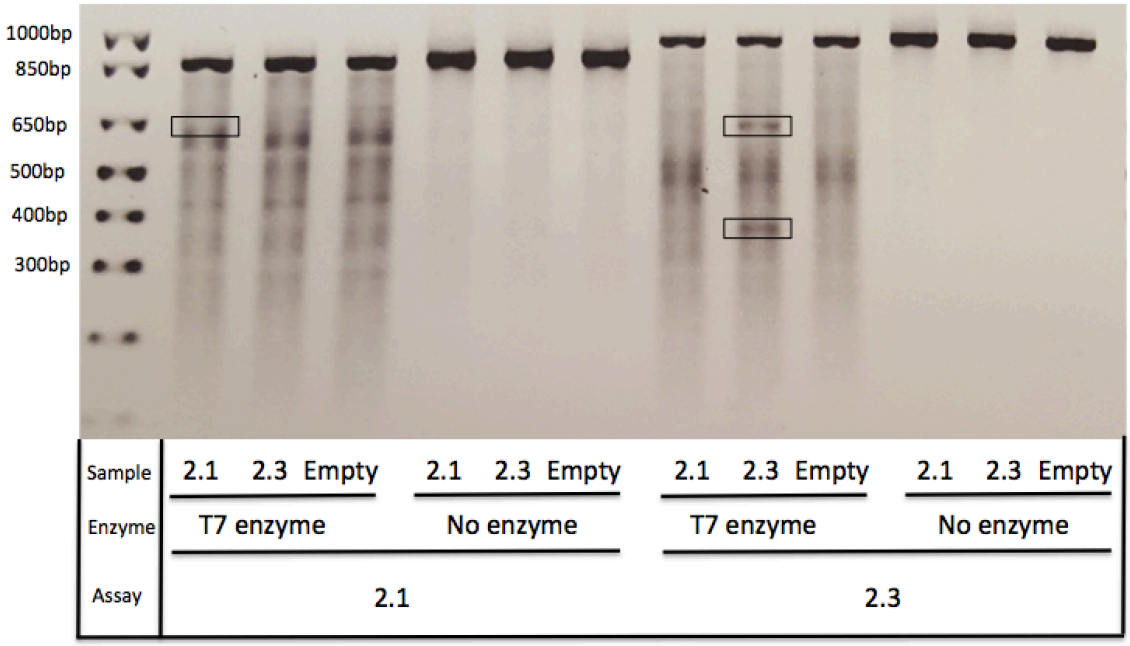

Figure 5.8 Msh3-JD2.1 and Msh3-JD2.3 sgRNA validation by T7 assay. DNA from NIH/3T3 cells transfected with pX458-guide constructs and FACS sorted for GFP was amplified by PCR, gel purified, denaturated, reannealed into heteroduplexes and incubated with T7 enzyme. The reaction product was resolved in a 1.5% agarose gel. As controls for sgRNA specificity, each assay included a sample from empty construct treated cells and a sample from cells treated with the other guide, without a cut site in the amplicon being assayed. Samples that were not treated with T7 enzyme were also included as controls for the specificity of the T7 enzyme. Potential cleavage products are indicated with boxes.

5.2.2 DNA validation: NGS assay

Primers were designed and optimised for NGS as described (table 5.3).

Table 5.3 Design of primers to amplify sgRNA targeted regions for NGS assay.

GENE	GUIDE ID	NGS Forward Primer	NGS Reverse Primer	Tm (°C)	Product Length (bp)
Mlh1	A1	ATCCTGCAGAAGGAAGATCTGG	TGCATGTGGTGAACAAGTACAAAG	67	216
	JD4	GTTCCTGGGCATCTGATAAGG	TACAGGAATGGGTGTGTGTTT	67	210
Msh3	JD2.3	TTTGGTTTCTAGTGCTCTACCCA	GGAAGCCCGCACTTACTTGG	65.8	216
Msh2	JD1	AAATGGCGGTGCAGCCTAAG	ATGTACTTGATCACGCCCTGG	70.4	202
Mlh3	JD2	TTATATGCAGGCCAAAGAATGGC	CTTCCTGAATACAAATCAACACGGT	66	198
Fan1	JD2.1	CCAAAACGCTGGTGTCTGG	CTCCGCAGGTAGTACGGGT	61.3	202
	JD2.2	ACAGTGCACCACCTGCTAAA	AGGTGTCACATTCTCTAAAGGACC	67.6	203
	JL1.8	TACTACCTGCGGAGCTTCCT	CCTCGTGGTGGTCATCAGAC	66	204
FancJ	JD2.1	TGAAGCTCACAACATTGAAGACTG	AAGACAATATGAACACACTGACAGA	65.8	200
	JD2.6	TGATGCGTTCTGAGAGATGTTCT	TGTGTCCGCGTCCCAAAATA	67.6	205
Pms2	JD2.1	GTAAGTTTAGCCATGCAGTTCCT	GCCAAACACAGACCCGATATTT	65.8	222
	JD2.2	AGAGTGTAGGCATGTCTGCTAA	TCTGAGACACGTGGAAATGACT	67.6	239
Exo1	JD2.1	CCTAGGAGTGGATTGCCTCG	ACGAAGATTTTAGCTCCTCAGCA	66	201
	JD2.2	CCCGATTCTGGGACTGCTC	GCAGGCTTCCGCAAATAGTGA	66	200
Ercc3	JD2.1	CCCACCCTGATGTTATCCAGC	ACCATGAACTGAAGCTGTGCT	66	202
	JD2.2	GCAGCTCATTGATATTCTTGCCT	AACACTGACCGCTGCATAGA	66	201
Pms1	JD2.1	CGACGCTATTATAATCTGAAGTGC	AAGGGAACAGTTAGGACAGCATA	66	207
	JD2.2	GTGAGGGCATCAAGGCTGTA	AAGCTACTCTGACAAACAATGAAT	66	240
Rrm2b	JD2.1	CAATTGTGTTCGGAGACAGAAGAA	TTCTGTAAGGCAAAGGTGAGTG	66	200
	JD2.2	CAATTGTGTTCGGAGACAGAAGAA	TTCTGTAAGGCAAAGGTGAGTG	66	200
Msh6	JD2.1	AACTCCCATTCTCTCAGAAACCAA	GGTGGGGTTAAATTCAGGGTGA	66	205
Msh6	JD2.2	GGATACTTGCCATACGCCCTTTG	CAGGGGAGACCCAACATTATGA	66	211

The PCR amplified target regions were submitted for sanger sequencing for a preliminary screening. As exemplified for Mlh1-A1 in figure 5.9, guide treated samples shift from a clear sequence with good resolution to an heterogeneous sequence with multiple overlapping signals as they pass by the targeted cut site. This is consistent with the presence of CRISPR induced mutations.

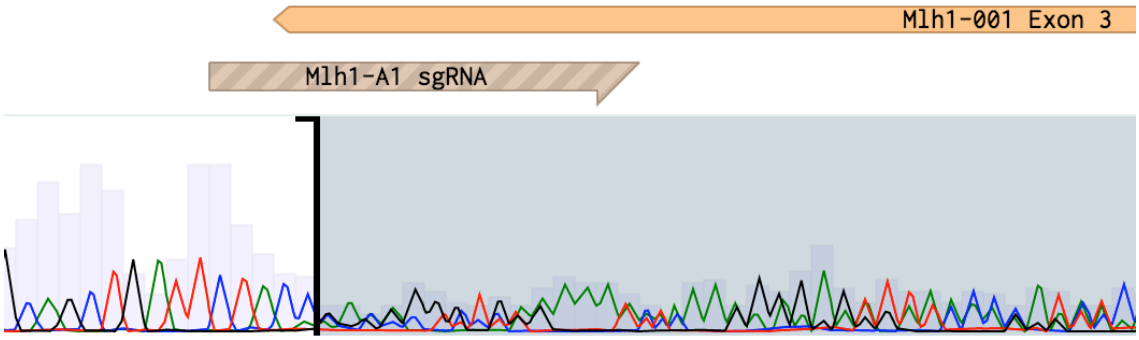

Figure 5.9 Sanger sequencing of the Mlh1-A1 target region for DNA of Mlh1-A1 treated and FACS sorted NIH/3T3 cells. The sanger sequencing results are presented for this sample for the part of the sequence that maps to the guide targeted genome sequence. The sequence is represented from left to right, in the same sense as the sequencing, with peaks in different colors representing the nucleotides guanine, cytosine, thymine and adenine. Peak hight represents the intensity of the sequencing signal and each peak is expected to represent a nucleotide. The bars in the background represent the confidence in determining each sequenced nucleotide. The Mlh1-A1 sgRNA is represented in brown, in alignment above the sequencing results. Mlh1's exon 3 is also represented aligned with the sequence. On the right of the black vertical line, where the background is a darker shade of grey, there is a very low confidence in the calling of base pairs. This low sequencing resolution is due to overlapping sequencing signals (several overlaping peaks for the same nucleotide). The black vertical line matches the sgRNA's targeted cut site. The alignment and sequencing result representation was performed on Benchling.

PCR products were then analysed by NGS sequencing (NGS assays for *in vitro* validation are summarized in the annexes).

Overall, 20 sgRNA's were tested by NGS assay: 16 guides using a puromycin selection protocol; 3 guides using a FACS selection protocol; Mlh1-A1 using both protocols and one repetition experiment through FACS.

For this dataset, total CRISPR induced genetic modification quantified by NGS, was not found to correlate strongly with JD version 2 *in silico* predicted on-target efficiency (figure 5.10). Designing guides with the best possible score and above 50% on-target predicted efficiency using JDv2, yielded guides with an associated modification rate by NGS that was also above 50%, but did not allow for a finer discrimination of guides for this metric.

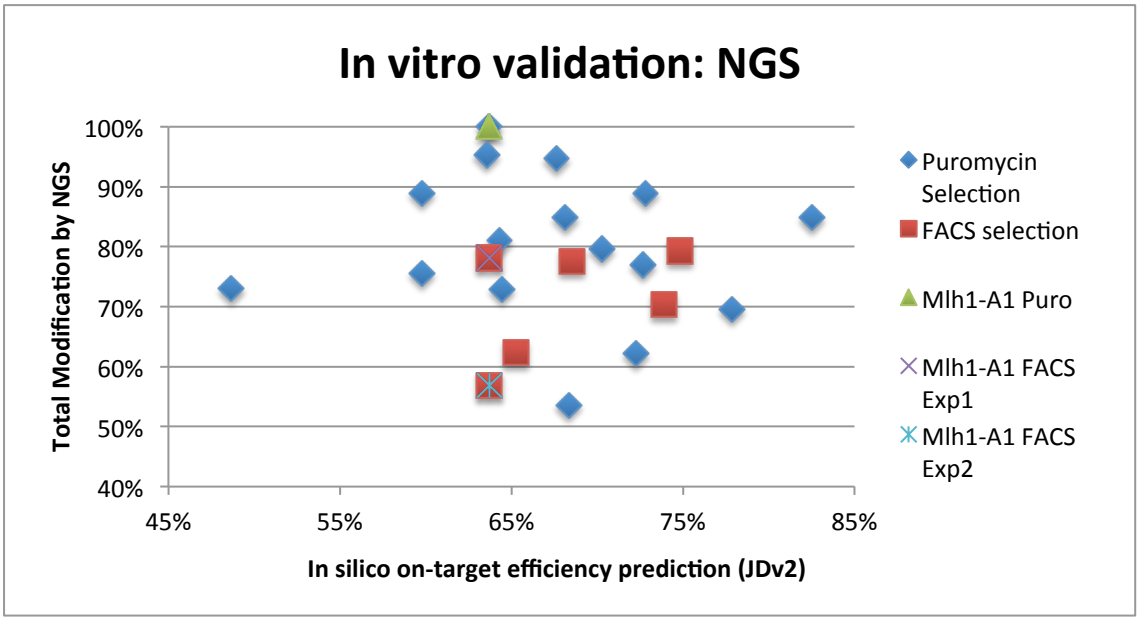

Figure 5.10 *In vitro* guide validation by NGS and in silico predictions. All guides tested by NGS assay were plotted in terms of their *in silico* predicted and NGS assessed modification rate. Guides were discriminated according to the transfected cell selection method.

A similar analysis was performed for NGS quantification of frame shift mutations, which did also not correlate strongly with the *in silico* predicted efficiency. (figure 5.11) Guides designed according to the JD algorithms yielded an above 40% frame shift mutation rate.

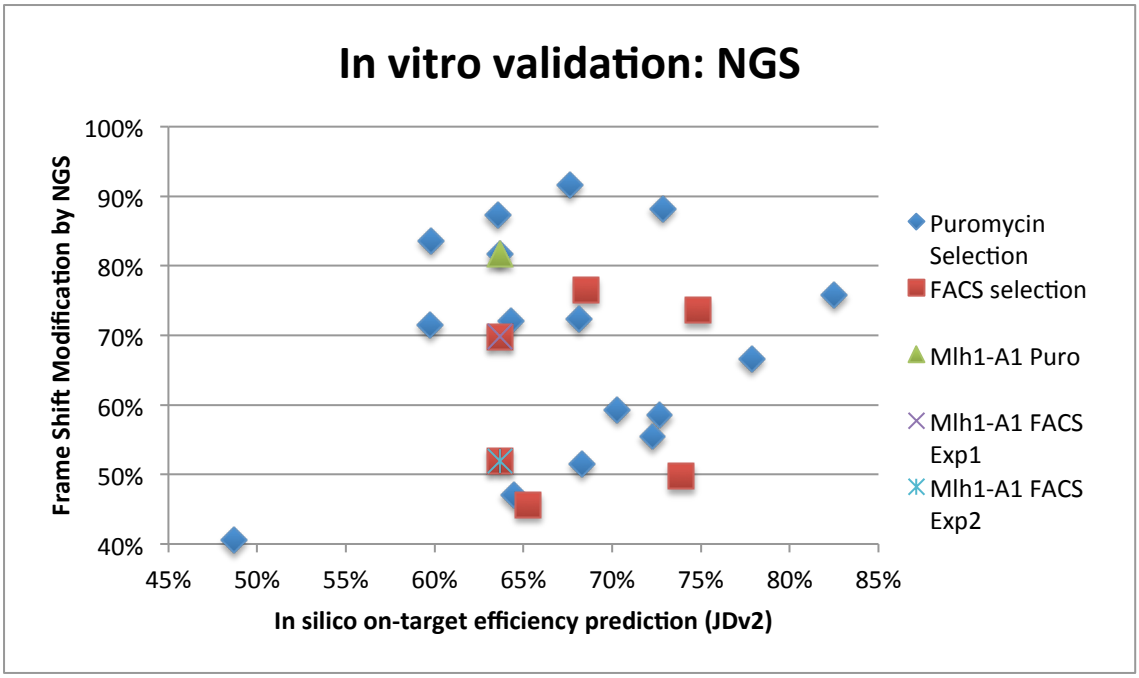

Figure 5.11 *In vitro* guide validation by NGS and *in silico* predictions: frame shift mutations. All guides tested by NGS assay were plotted in terms of their *in silico* predicted and NGS assessed frame shift modification rate. Guides were discriminated according to the transfected cell selection method.

