Gene editing approaches of microsatellite disorders : shortening expanded repeats

Lucie Poggi

Sorbonne Université

Ecole doctorale Complexité du Vivant

UMR3525 / Instabilités naturelles et synthétiques des génomes

Gene editing approaches of microsatellite disorders

Shortening expanded repeats

Lucie Poggi

Thèse de doctorat de Biologie

Dirigée par Guy-Franck Richard

Présentée et soutenue publiquement le 5 Juin 2020

Devant un jury composé de :

M. Bertrand Llorente	Rapporteur
M. Nicolas Charlet-Berguerand	Rapporteur
Mme. Carine Giovannangeli	Examinateur
Mme. Irena Draskovic	Examinateur
M. Bruno Dumas	Co-Directeur de thèse
M. Guy-Franck Richard	Directeur de thèse

A ma famille,

Acknowledgments

Je tiens à remercier Bertrand Llorente et Nicolas Charlet-Berguerand, rapporteurs de cette thèse ainsi que Carine Giovannangeli et Irena Draskovic, examinateurs, pour avoir accepté d'avoir fait partie de mon jury et pour avoir évalué mon travail. Merci également à Irena pour son suivi lors de mes comités de thèse.

Je remercie Guy-Franck Richard, mon directeur de thèse pour son encadrement durant ces quatre années, merci pour la liberté et la confiance que tu m'as accordée, et pour ta disponibilité tout au long de ces années. Merci pour tout ce soutien que tu m'as apporté !
Je remercie Bruno Dumas, mon co-directeur pour ses encouragements, son soutien et pour m'avoir permis de réaliser cette thèse en partenariat avec Sanofi. Merci pour ton optimisme à toute épreuve !

Un grand merci à Samia Miled pour son aide et son soutien constant, son énergie. Que de manips faites avec juste toi dans le labo tôt le matin comme tard le soir. Merci pour tout ce café bu avec moi !
Je voudrais remercier David Viterbo, on a passé beaucoup de temps ensemble dans la salle café, je ne me lasse pas de tes imitations, nos conversations et tes blagues me manqueront beaucoup ! Vous avez été d'un grand soutien Wilhelm et toi, vous m'avez fait beaucoup rire !!

Je remercie mes collègues du labo, ce fut un plaisir de travailler avec vous : Valentine Mosbach, Stéphane Descorps-Declère, Olivia Frenoy, Wilhelm Vaysse-Zinkhöfer (Coach !), Lisa Emmenegger, Astrid Marchal, Laureline Bétemps. Merci Stéphane pour m'avoir montré les analyses de séquences.
Je remercie Christelle Lenormand pour son aide et sa bonne humeur.

Je remercie l'équipe de Biologics Reasearch, en particulier Emmanuelle Vigne. Ce fut un plaisir également de croiser Micaela et Cristina lors de mes visites à Sanofi. Je remercie également Seng H. Cheng pour ses conseils et grâce à qui les AAV ont pu être produits à Genzyme par l'équipe de Nelson Yew que je remercie également. Merci également à Lisa Stanek pour son suivi lors de mon deuxième comité de thèse.

Je remercie l'équipe de Geneviève Gourdon de l'institut Imagine, en particulier Stéphanie Tomé, pour ton aide avec les amplifications de triplet et pour m'avoir montré ton protocole de Southern blot. Un grand merci à Aline Huguet pour sa patience et sa pédagogie, c'est toi qui nous as tout appris à Olivia et moi pour les souris. Ce fut un plaisir de tout apprendre à tes côtés Olivia !

Je remercie l'équipe de Denis Furling, en particulier Arnaud Klein pour m'avoir formée sur les techniques de culture cellulaire des cellules de patient utilisées lors de cette thèse ainsi que pour m'avoir montré les transductions lentivirales.

Je remercie le laboratoire Microfluidique Physique et Bio-ingénierie de Charles Baroud, en particulier Antoine Barizien et Nadia Vertti, pour m'avoir initiée à la microfluidique. Nos échanges entre physiciens et biologistes furent très enrichissants !

Je remercie tous les jeunes du quatrième étage pour les bons moments passés ensemble : Elodie Zhang, Alicia Nevers, Rostyslav Makarenko, Simon Malesys, Lena Audebert, Agathe Gilbert, Estelle Dacheux, Varun Khanna, Marine Dehecq, Thomas Damagnez. Merci pour tous les bons moments passés ensemble, à Pasteur et à l'extérieur ! Merci également à Antonin Piasco et Quentin Nevers.
Je remercie tout le quatrième étage Fernbach, en particulier Lucia Oreus pour tout le travail fourni et sans l'aide de qui aucune levure ne pourrait pousser à l'étage !

Merci à mes amis, mes deux amies d'enfance Marie et Aurélie, ainsi que mes amis de l'agro, en particulier Marie, Emilia, Gaëtan, Tania, Margaux et Alix, Pauline et Mathilde.
Je remercie ma famille pour leur soutien et encouragements, en particulier mes parents, Cécile et Jean-Pierre, ma sœur, Elise et ma mamie, Christiane.
Je remercie aussi ma deuxième famille : Pierre, Véronique et Juliane, et les remercie pour les beaux voyages que nous avons faits ensemble.
Un immense merci à Maxime sans qui je n'aurais certainement pas réussi à aller jusqu'au bout.

List of abbreviations

BER: **B**ase **E**xcision **R**epair

BFP: **B**lue **F**luorescent **P**rotein

BIR: **B**reak-**I**nduced **R**eplication

CRISPR: **c**lustered **r**egularly **i**nterspaced **s**hort **p**alindromic **r**epeats

Cas: **C**RISPR-**as**sociated protein

DM1: Myotonic dystrophy type 1 (**D**ystrophia **M**yotonica type 1)

DSB: **D**ouble-**s**trand **b**reak

DUE: **D**NA **u**nwinding **e**lement

FISH: **F**luorescence **i**n **s**itu **h**ybridization

FXS: **F**ragile **X** **s**yndrome

GFP: **G**reen **F**luorescent **P**rotein

HD: **H**untington **d**isease

HEK: **H**uman **e**mbryonic **k**idney

HR: **H**omologous **R**ecombination

ITR: **I**nverted **t**erminal **r**epeat

MMEJ: **M**icrohomology-**m**ediated **e**nd **j**oining

MMR: **M**is**m**atch **r**epair

MRN: **M**re11-**R**ad50-**N**bs1

MRX: **M**re11-**R**ad50-**X**rs2

NHEJ: **N**on-**h**omologous **e**nd **j**oining

NMR: **N**uclear **m**agnetic **r**esonance

PAM: **P**rotospacer **a**djacent **m**otif

rAAV: **R**ecombinant **a**deno-**a**ssociated **v**irus

RPA: **R**eplication **p**rotein **A**

RVD: **R**epeat **V**ariable **D**iresidue

SBMA: **S**pinal and **b**ulbar **m**uscular **a**trophy

SCA: **S**pino**c**erebellar **a**taxia

SSA: **S**ingle-**s**trand **a**nnealing

TA: **T**ibialis **a**nterior

TALEN: **T**ranscription **a**ctivator-**l**ike **e**ffector **n**ucleases

UTR: **u**n**t**ranslated **r**egion

ZFN: **Z**inc **F**inger **N**uclease

List of figures

List of tables

Table of content

Introduction

This introduction will give an outline of microsatellite sequences causing disorders in human. Their secondary structure, the way they trigger pathology and how they expand in the genome will be presented. In a second part, myotonic dystrophy type I will be described, and due to the lack of treatment for this disease, potent therapeutic approaches will be presented, including genome editing approaches. In the third part of this introduction, targeted genome editing techniques using highly specific nucleases will be presented through the successive discovery of four families of specific nucleases: meganucleases, Zinc Finger Nucleases, Transcription Activator Like Effector Nucleases and CRISPR-associated nucleases. The repair pathways that exist to repair double-strand breaks induced by nucleases will then be described. In a last part, the current advances of genome editing techniques for muscular dystrophies will be commented and challenged regarding the current hurdles impeding their implementation as a therapeutic approach.

Canonical structure and alternative forms of DNA

Right-handed DNA helices

Historically, fiber x-ray crystallography identified two distinct structural forms of DNA: A-DNA and B-DNA (Franklin and Gosling, 1953). A-DNA was isolated at 75% humidity and B-DNA at higher percentages. Concomitantly, Watson and Crick identified B-DNA as a double-helix structure and proposed a model of an anti-parallel double-stranded helix, formed by two linear sugar-phosphate backbones that run in opposite directions (Watson and Crick, 1953). The two strands are connected by hydrogen bonds between the purine and pyrimidine bases, consistent with the previously enounced Chargaff's rule, stating that purines and pyrimidines ratio should be 1:1 in all organisms (Chargaff et al., 1950). These bonds are called Watson-Crick bonds. Alternative Hoogsteen base pairings can also be observed in alternative forms of DNA, in which the purine is rotated in such a way that bonds are made between its other face and the pyrimidine (Hoogsteen, 1963).

A-DNA is a thicker right-handed duplex with a shorter distance between base pairs and has been described for RNA-DNA duplexes and RNA-RNA duplexes. RNA can only form A-type double helices because of the steric restrictions of the ribose 2′ hydroxyl residue (Arnott et al., 1968; Xiong and Sundaralingam, 2000).

Left-handed Z-DNA structure

Surprisingly, it was only in the late 1970s, when single crystal X-ray diffraction was available to validate the proposed model of the double helix, that the expected B-DNA structure was not the first to be observed. Instead, the first single-crystal X-ray structure of a DNA fragment showed a left-handed double helix (Wang et al., 1979). This unexpected result was already a hint toward the complexity of the different conformations that DNA can adopt. Since the ribose phosphate backbone followed a zig-zag, this form of DNA was called Z-DNA. It was later recognized that Z-DNA is formed due to negative supercoiling, produced behind a moving RNA polymerase during transcription (Liu and Wang, 1987) and that its formation is favored near transcription start sites (Schroth et al., 1992).

Alternative forms of DNA

The lowest energy level state of DNA in physiological conditions is B-DNA. However, formation of alternative structures can occur when the DNA duplex is unwound during metabolic DNA processes such as DNA replication and transcription. Other alternative forms of DNA were discovered later on, including G-quadruplexes, hairpins, H-DNA and palindromes. Many repeated sequences were shown to form DNA secondary structures *in vitro,* and were supposed to form *in vivo.*

Microsatellite disorders

Classification based on the nature of the repeated sequence

More than 40 disorders are due to expanded microsatellites throughout the human genome, and most of them are neuromuscular disorders (**Table 1**). Repeated sequences include tri-, tetra-, penta- or hexa- nucleotides. Specific aspects of microsatellite disorders are discussed above, divided in four categories for each microsatellite: secondary structure formation, disease association, pathogenic mechanisms and mechanisms underlying the instability.

CAG/CTG repeats

Secondary structure of the repeat

CAG/CTG form imperfect hairpins *in vitro* (**Figure 1.A**); this was demonstrated using ^{1}H nuclear magnetic resonance (NMR) which infers the position of T and G bases, carrying imino protons. *In vitro* experiments measuring the melting temperature (Tm) of oligos made of various numbers of CAG or CTG showed that CTG hairpins are more stable than CAG hairpins. It may be because purines occupy more space than pyrimidines and are most likely to interfere with hairpin stacking forces. Also, since the folds formed by 30 CTG or CAG repeats do not exhibit a higher Tm and are only 40% more stable in free energy than those formed by 10 repeats, it suggested that triplet expansions with higher repeat number may result from the formation of more hairpins with similar stability rather than larger hairpins (Petruska et al., 1996). Experiments of denaturation/renaturation of a plasmid containing 50 CAG/CTG repeats were carried out and length was resolved on 4% polyacrylamide gels. Upon renaturation, 60% of DNA was slipped-strand DNA. Observation of the same plasmid and plasmids carrying 255 CAG/CTG repeats under electron microscope revealed compacted and bend molecules, corresponding to structured DNA fragments. The heterogeneity and complexity of the molecules observed increased with the number of repeats (Pearson et al., 1998). Thus, biophysical studies showed that CAG/CTG form hairpins *in vitro,* and some evidence tend to confirm that they also form *in vivo* during replication; this will be discussed in the section "Mechanisms of instability of CAG/CTG repeats".

Sequence	Disease	Locus	Expansion length (bp)	Pathogenic mechanism	Clinical features
$(CAG)_n$	Huntington Disease	*HTT* exon	30-180	Gain of function	Chorea, dystonia, cognitive deficits, psychiatric problems
$(GCN)_n$	Synpolydactyly, type 1	*HOXD13* exon	15	Gain of function	Ataxia, cognitive impairments, fusion of digits) and production of supernumerary digits
$(CTG)_n$	Myotonic dystrophy type 1 (DM1)	*DMPK* 3'UTR	50-10,000	RNA-mediated pathogenesis	Myotonia, weakness, cardiac conduction defects, insulin resistance, cataracts, testicular atrophy, and mental retardation in congenital form
$(CGG)_n$	Fragile X syndrome	*FRAXA* 5'UTR	55-200 (premutation: FXTAS) >200 (full mutation: FXS)	Loss of function	Ataxia, tremor, Parkinsonism, dementia. Mental retardation macroorchidsm, connective tissue defects
$(GAA)_n$	Friedreich ataxia	*FRDA* exon	200-1700	Loss of function	Sensory ataxia, cardiomyopathy, diabetes
$(CCTG)_n$	Myotonic dystrophy (DM2)	*ZNF9* intron	75-11,000	RNA-mediated pathogenesis	Similar to DM1
$(ATTCT)_n$	Spinocerebellar ataxia, type 10 (SCA10)	*ATXN10* intron	500-4500	RNA-mediated pathogenesis	Ataxia, tremor, dementia
$(TGGAA)_n$	Spinocerebellar ataxia, type 31 (SCA31)	*TK2 / BEAN* intron	500-760	RNA-mediated pathogenesis	Similar to SCA10
$(GGCCTG)_n$	Spinocerebellar ataxia, type 36 (SCA36)	*NOP56* intron	>650	RNA-mediated pathogenesis	Similar to SCA10
$(GGGGCC)_n$	Amyotrophic lateral sclerosis	C9orf72 intron	700-1600	RNA-mediated pathogenesis	Muscle weakness, cognitive impairment

Table 1: Microsatellite disorders and associated disease, repeat length, locus and pathogenic mechanism mediated by the expanded microsatellite. For GCN sequence, N can be any base.

Diseases linked to CAG/CTG expansions

CAG expansions in coding regions result in polyglutamine disorders such as Huntington disease (HD), Spinobulbar Muscular Atrophy (SBMA) and Spinocerebellar Ataxias (SCA). SBMA is an X-linked disorder which mainly affects motor neuron and is characterized by proximal and bulbar muscle wasting (Arbizu et al., 1983). HD is also a neurodegenerative disorder which is otherwise known as Huntington chorea due to the characteristic initial physical symptoms including random and uncontrollable movements. These primary symptoms are followed by memory deficits, changes in personality, and depression (Vonsattel and DiFiglia, 1998). Many SCAs exist, SCA 1, 2, 3, 6, 7, and 17 are caused by a CAG expansion. Ataxia[1], tremor[2], and dysarthria[3] are common to all the SCAs, with each type of ataxia having additional symptoms (Paulson, 2009).

CTG expansions cause myotonic dystrophy type I (DM1). Congenital DM1 is characterized by hypotonia[4], facial diplegia[5], and mental retardation. Adult forms are milder, patients show myotonia[6], muscle degeneration, and may develop cataracts, cardiac conduction defects, insulin resistance, sleep disorders, testicular atrophy (Machuca-Tzili et al., 2005). SCA8 is also caused by an expanded CTG repeat (Koob et al., 1999).

Mechanism of pathogenicity associated to CAG/CTG expansions

Polyglutamine disorders are mediated by a gain-of function mechanism: polyglutamine proteins will form aggregate and will induce mortality of the cells, usually neurons (Ross and Poirier, 2004). The polyglutamine inclusions will either disrupt normal transcription (Dunah et al., 2002) (Lam et al., 2006) or impair the ubiquitin-proteasome system (Donaldson et al., 2003).

In DM1, CUG-expanded RNA form aggregates which are visible under the microscope as foci (Taneja et al., 1995). DM1 pathogenesis will be more extensively described in the section “RNA-mediated pathogenesis”. It was also recently demonstrated that CAG and CTG repeats are subjected to Repeat Associated Non-methionine translation (RAN translation), in either direction, even when located in non-translated regions (5’ UTR, introns). The resulting repeated peptides are toxic for the cells (Zu et al., 2011).

[1] Lack of muscle control or coordination of voluntary movements, resulting in slurred speech, trouble eating and swallowing, deterioration of fine motor skills, difficulty walking etc.

[2] Uncontrollable shaking

[3] Speech disorder caused by muscle weakness

[4] Lack of resistance to passive movement due to low muscle tone

[5] Paralysis of both sides of the face

[6] Delayed relaxation after voluntary contraction of skeletal muscle

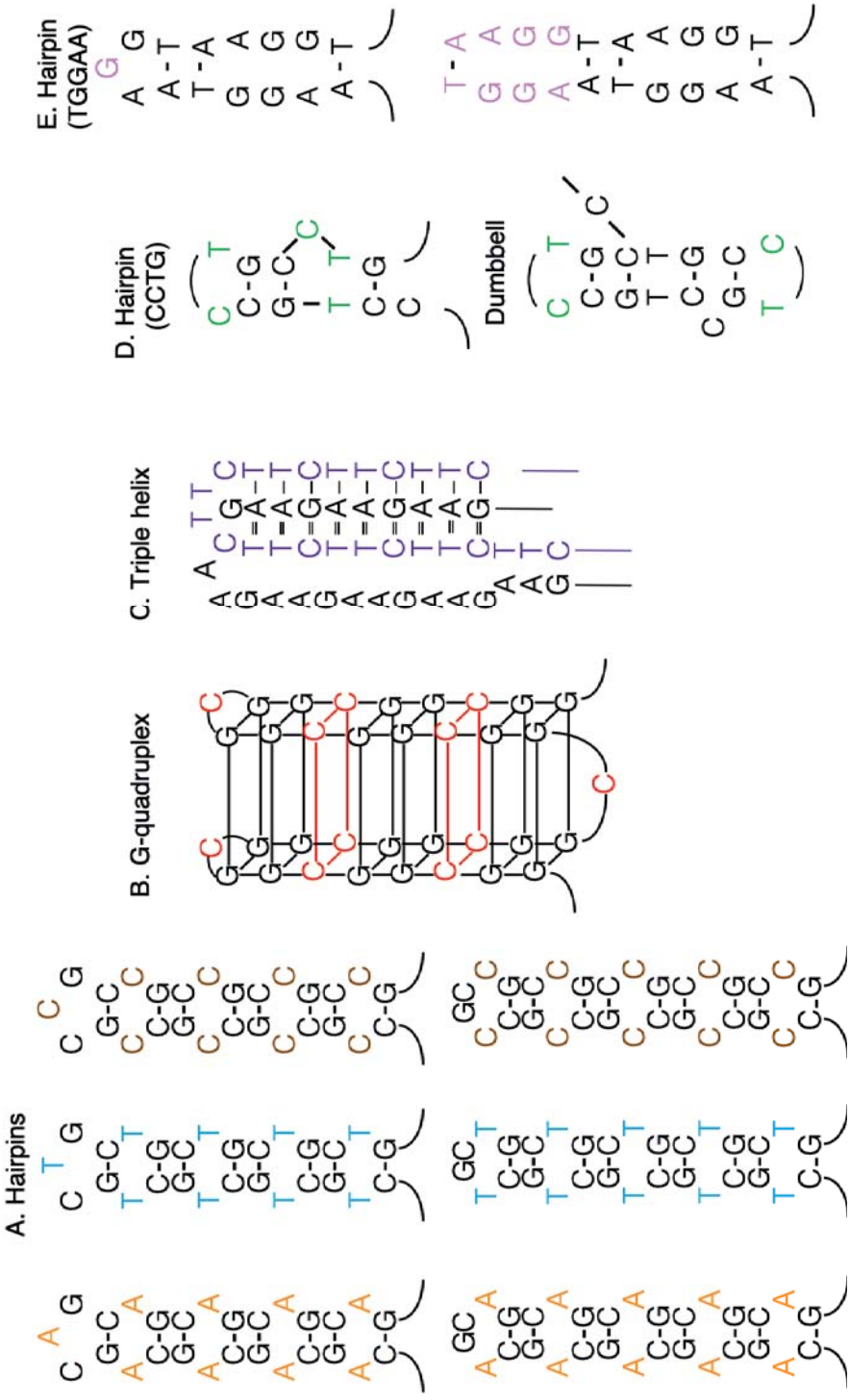


Figure 1: Secondary structures formed by microsatellites associated to human disorders. A. Hairpins. Hairpins. CAG, CTG, and CCG hairpins formed by an odd number of repeat units (top) or even number of repeat units (bottom) Bases making no pairing within the stem are colored. **B.G-quadruplex.** Structure formed by $(CGG)_n$ repeats. Cytosines and cytosine bonds are highlighted in red. **C. Triple-helix.** Structure formed by $(GAA)_n$ repeats. Watson-Crick pairings are shown by double lines, and Hoogsteen pairings are shown by single lines. **D. Hairpin and dumbbell formed by $(CCTG)_n$ repeats.** When the repeat length is equal to three, a hairpin with a two-residue CT loop is formed. When the repeat length is equal to four, a dumbbell structure with two CT-loops is formed. **E. Hairpins formed by $(TGGAA)_n$ repeats.** Odd numbers form an end to end conformation with a 1-nt loop (top) whereas even numbers can either form a 1-nt loop (top) or an octaloop (bottom).

Mechanisms of instability of CAG/CTG expansions

Impact of Mismatch Repair

Mismatch repair proteins are in charge of detecting mispairings resulting from DNA polymerase errors during replication, and signal them to the repair machinery (Larrea et al., 2010). Tumor cells from hereditary nonpolyposis colorectal cancer patients display high alteration of the length of microsatellites and this was linked to a defect in mismatch repair (MMR) (Parsons et al., 1993). In yeast, the absence of any of three genes (*PMS1*, *MLH1* and *MSH2*) involved in MMR leads to a 700-fold increase in instability of GT repeats (Strand et al., 1993). It suggests that following replication, a functional MMR is required to fix small slippage errors done by the polymerase.

However, MMR loss through *pms1* and *msh2* deletion in yeast strains did not affect large scale changes in CAG repeat tract or even prevented them, while promoting short scale changes of +1 or -1 repeat (Schweitzer and Livingston, 1997). It suggested that large scale expansions and small-scale expansions have different underlying origins.

In yeast, CAG/CTG trinucleotide repeats are more unstable when the CTG triplets are located on the lagging strand template, supposedly more prone to form single-stranded secondary structures, than the leading strand template. Additionally, using 2D gels to visualize replication intermediates, it was shown in *E. coli* that fork stalling during replication occurred for plasmids bearing CTG repeats, but not CAG repeats. Pausing was more frequent as the length of the repeat increased (Samadashwily et al., 1997). The same experiment was carried out in yeast, and showed a mild effect on fork stalling at 80 CAG or CTG repeated sequences (Pelletier et al., 2003).

CTG or CAG hairpins are likely to be recognized by MMR proteins which will mistakenly recognize them as mismatches. Such slipped stranded structures were generated by annealing stretches of repeats of CAG/CTG of either the same or different length and band-shift assay was performed to detect MSH2 binding. MSH2 was found to bind to these structures and its affinity increased with repeat length. Furthermore, MSH2 binds more efficiently to $(CAG)_{15}$ oligonucleotide than to $(CTG)_{15}$ oligonucleotide (Pearson et al., 1997).

Secondary structure formation during replication may stall fork progression. CTG repeats are able to form more complex molecules *in vitro* as the length of the repeat increases. It suggests that in humans, secondary structures will more likely form and be stabilized as the number of repeat increases. This may explain why larger CTG repeats are more prone to expansions than shorter ones (Pearson et al., 1998). More precisely, ATPase domain of Msh2 was found to be

essential in large scale changes of CTG repeats in a mouse model containing 1000 CTGs (Tomé et al., 2009). Finally, CAG/CTG repeats were shown to be bound by mismatch repair proteins and to trigger replication fork stalling (Viterbo et al., 2016). MMR is clearly involved in CAG/CTG repeat instability, but fails to explain how large expansions arise.

Impact of replication

Evidence of the involvement of replication in triggering expansions was confirmed by studying repeat instability in the presence of drugs altering cell replication. Mimosine which inhibits replication initiation did not have an effect on triplet repeat length. Aphidicoline, by inhibiting DNA polymerases affects both leading and lagging strand progression, triggered instability of the large DM1 allele. Emetine which preferentially blocks Okazaki fragments had the more dramatic effect on CTG instability (Yang et al., 2003). *In vivo*, in HeLa cells, evidence for hairpin formation was given by the use of a Zinc Finger Nuclease (ZFN) that recognizes CTG repeats. Only one ZFN arm was found to be able to cut DNA suggesting that a hairpin was formed and cut; this argument can be discussed as hairpins formed by CTG repeats are imperfect and do not mirror structurally Watson-Crick bounds. When the same cells were serum deprived and were not cycling, no cutting was found, suggesting that hairpin formation was replication-dependent. CAG/CTG repeats showed increased instability through multiple cycling division, suggesting that replication was implicated in their instability; in HeLa cells, instability of $(CAG)_{102}$ and $(CTG)_{102}$ was observed after 250 doublings. The instability was suppressed when close-by replication origin was inactivated (Liu et al., 2010a). These results confirmed the active role of replication in repeats instability. The mechanism explaining repeat instability during replication is that secondary structures may form on either strand of the DNA, slippage of the DNA polymerase and the nascent strand backwards on the template strand would result in the formation of structures containing an excess of repeats on the nascent strand, resulting in expansion products. The opposite would give rise to deletion products. CTG repeats may also induce fork stalling and collapse. Restart then involves repair and recombination machinery to pursue replication. The exact mechanism following restart is unknown. Fen1 homologue of Rad27 in yeast is a nuclease that removes displaced RNA-DNA primers on the lagging strand during replication and is also critical for progress of stalled forks. It was a good candidate as a major driver of CTG/CAG instability as the *rad27* strain exhibits enhanced instability of CAG/CTG repeats (Callahan et al., 2003). Partial knock-out of Fen1 in mice carrying DM1 DMPK gene did not exhibit any change in instability suggesting that Fen1 is not a major driver of human instability (van den Broek et al., 2006).

The impact of replication through DM1 CTG expansion was assessed in DM1 fibroblasts and in mice carrying the DM1 locus and 328 repeats. Nascent DNA were found both upstream and downstream the repeats although in mice they were detected only downstream of the repeat. Study of replication profile in mice at different ages and tissues did not show any clear correlation; nevertheless, in testis, replication origin downstream of the repeat showed reduced activity overtime while CTG repeat instability was higher. Authors suggest that CTCF binding at CTCF binding sites located downstream CTG repeats in DM1 may play a role in the stability of the repeat (Cleary et al., 2010). The exact role of replication in triggering instability in patients is still unclear.

However, replication slippage fails to explain how instability arises in non-dividing cells since all the experiments described here were carried out in actively dividing cells.

DSB-induced instability

In yeast, CTG repeats were integrated alongside a *URA3* gene between direct repeats. When exposed to hydroxyurea (HU), 5'FOA resistant colonies number increased with CTG repeat length. These repeats act as fragile sites and promote DSBs in their vicinity, resulting in the loss of *ura3* gene in this experimental assay (Freudenreich et al., 1998). However, so far, CTG/CAG repeats in patients have never been identified as fragile sites, challenging observations made in yeast. DSB repair slippage may occur in yeast during gene conversion and resulted in deletions of CAG repeats flanking an I-SceI cleavage site (Richard et al., 1999). HO-induced DSB that can be repaired using a donor plasmid carrying CAG_{98} was used to assess repair at repeated sequences. Repair greatly increased and expansions increased when overexpressing Mre11 and Rad50, both part of the MRX complex. It suggested that this complex was actively needed to remove secondary structures blocking repair (Richard et al., 2000) (**Figure 2**). ZFNs were used to induce DSBs into CAG/CTG repeats which mostly led to contractions in CHO cells (Mittelman et al., 2009) and in a HEK293 cells GFP reporter assay (Santillan et al., 2014). These experiments indicate that repairing a DSB inside a repeat leads to a high instability.

Break induced repair (BIR) and large-scale expansions

BIR is a repair pathway that resolves one-ended double-strand breaks, arising when replication forks collapse. If the fork is stalled at CAG/CTG repeats, this sequence will serve for homology search, and strand invasion will likely happen out of register in the repeats. This would lead to large expansions. In yeast, *pif1*Δ and *pol32*Δ mutants show reduced instability of CAG/CTG

repeat tracts. Both *pif1* and *pol32* are required for BIR, suggesting a role of this pathway in expansions (Kim et al., 2017). In non-dividing cells, BIR may process double-strand breaks arising at repeated sequences.

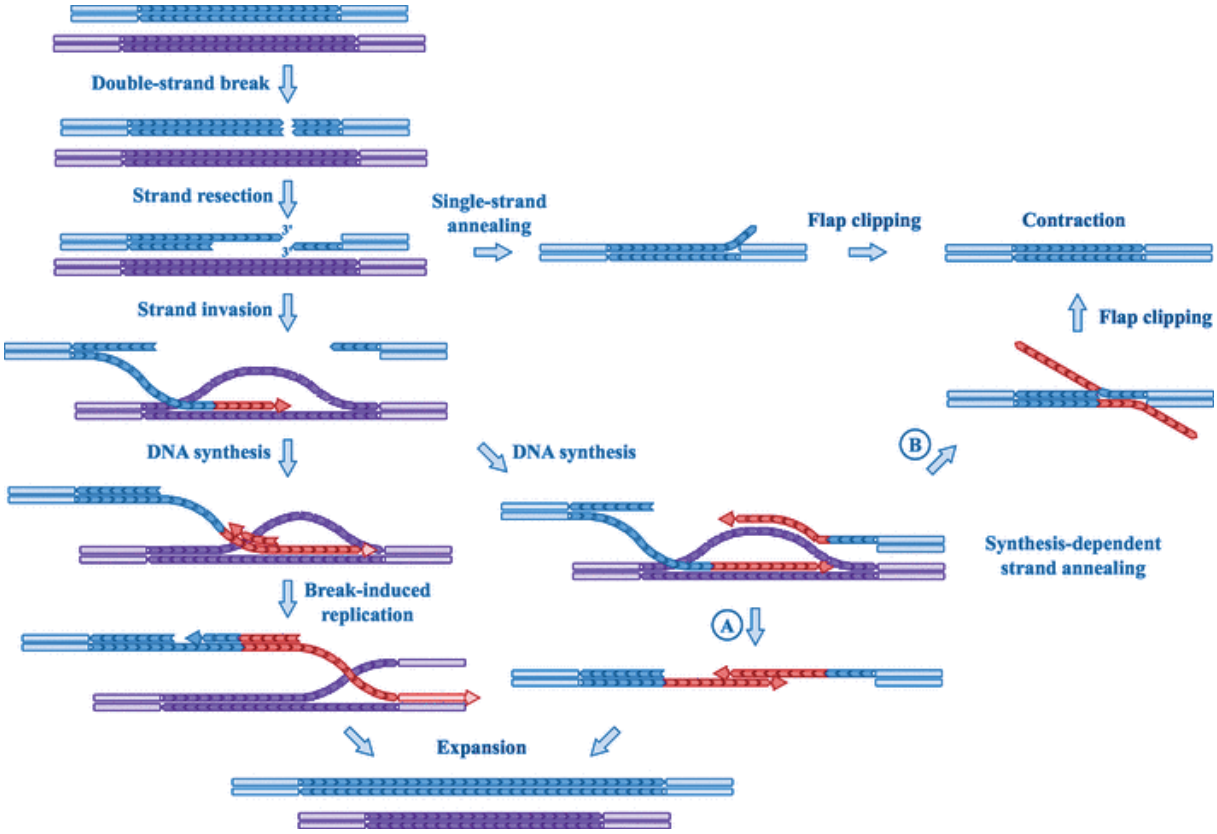


Figure 2: Double-strand break repair mechanisms leading to repeat contraction or expansion. Exogenous sources (radiation,…) or endogenous sources like replication stress, fork stalling, MMR recognition may induce a DSB in or near a CTG/CAG repeat tract. The broken molecule is resected by several nucleases and helicases leading to 3′-hydroxyl single-stranded ends. These ends may engage into different types of homologous recombination. Direct annealing of the two ends by SSA leads to repeat tract shortening (right). DNA synthesis during BIR generates repeat expansions (bottom). Synthesis-dependent strand annealing is resolved by unwinding and out-of-frame annealing of the recombination intermediate, possibly leading to repeat expansion (a) or repeat contraction (b). Figure from (Mosbach et al., 2019a).