The outlier datapoint in both figure 5.10 and figure 5.11, with a predicted efficiency below 50%, corresponds to Fan1-JL1.8, which was designed by an independent group and tested in the present thesis in parallel with the other 15 guides in vitro validated through the puromycin selection pipeline. Its poor *in silico* predicted score did not reflect a low score in terms of CRISPR induced modification rate (over 70% as represented in figure 5.10), but it did predict a relatively low CRISPR induced frameshift inducing mutation rate (just above 40% as represented in figure 5.11).

There was a considerable 21% difference in modification rate for 2 independent Mlh1-A1 transfections using FACS as the cell selection method, (represented in figure 5.10 and in more detail in figure 5.12). There is however an even more marked difference between these and the measured modification rate of 100% for Mlh1-A1 when analysed through the puromycin selection pipeline. It is also of note the difference in the number of contigs detected for Mlh1-A1 using FACS, 38 and 30 contigs for experiment 1 and 2 respectively, when compared with the 8 contigs detected through puromycin selection. This difference is nevertheless biased by a more than 10 fold difference in the number of specific reads between the assays. (figure 5.12)

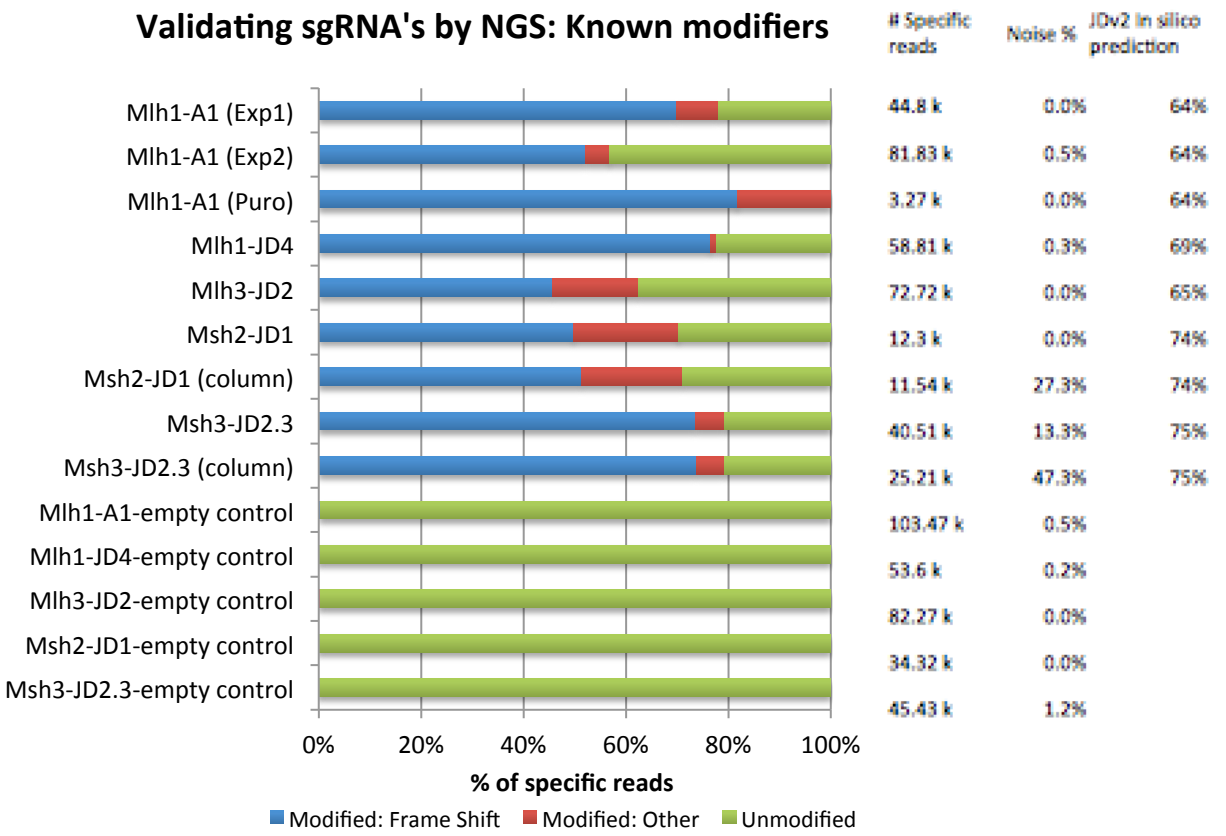

Figure 5.12 *In vitro* guide validation by NGS: known modifiers of CAG repeat instability. All guides targeting the 4 known modifiers of CAG repeat instability tested by NGS assay were characterized in terms of their frame shift and non-frame shift induction modification rate. For each assay, the number of specific reads (that map to the region of interest), the percentage of noise (non-specific reads) and the JDv2 *in silico* prediction of the on-target efficiency are presented. The number of specific reads characterized as frame shift mutated for each assay are also presented on the left side of the bars. Samples treated with pX458-guide constructs (above) and pX458-empty construct (below) are presented. Samples marked as "(column)" were only column purified rather than gel purified.

The reproducibility of the NGS assay using different PCR product purification methods, namely gel purification and column purification was directly compared by purifying the same sample with both methods for Msh2-JD1 and Msh3-JD2.3. As reflected in figure 5.12, the relative proportions of unmodified and frame shift modified reads was very close between the 2 purification methods for these samples. The reduction in the number of reads specific for the amplicon of interest when gel purification is not performed is proportional to the increase in number of non-specific reads, quantified as noise in figure 5.12.

The NGS results for the following batch of *in vitro* tested sgRNAs, targeting new candidate modifiers is presented in figure 5.13.

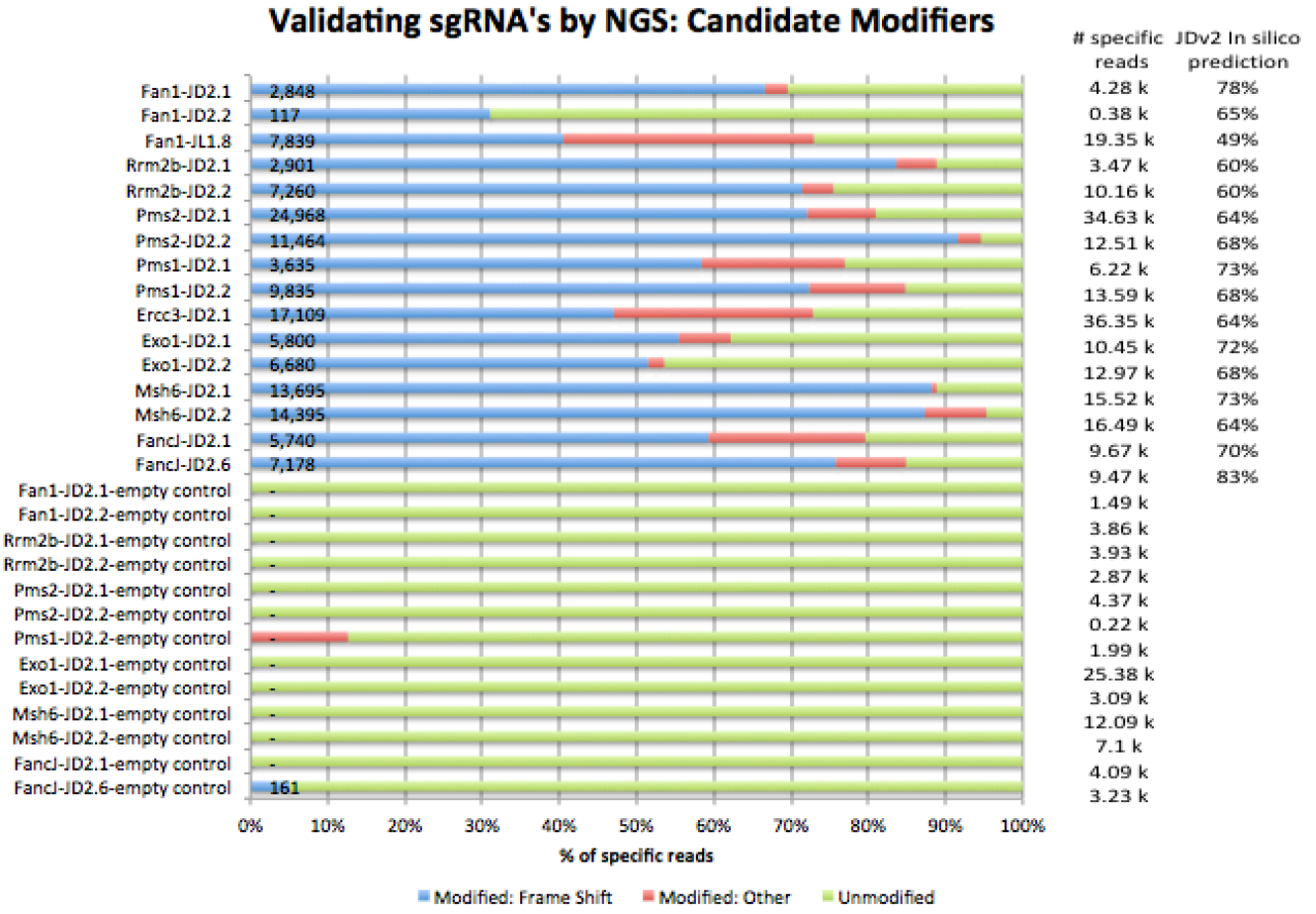

Figure 5.13 In vitro guide validation by NGS: candidate modifiers of CAG repeat instability. All guides targeting candidate modifiers of CAG repeat instability tested by NGS assay were characterized in terms of their frame shift and non-frame shift induction modification rate. For each assay, the number of specific reads (that map to the region of interest) and the JDv2 *in silico* prediction of the on-target efficiency are presented. The number of specific reads characterized as frame shift mutated for each assay are also presented on the left side of the bars. Samples treated with pLentiCRISPRv2-guide constructs (above) and pLentiCRISPRv2-empty construct (below) are presented.

While submitting pooled samples for NGS in groups of 2 or 6 significantly reduced the number of reads per sample, in this batch of *in vitro* tested sgRNAs (figure 5.13) most samples kept having several thousand reads therefore retaining some representation of less common reads. That was not the case for Fan1-JD2.2, which had only 380 reads.

Empty vector treated control samples assayed by NGS showed very little background noise in terms of modified reads (figure 5.12 and figure 5.13). They enabled in several cases the calling of SNP's, common to both treated and non-treated samples at the same high relative frequency close to 50%.

The empty vector treated 3T3 DNA assayed with the Pms1-JD2.2 NGS assay (figure 5.13), presented a T>C point mutation in only 12% of specific reads, that was not present in guide treated cells and was far from the sgRNA position and simultaneously not overlapping the primers. This could be consistent with a mutation during replication at an early stage of PCR

amplification or background noise of the assay due to inaccurate calling of base pairs.

A very similar circumstance is behind the background modification of the control sample for Mlh1-A1 (figure 5.12), which had an A>C modified contig and a contig with a 1bp deletion. Both were not present in guide treated samples and had low frequencies of 0.13% and 0.29% respectively.

The background mutation detected in empty vector treated 3T3 DNA assayed with the Pms1-JD2.2 NGS assay (figure 5.13), resembled a duplication and inversion of the amplicon which could be consistent with errors during preparation of the NGS assay at the core as samples from different submissions are multiplexed during their pipeline.

5.2.3 Protein validation: Western blot

DNA validation of sgRNAs was complemented with protein validation by western blot when validated specific antibodies were available. Protein levels were quantified and normalized for tubulin levels as described. Reduction in protein level of guide treated cells as compared with the best available control was calculated.

MLH1 protein quantification (figure 5.14) indicates a reduction of 70% and 77% of protein level for Mlh1-A1 and Mlh1-JD4 treated cells selected by FACS as compared to empty guide treated FACS selected cells.

These results were in accordance with a previous experiment (experiment 1) for which protein levels were also lowered for Mlh1-A1 treated cells. In experiment 1, Mlh1-A1 treated cells selected by FACS had a 72% reduction in protein levels as compared to empty vector treated non-FACS selected cells. In this experiment Mlh1-A1 treated cells not selected by FACS also showed a reduction in protein levels as compared with empty vector treated cells, though of only 30%.

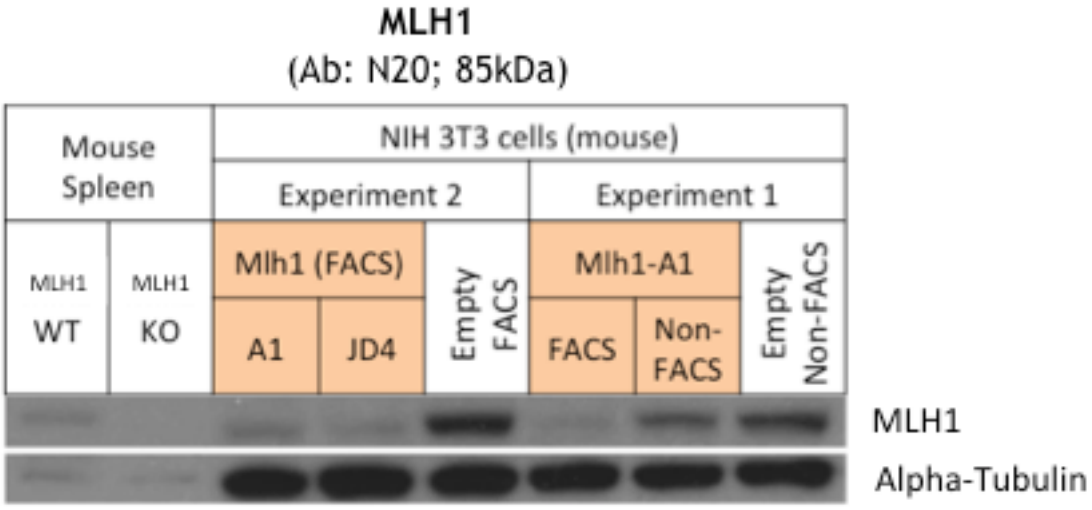

Figure 5.14 *In vitro* guide validation by Western Blot: MLH1. Protein was extracted from NIH/3T3 cells transfected with pX458-guide contructs and FACS sorted (or not) for GFP. 40µg/lane of protein were resolved in a 12% polyacrylamide gel, wet transfered to a membrane and probed for MLH1 with N20 antibody (1:200). Results for samples from 2 independente experiments and transfections (1 and 2) are presented. Protein lysate from tissue of wild type and knock out mice for *Mlh1* is presented as a control to better identify MLH1's specific band. In the pannel above, the results for the western blot probed for MLH1 protein are presented. Alpha-tubulin was also probed for each sample for normalization (pannel below).

For *Msh3* a total of 4 guides were compared, JD2.1, JD2.3, A1 and A2 (figure 5.15). When compared with empty vector treated cells selected by FASC, cells treated with these guides and selected by FACS showed a reduction in protein levels of 62%, 79%, 70% and 77% respectively. From the 2

most promising guides in term of protein analysis, Msh3-JD2.3 and A2, the first one has a better off-target ranking.

Samples from cells treated with guides for *Msh2* and *Mlh3* were also analysed for MSH3 expression levels. While *Mlh3* guide treated cells did not have a reduced MSH3 expression compared with the empty vector control, for Msh2 a 68% protein level reduction was detected.