Base Excision Repair (BER) -mediated repair and MMR

R6/1 mouse is a mouse model for Huntington disease and harbor a transgene containing exon 1 of human HD gene including around 100 CAG repeats. These repeats are stable until 11 weeks of age, after which they tend to expand (Kovtun and McMurray, 2001). Oxidative lesions tend to accumulate with age and oxidized bases are repaired by the BER pathway, involving DNA glycosylases such as OGG1. R6/1/*OGG1*$^{-/-}$ mice show a decreased CAG instability. On the other hand, MSH2/MSH3 MMR protein complex was previously found to be essential for somatic expansions in mice (Owen et al., 2005). The proposed model is that CAG repeats undergo base oxidation, repaired by BER OGG1 glycosylase. During the course of the repair, single stranded CAG repeats can form hairpins. These secondary structures are recognized by MMR proteins which trigger its instability (Kovtun et al., 2007) (**Figure 3**).

Conclusion

All mechanisms involving *de novo* DNA synthesis have been involved at some point in CAG/CTG instability. For example, transcription was also linked to instability and R-loops formation was proposed to induce repeat instability of CAG repeat tract (Reddy et al., 2014). Different mechanisms may be at stake and probably through different steps in dividing versus non-dividing cells.

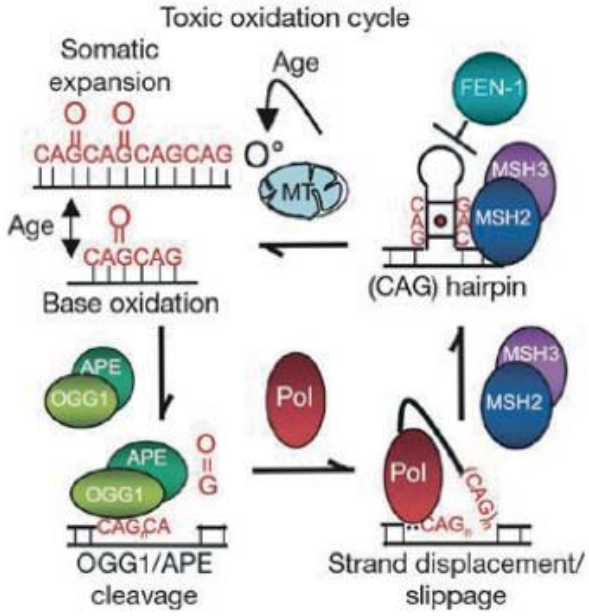


Figure 3: Toxic oxidation cycle model for age-dependent somatic expansion. Endogenous oxidative radicals (O°) arising from mitochondrial (MT) respiration creates oxidative DNA lesions on CAG/CTG repeat tract. These lesions are removed by the Base Excision Repair pathway. OGG1/APE cleavage produces a nick, and polymerase (Pol) facilitates hairpin formation during gap-filling synthesis. CAG hairpins are stabilized by MSH2/MSH3 binding (red dot is a mismatch in the stem) and escape FEN-1 loading and cleavage owing to a hidden 5′ end. Strand displacement leads to a longer CAG template, which is again subject to oxidative DNA damage. From (Kovtun et al., 2007).

CGG repeats

Secondary structure of the repeat

CGG repeats were shown to form tetraplexes (**Figure 1.B**); replication was blocked *in vitro* by 20 stretches of CGG, in a K^+ dependent manner. Tetraplexes require cations small enough to fit in the cavity created by guanine tetrads and stabilize the structure, so K^+ dependency is an indicator of tetraplex formation (Usdin and Woodford, 1995). Additionally, it was also shown that the stoichiometry of the structure is tetramolecular and that guanines are protected from dimethylsulfate (DMS) guanine-specific cleavage, further confirming tetraplex structure formation (Fry and Loeb, 1994). Also, CGG repeats could theoretically form hairpins (**Figure 1.A**); a dynamic change between tetraplexes and hairpins may occur *in vivo* (Zumwalt et al., 2007).

Diseases linked to CGG expansions

CGG repeat in the 5′ untranslated region (UTR) of FMR1 is the cause of Fragile X syndrome. It was first described in a family in which mental retardation segregated as an X-linked trait (Martin and Bell, 1943). Karyotypes examination of patients suffering the syndrome showed chromosome breakage at the X chromosome (Lubs, 1969). This breakage was later identified as fragile site (Richards et al., 1981), the disease was thus called Fragile X syndrome (FXS). Affected males have moderate-to-severe mental retardation, speech delay, and a variety of behavioral and social problems (Richards et al., 1981). Normal alleles contain 6–55 repeats, premutation alleles contain 55–200 and cause Fragile X tremor/ataxia syndrome (FXTAS), and alleles that have above 200 repeats cause FXS (Crawford et al., 2001). Fragile X tremor/ataxia syndrome (FXTAS) typically occurs in older males carrying a premutation allele. The main clinical features of FXTAS are late-onset ataxia, intention tremor, cognitive deficits, rigidity, peripheral neuropathy[7], lower limb muscle weakness, and autonomic dysfunction (Jacquemont et al., 2007). Finally, Fragile XE syndrome (FRAXE) is caused by the expansion of a CCG repeat in the 5′ UTR of the fragile X mental retardation 2 (FMR2) gene. Patients affected with FRAXE present mild mental retardation, learning deficits, and developmental delay (Gecz, 2000).

Mechanism of pathogenicity associated to CGG expansions

CGG repeats are prone to methylation and FXS is the result of hypermethylation of the promoter of FMR1 gene leading to a loss of function of the corresponding protein (Oberlé et al., 1991). FXTAS is caused by a CGG-containing RNA transcript which sequesters key proteins in neuron functioning, such as Pur-α (Aumiller et al., 2012). Finally, more than 200 CGG (full mutation) results in the silencing of the FMR1 promoter and the loss of the associated protein is the cause of the disease (Willemsen et al., 2011).

Mechanisms of instability of CGG expansions

CGG repeats from FMR1 were identified as fragile sites (Verkerk et al., 1991). Fragile sites are chromosomal loci susceptible to breakage, visible on metaphase chromosome. Common fragile sites are visible in standard cell culture conditions (Magenis et al., 1970). Rare fragile sites such as in FXS are visible under folic acid deprivation (Richards et al., 1981) which was later shown to impede DNA replication (Glover, 1981). Fragile sites breakage and instability were well

[7] Nerves affection resulting in numbness and tingling in the feet or hands, pain in affected areas, loss of balance and coordination, muscle weakness, especially in the feet

studied in humans. ATR, a checkpoint protein controlling late S-phase integrity of the genome before entering into mitosis, was shown to regulate fragile site stability (Casper et al., 2002). Fragile sites are late replicated and sometimes un-replicated, leaving single stranded DNA inducing chromosome breakage (Glover et al., 2005). Additionally, FRA16B breakage mechanism was elucidated: secondary structure formation leads to replication fork stalling and reversal leading to breakage upon restart (Burrow et al., 2010). The progression of the replication fork through CGG repeats cloned into a bacterial plasmid was followed by 2D gels. Replication fork stalling was visible in *E. coli* for plasmids bearing more than 30 repeats of either CGG or GCC. The stalling was abrogated when CGG repeats were interrupted by AGG repeats (Samadashwily et al., 1997). Similar experiments in yeast showed that CGG/CCG repeats block replication fork, in a length dependent manner starting with only 10 repeats (Pelletier et al., 2003). These two experiments suggest that CGG may act as a fragile site and block replication fork leading to DNA breaks which may be repaired by homologous recombination leading to large expansions. Msh2 KO mice showed significantly reduced intergenerational instability of CGG repeat suggesting that Msh2 is a key factor for CGG expansions (Lokanga et al., 2014). The same trend was observed for CAG/CTG repeats, suggesting that MMR proteins may also recognize secondary structures formed by CGG repeats, inducing replication fork stalling.

GAA repeats

Secondary structure of the repeat

GAA form triple helices *in vitro*, exhibiting specific melting curves (**Figure 1.C**) (Gacy et al., 1998). In triple helices, the third strand is provided by one of the strands of the same duplex DNA molecule at a mirror repeat sequence, and is bound by Hoogsteen hydrogen bond. First evidence of this non B-DNA form was given by the increased stability of polyU and polyA stretches in a 2:1 ratio, forming a three-stranded structure (Felsenfeld and Rich, 1957). Intramolecular triplexes can be formed of T-A*T or C-G*C^+ sequence where * is a Hoogsten bound, – a Watson-Crick bound and C^+ a protonated cytosine. Because of the requirement of the cytosine to be protonated, this structure is called H-DNA. On the contrary, *H-DNA is maintained by T-A*A or C-G*G base triplets and is not pH dependent (Malkov et al., 1993). Evidences of formation *in vivo* were given by antibodies targeting triple stranded DNA (Lee et al., 1987) and single stranded probes (Ohno et al., 2002). However, none of these experiments

were carried out in physiological conditions, their transient and dynamic nature may explain why they are so difficult to detect.

Diseases linked to GAA expansions

Friedreich's ataxia is a neurodegenerative disorder characterized by progressive ataxia, followed by uncoordinated movements, tremors, and muscles waste. The first symptoms are usually observed in childhood or the early teens; the disease is progressive and evolves throughout the life of the patient (Zeigelboim et al., 2017).

Mechanism of pathogenicity associated to GAA expansions

GAA/TTC repeats were shown to block transcription by T7 RNA polymerase *in vitro*, in a length-dependent manner, 88 repeats completely abolishing transcription. Hence, the formation of DNA-RNA triplex structures blocks transcription, which in patients results in the loss of frataxin protein (Grabczyk and Usdin, 2000). Frataxin is essential in mitochondrial iron storage and regulation of iron levels; its loss eventually leads to cell death (Puccio et al., 2001).

Mechanisms of instability of GAA expansions

In *S. cerevisiae,* using 2D gels to monitor replication fork stalling, GAA repeats located on the lagging strand were shown to arrest replication fork, while TTC repeats did not affect replication (Krasilnikova and Mirkin, 2004). GAA repeats were confirmed to be prone to expansions only when the GAA strand serves as a lagging strand template (Shishkin et al., 2009). It was thus postulated that DNA polymerase stalls on the lagging strand due to GAA triplexes, while the polymerase on the leading strand continues, leading to long stretches of single stranded DNA. The stalling region is bypassed when the stalled strand invades its sister chromatin. This pathway is called template switching and may account for large expansions at GAA repeats. Additionally, MMR suppression in mice by inactivation of *Msh2* and *Msh6* and *Psm2* prevents instability in GAA expansions in neurons (Bourn et al., 2012). Finally, H-DNA colocalizes with fragile sites such as c-Myc locus (Kinniburgh, 1989) and BCL-2 locus (Raghavan et al., 2005). In yeast, GAA trigger DSBs (Kim et al., 2008) and the GAA repeat expansion at FXN locus in lymphoblastoid cells was linked to chromosomal breakage (Kumari et al., 2015). DSBs induced by H-DNA may account for its instability, although formal evidence lacks to support this hypothesis.

GGGGCC repeats

Secondary structure of the repeat

In vitro, X-ray diffraction demonstrated that guanylic acids can assemble into tetrameric structures. In these tetramers, four guanine molecules form a square in which each guanine is hydrogen-bound to the two adjacent guanines (Gellert et al., 1962) (**Figure 1.B**). Later, G-quadruplexes were shown to be formed *in vivo.* The deletion of a specific helicase in *C. elegans* led to the systematic deletion of polyglutamine tracts. The helicase was renamed dog-1 for deletions of guanine-rich DNA. This helicase appears to be essential for the resolution of GGGGCC repeats during replication, indirectly showing the formation of secondary structures *in vivo* (Cheung et al., 2002). To confirm secondary structure formation *in vivo,* antibodies were isolated to target G-quadruplexes (Schaffitzel et al., 2001). In yeast, analysis of G4 in *pif1Δ* mutants revealed that this replication helicase prevented genomic instability in the G-rich human minisatellite CEB1 inserted in *Saccharomyces cerevisiae* genome (Ribeyre et al., 2009). ChiP-Seq revealed that Pif1 bound to G4 motifs (Paeschke et al., 2011). Through ChiP microarray experiments of DNA associated to DNA Polymerase 2 in a strain where PIF1 expression is reduced (*pif1-m2* allele) and 2D gels, it was revealed that Pol2 accumulated at G4 sequences and that replication forks were slowed down at these loci. Ligands of the PhenDC family (De Cian et al., 2007) were used to probe the formation of G-quadruplexes *in vivo*. *pif1Δ* cells treated with the Phen-DC3 ligand showed an increased CEB1 instability, confirming G4 formation *in vivo* (Piazza et al., 2010).

GGGCC repeat in c9orf72 was shown to form DNA G-quadruplex as well as RNA G-quadruplex and DNA-RNA hybrids through biophysical techniques: such as circular dichroism absorption measurement showed a characteristic spectrum for antiparallel G-quadruplexes. Dimethylsulfate protection assay revealed that guanines are protected from cleavage, indicating the formation of hydrogen bonding in a G-quartet conformation (Haeusler et al., 2014).

Diseases linked to GGGGCC expansions

GGGGCC expansions located in c9orf72 are the cause of amyotrophic lateral sclerosis (ALS) or Charcot disease and result in progressive motor dysfunction (Rowland and Shneider, 2001). The GGCCTG expansion in NOP56 causes SCA36 which is also a neurodegenerative disorder, characterized by a late-onset, progressive cerebellar ataxia typically associated with hearing loss (Kobayashi et al., 2011).

Mechanism of pathogenicity associated to GGGGCC expansions

DNA/RNA hybrids were shown to stall RNA polymerase, impeding transcription and leading to nucleolin mislocalization. Nucleolin is an essential nucleolar protein and colocalizes with both DNA and RNA containing GGGGCC repeats (Haeusler et al., 2014). The c9orf72 GGGGCC expansion was shown to form RNA G-quadruplex inclusions which sequesters hRNP, an important splicing factor in ALS brains (Conlon et al., 2016), and Pur α (Xu et al., 2013). Pathogenesis of SCA36 most probably involves an RNA-gain of function mechanism (Kobayashi et al., 2011).

Mechanisms of instability of GGGGCC expansions

GGGGCC was shown to disrupt replication fork progression by performing 2D gels on replicative HEK293T expressing a plasmid containing either 21 or 41 GGGGCC. Both lengths impair replication, although longer repeat show a more dramatic effect (Thys and Wang, 2015). As discussed above for CAG/CTG, replication fork stalling may also lead to expansions. Additionally, R-loops can form *in vitro* in GGGGCC repeat and contribute to its instability (Reddy et al., 2014).

CCTG repeats

Secondary structure of the repeat

By observing the behavior of oligos containing repeats after enzymatic or chemical treatments, it was inferred that CAGG form hairpins. No structure was observed for CCTG in the tested conditions, suggesting that CAGG hairpins are more stable (**Figure 1.D**) (Dere et al., 2004). However, using Nuclear Overhauser effect which is a more recent type of nuclear magnetic resonance, it was suggested that CCTG may form hairpins with a two-residue CT loop or a dumbbell (Lam et al., 2011).

Diseases linked to CCTG expansions

DM2 is caused by an expanded CCTG at the ZNF9 gene; it shares the same symptoms as DM1 and is in general milder than DM1 (Machuca-Tzili et al., 2005).

Mechanism of pathogenicity associated to CCTG expansions

The mechanism leading to DM2 resembles DM1, with CCUG-expanded RNA sequestering splicing factors similar to DM1 (Ranum and Cooper, 2006).