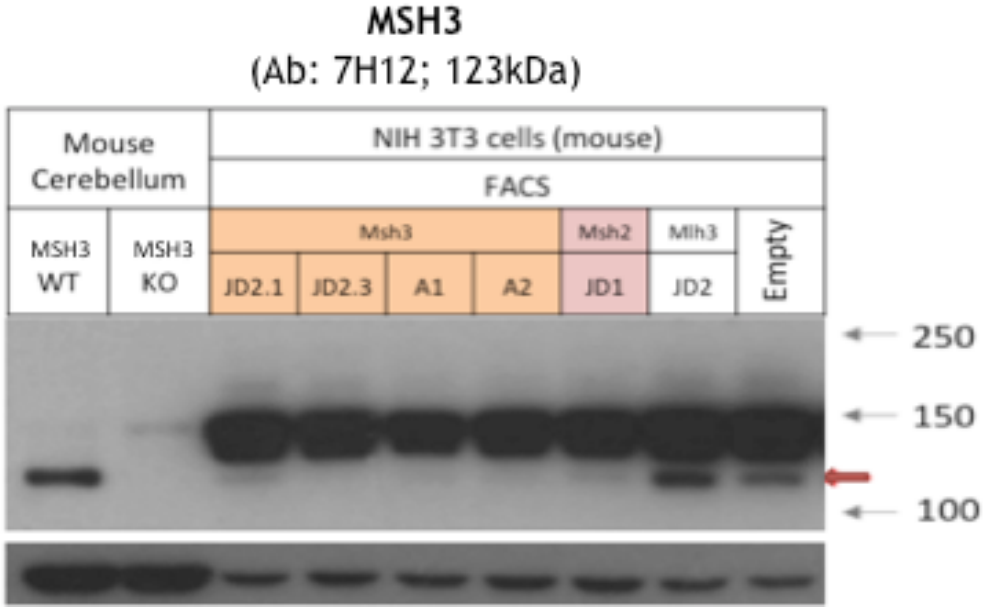

Figure 5.15 *In vitro* guide validation by Western Blot: MSH3. Protein was extracted from NIH/3T3 cells transfected with pX458-guide contructs and FACS sorted for GFP. 50µg/lane of protein were resolved in a 3-8% polyacrylamide gel, wet transfered to a membrane and probed for Msh3 with 7H12 antibody (1:300). Protein lysate from tissue of wild type and knock out mice for Msh3 is presented as a control to better identify Msh3's specific band. In the pannel above, the results for the western blot probed for MSH3 protein are presented. The molecular weight corresponding to MSH3's specific band is indicated by a red arrow. Alpha-tubulin was also probed for each sample for normalization (pannel below).

Looking at MSH2 protein levels, Msh2-JD1 treated cells selected by FACS showed a 89% reduction compared to the empty vector treated cells. There was no marked reduction in MSH2 protein levels for *Msh3* and *Mlh3* guide treated conditions.

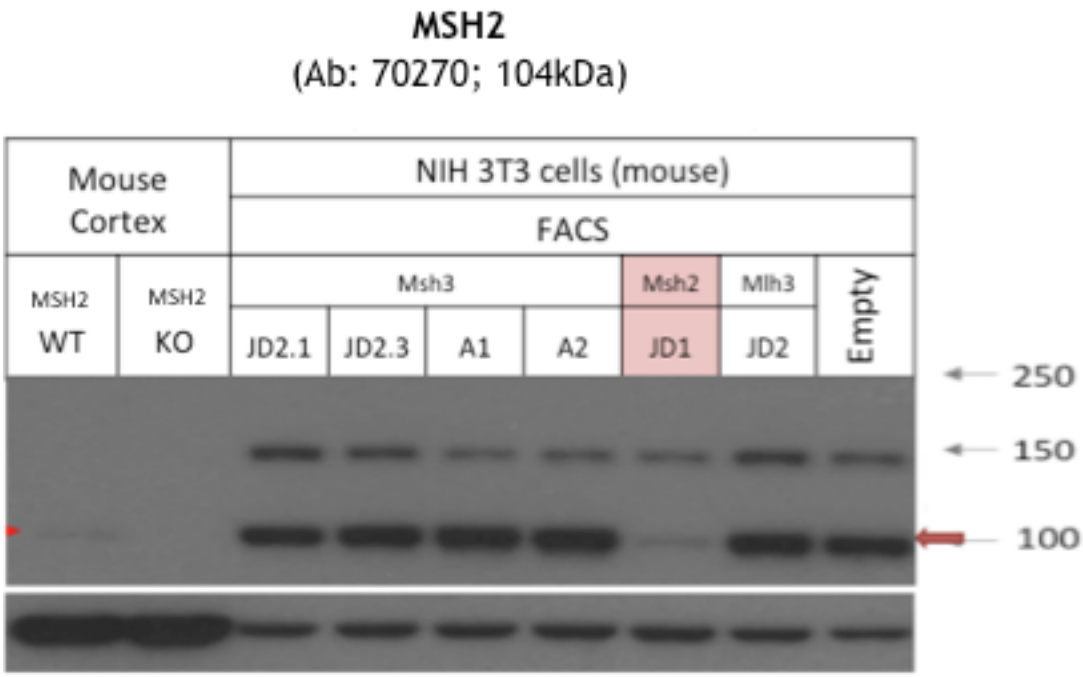

Figure 5.16 *In vitro* guide validation by Western Blot: MSH2. Protein was extracted from NIH/3T3 cells transfected with pX458-guide contructs and FACS sorted for GFP. 50µg/lane of protein were resolved in a 3-8% polyacrylamide gel, wet transfered to a membrane and probed for MSH2 with 70270 antibody (1:1000). Protein lysate from tissue of wild type and knock out mice for *Msh2* is presented as a control to better identify MSH2's specific band. In the pannel above, the results for the western blot probed for MSH2 protein are presented. The molecular weight corresponding to MSH2's specific band is indicated by red arrows. Alpha-tubulin was also probed for each sample for normalization (pannel below).

No strong correlation was found when comparing the *in vitro* quantified CRISPR induced protein reduction levels with the in sillico predictions for on-target efficiency using JD version 2. (figure 5.17) Overall, for an N=8 of measurements, guides with a predicted on-target efficiency above 50% induced a consistent above 60% reduction in protein levels, compared with empty vector treated samples. It should be noted that only FACS selected samples have been analysed by western blot so far.

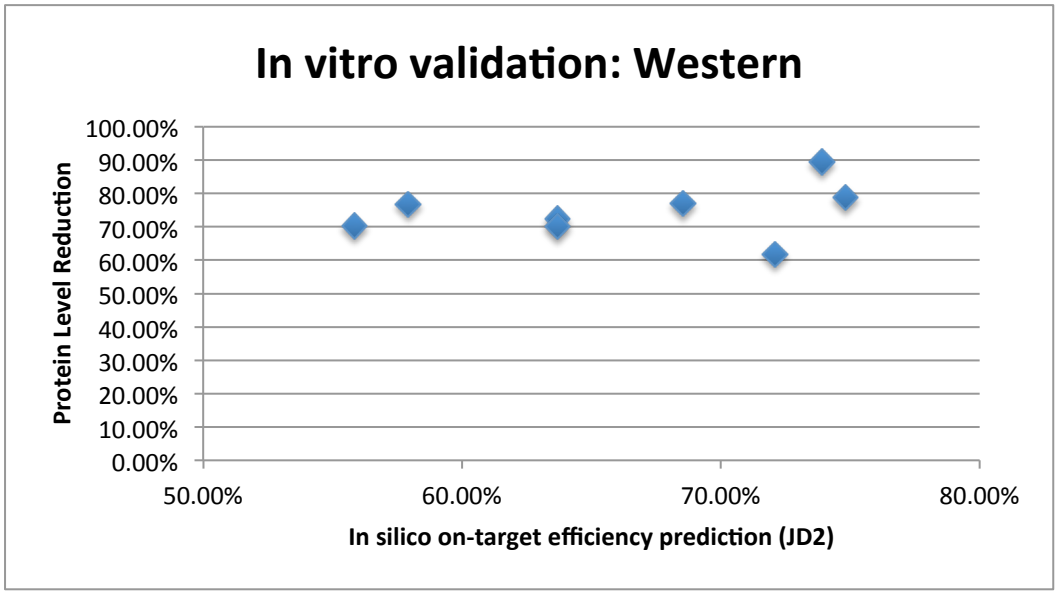

Figure 5.17 *In vitro* guide validation by Western and *in silico* predictions. All guides tested by Western Blot assay were plotted in terms of their *in silico* predicted and Western assessed protein level reduction.

5.3 GENERATING AAV FOR *IN VIVO* CRISPR

The pAAV construct was confirmed by restriction mapping (figure 5.18) and whole plasmid sequencing. The guides Mlh1-A1 and Mlh1-JD4 were cloned into the pAAV construct and their correct cloning successfully validated by sanger sequencing. The validated pAAV-guide constructs were maxipreped and validated in terms of ITR integrity (figure 5.19). The constructs were successfully validated *in vitro* in terms of mCherry expression (figure 5.20).

The validated maxipreped constructs were submitted for AAV8 production.

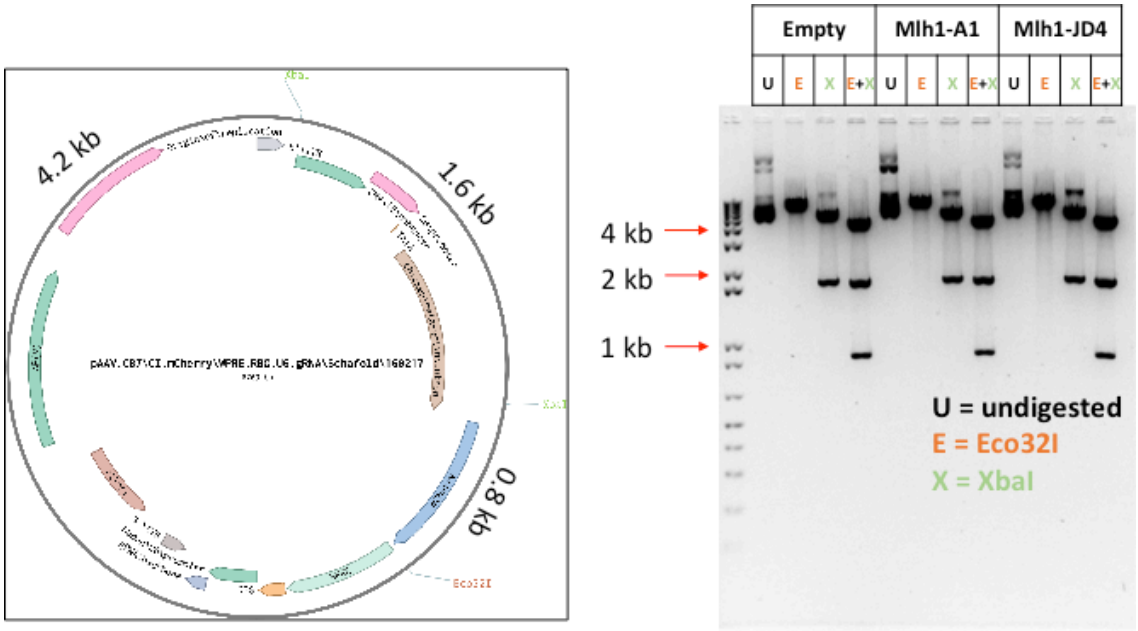

Figure 5.18 Restriction mapping of pAAV constructs. pAAV-guide constructs were diggested with 2 specific restriction enzymes, Eco321 and Xba1, separately and simultaneously. Digestion products were resolved in a 1% agarose gel and found to match the expected molecular weights for a successful assembly of the construct.Undigested product was used as a negative control for the digestion.

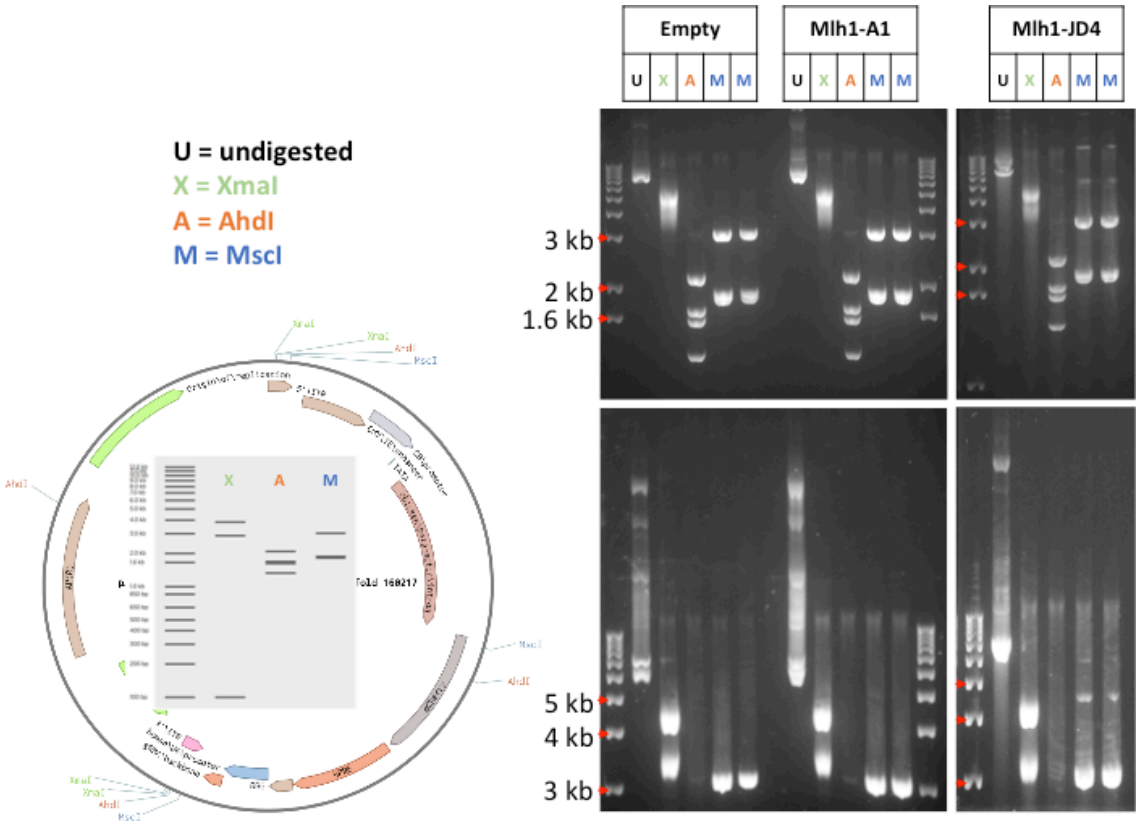

Figure 5.19 ITR integrity validation. pAAV-guide constructs were digested with 3 restriction enzymes specific for sequences within the ITR, XmaI, AhdhI and MscI. The digestion products were resolved in a 1% agorese gel and found to be successfully cut by these restriction enzymes indicating that the sequences they recognize in the ITR are intact. For Xma1 a longer resolution was required in order to detect its digestion products. Molecular weight of the digestion products matched *in silico* expectations. Undigested product was used as a negative control for the digestion.

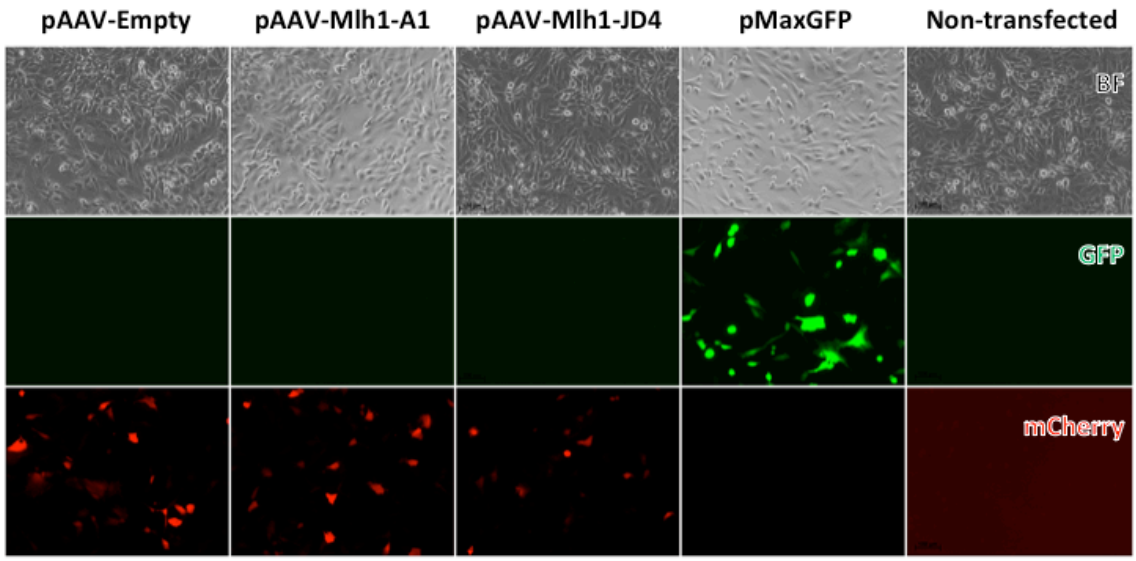

Figure 5.20 *In vitro* validation of pAAV induced mCherry expresssion. NIH/3T3 cells transfected by lipofection with pAAV-guide constructs were assessed in terms of fluorescence at 48h. Red channel fluorescent cells consistent with mCherry expression were found for the pAAV-guide and pAAV-empty transfected cells. The low percentage of fluorescent cells could derive from a low transfection efficiency as pMaxGFP transfected cells, used as a control for transfection, also had a low transfection rate as assessed by the overall number of cells fluorescent at wavelengths consistent with GFP. Non-transfected cells, used as a negative control did not present fluorescence.

5.4 *IN VIVO* VALIDATION OF CRISPR Constructs

In vitro validated guides Mlh1-A1 and Mlh1-JD4, successfully cloned to pAAV and packaged in AAV8 were delivered by tail vein injection (TVI) and intraperitoneal injection (IP) to Rosa26-Cas9 mice, heterozygous for Cas9. (table 5.3) Mice were sacrificed 10 days post injection and a liver sample collected, processed and analysed by NGS.

Table 5.4 Mice used in preliminary *in vivo* CRISPR experiments. Mice from 3 age groups, 2, 7 and 9 months, heterozygous for constitutive CAS9 expression were treated with *Mlh1* sgRNA's A1 and/or JD4, by tail vein injection (TVI) or intraperitoneal injection (IP). PBS treated wild type mice were used as negative controls. Mouse 9 was not included in this experiment.

Mouse #	Sex	Age (months)	Cas9	Delivery Route	Treatment
1	M	7	HET	TVI	Mlh1-A1
2	M	7	HET	TVI	Mlh1-JD4
3	M	7	WT	TVI	PBS
4	M	9	HET	TVI	Mlh1-A1
5	M	9	HET	TVI	Mlh1-JD4
6	M	9	HET	TVI*	Mlh1-JD4
7	M	2	HET	TVI	Mlh1-A1
8	M	2	HET	TVI	Mlh1-JD4
10	M	2	HET	IP	Mlh1-A1 & Mlh1-JD4

Both for Mlh1-A1 and Mlh1-JD4 treated mice, an above 50% rate of modification of the target DNA was achieved. CRISPR induced modification was target specific, with Mlh1-A1 mice presenting mutations in the Mlh1-A1 DNA target (NGS assay A1) but not on the Mlh1-JD4 DNA target (NGS assay JD4). The same was true for Mlh1-JD4.

Mlh1-A1 treatment delivered by tail vein injection was able to successfully mutate mice with an age of 2, 7 and 9months at the time of injection, respectively mice numbers 7, 1 and 4.

Delivery of this guide by tail vein injection induced similar rates of total and frame shift mutations in the 3 mice: 51%, 53% and 58.37% of total

mutation rate for the mice aged 2, 7 and 9 months respectively. Over 85% of mutations were frame shift mutations.

Mlh1-JD4 treatment delivered by tail vein injection was equally able to successfully mutate mice with an age of 2, 7 and 9months at the time of injection, respectively mice numbers 8, 2 and 5.

As for Mlh1-A1, delivery of this guide by tail vein injection induced similar rates of total and frame shift mutations in the 3 mice: 73%, 70% and 66% of total mutation rate for the mice aged 2, 7 and 9 months respectively. Over 97% of mutations were frame shift mutations.

Mouse 6, 9months of age at the time of the experiment, was injected with a low ill-defined amount of Mlh1-JD4 virus and was thought to have been unsuccessfully injected. However, when tested by NGS, this mouse was also successfully mutated, at similar levels with the other JD4 treated mice (74% of total mutation rate).

AAV delivery by intraperitoneal injection was attempted on mouse 10. Simultaneous treatment with both Mlh1-A1 and Mlh1-JD4 AAV successfully produced mutations in the mouse's liver with a total mutation rate of 56% for both guides, over 90% of which were frame shift mutations. As ill-defined amounts of Mlh1-A1 and Mlh1-JD4 AAV were delivered it is not possible to directly compare IP and TVI delivery in terms of liver cell mutation efficiency.

No induced mutations were found in the liver sample from mouse 3, a wild type mouse treated with PBS by TVI.

Unspecific reads accounted for less than 3% of total reads per each sample in Mlh1-A1 NGS assayed *in vivo* samples and less than 1% for Mlh1-JD4 NGS assayed *in vivo* samples.

As for some NGS assays *in vitro*, there were some very rare specific mutations detected in controls, in this case only for the Mlh1-A1 NGS assay. Mouse 2 (7months of age; JD4 treated by tail vein injection) presented a 0.37% of reads with a 1bp deletion. Mouse 6 (9months of age; JD4 treated by a tail vein injection, of ill-defined dosage) presented 0.56% of reads that had indels near the guide sequence consistent with Mlh1-A1 induced mutations, making it the only assay on a control sample either *in vitro* or *in vivo* in this thesis to present such type of mutations. It should be noted that this particular assay had 498k specific reads, a total of 648k reads combined with the assay for Mlh1-JD4 for the same mouse it was pooled with during NGS submission. NGS submissions typically yield 40k reads (values typically range from 30k to 50k reads).

NGS Assay	Treatment	Delivery	Mouse #
A1	A1	TVI	1
	JD4	TVI	2
	PBS	TVI	3
	A1	TVI	4
	JD4	TVI	5
	JD4	TVI*	6
	A1	TVI	7
	A1 + JD4	IP	10
JD4	A1	TVI	1
	JD4	TVI	2
	PBS	TVI	3
	A1	TVI	4
	JD4	TVI	5
	JD4	TVI*	6
	JD4	TVI	8
	A1 + JD4	IP	10

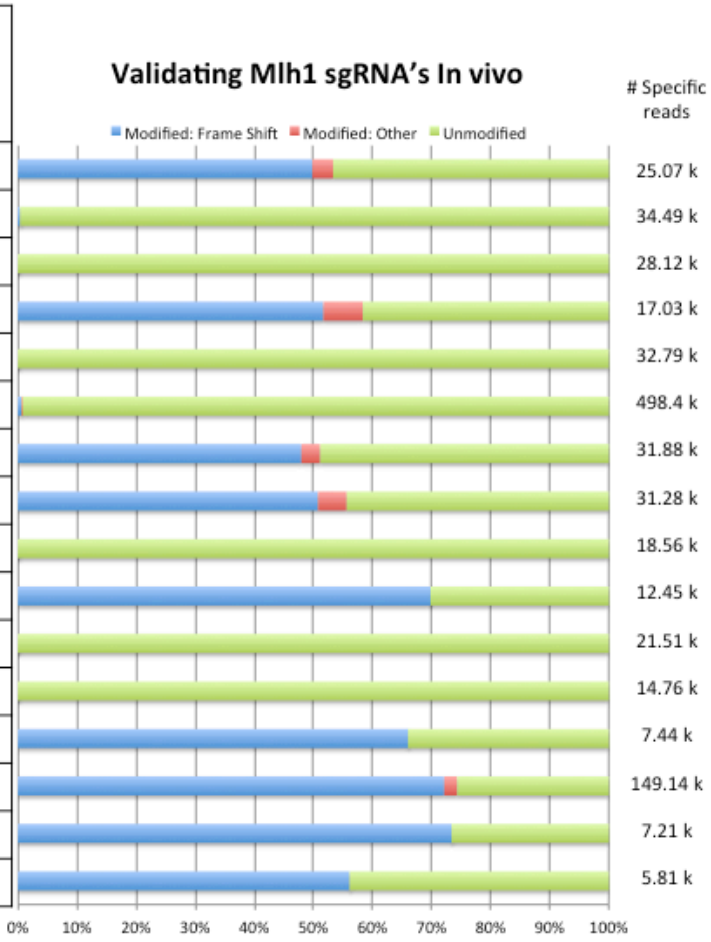

Figure 5.21 *In vivo* guide validation by NGS: Mlh1. Mice described in Table 5.3 were treated with Mlh1-A1 and/or Mlh1-JD4 AAV8 and liver tissue samples collected at day 10 post-treatment. For each mouse, DNA was extracted, amplyfied and assayed both for Mlh1-A1 and Mlh1-JD4 (except for mice 7 and 8, tested only for A1 and JD4 respectively). Mice treated by tail vein injection(TVI) received a total of 3x10^11 viral particles of AAV8. Mice treated by TVI* or IP received a ill defined dose of viral particles. Each NGS assay was characterized in terms of frame shift and non-frame shift induction modification rate. For each assay, the number of specific reads (that map to the region of interest) are presented.

5.5 SUMMARY OF PROGRESS SO FAR IN SGRNA VALIDATION FOR KNOWN AND CANDIDATE MODIFERS

Table 5.5 Summary of progress so far in sgRNA validation. Stage of validation for all sgRNA's being tested.

GENE	GUIDE ID	CLONED	VALIDATED		AAV	VALIDATED *IN VIVO*
			NGS ASSAY	PROTEIN		
MLH1	**A1**	**+**	**+**	**+**	***+ (IN AAV8)***	***+***
	JD4	**+**	**+**	**+**	***+ (IN AAV8)***	***+***
MSH3	**JD2.3**	**+**	***+***	**+**	***CLONED TO PAAV***	
	A1 & A2	**+**		**+**		
	JD2.1	**+**		**+**		
	JD1	**+**				
MSH2	**JD1**	**+**	***+***	**+**	***CLONED TO PAAV***	
MLH3	**JD2**	**+**	**+**	***FAIL****	***CLONED TO PAAV***	
FAN1	**JD2.1 &JL1.8**	**+**	**+**			
	JD2.2	**+**	**FEW READS**			
	JL1.4	**+**	***IN PROGRESS***			
RRM2B *PMS2* *PMS1* *EXO1* *MSH6* *FANCJ*	**JD2.1 & JD2.2**	**+**	**+**			
ERCC3	**JD2.1**	**+**	***+***			
	JD2.2	**+**	***IN PROGRESS***			

***NO SPECIFIC ANTIBODY AVAILABLE**

5.6 CAG INSTABILITY *IN VIVO IN CONSTITUTIVE CAS9 EXPRESSING MICE*

Somatic instability of the CAG repeat mutation in the humanized genetic region of Htt exon1 was assessed in a total of 7 Q111 mice at the age of 7 an 9 months with different Cas9 genotypes. A representative panel of gene scans depicting the instability phenotype for different CAS9 genotypes is presented in figure 5.20.

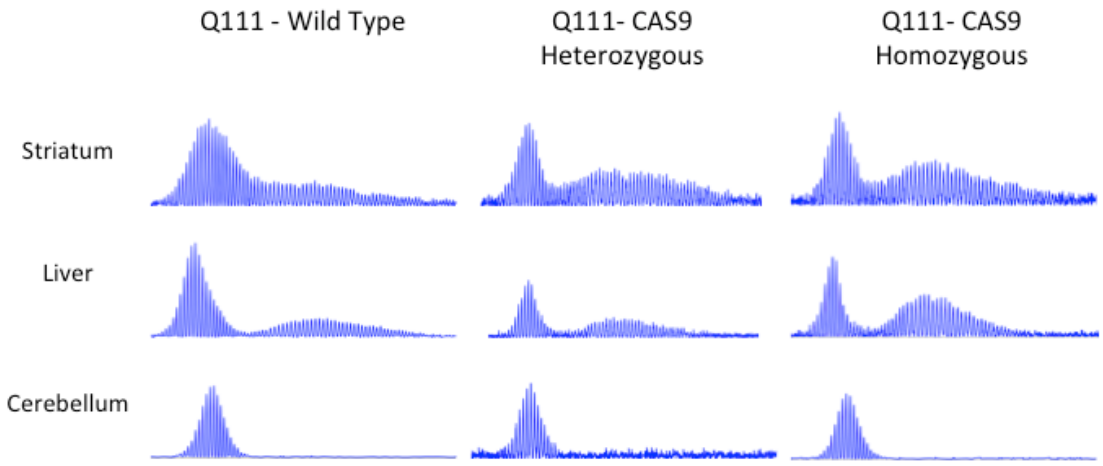

Figure 5.22 Somatic CAG instability in Q111-Cas9. A representative panel of gene scans depicting the instability phenotype for different CAS9 genotypes is presented based on results from 9 month old mice 3 (wild type), 2(heterozygous) and 1(homozygous). For each genotype, striatum, liver and cerebellum gene scans are presented.

Somatic CAG instability, suggested by the presence of a variable length of CAG repeats, was found to be present in unstable tissues (striatum and liver) of Q111 mice for all Cas9 genotypes (wild type, CAS9 heterozygous and CAS9 homozygous). The cerebellum, a stable tissue for this phenotype, was also found to be stable for all CAS9 genotypes.

A finer comparison between genotypes and the 2 age groups is on going through proper quantification of gene scans.

6. DISCUSSION

In this section, the progress achieved at different stages in the design and validation of reagents for the identification of genetic modifiers of somatic CAG instability in Huntington's disease is discussed.

6.1 Guide design: on-target and off-target prediction

For the purpose of the current thesis, of developing a platform for higher throughput *in vivo* screening of somatic CAG repeat instability, the sgRNA on-target efficiency is the most important criterion, as modifier gene disruption in a low percentage of cells could mask biologically relevant effect sizes in somatic instability modification. As instability modifiers were mainly tested in conventional knock out models, in which protein expression is disrupted even before birth, it is hard to predict the effect size that a known modifier will have when it is only disrupted after birth and only in transfected cells. For known modifiers such as *Mlh1* and *Mlh3*, changes in the instability phenotype were substantially more pronounced in homozygous knock out mice compared to heterozygous knock out mice (Pinto et al. 2013) further indicating the prudence of aiming for a highly on-target efficient guide. Particularly for the initial validation of the platform and this type of models, in which many methodological issues such as vector delivery are still being optimised, it is therefore critical to have the highest possible sgRNA on-target efficiency.