Mechanisms of instability of CCTG expansions

CCTG/CAGG repeats were transfected in green monkey kidney cell line COS-7. Instability was greater when CAGG was on the leading strand, and instability was length dependent (Dere et al., 2004). As discussed above for CAG/CTG, replication fork stalling may lead to expansions. The involvement of BIR or MMR, in the instability still has to be proven.

ATTCT repeats

Secondary structure of the repeat

DNA unpairing at ATTCT repeats in supercoiled DNA was detected by 2D gels, indicating that ATTCT form structures similar to DNA unwinding elements (DUE) (Potaman et al., 2003). DUE were first observed in *S. cerevisiae* as easy to unwound sequences, located at replication origins (Umek and Kowalski, 1990). DUE are a common feature of prokaryotic and eukaryotic replication origins and act as a start point for strand separation and unwinding of the DNA double helix.

Diseases linked to ATTCT expansions

SCA10 is due to an expanded ATTCT in the ataxin10 gene. The expansion can be very large, up to 22.5kb (Matsuura et al., 2000). Symptoms are similar to other SCA (Paulson, 2009).

Mechanism of pathogenicity associated to ATTCT expansions

The pathogenesis mechanism is unknown although studies in mice revealed that either loss or gain of function of ataxin fail to recapitulate the disease features (Wakamiya et al., 2006). As ATTCT expansion is intronic, it was postulated that the expanded AUUCU-containing mRNA was toxic. RNA aggregates were also observed in SCA10 patient cells, colocalizing with hnRNP K protein aggregates, previously identified by coimmunoprecipitation assays with AUUCU repeats (White et al., 2010). To confirm this hypothesis, a new transgenic mouse model was constructed in which the pentanucleotide repeats were integrated in 3'UTR of the LacZ gene to ensure transcription without translation. The resulting mice showed irregular gait, increased seizure and neuronal loss, resembling SCA10 patient phenotype. Pathogenesis mechanism of SCA10 is thus more likely an RNA gain of function mechanism (White et al., 2012).

Mechanisms of instability of ATTCT expansions

ATTCT repeats were able to trigger aberrant replication initiation in Hela cells. Instability of the repeat may come from the firing of replication after the replication fork has already passed through the repeat leading to further replication and massive expansions (Potaman et al., 2003) (**Figure 4**). ATTCT repeats were later linked to fragility and to expansions in a yeast reporter assay where ATTCT repeats were integrated in the middle of a *URA3* gene. *rad5* mutants did not showed a decrease in expansion rate and fragility of the repeat tract. Rad5 is involved in template switching, which appears to be another determinant step for ATTCT instability (Cherng et al., 2011).

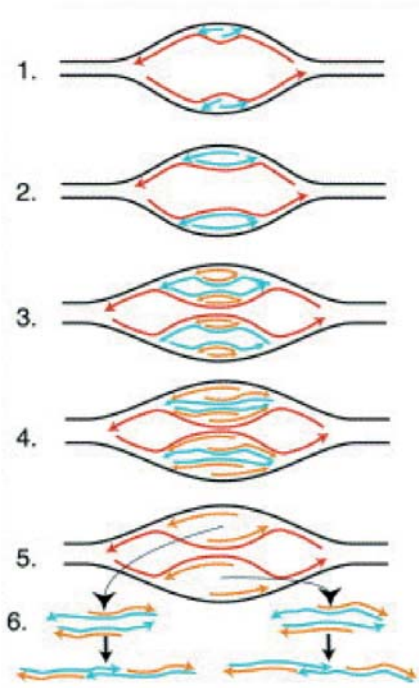


Figure 4 : Model for repeat instability based on aberrant replication origin activity triggered by $(ATTCT)_n$ repeats. Onion skin replication can occur via continued initiation within the A+T-rich sequence. Continued replication will lead to a displacement of four DNA molecules from the original DNA molecule (steps 1 to 5). These molecules can be joined into very long strands (step 6) and become incorporated into the repeat tract leading to massive expansion. From (Potaman et al., 2003).

TGGAA repeats

Secondary structure of the repeat

It was shown using NMR that TGGAA repeats were able to form hairpins with a 1-nt loop (Zhu et al., 1996). Later, it was demonstrated that TGGAA actually form two different hairpin structures depending on repeat number parity, by applying FRET and X-ray crystallography on $(TGGAA)_3$ and $(TGGAA)_4$ oligonucleotides. Odd numbers form an end to end conformation with a 1-nt loop whereas even numbers can either form a 1-nt loop or an octaloop (**Figure 1.E**) (Huang et al., 2017).

Diseases linked to TGGAA expansions

TGGAA expansion in the BEAN and TK2 genes is the cause of SCA31, a late-onset of cerebral ataxia (Sato et al., 2009).

Mechanism of pathogenicity associated to TGGAA expansions

The disease mostly affects Purkinje cells. RNA foci were found using a probe recognizing gene transcript containing TGGAA sequence. Electrophoretic mobility shift assay revealed proteins that bind to UCCAA RNA. TGGAA repeats are most likely to act in an RNA-mediated mechanism which leads to sequestration of splicing factors (Sato et al., 2009).

Stability of TGGAA expansions

In normal individuals, no TGGAA sequence can be found in place of the TGGAA expansion in SCA31 patients. This means that the expansion was the result of an insertion. Since TGGAA sequence can be found in the centromeres of 8 chromosomes in humans, this mutation may have arisen from an heterochromatin insertion (Sato et al., 2009). A study on 17 parent-child pairs suggested that TGGAA repeat tracts do not exhibit a marked inter-generational instability. Small changes can occur as the mean length of change was 12.2 bp during transmission. However the change was not significantly different between two generations (Yoshida et al., 2017).

GCN repeats

Secondary structure of the repeat

No secondary structure has been reported for GCN repeats. This may be because it was not studied or because the impurity of the repeat prevents secondary structure to form.

Diseases linked to GCN expansions

The GCN expansion in HOXD13 causes synpolydactyly type I (SPD), and results in congenital limb malformation (Muragaki et al., 1996)(Goodman et al., 1997). Eight other polyalanine tract disorders exist: oculo-pharyngeal myotonic dystrophy[8] (Calado et al., 2000) or holoprosencephaly[9] (Roessler et al., 2009), hand–foot–genital syndrome[10], congenital central hypoventilation syndrome[11], X-linked mental retardation, X-linked mental retardation and growth hormone deficit[12], blepharophimosis-ptosis-epicanthus inversus syndrome[13], cleidocranial dysplasia[14] (Amiel et al., 2004).

Mechanism of pathogenicity associated to GCN expansions

GCN repeat translated in alanine are implicated in the gain of function of the HOXD13 protein which is an essential transcription factor during development (Muragaki et al., 1996).

Mechanisms of instability of GCN expansions

Stability over at least seven generations was observed in a large SPD family (Akarsu et al., 1996). The repeats are not pure and are made of cryptic alanine codon of either GCC, GCA, GCT, GCG. These interruptions may stabilize the mutation dynamic. It was hypothesized that the first expansion of SPD has arisen from crossover: careful examination of three mutant alleles revealed a plausible event of recombination between two mispaired normal alleles by crossing over within a short trinucleotide tract (Warren, 1997). However, this mechanism might not be true for other polyalanine disorders such as congenital central hypoventilation syndrome. Over 161 patients affected with this disease, 3 showed somatic mosaicism. If unequal crossing over was leading to this instability, wild-type, expanded and unexpanded alleles would be

[8] Slowly progressing myopathy affecting the muscles of the upper eyelids and the throat
[9] Cephalic disorder in which the forebrain of the embryo fails to develop into two hemispheres
[10] Impaired development of the hands and feet, the urinary tract, and the reproductive system
[11] Breathing affection causing hypoventilation
[12] Mental deficiency associated with specific abnormalities due to lack of growth hormone
[13] Condition affecting the development of the eyelids
[14] condition affecting the development of the bones and teeth

observed. Only expanded and wild-type alleles were observed, suggesting that another mechanism may lead to expansions (Trochet et al., 2007). Finally, a spontaneous expansion in mice in the Hoxd13gene, arose probably by replication fork stalling and template switching resulting in the duplication of the normal polyalanine tract (Johnson et al., 1998).

Myotonic Dystrophy type I

CTG expansion length and purity impact on disease severity

In 1909, Steinert and colleagues first described myotonic dystrophy type I (DM1) which was called Steinert disease; in 1992 the DMPK gene and CTG expansions were found to be the cause of this disorder (Brook et al., 1992). The gene is located on chromosome 19q13.3 and the CTG expansion is located in the 3'UTR. The disease is an autosomal dominant and multisystemic disorder, mainly characterized by progressive muscle weakness and myotonia, sometimes associated to cataract, cardiac arrhythmia and cognitive impairments. A positive correlation between repeat size and severity of the disease exists. As the length of the repeat increases during transmission from parents to children, the severity of the disease increases from one generation to the next, resulting in an earlier age of onset. This phenomenon is called anticipation (Richards and Sutherland, 1992). Repeat expansions are called "dynamic mutations" which means that the longer the repeat tract the more likely the expansion (Richards and Sutherland, 1992). In patients affected by DM1 the repeat size ranges from 50 to 4,000 triplets (150–12,000 bp). Healthy genomes contain from 5 to 37 repeats. The DM1 mutation length above 2,000 repeats causes the congenital form of the disease with the most severe symptoms and earliest age of onset (Ashizawa et al., 2000). Somatic cell mosaicism is often observed in patient cells, larger expansions being found in muscle as compared to blood cells (Higham et al., 2012)(Thornton et al., 1994). Repeat purity seems to affect the stability of the sequence. Interrupted repeats such as CCG, CTC and GGC stabilize the repeat tract and reduce its instability repeat during patient lifetime, resulting in milder symptoms. Additionally, interrupted repeats may fold in different structures, affecting affinity for downstream DM1 effectors (Cumming et al., 2018).

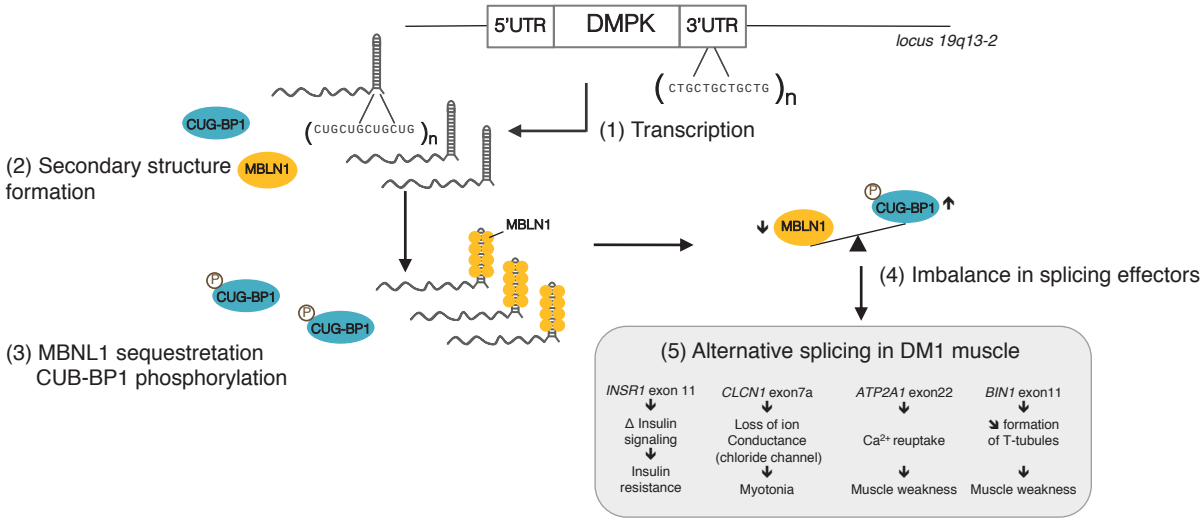


Figure 5: RNA-mediated mechanisms leading to Myotonic Dystrophy type I. CTG repeat expansions are transcribed into $(CUG)_n$ expanded RNA (1). These RNAs form secondary structures and aggregates (2) which are going mediate to MBNL1 sequestration and CUG-BP1 phosphorylation (3). The imbalance between these two splicing factors (4) is going to cause major splicing defects in DM1 muscle cells. Each misspliced transcript will result in cellular defects leading to particular symptoms for the DM1 patients (5).

RNA-mediated pathogenesis

Identification of RNA foci and CUG-bound proteins

Pioneering works revealed the existence of RNA foci by RNA-FISH using a CAG probe, in DM1 cells. The probe bound to CUG-expanded RNA, and up to 13 foci per nucleus were observed in skin fibroblasts (Taneja et al., 1995). The identification of a CUG binding protein (CUGPB1) by band-shift analysis was a first step toward the understanding of the molecular pathogenesis mechanism responsible for DM1 (Timchenko et al., 1996). It was later shown that CUG-BP1 does not colocalize with foci in DM1 cells (Jiang et al., 2004) and is more stable due to hyperphosphorylation in patient muscles and heart (Kuyumcu-Martinez et al., 2007). Proteins bound to CUG expanded RNAs transfected into HeLa cells were purified and analyzed by mass spectrometry, allowing to identify the MBNL protein (Miller et al., 2000). Each of the three MBNL isoforms were subsequently shown to be sequestered by CUG RNA in DM1 nuclei. MBNL proteins have extensive roles in muscle cells especially in regulating of the expression and splicing of several mRNA, making their sequestration harmful for the cell (Wang et al., 2012) (**Figure 5**).

Protein sequestration by expanded CUG containing RNA result in splicing defects

To understand MBNL role in DM1 pathogenesis, knock-out mice were produced. Mbnl1 knock-out mice recapitulate most features of DM1: muscle abnormalities, cataract and RNA miss splicing (Kanadia et al., 2006). Splicing defects of cTNT gene was identified in patients; and in muscle cells, missplicing was increased when CUG-BP1 level increased (Philips et al., 1998). Many splicing defects were then linked to various disease symptoms usually resulting in the expression of embryonic or alternative forms, for example insulin receptor splicing defects results in the expression of the isoform A, which is less responsive to insulin, inducing insulin resistance in DM1 patients (Savkur et al., 2001) (**Figure 5**). However studies performed in animals show that mRNA splicing, muscle pathogenesis and RNA foci can be dissociated in mice models, with one model only recapitulating part of the disease (Gomes-Pereira et al., 2011). It suggests that other elements contribute to DM1 phenotype.

Other factors contribute to DM1 pathogenesis

Reduced DMPK levels were observed in patients; this reduction is currently thought to be a consequence of RNA sequestration in the nuclei rather than playing a central role in pathogenesis (Furling et al., 2003). As a confirmation, downregulation of the DMPK gene in mice (Dm15) does not result in myotonic dystrophy-like disease (Jansen et al., 1996).

RAN translation may also play a role in the pathogenic mechanism (Zu et al., 2011) as peptides resulting from this alternative translation are toxic. Also, misregulation of miRNAs which are implicated in gene transcription regulation may also participate in the disease pathogenicity; miR-1 processing is altered in heart samples from people with myotonic dystrophy. MBNL1 was identified as a regulator of pre-miR-1 biogenesis (Rau et al., 2011). Finally, the surrounding of CTG repeat expansion were shown to be less amenable to DNaseI digestion and PvuII cutting probably because of a more condensed chromatin state (Otten and Tapscott, 1995). Consistent with this, SIX5 which is a downstream neighbor of DMPK shows reduced mRNA levels in patients (Thornton et al., 1997). SIX5 knockout mice only develop cataracts, suggesting that SIX5 reduced expression in patients contribute also to DM1 pathogenesis (Klesert et al., 2000).

To conclude, DM1 pathogenesis is a multifactorial and complex process, involving not only alternative splicing but also changes in gene expression, unconventional translation, and microRNA deregulation.

In vitro and *in vivo* models of DM1

Patient cell models

ASA cells are fibroblasts from patient skin harboring over 1000 CTGs (Arandel et al., 2017). They were immortalized by constitutive expression of the hTERT gene, encoding the human telomerase reverse transcriptase. They contain a doxycycline-inducible MyoD gene. When doxycycline and insulin are added to the culture media, cells differentiate into myotubes. Since myotubes are the cells exhibiting the most extreme disease phenotype, cell differentiation is used to study phenotype changes. ASA cells exhibit numerous RNA foci and show many splicing defects as in DM1 patients. Phenotype reversal induced by a drug can be easily quantified, making this cell model an attractive tool for drug testing (**Figure 6**).