Guides designed and chosen for further validation were thus picked based on the best *in vitro* evidence backed predictions available at the time for on-target loss of function mutation inducing efficiency.

Off-target efficiency should, however, not be disregarded as modification at off-target sites, even if less efficient, could also affect the measured phenotype, leading to inconclusive results.

This consideration is particularly relevant for studies using multiple guides (therefore more off-target sites to control for) and with more

constraints to guide selection such as a narrow target sequence (less flexibility to select guides with low off-target efficiency). In this regard, the current study using only one sgRNA at a time and having the flexibility to choose any promising target site within the gene of interest seems comparatively less vulnerable to off-target bias.

One aspect of the model that should be considered in terms of expected on and off target efficiency is the fact that the Cas9 mouse line selected constitutively expresses Cas9. In the presence of sgRNA's, DNA is thus expected to be continuously exposed to the CAS9-guide complex increasing the chance that even low predicted efficiency off-target sites might be modified.

Off–target efficiency was predicted using the most recent version of the Broad's sgRNA designer tool that enables the simultaneous ranking of guides in terms of on- and off-target efficiency. Off–target efficiency ranking considerations could therefore be incorporated on the choice of more recently designed guides. (Doench et al. 2016)

As the off-target rank in this model is presented not as an absolute metric but rather relative to the other sgRNAs targeting the same gene it is hard to determine the weight it should have when picking a new guide. Considering the dataset used to generate the off-target ranking model was based on *in vitro* studies with human cell lines it is also not possible to directly infer the extent of its reproducibility in predicting off-target sites *in vivo* mouse studies. (Doench et al. 2016)

In on going and future studies, off-target cut sites will be sequenced to ensure they are not a confounding factor, so as to better interpret results.

Several approaches could be pursued to further confirm a positive result from an *in vivo* crispr study. One such approach would be to test reproducibility using a different guide targeting the same gene, as it would be much more likely that the difference in phenotype produced in both instances would be due to the specific high efficiency disruption of the gene of interest rather than to the low efficiency disruption at guide specific off-target sites.

A follow up experiment with a double nicking strategy could also be employed to minimize off-target activity.(Chiang et al. 2016) In this strategy, an adaptation of Cas9, the D10A mutant Cas9 nickase (Cas9n), is used to induce only single strand breaks that do not result in indels during repair. A pair of appropriately spaced and oriented sgRNAs targeting the same locus can produce in this strategy double stranded breaks leading to indels during repair by non-homologous end joining. The double stranded breaks would be

very specific of the targeted locus as off targets would have to be recognised by both guides in the correct spacing and orientation to also be subjected to double strand breaks and indels. (Ran et al. 2013) The double nickase strategy would however possibly not be the best-suited strategy to initially screen several new candidate modifiers as its added guide design restrictions make it harder to find high on-target efficient guides.

Another option to reduce off-targets would be to use a mouse model expressing a more faithful rationally-engineered Cas9 (Slaymaker et al. 2016)(Chiang et al. 2016)

To summarise, the on- and off-target prediction models used for guide design in this study do not directly relate to efficiency in *in vivo* mouse models but they are the best evidence based *in silico* tools available to select guides that might be more effective and specific in general, and hopefully also in this particular study. For the purpose of the current study, particularly during the first preliminary experiments, on-target efficiency was the main selection criterion as a higher cell mutation rate is expected to maximize detectable changes in phenotype. Off-targets could affect result interpretation but can be controlled for in follow up experiments, not being as critical in an initial screening stage.

6.2 Guide design: cut site position within the targeted gene

In the present thesis sites to target within particular candidate modifier genes were selected so as to most confidently induce loss of function mutations.

The chosen approach was to design sgRNA guides targeting the upstream portion of protein coding sequences such that CRISPR induced frame shift mutations might compromise as many downstream coded protein elements as possible and therefore impair protein function.

While targeting protein coding sequences within the 5-65% of the associated protein length is recommended in the literature as a general guidance for this purpose (Doench et al. 2016), it does not account for a lot of gene and cell type specific variability for instance due to local chromatin structure and its impact on CAS9 access to PAM sequences (Wu et al. 2014) and due to not all exons from the reference protein isoform being expressed in specific cell type relevant isoforms (Doench et al. 2014). It is therefore common to test several guides trying to cover different gene regions so as to have a greater chance of targeting a cell type relevant and accessible gene region. (Doench et al. 2016)

In this thesis, in accordance with the insight from the literature, a general 5-65% rule was used as guidance, and exceptions made for particularly promising guides, namely for Msh2-JD1 which turned out to be very successful in terms of gene modification rate and protein level lowering. The JD version 2 algorithm incorporates a 5% spacing between target sites of picked guides which further adds to target gene region diversity and greater chances of successful gene disruption.

6.3 *In vitro* validation: evolving methodology for a higher throughput in selecting transfected cells

For low transfection efficiencies, a great sensitivity is required in order to detect induced changes as transfected and potentially modified cells are diluted within all the non-transfected cells. Selecting just the transfected cells is thus key in order to confidently detect induced mutations in these cells.

For instance for Mlh1-A1 treated cells, a CRISPR induced decrease in protein levels was considerably more evident when transfected cells were selected, in this case by FACS, than when they were not selected.

Ideally selection of transfected cells would just exclude non-transfected cells from the total cell population. However, depending on the method used for selection, this process will also have other impacts on the new resulting cell population.

Using FACS sorting, a low number of cells that harbour the fluorescent reporter-expressing vector is selected and then cultured to grow for several cell divisions until enough cell material is generated for further testing. Selecting a low number of cells, implies that these might not be representative of the original population of transfected cells but rather a randomly selected set of these. Though FACS can select cells based on fluorescence above thresholds found to exclude control samples there can always be some residual carry over of non-transfected cells.

Using antibiotic selection, such as puromycin selection, the cell population suffers a strong selective pressure, with transfected cells carrying the resistance conferring vector having a much greater chance of survival. After the treatment period, the cell population will be composed of a fraction of the vector carrying cells that remained viable and possibly a much rarer subpopulation of surviving non-transfected cells if the antibiotic dosage is not strong enough. The fraction of surviving cells might not be an entirely random sampling of the transfected cell population, but possibly also biased by advantages or disadvantages conferred by different CRISPR induced mutations. As the total cell population is severely reduced in numbers during selection, the resulting cells are again cultured to grow for several cell divisions until enough cell material is generated for further testing.

Expansion of a heterogeneous population of cells, regardless of selection methods used prior or not, is inherently a selective process as cells

with different types of mutations will have different selective advantages in terms of multiplication rate and survival. This consideration becomes particularly relevant in the context of mutations in genes known to be associated with cancer and cell division as is the case of the DNA repair associated genes being targeted in this thesis.

Using a naturally bias inducing selection method followed by cell expansion that adds to the bias, thus produces a cell population that while enriched for transfected cells is not necessarily representative of the initial population in terms of induced mutation rate and relative frequency of different mutations.

The variability that characterises the transfection and selection process, is captured in the considerable 21% difference in total mutation rate for cells transfected with Mlh1-A1 and selected by FACS from 2 independent transfections.

This variability, together with differences in the nature of the chosen selection method, might explain the even more striking difference in mutation rates detected between Mlh1-A1 treated cells selected by FACS and cells treated with the same guide and selected by puromycin. Given different constructs pX458 and pLentiCrispr_v2 were used for each selection method, to deliver Cas9 and the sgRNA, differences in cell transfection efficiency and in Cas9 and sgRNA expression levels could also contribute to the measured difference in mutation rate.

Mlh1-A1 treated cells selected with puromycin presented 100% of mutation rate when analysed by NGS. This is a very high mutation rate not predicted *in silico* or *in vitro* using FACS as the selection method. This result could be due to a strong advantage of mutated cells in terms of proliferation, and their consecutive selective enrichment in the cell population over time and passages. Such a high mutation rate might limit the ability to detect differences between guides in terms of on-target efficiency. However, it does not hinder the ability to validate the presence of successfully CRISPR induced mutations, and thus of validating guides as effective *in vitro*.

The main priority of *in vitro* experiments within the context of this thesis is to validate reagents capable of inducing targeted mutations rather than differentiating between sgRNAs with similar on-target efficiencies, as sgRNA relative efficiency *in vitro* does not necessarily translate to the *in vivo* context in which they must also be validated in a preliminary stage of *in vivo* experiments. Puromycin selection, having been optimized for this cell line and protocol and having successfully validated Mlh1-A1 was thus adopted for

more recent experiments as it enables a higher throughput by being more scalable and by removing time restrictions and costs associated with using shared equipment.

For the following batch of guides *in vitro* validated using puromycin selection (targeting the candidate modifiers), transfection was performed in a bigger well format (6well plates as opposed to 24 wells) so as to have more cells at the end of the antibiotic selection, thus requiring less cell divisions in order to have enough cell material for further analysis. Reducing the number of required cell divisions before analysis, other than hastening sample acquisition, should also allow for a less pronounced drift from the original frequency of the different induced mutations and for lower chances of having efficiencies too high to be able to differentiate between guides. In the new batch of *in vitro* validations, in fact, no guide had a 100% efficiency detected, but it is not possible to establish a direct comparison as different guides were tested. Mlh1-A1 was also transfected in this more recent batch of validations for comparison but has not yet been quantified. Reproducibility of different *in vitro* assays and metrics in different transfections for the same guide using the puromycin pipeline should be tested so as to be able to better interpret future results. Quantification of the mutation rate for consecutive passages would enable a better understanding of the enrichment in mutated cells over time, however it would only yield gene specific data, as mutating different genes is expected to have different effects of selective advantage in this context.

The consecutive methodological adaptations during this thesis in terms of transfected cell selection greatly increased the sgRNA *in vitro* validation throughput, from an initial guide by guide analysis stage to the more recent batch of 16 guides validated in a single experiment.

6.4 *In vitro* validation: analysing a growing number of CRISPR treated samples

A total of 24 sgRNA's were *in vitro* validated during this thesis. At an early stage, while sgRNA's targeting known modifiers were being developed and tested, *in vitro* validation was performed on a matched almost guide by guide pace. Since then, as discussed, throughput has increased in terms of producing guide transfected and selected samples for analysis. The throughput in analysing these samples was similarly increased by adaptation of the *in vitro* validation pipeline based on accumulated experience. These adaptations enabled a more efficient analysis of the recent 16 guide experiment.

6.4.1 T7 assay

Initially sgRNA's were validated by T7 assay analysis of DNA from *in vitro* treated cells, as detection and quantification of CRISPR indused indels by this method is well described in the literature (Chiang et al. 2016)

The T7 assay produced specific results with control DNA samples from a kit and suggested the validation of Mlh1-A1 and Msh3-JD2.3 while being inconclusive for Msh3-JD2.1. This method was found to be hard to interpret due to its high level of background bands in controls as well as guide treated samples.

These bands could derive from unspecific cleavage by the T7 enzyme, though this level of unspecific cleavage was not apparent in previous experiments when validating the enzyme using other samples.

They could also derive from the cleavage of real mismatched duplexes formed during annealing as a consequence of the presence of polymorphisms in the targeted locus or of contaminant unspecific PCR products mapping to other parts of the genome.

The former hypothesis can be probed considering the NGS results for part of these same samples (Mlh1-A1, Msh3-JD2.3 and their empty controls) and considering NGS results overall.

The presence of polymorphisms such as SNPs was not found for either the Mlh1-A1 or Msh3-JD2.3 samples analysed by T7, nevertheless they should

not be disregarded as SNPs can be present in other target loci as is the case for instance of Msh2-JD1.

Contaminant unspecific PCR products were found in both Mlh1-A1 and Msh3-JD2.3 samples even after gel purification, which could explain the presence of the background bands. Considering the direct comparison of gel purified and column purified samples for Msh3-JD2.3 it is clear that gel purification reduces contaminant concentration but does not necessarily eliminate all contaminants, especially those with a molecular weight matching that of the amplicon of interest.
Mutations during replication in the PCR amplification of these products could also have contributed to the formation of heteroduplexes but are partly prevented by the use of a high fidelity polymerase.

Considering the 21 guides validated by NGS, almost all have either polymorphisms or unspecific contaminant PCR products which indicates the limitations of the T7 assay in this context.

One can therefore infer that the T7 assay is not optimal for confident indel detection and quantification, when using this cell line (NIH/3T3), that harbours polymorphisms, and this primer design algorithm (primer-blast), high fidelity polymerase (Phusion; thermo scientific) and PCR product purification strategy (gel purification) that do not confidently guaranty exclusion of unspecific PCR products.

The T7 assay is also not ideal for a higher throughput *in vitro* validation pipeline as it implies gel purification, which can be time consuming, and assay optimization for primers whose amplicon (about 1000bp long to allow asymetric cleavage products to be well resolved and detected in a standard agarose gel electrophoresis) is too long to be compatible with the current NGS assay criteria.

This enzyme cleavage assay, other than not being ideal at indel detection in this system, also does not replace NGS validation as it is less informative of the types of mutations present. For instance it cannot differentiate frame shift and non-frame shift mutations which have different expected potentials in terms of impact on protein function. For these reasons, NGS analysis is the quantitative and informative method of choice to evaluate candidate modifier mutations in the *in vivo* stage of the project, independently of the strategy for *in vitro* validation.

Given the high rate of success in *in vitro* validation using the current guide design strategy and the limitations of the T7 assay in reliably pre-screening guides, a strategy of directly analysing guide on-target efficiency by

NGS was adopted. Pre-screening of samples by sanger sequencing was performed so as to have a preliminary validation prior to NGS submission.

6.4.2 NGS

Quantification and characterization of CRISPR induced mutations by NGS, was initially performed by submitting gel purified PCR products corresponding to sgRNA targeted genome regions, one PCR product per submission.

Direct comparison by NGS of gel and non gel purified samples for Msh3-JD2.3 and Msh2-JD1 treated samples, indicates that the NGS assay is resilient to unspecific noise reads, that can be easily excluded during the analysis of results. The increase in unspecific reads is associated with a reduction of the total number of specific reads, but even a reduction of 38% of the number of reads does not significantly impact the ratios of different types of mutations. Considering these results, gel purification of samples prior to NGS submission was supressed.

From the previous experiment, one can infer the average 40k reads per submission might not be required for an accurate relative quantification of different types of mutations. It was also shown that from the total population of detected contigs, subpopulations of contigs could be selected and analysed independently.

The possibility of pooling multiple samples per NGS submission was thus considered, for samples that would be easily separated during contig analysis, as is the case of samples amplified using a different set of primers.

So as to still retain a good read resolution for guide treated samples, these samples were only pooled in groups of 2. For control samples, mostly consisting of the reference contig and unspecific products, samples were pooled in groups of 6 prior to submission. This protocol adaptation allowed for a higher number of samples to be processed simultaneously, therefore significantly reducing the associated cost.

For the 16 guide experiment only one guide, Fan1-JD2.2, had less than 3000 specific reads, in this case only 380 reads, being considered to require a new NGS validation for better confidence in results. The failure to achieve enough reads for this particular sample can be overcome in a future resubmission for NGS of a more concentrated PCR product, thus not detracting significantly from the overall increase in throughput.