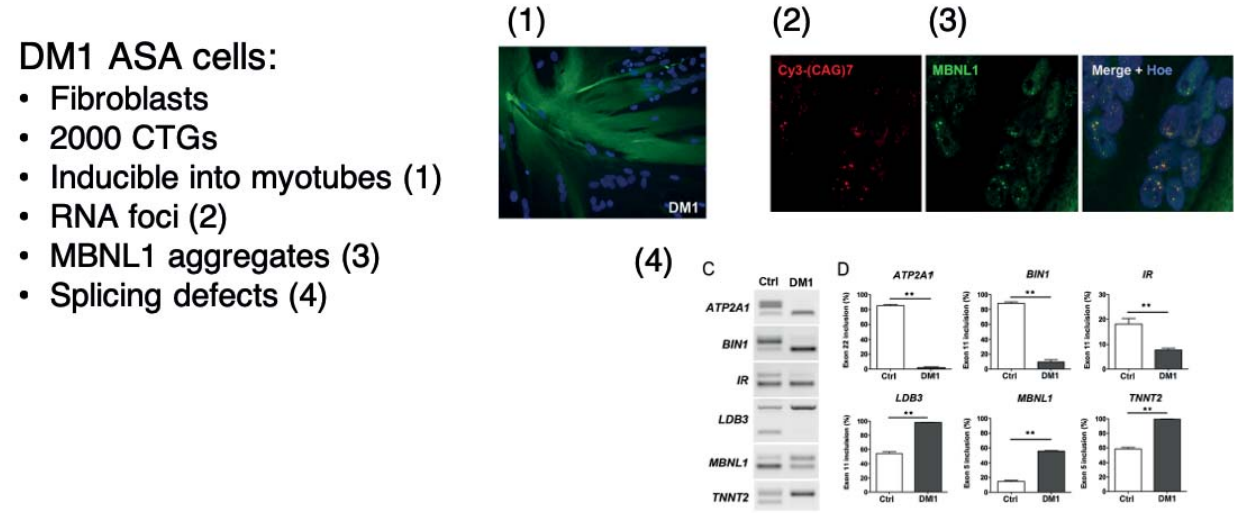


Figure 6: Main features of DM1 ASA cell model developed by the Institut de Myologie. DM1 patient skin fibroblasts were modified to obtain an immortalized, inducible model. (1) Myotubes formed by the fusion of myoblasts. Nuclei are stained with Hoechst and Desmin is stained in green using an anti-desmin antibody. Desmin is a muscle-specific protein implicated into sarcolemma formation. (2) RNA foci are visualized using RNA-FISH with a Cy3-$(CAG)_5$ probe. (3) MBNL1 expression is revealed using an antibodyanti-MBNL1. (4) Splicing defects measured by RT-PCR. Significant differences with control are found for these 6 transcripts. From (Arandel et al., 2017).

Mice models

Several mice model exist for DM1, including MBNL1 knock outs (Suenaga et al., 2012), CELF1 overexpression (Ward et al., 2010) or untranslated CUG expression (Mankodi et al., 2000). This last mouse model is called HSA^{LR}. It carries a genomic fragment containing the human skeletal actin (HSA) gene alongside 250 repeats. The model was extensively used to test drugs targeting CUG expanded RNA (Sobczak et al., 2013) and to decipher the basis of the RNA mediated pathogenesis in DM1 (Mankodi et al., 2002).

However, a model harboring the human genomic locus of DM1 would be the most valuable for the evaluation of gene editing approaches. Such a mice model was developed by Genevieve Gourdon and was called DM55 (Gourdon et al., 1997). Large genomic fragment of 45 kb of the DMPK gene from a patient with an affected allele containing ~50 CTGs was integrated into C57BL/6 mice. Later, other C57BL/6 mice were modified with a genomic fragment from the daughter of the patient whose genome was used to generate DM55 mice. Her genome contained one normal allele with 20 CTGs and an affected allele with ~500 CTGs. Several mice lines were isolated, carrying either 20 CTGs or 300 CTGs, respectively called DM20 and DM300. An increased instability in DM300 mice was observed as compared to lines bearing less repeats, exhibiting additional repeats ranging from 1 to 60 per generation, as compared to DM55 showing at most a +6 CTGs increase in repeat size. Somatic instability increased with age, most notably into liver, pancreas and kidney with up to +100 CTGs after 10 months where DM55 showed at most an increase of +12 CTGs. Germline cells also showed expansions. However, no instability was observed in the blood (Seznec et al., 2000). A major onset of DM1 is the large intergenerational instability towards expansion often observed in patients also called “big jumps” and that are responsible for the anticipation phenomenon. These “big jump” were recapitulated for the first time in DM300 mice leading to the establishment of the DMSXL mouse lineage carrying over 1000 CTGs. Two mice derived from DM300 mouse line, one female with 460 CTG and a male with 430 CTG transmitted big jumps of +250 and +480 to offsprings. Mild phenotype was observed in hemizygous mice. However, in homozygous mice a severe phenotype was observed, suggesting that transgene expression is too low when one copy is present and that there is a dose effect of the toxic RNA. Phenotypic alterations of homozygous mice are: high mortality, growth retardation splicing abnormalities in muscle and the central nervous system (Gomes-Pereira et al., 2007). In addition, DMPK transgene expression localization in DMSXL homozygous mice is similar to human patients, and located in heart and muscles. Sense and antisense transcripts are detected, as well as foci (up to 30% cells containing foci) and mild splicing defects -shift up to +40% for Lsdb3 exon 11

missplicing. Finally, muscle strength of DMSXL homozygous mice is impaired, retaining 62% of the WT tetanic force (Huguet et al., 2012) (**Figure 7**). One limitation of using DMSXL mice is the breeding strategy which is time consuming as homozygous cannot reproduce: to obtain homozygous mice, hemizygous mice are bred.

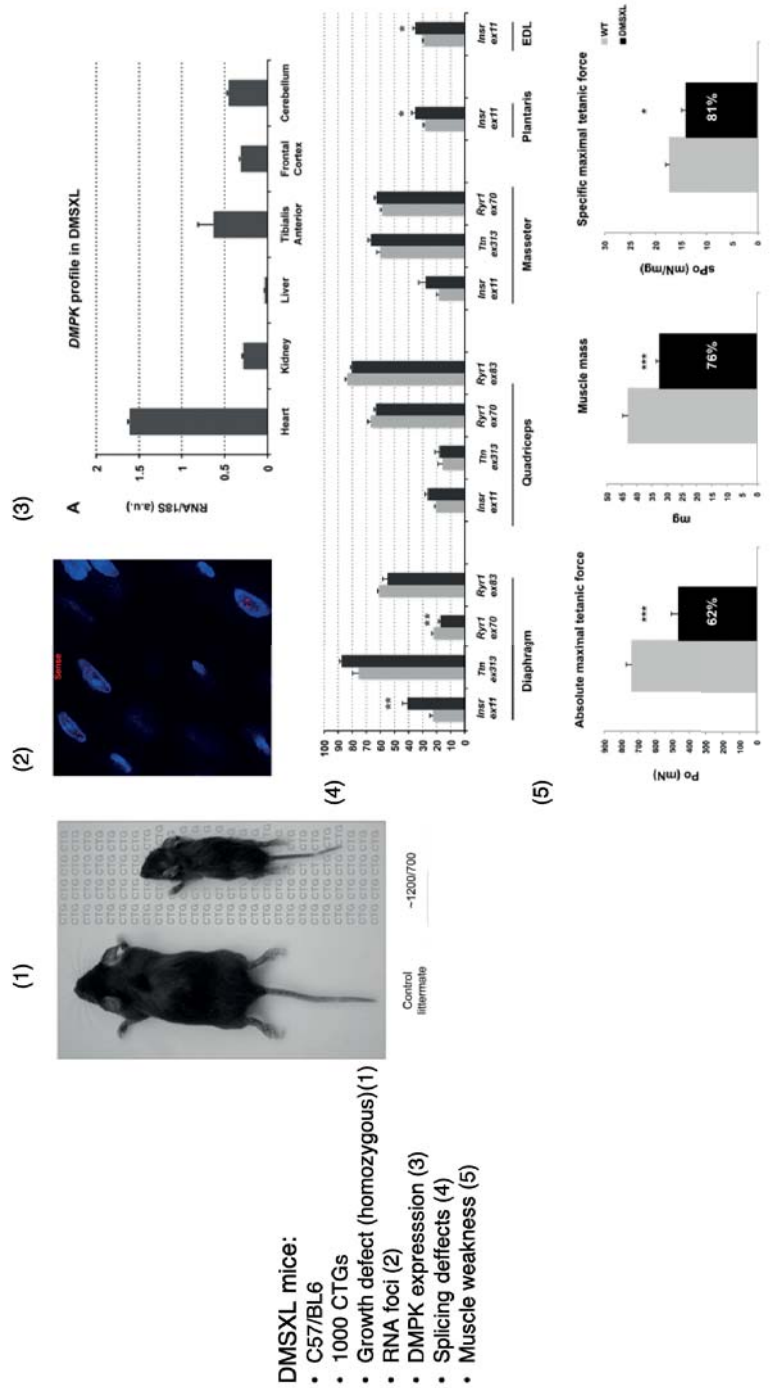


Figure 7: DMSXL mice model. These mice carry a long $(CTG)_{\sim 1000}$ fragment. (1) Homozygous DMSXL mice show a severe macroscopic phenotype with shorter size and higher mortality rate. (2) Muscle cells exhibit RNA foci visible using a Cy3-$(CAG)_{15}$ probe. (3) The DMPK gene from human genomic fragment is expressed in muscle tissues and mimic what is observed in DM1patients. (4) Splicing defects are visible in various muscle tissues. (5) Muscle weakness is lowered as shown by tetanic force quantification and muscle mass measurement. From (Huguet et al., 2012).

Recent development of therapeutics for DM1

Targeting toxic CUG-expanded RNA

For now, only symptomatic treatments exist for DM1. Most promising drugs developed so far are trying to prevent CUG-RNA and MBNL1 interactions: small molecules (Rzuczek et al., 2017), or morpholinos antisense oligonucleotide CAG25 (Wheeler et al., 2009). CRISPR-Cas mediated degradation of CUG-RNA showed promising results (Batra et al., 2017). Antisense oligonucleotide mediating the specific degradation of CUG-RNA by recruiting RNAse-H (Wheeler et al., 2012) was brought to the clinic by Ionis which tested their IONIS-DMPKRx. But the lack of potency of the drug did not allow them to pursue the clinical trial after phase 1/2 as the dose to reach objectives would have been above authorized thresholds. The company is now trying to improve the chemistry of its oligonucleotide using LICA (Ligand-Conjugated Antisense) technology.

Targeting CTG expansions

As the cause of the disease is the expanded CTG, it is possible to think that eliminating or shortening the CTG expansion to non-pathological length could suppress symptoms of the pathology and could be used as a new gene therapy approach (Richard, 2015). Natural contraction of an expanded CTG repeat tract (600 repeats) was observed during transmission from father to daughter of an expanded myotonic dystrophy allele and clinical examination of the daughter showed no symptom of the disease (O'Hoy et al., 1993). Another similar observation of the transmission of a contracted allele from father to son was also reported. At the age of 35, the patient was still asymptomatic (Shelbourne et al., 1992).

Inducing double-strand breaks at specific locations in the genome is currently the main approach to induce targeted genome editions. Cells can either rejoin the extremities of the DSB with the possibility to introduce insertions or deletions (indels), or repair using a donor template and integrate exogenous DNA at the break site. Four large families of specific nucleases exist: meganucleases, ZFN, TALEN and CRISPR-Cas nucleases. In the following chapter, each of these families will be described.

Programmable nucleases and genome modifications

Highly specific nucleases

Meganucleases

HO (Kostriken et al., 1983) and I-SceI (Colleaux et al., 1986) were the first meganucleases discovered. I-*Sce*I was shown to be encoded by a yeast mobile genetic element. Through homing process, meganucleases cleave their target gene, initiating a homologous recombination event that results in the transfer of the mobile element into the cleaved allele and the spreading of the meganuclease encoding gene. I-*Sce*I cleaves with a high specificity an 18-bp sequence (Colleaux et al., 1988) (Nickoloff et al., 1986). Following this discovery, I-*Sce*I recognition sequence was integrated in various genomes in order to induce localized double-strand breaks: in mouse cells (Rouet et al., 1994), in human cells (Porteus and Baltimore, 2003). However, the use of I-*Sce*I was dependent on the prior introduction of an 18 bp target sequence in the gene of interest, which limited any therapeutic application. The first report of genome edition using I-*Sce*I in mouse cells disrupting a ß-galactosidase-positive phenotype was made in 1995 (Choulika et al., 1995). Engineering meganucleases was proved to be a daunting task even when crystal structures were available, I-*Cre*I being the first to be crystalized (Heath et al., 1997). Although the making of entirely synthetic and rationally designed homing endonucleases cleaving the human RAG1 gene was a success (Smith et al., 2006), easier alternatives were favored for genome editing. I-*Sce*I was extensively used in various organisms to induce DSBs and to study DNA repair (**Figure 8**). Cellectis was a pioneer into the design and generation of novel meganucleases with new specificities (Arnould et al., 2006).

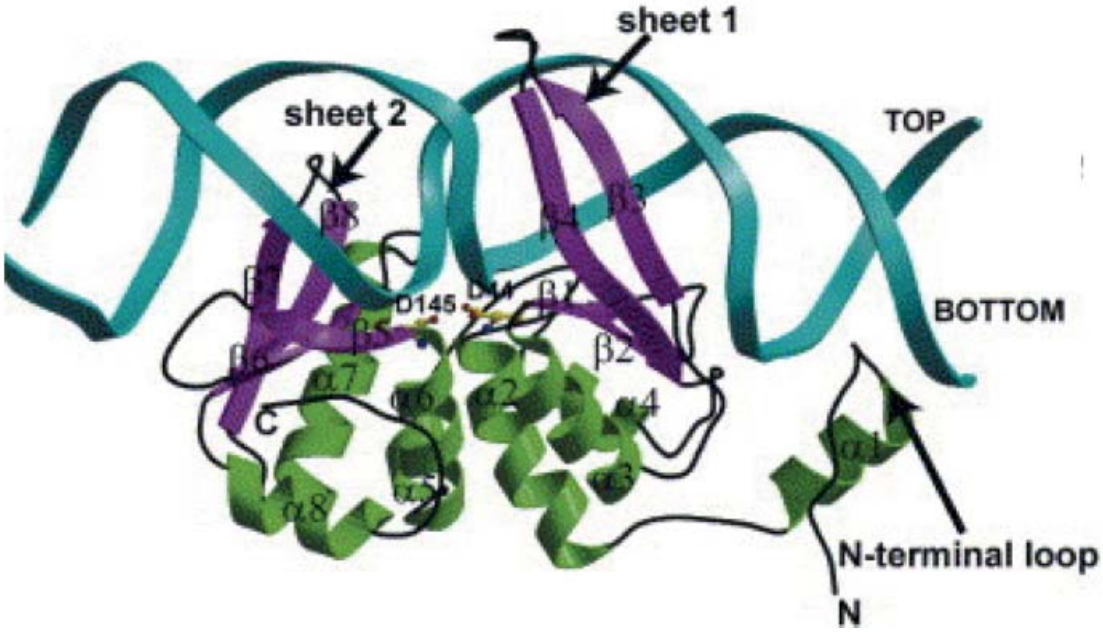


Figure 8: Overall structure of the I-SceI–DNA complex determined by crystallography. DNA (cyan ribbon) is in contact with sheet 1 and sheet 2 (major groove contacts) and the N-terminal loop (minor groove contacts). α-Helices are depicted in green and β-strands in magenta. The catalytic aspartate residues Asp44 and Asp145 are labeled. It shows a ααββαββα topology characteristic of LAGLIDADG endonuclease structures. From (Moure et al., 2003).

Zinc-Finger Nucleases

Zinc-finger proteins (ZFPs) are involved in several cellular processes acting through different molecular mechanisms, able to bind DNA, RNA, proteins. The crystallization of the mouse immediate early protein Zif268, a well-studied master regulator of cell development and function (Lemaire et al., 1988), was a key step toward understanding ZFP motif binding to DNA (Pavletich and Pabo, 1991) (Elrod-Erickson et al., 1996). In classic Cy2His2 zinc fingers, two cysteine and two histidine residues are bound to a zinc atom, which stabilizes the structure. An individual zinc-finger consists of approximately 30 amino acids in a conserved ββα order, contacting one DNA triplet. Each ZF/DNA interaction is mediated by four residues. A library of ZF variants was created by randomizing these four residues; using phage display, the affinity of these ZF variants was tested by affinity selection method on different DNA sequences. ZF amino acid sequence was thus associated to a specific DNA sequence (Kim et al., 1996). The recognition of a particular sequence was then achieved by assembling specific ZF motifs (Liu et al., 1997). However, a crosstalk exists between adjacent residues, with some recognizing a 4bp-overlapping region instead of 3-bp motifs, making the engineering of specific ZFP challenging (Isalan et al., 1997).

ZFP were fused to the FokI catalytic domain and were able to cleave a specific sequence in a test tube (Kim et al., 1996). FokI is a type IIs restriction endonuclease and was discovered in *Flavobacterium okeanokoites*. It belongs to a class of restriction endonucleases that recognize specific nucleotide sequences and introduce staggered cleavages at positions away from the recognition sequence (Sugisaki and Kanazawa, 1981). Its crystal structure shows great similarities with the BamHI catalytic domain. The catalytic domain must dimerize for the catalytic site to be complete and for DNA cleavage to occur (Bitinaite et al., 1998) (Wah et al., 1998). Further characterizations showed that the catalytic domain of FokI needs to dimerize to cut, making the recognition sequence of 18 bp in total, with a 4bp spacer in between (Smith et al., 2000). This chimeric nuclease was made of two sets of ZFPs each composed of three Cys2His2 zinc fingers recognizing 9 bp and was called Zinc Finger Nuclease (ZFN) (**Figure 9**).

ZFN were first used in *Xenopus laevis* oocytes in which they induced tandem repeat recombination (Bibikova et al., 2001). The same team later gave first evidence of gene editing in drosophila and achieved the modification of the yellow eye gene in 1% of offsprings expressing ZFN and a linear extrachromosomal donor expressing the mutated yellow gene (Bibikova et al., 2003). Then ZFN were shown to induce homologous recombination at levels comparable to I-*Sce*I in HEK293 using a GFP reporter assay. However, after effective edition,

cells expressing I-*Sce*I remained stable while cells expressing ZFN decreased, suggesting that long term expression of ZFN may be toxic due to off target cleavage (Porteus and Baltimore, 2003). FokI was found to be able to form homodimer leading to a low specificity as one arm of ZFN was able to induce a DSB by itself. Modification of FokI led to a new architecture of the protein which has to form an heterodimer to be able to induce a DSB (Miller et al., 2007).
Nevertheless, Sangamo, a pharma company, is still using ZFNs as a way to edit genome for therapeutic purposes. Engineering the FokI domain by a single base mutation (Q481A) located in the catalytic site resulted in no detectable off-target activity. The proposed mechanism underlying this improvement is probably that the catalysis is slowed and more time is left for the nuclease to discriminate between off and on target. This improvement can be used in addition to modifications of the ZF binding domain (Miller et al., 2019). It is now possible to induce with high precision a DSB at a desired position (Paschon et al., 2019). The use of ZFN for *in vivo* genome modification was tested for example in hepatocytes to remove Hepatitis B Virus (Weber et al., 2014), or in mice to cure hemophilia (Li et al., 2011). In 2017, Sangamo had three phase 1 /2 clinical trials for genome editing of Hemophilia B, Mucopolysaccharidosis I and II (Laoharawee et al., 2018), all of which implicating *ex vivo* modifications of target cells using ZFNs.

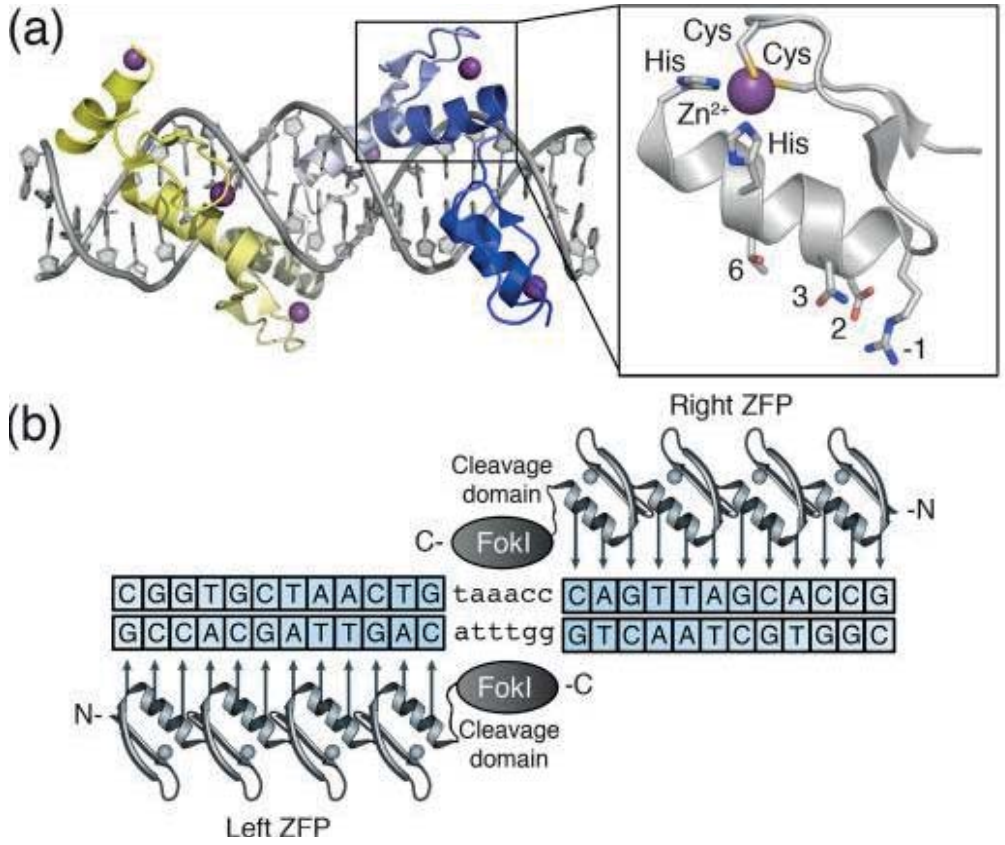


Figure 9: Structure of zinc-finger Nucleases. (a) ZFP in complex with target DNA (grey). Each zinc-finger consists of approximately 30 amino acids in a ββα arrangement (inset). Surface residues (−1, 2, 3 and 6) that contact DNA are shown in the insert (inset). The side chains of the conserved Cys and His residues are depicted as sticks in complex with a Zn2+ ion (purple). (b) Cartoon of ZFN dimer bound to DNA. It is made of two zinc-finger binding sites separated by a spacer sequence cleaved by the FokI cleavage domain. From (Gaj et al., 2013).

Transcription Activator Like Effector Nucleases

Xanthomonas is a pathogenic agent of several crop plants. These bacteria use Transcription Activator Like effectors (TALE) to reprogram host cells by mimicking eukaryotic transcription factors to their own benefit. Molecular analysis of the avirulence gene avrBs3 revealed that TALE are made of 34bp nearly identical repetitive sequences repeated 17.5 times (Herbers et al., 1992). TALE effectors are made of a repeated domain, a nuclear localization signal and an acidic transcriptional domain (Schornack et al., 2006). These TALE were later identified as DNA binding proteins, acting as transcription factors. Cell reprogramming results in developmental changes. AvrBs3 activates a regulator of cell size: upa20 by binding to its promoter, inducing hypertrophy of plant leaves (Kay et al., 2007). The repeated domains are identical except for two hypervariable residues: amino acids 12 and 13. Knowing the sequence of the target of AvsB3 (UPA box), amino acids 12 and 13 were linked to their cognate base on the UPA box. This code was then confirmed on other known *in vivo* targets of different TALE. Experimental validation of this model was achieved by predicting DNA sequences of known TALE *in vivo* (Boch et al., 2009). Amino acids 12 and 13 were called repeat variable diresidues (RVD). Co-crystal structures of TALE DNA-binding domains revealed that residue 8 and 12 within the same repeat make a contact with each other that may stabilize the structure of the domain while the residue at position 13 makes base-specific contacts with the DNA (Deng et al., 2012). Additional work optimized the RVD/DNA code using a Systematic evolution of ligands by exponential enrichment (SELEX) assay. This method relies on the assessment of the binding affinity of oligonucleotides library, and was used to determine the binding preferences of four engineered TALE. It led to the validation of a new RVD for improved recognition of guanine (Miller et al., 2011).

Similarly to the engineering of ZFN, TALE were fused to the FokI catalytic domain, adjustments were made to determine the optimum spacer length which was between 15 and 21bp, due to the large C-terminal region of TALE (Christian et al., 2010). This new protein was called Transcription Activator-Like Effector Nuclease (TALEN). The architecture of the TALEN was then refined by truncating TALE C-terminal region to various length and analyzing their nuclease activity *in vitro* on the HEK293 genome. The optimized length was 28, out of the 278 original TALE C-terminal residues (Miller et al., 2011) (**Figure 10**). The main issue using TALEN is their tedious assembly due to extensive identical repeat sequences. Methods were developed to overcome this problem, by constructing plasmids containing TAL effectors that can be easily assembled by Golden Gate cloning to form a full TALEN (Cermak

et al., 2011). Nevertheless, their assembly remains challenging and TALEN remain recombinogenic.

TALENs were widely used in various organisms, including in macaques in which TALEN containing plasmids were injected into embryos (Liu et al., 2014). Cellectis is still using TALENs and has a clinical trial in cancerology using CAR-T-cells which are T-cells modified using a TALEN to integrate a modified T-cell receptor (Chimeric Antigen Receptor – CAR) recognizing CD19, a surface protein often expressed by tumor cells (Gautron et al., 2017) (Valton et al., 2015).

A recent report in bioarchive stated that cattle modified using TALEN bear plasmid backbone in its genome. Authors analyzed two publicly available whole genome sequencing data from genome edited calves. These calves were modified to bear a duplication of the celtic polled allele, a variant that produces hornless cattle; the intended modification was that TALEN induced DSB was repaired by homologous recombination using a provided template. Authors believed that this unintended modification went undetected because authors did not include plasmid backbone in their sequence alignment (Norris et al., 2019).

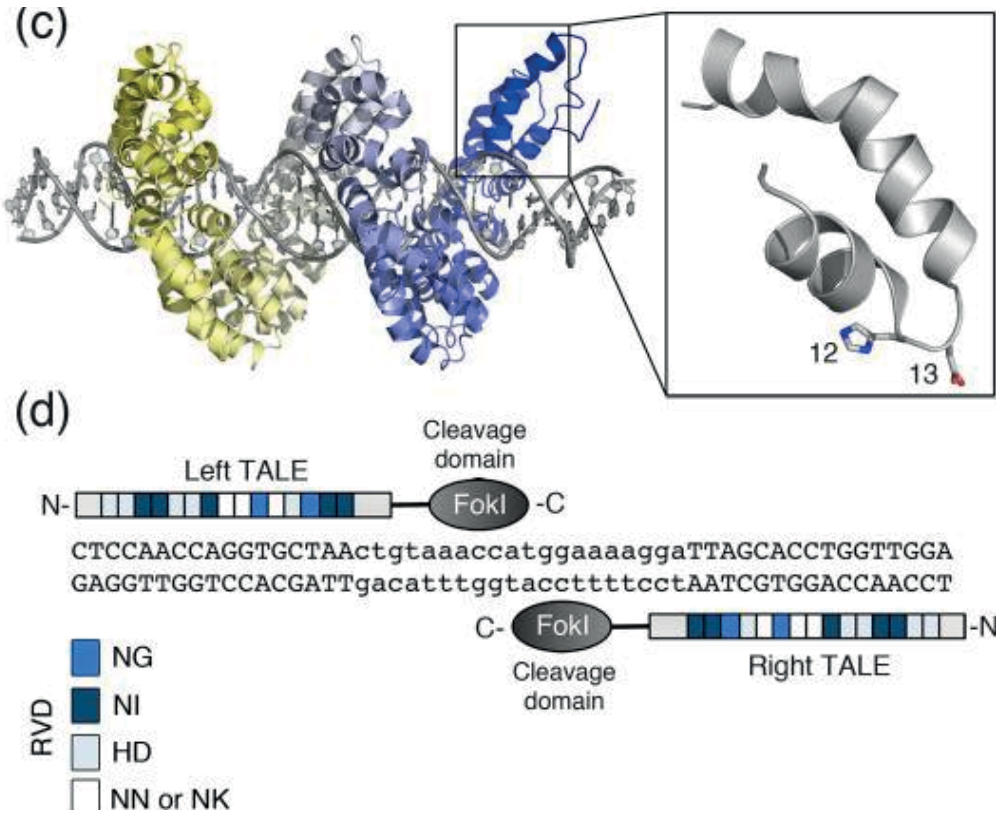


Figure 10: Transcription-activator-like-effectors Nuclease structure. (a) TALE in complex with target DNA (grey). Individual TALE repeats contain 33–35 amino acids that recognize a single bp via two RVDs shown as sticks (inset). (d) Cartoon of a TALEN dimer bound to DNA. TALEN target sites consist of two TALE binding sites separated by a spacer sequence of varying length (12-to 20-bp), cleaved by the FokI cleavage domain. From (Gaj et al., 2013).

CRISPR-Cas nucleases

A new genome editing platform revolutionized the genome editing field due to its simplicity and adaptability: Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR) Cas system.

Discovery of CRISPR system: a prokaryotic adaptive immune system

The CRISPR acronym was first employed to describe a novel family of repeated DNA sequences spanning in over 40 prokaryotic species and absent in viral and eukaryotic genomes (Jansen et al., 2002). The first description of a CRISPR was made in 1987 in *E. coli* (Ishino et al., 1987). It then took a decade and the sequencing of various phages to realize that in between repeated sequences (called spacers) phage genomic DNA was inserted (protospacer) (Mojica et al., 2005). Phage 858 resistance in *S. thermophilus* was lost when protospacers S1 S2 homologous to this phage were removed. Similarly, adding protospacers S1 and S2 to wild type strain conferred resistance to phage 858. Consequently, the CRISPR system was described as an adaptive immune system for archaea and bacteria against invasive phages (Barrangou et al., 2007). CRISPR-associated (Cas) enzymes are guided by short CRISPR RNAs (crRNA) transcribed from the spacer sequences (Brouns et al., 2008). Later, *S. thermophilus* transformed with a plasmid carrying an antibiotic resistance gene were grown for 60 generations without selection. The colonies that had lost antibiotic resistance were analyzed and revealed that 55% of the colonies had acquired a spacer whose sequences were homologous to the transformed plasmid. Strains mutated for csn1 were not able to lose the plasmid, indicating that csn1 was responsible for plasmid loss. Authors observed that a linear version of the plasmid was retrieved, cut next to the acquired spacer sequence, showing that plasmid DNA was targeted and cleaved by the CRISPR/Cas machinery. Further experiments showed that bacteriophage DNA was cleaved when infecting bacteria containing corresponding protospacers. Thus, one protein called csn1 was discovered *in S. thermophilus*, and shown to be the sole protein needed to carry out recognition and cleavage of invasive DNA (Garneau et al., 2010). This protein was later called Cas9. Another key discovery was the observation that the acquired spacer sequences were highly similar to each other at regions called protospacer-adjacent motifs (PAMs) and that this sequence was required for efficient cutting (Deveau et al., 2008). Finally, the endogenous CRISPR system was shown to require two short RNAs for guiding Cas9: the mature crRNA and a trans-activating crRNA (tracrRNA) (Karvelis et al., 2013).