CRISPR induced mutations detected during *in vitro* validation were found to be specific, as they were not present in empty vector treated samples. The rare mutations found in controls were consistent with mutations during replication at an early stage of PCR amplification or inaccurate base calling and were found to be present only on the control sample and not the guide treated sample. For one of the control samples there was also an amplicon duplication and inversion, which could be due to an error during sample processing at the NGS core.

SNPs were identified as point mutations at frequencies close to 50% and present in both control and guide treated samples. For SNP's near the cut site it was helpful to have a control sample to compare to so as not to count the SNP as an induced mutation.

Even though the total number of specific reads differed between samples, as mutation rates are relative measurements it was still possible to compare them, for instance for guides targeting the same gene. However, the analysis of reproducibility of NGS quantified mutation rates for the same guide tested in different transfections is still on going. Suggestive differences in mutation rate might be informative towards the selection of guides to move forward to *in vivo* testing, but at this point it is impossible to know with confidence if they are reproducible or predictive of efficiency in the *in vivo* setting. Preliminary data from the *in vivo* validation of Mlh1-A1 and Mlh1-JD4, discussed in more detail below, seems to indicate differences between guides in NGS analysis *in vitro* might also apply to the *in vivo* setting.

6.4.3 Western Blot

While NGS can validate guides capable of inducing a high rate of mutations in targeted sites of genes of interest, it does not necessarily reflect the actual level of compromised protein expression as discussed during sgRNA design. Protein quantification by western blot is therefore a more informative method and gives greater confidence in the guide's potential to work *in vivo*, though it still does not guaranty reproducibility *in vivo* and in different tissues. Protein quantification by WB is also intrinsically dependent on availability of specific antibodies. It might thus not be readily available in the context of screening new less well-described candidate modifiers for instance suggested by GWAS evidence such as *Fan1*.

In the scope of this thesis, that aims to establish a platform for CAG instability genetic modifier screening it therefore makes better sense to

combine the best of both techniques, using NGS to validate guides for a wider set of genes and WB for cross validation when specific antibodies are available, for a more confident validation for specific guides.

Western blot analysis was performed for validation of *Mlh1*, *Msh3* and *Msh2* targeting guides. *Mlh3* analysis failed due to lack of specific antibodies.

An above 70% MLH1 protein level reduction was verified for Mlh1-A1 and Mlh1-JD4 treated and FACS sorted cells relative to empty vector treated and FACS sorted cells. Consistent with NGS results, Mlh1-JD4 had slightly higher efficiency in lowering *Mlh1* expression.

Protein lowering levels seem to be consistent for Mlh1-A1 across different transfection experiments, though a direct comparison cannot be established as different types of controls were used in each experiment. Mlh1-A1 induced protein lowering was substantially easier to discern when successfully transfected cells were selected by FACS, indicating this could be a critical step when validating guides. Particularly as there could be differences in transfection efficiency between conditions.

For *Msh3,* the 4 guides tested (JD2.1, JD2.3, A1 and A2) induced an above 60% MSH3 protein level reduction relative to the empty control, most pronounced for JD2.3 and A2. These results point to the potential and limitations of the *in silico* JDv2 predictions as they could overall predict these guides as efficient but could not predict their relative ranking in terms of efficiency. The results also point to the already mentioned limitations of the T7 assay, which was inconclusive when assessing the now protein validated Msh3-JD2.1 sgRNA.

When analysing MSH2 protein levels, Msh2-JD1 guide treated cells presented a high MSH2 protein level reduction, above 80% reduction relative to the empty construct treated cells. MSH2 expression levels were not substantially reduced by the *Msh3* guides, in accordance with the literature. (Halabi et al. 2012)

Msh2-JD1 guide treated cells, shown to have reduced MSH2 protein levels, presented a reduction in MSH3 protein levels comparable with that induced by *Msh3* guides. Reduction of MSH3 protein levels induced by reduction of MSH2 levels is in accordance with the literature as heterodimerization with MSH2 is thought to have a stabilizing effect on MSH3.(Halabi et al. 2012) (Genschel et al. 1998)

Mlh3-JD2 guide treated cells, predicted to have a lowered MLH3 expression from *in silico* JDv2 and NGS analysis, were not affected in terms of MSH3 or MSH2 expression relative to empty construct treated cells.

The current western blot results confirm that CRISPR induced protein level reduction is specific and specified by the sgRNA, as guides specific for other genes do not induce protein lowering, for instance Mlh3-JD2 does not affect MSH2 and MSH3 expression. An exception is made when the lowering of expression of another target gene is biologically expected to affect the levels of the protein being studied, as is the case for the *Msh2* guide treated cells whose reduced levels of MSH2 are expected to induce lowered levels of MSH3.

The indirect effect that Msh2-JD1 has on MSH3 protein levels is a strong indication of its potential to induce loss of function mutations in *Msh2*.

6.4.4 Guide design, in silico JDv2 predictions and in vitro validation

Considering the dataset of 24 sgRNA's *in vitro* validated during this thesis (4 by NGS and western blot, 17 just by NGS and 3 just by western blot), *in silico* JD version 2 on-target efficiency predictions did not strongly correlate with *in vitro* tested guide efficiency, for the studied range of 45 to 85% of on-target predicted efficiency. This was true for total and frame shift induced mutation ratios measured by NGS and for protein level reduction in western blots. This weak correlation with NGS results was verified both when performing transfected cell selection by FACS and when performing it by puromycin treatment.

Although this is a very limited dataset, and reproducibility of results for different transfections is still being validated, JD version 2 *in silico* predictions do not seem to strongly predict the *in vitro* based ranking of guides in terms of on-target efficiency.

Overall however, the sgRNA's tested in this thesis, which were designed or re-evaluated with JD version 2 tools and had an above 50% on-target predicted efficiency, were shown to have an above 50% total mutation rate and 40% frame shift mutation rate as evaluated by NGS. The guides tested by western blot also had above 60% of protein level reduction.

The design and selection of sgRNAs based on JD version 2 used in this thesis seems thus to consistently yield *in vitro* validated sgRNAs.

Since all tested guides had a predicted efficiency above the 50% threshold, it is not possible to know if this could be an important criterion. The independently designed guide Fan1-JL1.8, found to have predicted on-target efficiency just below 50%, was also successfully *in vitro* validated by NGS. It is not possible to discern if the protocol followed in this thesis for

guide design presents significant improvements over other protocols using different tools and algorithms.

In future works, the number of sgRNA's tested will be increased and reproducibility across transfections of *in vitro* tested on-target efficiency will be further evaluated. Guide *in vitro* validation could also be expanded so as to accommodate off-target efficiency, for instance by sanger sequencing of the highest ranking off-targets in exon coding sequences predicted by algorithms such as the one used by crispr.mit (http://crispr.mit.edu/).

As *in vitro* validated guides start to be validated *in vivo*, the potential of *in vitro* validation to predict successful *in vivo* validation will be assessed. The power of *in vitro* differences to predict *in vivo* differences in on-target efficiency for different guides targeting the same gene will also be assessed, so as to determine if *in vitro* results could be used to prioritize the choice of guides to be tested *in vivo*.

The ability of *in vitro* testing to predict successful *in vivo* validation will be compared to that of *in silico* JDv2 predictions. The pertinence of performing *in vitro* sgRNA testing might need to be re-evaluated if the current guide design protocol yields a rate of successful validation *in vivo* similar to that which was found *in vitro*.

6.5 *In vivo* sgRNA validation

Although *in vitro* sgRNA validation provides more confidence in the reagent's potential to induce mutations *in vivo*, *in vivo* efficiency is not necessarily expected to reflect *in vitro* results. Substantial differences in induced mutation rates are expected to stem from differences both in the sgRNA delivery system and in its efficiency to induce mutations in particular DNA targets in transfected cells.

The sgRNA *in vitro* delivery, as described, relies on the lipofection of sgRNA and Cas expressing constructs. The sgRNA *in vivo* delivery in the current approach relies in AAV delivery, which has serotype specific affinities to different tissues. The selected serotype AAV8, is reported to successfully produce extensive liver cell transfection through different routes, namely intraperitoneal injection (IP) and tail vein injection (TVI). (Wang et al. 2006)

The efficiency of the system in inducing mutations when there is successful sgRNA delivery is also expected to depend on the targeted cells, as the DNA epigenetic landscape for each locus and its acessibility by Cas9 are cell type dependent, as well as the exons present in the main protein isoforms expressed, which can lead to substantial differences in terms of success to induce mutations and of mutations to induce protein loss of function. (Doench et al. 2014)

For these reasons, the current study performed preliminary *in vivo* experiments to validate the previously *in vitro* validated guides in the context of *in vivo* gene editing of liver cells. As mentioned, liver cells were chosen as targets since, much as the striatum, they are somatically unstable tissues where CAG repeat instability can be quantified and its modification assessed. Delivery to liver cells is easier as its not protected by the blood brain barrier and the liver is also more easily dissected than the striatum. Liver cells are thus very promising to model this phenotype in higher throughput and to identify with more confidence candidate modifiers that might be validated as effective on the striatum in follow up studies.

Mlh1 guides A1 and JD4 successfully cloned to pAAV and packaged in AAV8 were able to induce an above 50% rate of modification, at day 10 after treatment, of their target DNA in liver cells of Rosa26-Cas9 heterozygous mice. This was verified for mice of 2, 7 and 9 months of age treated with 3x10^11 viral copies by tail vein delivery, as well as for a 9month old mouse treated with a lower ill-defined amount of Mlh1-JD4 AAV.

Both for A1 and JD4 treated mice, differences in total mutation rate and frame shift mutation rate did not vary substancially for the differently aged mice with a maximum difference in total mutation rate of 9% being registered. It is not possible to infer about potential differences in induced mutation rates for these guides according to age at time of treatment without increasing the N of mice.

As for *in vitro* validation experiments, JD4 produced a higher mutation rate with a higher percentage of frame shift mutations (70% mutation rate, above 94% of which frame shift, as opposed to 54%, above 85% of which frame shift for Mlh1-A1). It is not possible to determine if this *in vitro* prediction of the most promissing *in vivo* guide was a mere coincidence without comparing a greater number of guides both *in vivo* and *in vitro*.

Since for Mlh1-JD4 AAV, a mouse treated with a lower ill-defined number of viral copies had a comparable total mutation rate (74%), it is possible that the current viral dose might be optimised so as to achieve similar results with a lower viral dosage in future experiments.

Simultaneous treament of a cas9 heterozygous mouse with Mlh1-A1 and JD4 AAV by IP also achieved an above 50% mutation rate in liver cells at day 10 post-treatment (56% total mutation rate, over 90% of which frame shift for both guides). As an ill-defined viral dosage of each guide was delivered, it is not possible to directly compare mutation rate in liver cells by TVI and IP for the same guide. Future studies comparing the 2 delivery routes for a matched viral dosage will allow for a better comparison. The IP route of delivery, that requires less operator experience than TVI, could be an interesting alternative especially during preliminary *in vivo* study experiments in which the emphasis is still on guide validation rather than optimization of the mutation rate and phenotipic comparison of mutated and non mutated mice. The added advantage of the IP route of more precision in the delivered volumes by less experienced operators could enable better comparisons of dose matched treatments.

The detected mutations were specific as they were not present in the PBS injected wild type mouse, nor at significant frequency in the Cas9 heterozygous mice treated with with sgRNA's targeting a different locus. Some very rare below 0.5% 1bp mutations were found in a cas9 heterozygous mouse in its non-targeted locus, but these were consistent with mutations during replication in the PCR amplification stage, being distant from the cut site and at a very low frequency. For the non-targeted locus of another cas9 heterozygous mouse, at a frequency below 1%, indels were found near the cut

site in a manner consistent with non-homologous endjoining CRISPR induced mutations. This is consitent with contamination either at the *in vivo* stage with a very low dosage of AAV, possibly due to contact of tail puncture wounds with residues from previously injected mice, or during sample processing and analysis by cross sample contamination. Extra care should be taken in future experiments to avoid such types of contamination.

In future studies, the ability to mutate specific cell populations of interest could be further investigated. Hepatocytes could be a particularly interesting cell type to study as they have been shown to be the cell population in the liver with the most prominent CAG somatic instability phenotype (Lee et al. 2011)The possibility of studying the impact of genetic modifiers in the CAG somatic instability of this particular cell population, where a stronger effect size is expected, could be evaluated by assessing genetic modification efficiency in liver samples of treated mice enriched for hepatocytes. (Lee et al. 2011)

A better understanding of the number of successfully transfected cells and of the cell types preferentially affected could be assessed in future studies in terms of histology, either by evaluating fluorescence of fresh tissue or by probing for mCherry by immunohistochemistry.

Future studies might also explore the possibility of FACS sorting mCherry positive cells (expected to have been successfully transfected by AAV) as a method to enrich for genetically modified cells. Enriching for modified cells could be important in order to detect modifiers with lower effect sizes, as non-modified cells are expected to have no directly induced change of phenotype, therefore diluting the measurement of real genetic modification induced phenotypic changes.

Given the validation of candidate CAG instability modifiers in the liver is meant as a higher throughput way to screen for potential striatum CAG instability modifiers, future studies should also seek to validate the ability to produce targeted loss of function mutations in the striatum.

While efficient striatum transfection might be possible by stereotaxic delivery of AAV9, (Aschauer et al. 2013) pAAV packaging in AAV-PHP.B, (a recently developed recombinant capsid capable of permeating the blood brain barrier), might enable efficient striatum transfection by tail vein injection, presenting an alternative more compatible with a higher throughput when studying modifiers in the striatum. (Deverman et al. 2016)

6.6 CAG instability in constitutive Cas9 expressing mice

In a preliminary experiment the CAG somatic instability phenotype was found to be present in Q111-CAS9 mice at 7 and 9months of age, both in mice heterozygous and homozygous for CAS9. On going studies with a similar design are being performed in parallel to validate other HD associated phenotypes such as EM48 staining (Pinto et al. 2013).

It is not possible to determine if there are significant differences in instability between the different Cas9 genotypes without quantifying gene scans and comparing instability indexes for a larger N. Differences in instability between genotypes would however not compromise the ability to validate and study instability modifiers for a chosen genotype.

The clear presence of an instability phenotype in CAS9 expressing mice is very promising as this phenotype could be assessed for the *in vivo* crispr validation of genetic modifiers in on going and future experiments.

Modification of the instability phenotype in CAS9 expressing Q111 mice by *in vivo* somatic crispr knock out of known instability modifiers is expected to reflect the non-embryonic and non-developmental dependent component of the effects reported in the literature for conventional knock out models of the same known modifiers. On going studies with the *in vivo* validated Mlh1 sgRNA's will try to address this question.

Having validated the ability to modify somatic CAG instability in Cas9-Q111 mice by delivery of genetic modifier targeting sgRNAs for a known modifier, this method could be used to test the impact of new candidate modifiers suggested by GWAS or *in vitro* screening assays have in this phenotype *in vivo*. As validation *in vivo* of the growing number of candidates genetic modifiers in terms of their effect in CAG repeat instability would not be compatible with studies using conventional knock-out models, the proposed alternative could possibly be a better option to address this growing need.