Diversity of CRISPR systems and classification

CRISPR-Cas adaptive immune systems are found in roughly 50% of bacteria and 90% of archaea (Makarova et al., 2015)(Wright et al., 2015). They are extremely diverse, and were classified into six types based on the presence of “signature genes”, indicated on **Figure 11**. CRISPR-Cas systems fall into two classes. Class 1 systems use a large multi-Cas protein complex capable of recognizing and cleaving nucleic acids complementary to the crRNA, while class 2 systems use a single large Cas protein. Class 2 system and more precisely Cas9 from the type II system raised the possibility of engineering it and use it to induce targeted genome modifications.

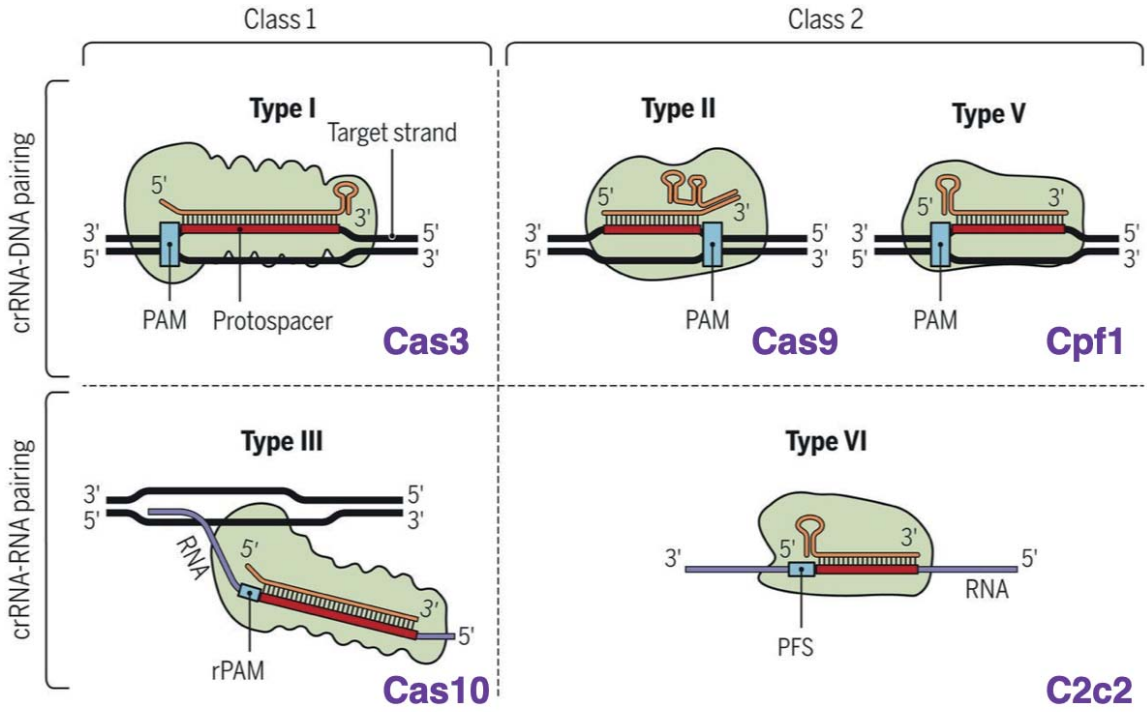


Figure 11 Diversity of CRISPR-Cas types. Recognition of the invading DNA target is made by the Cas nuclease (in green) and a guide RNA (in orange) in complex. It results in the formation of an RNA-DNA hybrid in which the nontarget DNA strand is displaced. The target strand contains the protospacer (red), which is complementary to the spacer sequence in the crRNA (orange). The protospacer-adjacent motif (PAM, in blue) is located at either the 3′ end (types I and V) or the 5′ end (type II) of the protospacer. Type I is made of a multiprotein complex. Type II and V rely on a single protein to mediate cleavage. Signature genes are indicated in purple. Up: CRISPR system mediating DNA cleavage, bottom: CRISPR system mediating RNA cleavage. Figure from (Shmakov et al., 2017).

Cas9 is an RNA-guided nuclease

SpCas9 structure and catalytic activity

The purification of Cas9 from *Streptococcus pyogenes* allowed to study cleavage requirements *in vitro,* and demonstrated that tracrRNA and crRNA can be fused as a crRNA:tracrRNA so that only one guide RNA mediates cleavage. The PAM of SpCas9 is -NGG and leaves a blunt cut 3 to 4 bp away from it, through concerted activation of two catalytical domains, RuvC and HNH each catalyzing one single-strand break (**Figure 12.A.1-2**). Quickly, other publications followed showing that the CRISPR-Cas9 system can achieve genome modification *in vivo* in human cells (Jinek et al., 2013)(Cong et al., 2013) (Mali et al., 2013). Crystallographic structure and molecular analysis extensively characterized how Cas9 works. Single molecule analyses enabled the precise determination of Cas9 binding and cleavage: first, the nuclease scrolls the genome for a PAM, then sequentially unwinds DNA starting from it (Sternberg et al., 2014). It was later explained that a HNH conformation change was required for efficient cleavage by the RuvC domain (Sternberg et al., 2015). Surprisingly, Cas9 tolerated mismatches in PAM distal region while the 10 -12 nucleotides proximal to the PAM were more conserved (Jinek et al., 2012) confirming the existence of a previously experimentally identified seed region (Semenova et al., 2011).

Nickases

Each catalytic sites RuvC and HNH can be knocked down by D10A and H840A mutations respectively, resulting in the cleavage of only one DNA strand (Jinek et al., 2012). N863A is another HNH mutant exhibiting a higher catalytic activity than H840A (Nishimasu et al., 2014). D10A retains a higher edition efficiency than H840A or N863A. The fact that a HNH conformation change was required for efficient cleavage by the RuvC domain explains why a HNH mutation has an impact on both catalytic domains resulting in lower activity of non-functional HNH nickases such as H840A and N863A (Sternberg et al., 2015).

dCas9

Dead Cas9 (dCas9), cumulating both D10A and H840A mutations, only retains its DNA binding capacity (Guilinger et al., 2014). dCas9 was then used for many different applications. It was fused to a cytosine deaminase to function as a base editor, catalyzing C→T or G→A modifications without inducing a double-strand break (Komor et al., 2016). It was also fused to

transcription activator like VP16, or repressors like KRAB, to modulate gene expression of a GFP reporter gene. The technique can also be used at any desired locus (Gilbert et al., 2013).

Enhanced specificity

When targeted to a particular locus, so-called "on-target", unintended edition may arise at other loci, called "off-targets". SpCas9 specificity was evaluated by generating a set of 57 sgRNAs containing all possible single nucleotide substitution in positions 1 to 19 to target one specific sequence. SpCas9 tolerated mismatches between guide RNA and target DNA at different positions particularly outside the seed region, and generated double-strand breaks even in the presence of mismatches (Hsu et al., 2013). Additionally, *in silico* predicted off-targets based on homology to the gRNA sequence, were showing detectable indel frequencies. It raised the issue about CRISPR-Cas9 specificity. In parallel, tools were developed to monitor off-targets genome-wide: GUIDE-Seq relies on the cotransfection of Cas9 and double-stranded oligodeoxynucleotide (dsODN) that act as a DNA tag. This oligo was designed and optimized to provide optimal stability in cells; they carry two phosphothiorate linkages at the 5′ ends of both DNA strands. When a cut is made, cells repair by integrating a dsODN. The location of this tag can be retrieved by sequencing using complementary primers to amplify a genomic library (Tsai et al., 2015).

CIRCLE-Seq evaluates off-targets in a test tube containing genomic DNA and a purified Cas9. Cut molecules are circularized and sequenced (Tsai et al., 2017). This last technique would be less relevant since off targets are assessed on DNA in a test tube which does not have the same conformation and chromatinian environment. It can nevertheless serve as an initial screen to identify potential off-target sites that can then be verified *in vivo*.

VIVO, was developed to monitor off targets in living animals. This method relies on CIRCLE seq which is first used on mouse cells and candidates are then assessed for mutations in transduced mice tissues (Akcakaya et al., 2018). Softwares were built in order to predict off-target effects of gRNA and help designing more specific gRNAs, based on off-target studies, such as the web tool CRISPOR (Haeussler et al., 2016)(Concordet and Haeussler, 2018).

To circumvent off-target issues, many variants were engineered to make SpCas9 more specific, such as enhanced Spcas9 (eSpCas9) (Slaymaker et al., 2016) and Cas9-HF1 (Kleinstiver et al., 2016a). In eSpCas9, three positively charged residues interacting with the phosphate backbone of the non-target strand were neutralized, conferring an increased specificity (Kleinstiver et al., 2016a), (**Figure 12.A.4)**. Similarly, Cas9-HF1 was mutated on 4 residues interacting through

hydrogen bonds with the target strand (Slaymaker et al., 2016). Both variants showed increased specificity while retaining most of its on-target activity (**Figure 12.A.3)**.

Altered PAMs

Finally, SpCas9 was modified to allow other PAM requirements: -NAG and -NGA. These variants were isolated using a bacterial selection system in which survival was conditional to efficient Cas9 cutting of a plasmid encoding a toxic gene (Kleinstiver et al., 2015a).

Other nucleases from the CRISPR system can be used to modify genomes

Other Cas9 were characterized including *Staphylococcus aureus* (SaCas9), a smaller Cas9 that can be packaged alongside its gRNA in AAV vectors (cargo size up to 4.7 kb). The smaller size compared to SpCas9 is due to a missing REC lobe, resulting in a shorter recognition domain (**Figure 12.A.1 / 12.B**). SaCas9 was used in many *in vivo* studies, its PAM is NNGRRT, and has a similar structure to SpCas9, with two catalytic sites (Ran et al., 2015a). SaCas9 was optimized by rational engineering of residues forming polar contacts within a 3.0 Å distance from the target DNA strand. By testing the efficacy and specificity of the resulting mutants by deep sequencing, the R245A/N413A/N419A/R654A quadruple mutant was found to be the most efficient at inducing gene edition and was called SaCas9-HF. A high fidelity version of SaCas9-KKH was also engineered (Tan et al., 2019). Using the same experimental setting as for the establishment of Cas9-HF1, SaCas9 was modified to recognize a NNNRRT PAM; this variant bearing the mutations E782K/N968K/R1015H was called SaCas9-KKH (Kleinstiver et al., 2015b).

Finally, type V CRISPR-Cas or Cpf1 nucleases, also called Cas12a exhibit very different features including a T-rich PAM located 3’ of target DNA and making staggered cuts leaving five-nucleotide overhangs by iterative activation of a single RuvC catalytic site (Zetsche et al., 2015), (**Figure 12.C**). Many other Cas9 are being discovered, including in uncultivated microbes. Metagenomics analysis of samples from groundwater and acid mine drainage revealed the existence of new Cas, called CasX and CasY from bacteria, and are among the most compact CRISPR systems, as well as an archaeal Cas9 (Burstein et al., 2017).

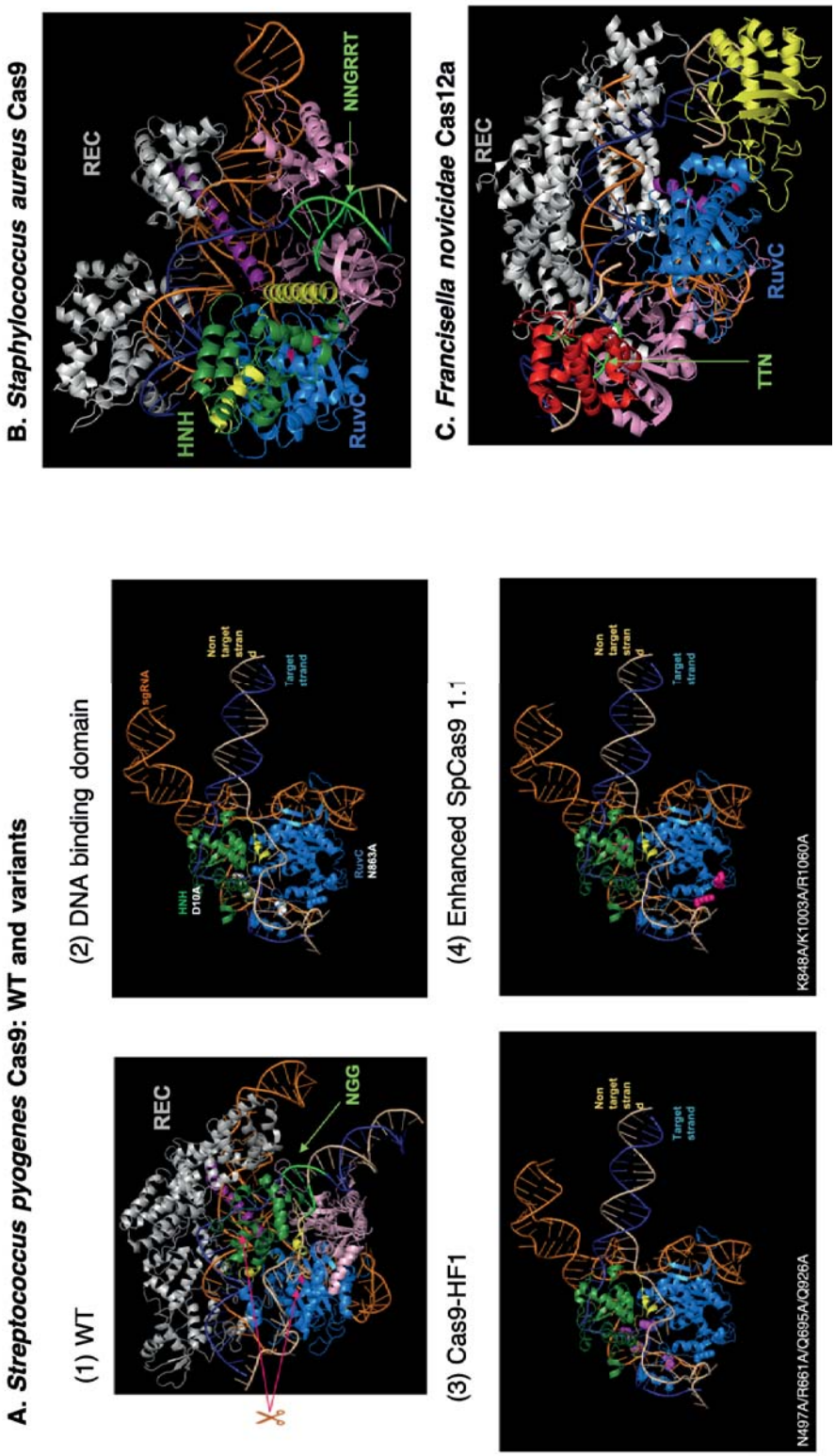


Figure 12: CRISPR associated nucleases exhibit different structures. A. SpCas9 and variants. (1) Crystallographic structure of WT SpCas9 bound to target DNA. PAM is indicated in green. In grey: REC domain. Catalytic site is indicated by scissors. (2) Zoom on catalytic domain with HNH and RuvC and corresponding neutralizing mutations. (3) and (4): Two variants with an enhanced specificity. Cas9-HF1 exhibits less interaction with the target strand (mutations highlighted in pink) while eSpCas9 shows less interaction with the non-target strand (mutations highlighted in dark pink). Both result in a weaker interaction with DNA. **B. SaCas9.** This smaller Cas9 lacks one REC domain and has an NGRRT PAM. **C. FnCas12a.** Also called FnCpf1. This nuclease belongs to the Type V and has a different structure compared to Type II with one catalytic site and a T-rich PAM.

CRISPR-based growing toolkit

CRISPR-Cas system was used to set up an impressive number of experimental assays; a few examples are listed in this last section. It has been extensively used in many organisms to induce specific modifications and also to achieve genome-wide screening based on gRNA libraries (Koike-Yusa et al., 2014). It was used to inactivate porcine endogenous retroviruses (PERV) inserted in the pig genome, which is useful since the risk of cross-species transmission to human impedes the clinical application of organ transplantation to humans. A PERV-inactivated pig was successfully produced by modifying embryos, using SpCas9 and two gRNAs targeting the catalytic core of the PERV virus: the *pol* gene (Niu et al., 2017). dCas9 fused to transcription activator or repressor elements was used to study the regulatory elements for embryo development in chicken (Williams et al., 2018). dCas9 fused to a GFP reporter gene was also used to track telomere dynamics in live mice (Duan et al., 2018). Other proteins from the CRISPR system are also of interest. It was recently shown that *Vibrio cholerae* transposable element Tn*6677* can be modified to integrate into the *E. coli* genome at a specific locus dictated by CRISPR-Cas RNA guided proteins (Klompe et al., 2019). This new system could be used to insert any sequence at any desired locus without the need for a donor DNA and a double-strand break.