The post-natal nature of the genetic intervention in this approach, makes it much closer and more relevant in terms of the context of gene therapy or of finding potential pathways to target using small molecules as it can be used to determine if intervention in the post-natal period could still be promising for therapy. Candidate modifiers of instability capable of inducing phenotypic changes *in vivo* by intervention in young and adult mice could hence be more compelling potential targets for silencing or over expression

therapeutic strategies in symptomatic patients than those identified by conventional models. Intervention at different time points of the disease progression could be simulated by sgRNA administration at different days.

At the same time, the age of onset associated polymorphisms found in humans are present since birth and could make most of their contribution to phenotypic change at this stage. A combined comparative approach with conventional and somatic *in vivo* crispr models could be promising to address such questions.

Adenovirus-mediated somatic genome editing could also enable a combined modification of several genes *in vivo* in order to better dissect phenotypic modification mechanisms. It could also be used to more directly compare the impact of different modifier and whether their effect is synergistic.

Insight from somatic CAG repeat instability in this model could potentially be translatable to some extent to other polyglutamine disorders such as SCA's, since it was shown in a independent cohort of patients that the SNP's from the HD GWAS in DNA repair genes potentially involved in somatic instability could predict age of onset in these disorders.

7. CONCLUSION

In this thesis, candidate genetic modifiers of somatic CAG instability in Huntington's disease were identified from the literature and reagents designed and *in vitro* tested for candidate validation in a faithful HD knock in mouse model through adenovirus-mediated somatic genome editing.

Reagents were successfully *in vitro* validated for a total of 12 genes. The design and validation protocol was also substantially optimized allowing for a 16 fold increase in the number of reagents *in vitro* validated at a time.

Preliminary *in vivo* validation of reagents was performed for *Mlh1*, a known CAG instability genetic modifier, so as to allow in on going experiments the validation of adenovirus-mediated somatic genome editing as a promising method for the identification of modifiers of this phenotype *in vivo*.

Future work will expand the number of reagents *in vitro* and *in vivo* validated for the study of candidate modifiers genes and evaluate their phenotypic effect on HD knock in mice to identify novel somatic CAG instability genetic modifiers. While in the scope of this project only the liver could be analysed, *in vivo* delivery of reagents will also be optimized for the HD pathology relevant striatum at a later stage.

References

American Psychiatric Association, 2013. *Diagnostic and Statistical Manual of Mental Disorders*, Available at: http://encore.llu.edu/iii/encore/record/C__Rb1280248__SDSM-V__P0,2__Orightresult__X3;jsessionid=ABB7428ECBC4BA66625EDD0E0C5AAFA5?lang=eng&suite=cobalt%5Cnhttp://books.google.com/books?id=EIbMlwEACAAJ&pgis=1.

Anon, T7E1 protocol. Available at: http://www.tools-biotech.com/image/2014/05/27/20140527101154.pdf.

Arsenic, R. et al., 2015. Comparison of targeted next-generation sequencing and Sanger sequencing for the detection of PIK3CA mutations in breast cancer. *BMC clinical pathology*, 15, p.20. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4652376&tool=pmcentrez&rendertype=abstract.

Aschauer, D.F., Kreuz, S. & Rumpel, S., 2013. Analysis of Transduction Efficiency, Tropism and Axonal Transport of AAV Serotypes 1, 2, 5, 6, 8 and 9 in the Mouse Brain. *PLoS ONE*, 8(9), pp.1–16.

Asokan, A., Schaffer, D. V & Jude Samulski, R., 2012. The AAV Vector Toolkit: Poised at the Clinical Crossroads. *Molecular Therapy*, 20(4), pp.699–708. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3321598&tool=pmcentrez&rendertype=abstract.

Bano, D. et al., 2011. Neurodegenerative processes in Huntington's disease. *Cell Death and Disease*, 2(11), p.e228. Available at: http://www.nature.com/doifinder/10.1038/cddis.2011.112.

Barrangou, R. et al., 2015. Advances in CRISPR-Cas9 genome engineering: lessons learned from RNA interference. *Nucleic acids research*, 43(7), pp.3407–3419. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv226%5Cnhttp://nar.oxfordjournals.org/content/43/7/3407.full.

Bean, L. & Bayrak-Toydemir, P., 2014. American College of Medical Genetics and Genomics Standards and Guidelines for Clinical Genetics Laboratories, 2014 edition: technical standards and guidelines for Huntington disease. *Genetics in medicine : official journal of the American College of Medical Genetics*, 16(12), p.e2. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25356969.

Bettencourt, C. et al., 2016. DNA repair pathways underlie a common genetic mechanism modulating onset in polyglutamine diseases. *Annals of Neurology*, 79(6), pp.983–990.

Boettcher, M. & McManus, M.T., 2015. Choosing the Right Tool for the Job: RNAi, TALEN, or CRISPR. *Molecular Cell*, 58(4), pp.575–585. Available at: http://www.sciencedirect.com/science/article/pii/S109727651500310X.

Bombard, Y. et al., 2009. Perceptions of genetic discrimination among people at risk for Huntington's disease: a cross sectional survey. *BMJ (Clinical research ed.)*, 338, p.b2175.

Bryant, L.M. et al., 2013. Lessons learned from the clinical development and market authorization of Glybera. *Human gene therapy. Clinical development*, 24(2), pp.55–64. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3992977&tool=pmcentrez&rendertype=abstract.

Calcedo, R. et al., 2009. Worldwide epidemiology of neutralizing antibodies to adeno-associated viruses. *J Infect Dis*, 199(3), pp.381–390. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19133809.

Carroll, J.B. et al., 2015. Treating the whole body in Huntington's disease. *The Lancet Neurology*, 14(14), pp.1135–1142.

Cetin, A. et al., 2006. Stereotaxic gene delivery in the rodent brain. *Nature protocols*, 1(6), pp.3166–73. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17406580.

Chang, R. et al., 2015. Transgenic animal models for study of the pathogenesis of Huntington's disease and therapy. *Drug design, development and therapy*, 9, pp.2179–88. Available at:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4404937&tool=pmcentrez&rendertype=abstract.

Chiang, T.W. et al., 2016. CRISPR-Cas9(D10A) nickase-based genotypic and phenotypic screening to enhance genome editing. *Sci Rep*, 6(January), p.24356. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27079678%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC4832145/pdf/srep24356.pdf.

Cong, L. et al., 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, 339(February), pp.819–822.

Craufurd, D. et al., 2014. Diagnostic genetic testing for Huntington's disease. *Practical Neurology*, 15(1), pp.80–84. Available at: http://pn.bmj.com/lookup/doi/10.1136/practneurol-2013-000790.

Crowley, S.T. & Rice, K.G., 2015. "evolving nanoparticle gene delivery vectors for the liver: What has been learned in 30 years." *Journal of Controlled Release*, 219, pp.457–470. Available at: http://dx.doi.org/10.1016/j.jconrel.2015.10.008.

Deng, Y.P. et al., 2004. Differential loss of striatal projection systems in Huntington's disease: A quantitative immunohistochemical study. *Journal of Chemical Neuroanatomy*, 27(3), pp.143–164.

Deverman, B. et al., 2016. Cre-dependent selection yields AAV variants for widespread gene transfer to the adult brain. *Nature Biotechnology*.

Divino, V. et al., 2013. The direct medical costs of Huntington's disease by stage. A retrospective commercial and Medicaid claims data analysis. *Journal of medical economics*, 16(8), pp.1043–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23789925.

Doench, J.G. et al., 2016. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology*, 34(2), pp.184–191. Available at: http://www.nature.com/doifinder/10.1038/nbt.3437.

Doench, J.G. et al., 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol*, 32(12), pp.1–13. Available at: http://dx.doi.org/10.1038/nbt.3026.

Dow, L.E., 2015. Modeling Disease In Vivo With CRISPR/Cas9. *Trends in Molecular Medicine*, 21(10), pp.609–621. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1471491415001562.

Dragileva, E. et al., 2009. Intergenerational and striatal CAG repeat instability in

Huntington's disease knock-in mice involve different DNA repair genes. *Neurobiology of disease*, 33(1), pp.37–47. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2811282&tool=pmcentrez&rendertype=abstract.

Duyao, M. et al., 1993. Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature genetics*.

Eguchi, A. et al., 2015. Liver Bid-suppression for treatment of fibrosis associated with nonalcoholic steatohepatitis. *Journal of Hepatology*, xxx, pp.1–27. Available at: http://dx.doi.org/10.1016/j.jhep.2015.11.002.

Evans, S.J.W. et al., 2013. Prevalence of adult Huntington's disease in the UK based on diagnoses recorded in general practice records. *Journal of neurology, neurosurgery, and psychiatry*, 84(10), pp.1156–60. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3786631&tool=pmcentrez&rendertype=abstract.

Ezzatizadeh, V. et al., 2014. MutLα Heterodimers Modify the Molecular Phenotype of Friedreich Ataxia. *PLoS ONE*, 9(6), p.e100523. Available at: http://dx.plos.org/10.1371/journal.pone.0100523.

Fishel, M.L., Vasko, M.R. & Kelley, M.R., 2007. DNA repair in neurons: So if they don't divide what's to repair? *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 614(1–2), pp.24–36.

Foiry, L. et al., 2006. Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. *Human Genetics*, 119(5), pp.520–526.

Freeman, A.D.J., Déclais, A.C. & Lilley, D.M.J., 2013. The importance of the N-terminus of T7 endonuclease i in the interaction with DNA junctions. *Journal of Molecular Biology*, 425(2), pp.395–410. Available at: http://dx.doi.org/10.1016/j.jmb.2012.11.029.

GBI Research, 2012. *Orphan Diseases Therapeutics in CNS to 2017 - Novel Agents such as AMR101 and ACR16 to Provide Treatment Options and Boost the Huntington's Disease Segment*,

Genetic Modifiers of Huntington's Disease (GeM-HD) Consortium, 2015. Identification of Genetic Factors that Modify Clinical Onset of Huntington's Disease. *Cell*, 162(3), pp.516–526. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867415008405.

Genschel, J. et al., 1998. Isolation of MutSbeta from human cells and comparison of the

mismatch repair specificities of MutSbeta and MutSalpha. *J Biol Chem*, 273(31), pp.19895–19901. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9677427%5Cnhttp://www.jbc.org/content/273/31/19895.full.pdf.

Goellner, E.M., Putnam, C.D. & Kolodner, R.D., 2015. Exonuclease 1-dependent and independent mismatch repair. *DNA repair*, 32, pp.24–32. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1568786415001020%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/25956862%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4522362.

Gohlke, C. et al., 1994. Kinking of DNA and RNA helices by bulged nucleotides observed by fluorescence resonance energy transfer. *Proceedings of the National Academy of Sciences of the United States of America*, 91(24), pp.11660–4. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=45291&tool=pmcentrez&rendertype=abstract.

Gusella, J.F. et al., 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*.

Gusella, J.F., MacDonald, M.E. & Lee, J.-M., 2014. Genetic modifiers of Huntington's disease. *Movement Disorders*, 29(11), pp.1359–1365. Available at: http://doi.wiley.com/10.1002/mds.26001.

Halabi, A. et al., 2012. DNA mismatch repair complex MutS?? promotes GAA??TTC repeat expansion in human cells. *Journal of Biological Chemistry*, 287(35), pp.29958–29967.

Harper, P., 2002. The epidemiology of Huntington's disease. In *Bates G, Harper P, Jones L, editors. Huntington's disease. Oxford: Monographs on Medical Genetics*.

Harvey, B.K. et al., 2011. Transgenic animal models of neurodegeneration based on human genetic studies. *J Neural Transm*, 118(1), pp.27–45.

Hoffner, G. & Djian, P., 2014. Monomeric, Oligomeric and Polymeric Proteins in Huntington Disease and Other Diseases of Polyglutamine Expansion. *Brain Sciences*, 4(1), pp.91–122. Available at: http://www.mdpi.com/2076-3425/4/1/91/.

Huang, M.C. et al., 2012. A simple, high sensitivity mutation screening using Ampligase mediated T7 endonuclease I and Surveyor nuclease with microfluidic capillary electrophoresis. *Electrophoresis*, 33(5), pp.788–796.

Huntington's, G.M. of & Consortium, D. (GeM-H., 2015. Identification of Genetic

Factors that Modify Clinical Onset of Huntington's Disease. *Cell*, 162(3), pp.516–526. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867415008405.

Iyer, R.R. et al., 2015. DNA Triplet Repeat Expansion and Mismatch Repair. *Annual Review of Biochemistry*, 84(1), pp.199–226. Available at: http://www.annualreviews.org/doi/abs/10.1146/annurev-biochem-060614-034010.

Kennedy, L. et al., 2003. Dramatic tissue-specific mutation length increases are an early molecular event in Huntington disease pathogenesis. *Human molecular genetics*, 12(24), pp.3359–67. Available at: http://www.ncbi.nlm.nih.gov/pubmed/14570710.

Kieburtz, K. & Olanow, C.W., 2015. Advances in clinical trials for movement disorders. *Movement disorders : official journal of the Movement Disorder Society*, 30(11), pp.1580–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26307591.

Kolodner, R.D., 2015. A personal historical view of DNA mismatch repair with an emphasis on eukaryotic DNA mismatch repair. *DNA Repair*. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1568786415300951.

Kumar, A. et al., 2015. Huntington's disease: An update of therapeutic strategies. *Gene*, 556(2), pp.91–97. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0378111914012839.

LaFountaine, J.S., Fathe, K. & Smyth, H.D.C., 2015. Delivery and therapeutic applications of gene editing technologies ZFNs, TALENs, and CRISPR/Cas9. *International Journal of Pharmaceutics*, 494(1), pp.180–194. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0378517315301265.

Langbehn, D.R. et al., 2004. A new model for prediction of the age of onset and penetrance for Huntington's disease based on CAG length. *Clinical Genetics*, 65(4), pp.267–277.

Lee, D.-Y. & McMurray, C.T., 2014. Trinucleotide expansion in disease: why is there a length threshold? *Current opinion in genetics & development*, 26C(Table 1), pp.131–140. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25282113.

Lee, J.M. et al., 2011. Quantification of age-dependent somatic CAG repeat instability in Hdh CAG knock-in mice reveals different expansion dynamics in striatum and liver. *PloS one*, 6(8), p.e23647. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21897851.

Liu, L. et al., 2015. Induced Pluripotent Stem Cells in Huntington's Disease: Disease Modeling and the Potential for Cell-Based Therapy. *Molecular Neurobiology*.

Available at: http://link.springer.com/10.1007/s12035-015-9601-8.

Lloret, A. et al., 2006. Genetic background modifies nuclear mutant huntingtin accumulation and HD CAG repeat instability in Huntington's disease knock-in mice. *Human Molecular Genetics*, 15(12), pp.2015–2024.

MacDonald, M.E. et al., 1993. A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6), pp.971–983.

Mason, A.G. et al., 2014. Expression levels of DNA replication and repair genes predict regional somatic repeat instability in the brain but are not altered by polyglutamine disease protein expression or age. *Human Molecular Genetics*, 23(6), pp.1606–1618.

Matsuzaki, K. et al., 2015. FANCJ suppresses microsatellite instability and lymphomagenesis independent of the fanconi anemia pathway. *Genes and Development*, 29(24), pp.2532–2546.

Moore, J.D., 2015. The impact of CRISPR–Cas9 on target identification and validation. *Drug Discovery Today*, 20(4), pp.450–457. Available at: http://www.sciencedirect.com/science/article/pii/S1359644614004875.

Nageshwaran, S. & Festenstein, R., 2015. Epigenetics and Triplet-Repeat Neurological Diseases. *Frontiers in Neurology*, 6(December), pp.1–9. Available at: http://journal.frontiersin.org/Article/10.3389/fneur.2015.00262/abstract.

Ngeow, J. & Eng, C., 2015. New Genetic and Genomic Approaches in the Post-GWAS Era – Back to the Future. *Gastroenterology*, 149(5), pp.1138–1141. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0016508515008239.

O'Geen, H., Yu, A.S. & Segal, D.J., 2015. How specific is CRISPR/Cas9 really? *Current Opinion in Chemical Biology*, 29, pp.72–78. Available at: http://linkinghub.elsevier.com/retrieve/pii/S136759311500109X.

Orr, H.T. & Zoghbi, H.Y., 2007. Trinucleotide Repeat Disorders. , pp.575–623. Available at: http://www.annualreviews.org/doi/pdf/10.1146/annurev.neuro.29.051605.113042.

Paulsen, J.S. & Long, J.D., 2014. Onset of Huntington's disease: Can it be purely cognitive? *Movement Disorders*, 29(11), pp.1342–1350.

Pearson, C.E., Edamura, K.N. & Cleary, J.D., 2005. Repeat instability: mechanisms of dynamic mutations. *Nature Reviews Genetics*, 6(10), pp.729–742. Available at: http://www.nature.com/doifinder/10.1038/nrg1689.

Pekny, M. et al., 2015. Astrocytes: a central element in neurological diseases. *Acta Neuropathologica*. Available at: http://link.springer.com/10.1007/s00401-015-1513-1.

Pinto, R.M. et al., 2013. Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. *PLoS genetics*, 9(10), p.e1003930. Available at: http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003930.

Platt, R.J. et al., 2014. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. *Cell*, 159(2), pp.440–455. Available at: http://dx.doi.org/10.1016/j.cell.2014.09.014.

Project, T.U.S. –Venezuel. C.R. & Wexler, N.S., 2004. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proceedings of the National Academy of Sciences of the United States of America* , 101(10), pp.3498–3503. Available at: http://www.pnas.org/content/101/10/3498.abstract.

Purves, D., 2004. *Neuroscience Third Edition*, Available at: http://www.nature.com/nrn/journal/v3/n8/abs/nrn896.html.

Quarrell, O. et al., 2012. The prevalence of juvenile huntington's disease: A review of the literature and meta-analysis. *PLoS Currents*, (JULY 2012). Available at: http://www.embase.com/search/results?subaction=viewrecord&from=export&id=L366286884%5Cnhttp://currents.plos.org/hd/article/the-prevalence-of-juvenile-huntingtons-disease-a-review-of-the-literature-and-meta-analysis/pdf/%5Cnhttp://sfx.library.uu.nl/utrecht?sid=E.

Rae, D. et al., A Standard of Care in Huntington ' s Disease. , p.3.

Ran, F.A. et al., 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*, 8(11), pp.2281–2308.

Ran, F.A. et al., 2015. In vivo genome editing using Staphylococcus aureus Cas9. *Nature*, 520(7546), pp.186–190. Available at: http://www.nature.com/doifinder/10.1038/nature14299.

Rangel-barajas, C., Coronel, I. & Florán, B., 2015. Dopamine Receptors and Neurodegeneration. , 6(5), pp.349–368.

Rawlins, M.D. et al., 2016. The Prevalence of Huntington's Disease. *Neuroepidemiology*, 46(2), pp.144–153. Available at: http://www.karger.com/?doi=10.1159/000443738.

Roos, R.A.C., 2010. Huntington's disease: a clinical review. *Orphanet Journal of Rare Diseases*, 5(1), p.40. Available at: http://www.ojrd.com/content/5/1/40.

Ross, C.A. & Tabrizi, S.J., 2011. Huntington's disease: From molecular pathogenesis to clinical treatment. *The Lancet Neurology*, 10(1), pp.83–98. Available at: http://dx.doi.org/10.1016/S1474-4422(10)70245-3.

Ross, C. a et al., 2014. Huntington disease: natural history, biomarkers and prospects for therapeutics. *Nature reviews. Neurology*, 10(4), pp.204–16. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24614516.

Samulski, R.J. & Muzyczka, N., 2014. AAV-Mediated Gene Therapy for Research and Therapeutic Purposes. *Annual Review of Virology*, 1(1), pp.427–451. Available at: http://dx.doi.org/10.1146/annurev-virology-031413-085355.

Sanjana, N.E., Shalem, O. & Zhang, F., 2014. Improved vectors and genome-wide libraries for CRISPR screening. *Nat Methods*, 11(8), pp.783–784.

Schmidt, F. & Grimm, D., 2015. CRISPR genome engineering and viral gene delivery: A case of mutual attraction. *Biotechnology Journal*, 10(2), pp.258–272.

Schmidt, M.H.M. & Pearson, C.E., 2015. Disease-associated repeat instability and mismatch repair. *DNA Repair*, 38, pp.117–126. Available at: http://dx.doi.org/10.1016/j.dnarep.2015.11.008.

Schwab, L.C. et al., 2015. Dopamine and Huntington's disease. *Expert Review of Neurotherapeutics*, 15(4), pp.445–458. Available at: http://www.tandfonline.com/doi/full/10.1586/14737175.2015.1025383.

Shalem, O. et al., 2014. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science*, 343(6166), pp.84–87. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1247005.

Shalem, O. et al., 2014. Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. , 343(6166), pp.84–87.

Simpson, S. a & Rae, D., 2012. A standard of care for Huntington's disease: who, what and why. *Neurodegenerative Disease Management*, 2(1), pp.1–5.

Singh, P., Schimenti, J.C. & Bolcun-filas, E., 2014. A Mouse Geneticist ' s practical guide to CRISPR applications. *Genetics: Early Online*, 199(January), pp.1–15.

Slaymaker, I.M. et al., 2016. Rationally engineered Cas9 nucleases with improved specificity. *Science*, 351(6268), pp.84–88. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.aac8608%5Cnhttp://www.sciencemag.org/cgi/doi/10.1126/science.aad5227.

Sleana, M.M. et al., 2008. Mutagenic roles of DNA “repair” proteins in antibody diversity and disease-associated trinucleotide repeat instability. *DNA Repair*.

Squitieri, F. et al., 2015. Epidemiology of Huntington disease: First post-HTT gene analysis of prevalence in Italy. *Clinical Genetics*, pp.1–4.

Swami, M. et al., 2009. Somatic expansion of the Huntington’s disease CAG repeat in the brain is associated with an earlier age of disease onset. *Human Molecular Genetics*, 18(16), pp.3039–3047.

Swiech, L. et al., 2015. In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat Biotechnol*, 33(1), pp.102–106. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25326897%5Cnhttp://www.nature.com/nbt/journal/v33/n1/pdf/nbt.3055.pdf.

Tamura, M., Gu, J. & Takino, T., 1999. Tumor Suppressor PTEN Inhibition of Cell Invasion , Migration , and Growth : Differential Involvement of Focal Adhesion Kinase and p130 Cas Tumor Suppressor PTEN Inhibition of Cell Invasion , Migration , and Growth : , 4370(21), pp.442–449.

Teimourian, S. & Abdollahzadeh, R., 2014. Technology developments in biological tools for targeted genome surgery. *Biotechnology Letters*, 37(1), pp.29–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25257583.

Telenius, H. et al., 1994. Somatic and gonadal mosaicism of the Huntington disease gene CAG repeat in brain and sperm. *Nature genetics*.

Tomé, S. et al., 2013. Tissue-specific mismatch repair protein expression: MSH3 is higher than MSH6 in multiple mouse tissues. *DNA repair*, 12(1), pp.46–52. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23228367.

Usdin, K., House, N.C.M. & Freudenreich, C.H., 2015. Repeat instability during DNA repair: Insights from model systems. *Critical Reviews in Biochemistry and Molecular Biology*, 50(2), pp.142–167. Available at: http://www.tandfonline.com/doi/full/10.3109/10409238.2014.999192.

Vale, T.C. & Cardoso, F., 2015. Chorea: A Journey through History. *Tremor and other hyperkinetic movements (New York, N.Y.)*, 5, pp.1–6. Available at: http://www.tremorjournal.org/index.php/tremor/article/view/296/html.

Vouillot, L., Thelie, A. & Pollet, N., 2015. Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. *G3 (Bethesda)*, 5(3), pp.407–415. Available at:

http://www.ncbi.nlm.nih.gov/pubmed/25566793%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349094/pdf/407.pdf.

Walker, F.O., 2007. Huntington's disease. *Lancet*, 369(9557), pp.218–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17240289.

Walsh, S., 2015. The pathology of Lynch Syndrome. *Diagnostic Histopathology*, 21(4), pp.161–164. Available at: http://dx.doi.org/10.1016/j.mpdhp.2015.04.007.

Wang, Z. et al., 2006. Widespread and Stable Pancreatic Gene Transfer by Adeno-Associated Virus Vectors via Different Routes. , 55(April).

Wexler, A., 2012. Huntington's Disease – A Brief Historical Perspective. *Journal of Huntginton's Disease*, 1, pp.3–4.

Whalen, K.A. et al., 2005. Genetic analysis of the polyomavirus DnaJ domain. *Journal of virology*, 79(15), pp.9982–90. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1181550&tool=pmcentrez&rendertype=abstract.

Wheeler, V.C. et al., 2002. Early phenotypes that presage late-onset neurodegenerative disease allow testing of modifiers in Hdh CAG knock-in mice. *Human molecular genetics*, 11(6), pp.633–40. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11912178.

Wheeler, V.C. et al., 2003. Mismatch repair gene Msh2 modifies the timing of early disease in HdhQ111 striatum. *Human Molecular Genetics*, 12(3), pp.273–281.

Wichmann, T. & Delong, M.R., 1996. Functional and pathophysiological models of the basal ganglia. *Current Opinion in Neurobiology*, 6(6), pp.751–758.

Wu, X. et al., 2014. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. , 32(7), pp.670–676.

Xue, W. et al., 2014. CRISPR-mediated direct mutation of cancer genes in the mouse liver. *Nature*, 514(7522), pp.3–7. Available at: http://www.nature.com/doifinder/10.1038/nature13589.

Zhang, N. et al., 2015. iPSC-Based Drug Screening for Huntington's Disease. *Brain Research*, pp.1–15. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0006899315007180.

Zielonka, D., Mielcarek, M. & Landwehrmeyer, G.B., 2015. Update on Huntington's disease: Advances in care and emerging therapeutic options. *Parkinsonism & Related Disorders*, 21(3), pp.169–178. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1353802014004878.

8. Annexes

8.1 Summary of NGS assays for *in vitro* sgRNA validation

NGS Assay (JDv2 in sillico mutation rate prediction)	Analysed sample	Cell Selection Method	PCR Product Purification	Noise Reads %	Mutation Rate: Total	Mutation Rate: Frame Shift	# Specific Reads
Mlh1-A1 (63.70%)	Guide in pX458 Experiment 1	FACS	Gel Purified	0.00%	78.11%	69.83%	44800
	Guide in pX458 Experiment 2	FACS	Gel Purified	0.50%	56.83%	52.00%	81833
	Empty pX458	FACS	Gel Purified	0.53%	0.42%	0.29%	103472
	Guide in pLentiCRISPR v2	Puromycin	Gel Purified	0.00%	100.00%	81.72%	3271
Mlh1-JD4 (68.53%)	Guide in pX458	FACS	Gel Purified	0.31%	77.62%	76.49%	58811
	Empty pX458	FACS	Gel Purified	0.19%	0.00%	0.00%	53600
Mlh3-JD2 (65.28%)	Guide in pX458	FACS	Gel Purified	0.00%	62.37%	45.57%	72722
	Different guide (Msh2-JD1 as control) in pX458	FACS	Gel Purified	0.00%	0.33%	0.00%	82270
Msh2-JD1 (73.91%)	Guide in pX458	FACS	Gel Purified	0.00%	70.28%	49.73%	12302
	Empty pX458	FACS	Gel Purified	0.00%	0.00%	0.00%	34317
	Guide in pX458	FACS	Column Purified	27.30%	71.13%	51.40%	11537
Msh3-JD2.3 (74.82%)	Guide in pX458	FACS	Gel Purified	13.27%	79.28%	73.58%	40510

	Empty pX458	FACS	Gel Purified	1.24%	0.00%	0.00%	45427
	Guide in pX458	FACS	Column Purified	47.32%	79.19%	73.68%	25213
Fan1-JD2.1 (77.86%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	69.57%	66.62%	4275
	Empty pLentiCRISPR v2	Puromycin	Column Purified	0%	0.00%	0%	1494
Fan1-JD2.2 (64.65%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	31.03%	31.03%	377
	Empty pLentiCRISPR v2	Puromycin	Column Purified	23.95%	0.00%	0%	3864
Fan1-JL1.8 (48.67%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	72.97%	40.51%	19352
Rrm2b-JD2.1 (59.78%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	88.87%	83.65%	3468
	Empty pLentiCRISPR v2	Puromycin	Column Purified	0.00%	0.00%	0.00%	3928
Rrm2b-JD2.2 (59.77%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	75.52%	71.46%	10160
	Empty pLentiCRISPR v2	Puromycin	Column Purified	0%	0.00%	0.00%	2866
Pms2-JD2.1 (64.31%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	80.97%	72.10%	34629
	Empty pLentiCRISPR v2	Puromycin	Column Purified	11.17%	0.00%	0%	4367
Pms2-JD2.2 (67.63%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	17.86%	94.65%	91.62%	12512
	Empty pLentiCRISPR v2	Puromycin	Column Purified	0%	0.00%	0%	221
Pms1-JD2.1 (72.66%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	76.97%	58.46%	6218
Pms1-JD2.2 (68.14%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	0%	84.84%	72.38%	13588
	Empty pLentiCRISPR	Puromycin	Column	0.00%	12.62%	0.00%	1989

	v2		Purified				
Ercc3-JD2.1 (64.45%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	2.16%	72.89%	47.06%	36353
Exo1-JD2.1 (72.26%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	6.99%	62.17%	55.48%	10454
	Empty pLentiCRISPR v2	Puromycin	Column Purified	18.64%	0.00%	0.00%	25382
Exo1-JD2.2 (68.32%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	7.04%	53.61%	51.51%	12968
	Empty pLentiCRISPR v2	Puromycin	Column Purified	0%	0.00%	0%	3090
Msh6-JD2.1 (72.83%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	2.74%	88.87%	88.22%	15523
	Empty pLentiCRISPR v2	Puromycin	Column Purified	9.71%	0.00%	0.00%	12091
Msh6-JD2.2 (63.55%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	32.26%	95.34%	87.31%	16487
	Empty pLentiCRISPR v2	Puromycin	Column Purified	23.62%	0.00%	0.00%	7097
FancJ-JD2.1 (70.28%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	1.26%	79.63%	59.35%	9671
	Empty pLentiCRISPR v2	Puromycin	Column Purified	8.39%	0.00%	0%	4093
FancJ-JD2.6 (82.51%)	Guide in pLentiCRISPR v2	Puromycin	Column Purified	6.53%	84.89%	75.78%	9472
	Empty pLentiCRISPR v2	Puromycin	Column Purified	24.81%	4.99%	4.99%	3229