Cas13 belongs to type VI-B CRISPR system and is a nuclease that specifically cleaves RNA. It was used to knockdown the expression of specific RNAs (Cox et al., 2017). It has many potential applications such as splicing modification, targeted localization of transcripts, change of RNA-binding proteins affinity.

Very recently, CRISPR-Cas9 system was modified to induce edition without double-strand breaks. H840A nickase was fused to an enhanced reverse transcriptase. This nickase was coexpressed with a prime editing gRNA (pegRNA) which was the fusion between a gRNA that specifies the location to be targeted by Cas9 and a primer sequence containing the modification for reverse transcription. After recognition by Cas9, the nicked strand was elongated by the reverse transcriptase to harbor the desired edition. Then, to facilitate repair, alongside nickases+pegRNA, a gRNA targeting the edited strand but inducing a nick on the opposite unedited strand will drive repair event toward the suppression of the unedited strand via FEN1 endonuclease which excises 5’ flap strands. This new technology led to successful edition event in HEK293T cells with 44% efficiency and in various cell models including HeLa cells (12% efficiency) and terminally differentiated mice neurons (7% efficiency). Efficiencies were measured by Illumina sequencing of transfected cells. However, two main limitations of this system are that the process leading to desired edition is not fully controlled and understood,

with efficiencies still being low. In addition, this system is large, twice the size of SpCas9 which can hinder further applications such as *in vivo* expression (Anzalone et al., 2019).

Once a DSB is made in a chromosome, cells may use different pathways to repair a DSB. The main features of DSB repair will be explained in the next part.

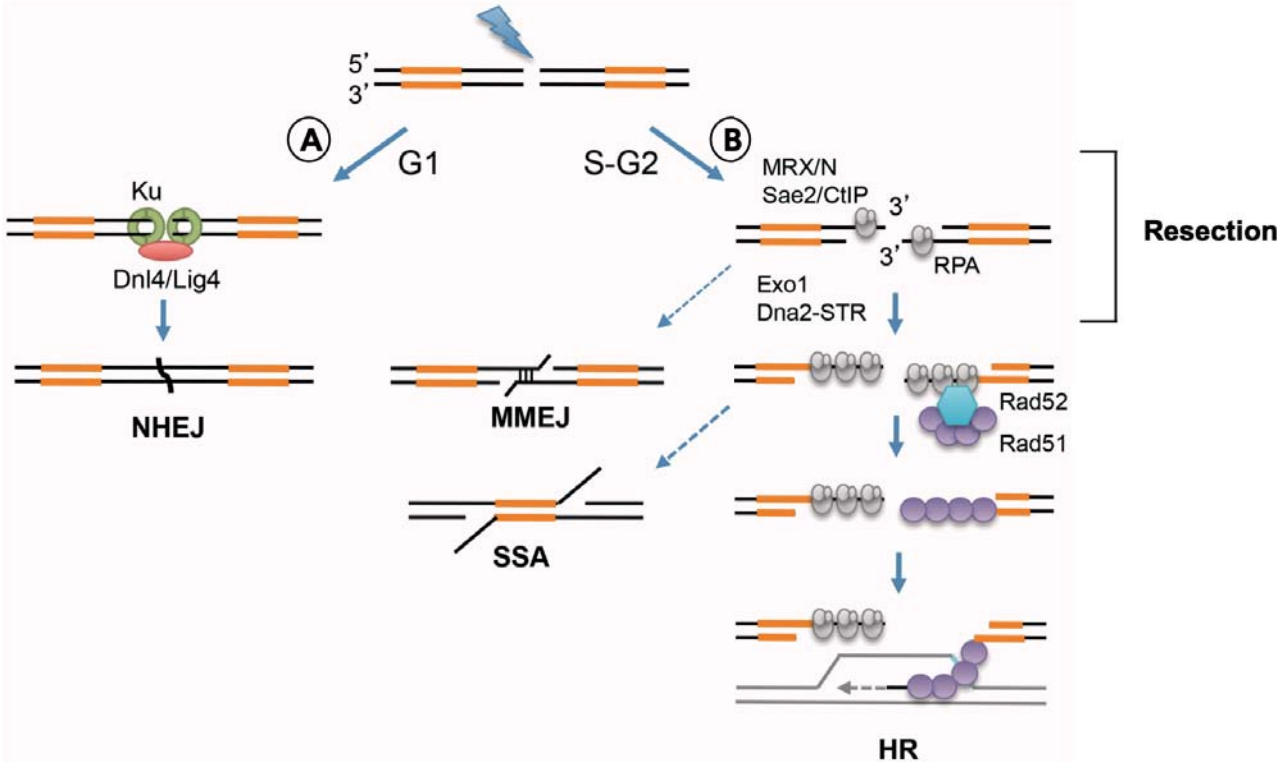


Figure 13: DSB repair pathways. A. Non-Homologous End Joining results in the direct ligation of DNA ends and is favored in G1 phase cells. It relies on Ku and Dnl4/Lig4**. B. Homologous Recombination.** End resection relies on the action of the MRX complex (in yeast), (MRN in humans), Sae2 (CtIP in humans), Exo1 and DNA2. It is favored in S-G2 phase cells and creates 3′-ssDNA overhangs that are bound by RPA. Rad52 (and its functional equivalent BRCA2 in human) mediates Rad51 filament assembly onto ssDNA. Rad51 catalyzes homologous pairing and strand invasion (HR). Alternatively, annealing of microhomologies internal to the ends results in repair by MMEJ, or by SSA if longer direct repeats flank the DSB. Rad1-Rad10 nuclease (human XPF-ERCC1) cleaves end flaps. Thick orange lines represent repeated regions. From (Symington, 2016)

Repairing nuclease-induced DSB

This part describes DNA repair pathways in mitotic cells, repair in meiotic cells will not be discussed. Meiosis nevertheless plays an important role in repetitive sequence instability. Early, some minisatellites were described as being hypervariable, showing extensive and frequent length polymorphism during meiosis (Buard and Vergnaud, 1994). This meiotic instability was later on shown to be triggered by homologous recombination, following double-strand breaks occurring at meiosis near the minisatellite (Jeffreys et al., 1998). Similar experiments were reproduced in yeast, in which CEB1 was integrated in a chromosome, near a known meiotic hotspot. Instability of the minisatellite was shown to depend on hotspot activity and on the product of the SPO11 gene, the type VI topoisomerase responsible for meiotic double-strand breaks in yeast (Debrauwère et al., 1999). Interestingly, PR domain–containing 9 (PRDM9) is a meiosis-specific histone H3 methyltransferase with a C-terminal ZF domain encoded by a minisatellite. The instability of the minisatellite induces a change of recombination hotspot, further triggering rapid evolution of the genome, probably promoting its evolution (Berg et al., 2010).

Non-Homologous End Joining

The most straightforward way to repair a broken chromosome is to rejoin the ends. This process is called Non-Homologous End Joining (NHEJ) and is subdivided into classical-NHEJ (C-NHEJ) and alternative-NHEJ in mammalian cells (Alt-NHEJ). Alt-NHEJ is called Microhomology Mediated End Joining (MMEJ) in *S. cerevisiae*. C-NHEJ was most studied for its role in V(D)J recombination in pre-B cells. Rag1 and Rag2 proteins cleave recognition sequences adjacent to V, D and J segments; rejoining of the segments by NHEJ results in a wide diversity of sequences coding for antibodies and T-cell receptors (Dai et al., 2003).

Classical-NHEJ

The heterodimer Ku70/Ku80 (hereafter called Ku) is the first complex to bind to DNA ends after a DSB, since its affinity to DNA is very strong (Blier et al., 1993). Nucleases, polymerases and ligases of NHEJ can then dock to the Ku:DNA complex in different orders. NHEJ reactions were reconstituted *in vitro* and confirmed that DNA-PKcs phosphorylate the Artemis protein which gains an endonuclease activity to trim damaged 3’ overhangs of dsDNA (Ma et al., 2002). Then the binding of the ligase complex XRCC4:DNA ligase IV completes the joining (Ma et

al., 2004) (**Figure 13.A**). XLF acts as a scaffold protein to stabilize the ligase complex at broken DNA ends and promotes end bridging (Riballo et al., 2009).

Alternative-NHEJ

Embryonic lethality of Ligase IV deficient mice is rescued by Ku deletion suggesting alternative end joining pathways exist, not only in cells able to efficiently carry out homologous recombination but also in non-dividing cells (Karanjawala et al., 2002). Yeast lacking Kup are still able to carry out NHEJ with an increased need for homology at the repair site for 5 to 9 nt between different plasmid donors (Boulton and Jackson, 1996). This suggests the existence of an alternative pathway relying on microhomologies later called as MMEJ.

Preference toward Homologous recombination in yeast

S. cerevisiae does not have end-processing nucleases among its NHEJ proteins as versatile as mammalian cells, which makes them poor at blunt-end ligation and processing of damaged DSB ends (Daley et al., 2005), leading them to preferentially repair a DSB by homologous recombination. It can nevertheless be argued that the efficiency of resection in yeast is the major driver toward homologous recombination (Symington, 2016).

Homologous Recombination

Donor template for homologous recombination

Homologous recombination (HR) is an accurate repair involving the invasion of a homologous donor by the ends of the broken chromosome to prime DNA synthesis and restore sequence at the DSB. In mitotic cells, the donor can be the sister chromatid after DNA synthesis, homologous chromosome in diploids, or any ectopic repeated sequence elsewhere in the genome. HR is a template-dependent process and studies in *S. cerevisiae* have demonstrated that the sister chromatid was the preferred template over a homologue, when given the choice. Sister-chromatid recombination and homologous recombination after X-ray induced damage were followed by selecting cells on appropriate media, each DNA repair event leading to the reconstitution of a specific auxotrophic marker (**Figure 14**). Sister chromatid was a preferred template over homologous recombination template elsewhere in the genome. Authors hypothesized that this preference was due to physical proximity from one another rather than the perfect homology between sister chromatids. Indeed, even on diploids sharing 100%

homology due to self-mating, the sister chromatid was still the preferred substrate (Kadyk and Hartwell, 1992).

DSB repair can be followed through *in vivo* biochemistry assays in yeast by monitoring HO-induced DSB repair by Southern blots (Sugawara and Haber, 2006). The main steps of homologous recombination are described in the next paragraphs.

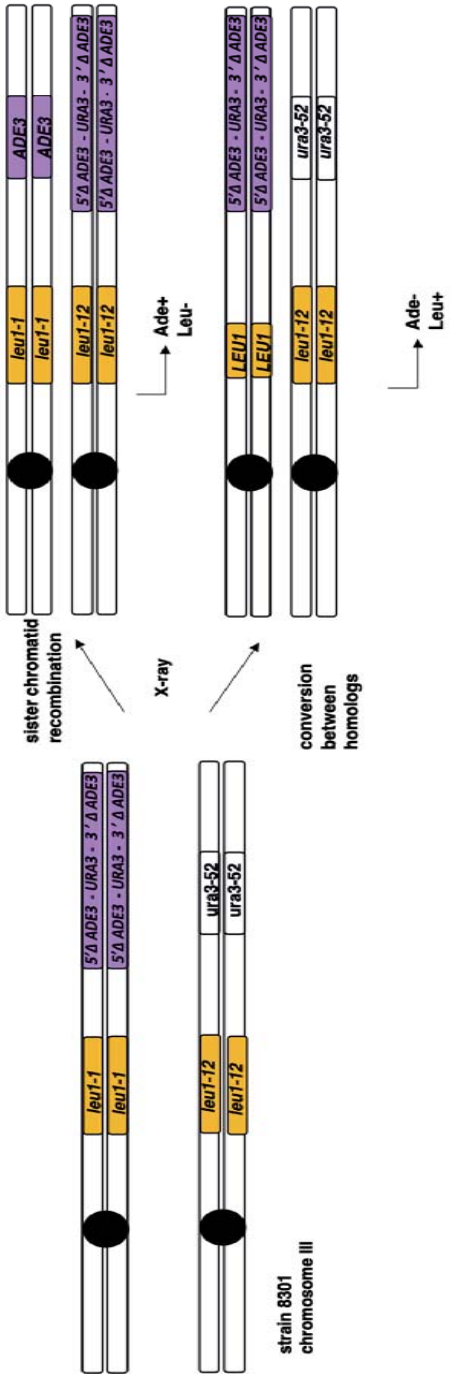


Figure 14: Experimental assay for template preference in *S. cerevisae*. Drawing of the chromosome III of strain 8301.This strain is heterozygous for the Sister Chromatid Recombination (SCR) substrate so that unequal sister chromatid recombination can be monitored. *ADE3* gene is split into two non-functional sequences that only intra chromosomal recombination event can reconstitute, by recombination between 5'Δ*ADE3* and 3'Δ*ADE3*. The strain is heteroallelic for the mutations leul-1 and leul-12, so that gene conversion between homologs can be monitored. After X-ray treatment, cells are plated on SC-leu plate and SC-ade plates to quantify conversion between homologs and conversion between sister chromatid respectively. Here, only one possible outcome leading to each phenotype is represented as an example.

Resection

First, end resection is initiated by the Mre11-Rad50-Xrs2 (MRX) complex in *S.cerevisiae* which together with Sae2 removes oligonucleotides from the 5' end (Mimitou and Symington, 2008). Extensive resection is then carried out either by Exo1 or the Sgs1-Top3-Rmi1-Dna2 ensemble (Cejka, 2015). Resection can be measured by qPCR assays relying on the selective amplification of digested DNA in which ssDNA was not digested and will be amplified. Difference between digested and undigested fraction will give the percentage of resected molecules (Chen et al., 2013). In mammalian cells, the resection genes are Mre11-Rad50-Nbs1 (MRN complex) and CtIp (Sae2 homologue), and the role of Exo1 is conserved (Syed and Tainer, 2018) (**Figure 13.B**). End resection also needs chromatin remodelers such as Fun30 in yeast and SMARCAD1 in human to effectively proceed through compact chromatin (Costelloe et al., 2012). ssDNA is then coated with ubiquitous and abundant ssDNA binding protein Replication Protein A (RPA) which protects this fragile intermediate (Ruff et al., 2016).

Mediators facilitate Rad51 loading

To facilitate Rad51 loading on ssDNA filaments, a mediator like BRCA2 (which function is similar to *RAD52* in yeast) is needed to remove RPA and promote assembly of the Rad51 filament (Jensen et al., 2010; Liu et al., 2010b). BRCA1 counteracts NHEJ factor 53BP1 to facilitate end resection and interacts with PALB2 proteins which in turn binds BRCA1 to facilitate Rad51 filament formation (Sy et al., 2009) (Zhang et al., 2009). Mutations in both *BRCA1* and *BRCA2* genes are associated to predisposition to breast and ovary and to a lesser extent prostate, pancreas and other cancers. The paralogs RAD51B, C, D, XRCC2, XRCC3 and DMC1 share limited sequence homology to each other or to RAD51 except for their critical ATP-binding domains. DMC1 has meiosis-specific functions. The mitotic paralogs form two major multi-protein complexes: the RAD51B/C/D/XRCC2 (BCDX2) and the RAD51C/XRCC3 (CX3) complexes (Masson et al., 2001). The paralogs promote phosphorylation of checkpoint effector kinase to stimulate DSB repair, however their precise role is still elusive (Suwaki et al., 2011).

Rad51 mediates the search for homology

Then the search for homology relies on the Rad51 nucleofilament that assembles on ssDNA. By following the change in length of fluorescently end-labeled lambda DNA, Rad51 filament assembly kinetics was measured. Using fluorescin-labeled Rad51, nucleation and growth assembly of the Rad51 nucleofilament was characterized. First, nuclei are formed, made of 2-

3 Rad51 protein, then the filament is extended starting from these nuclei. Finally, by comparing Rad51 disassembly in presence of either Mg^{2+} ATP or Ca^{2+} ATP (non hydrolysable), they show that disassembly is faster when ATP is not hydrolyzed, suggesting the conversion of an ADP bound Rad51 to DNA. In both cases it remains slow and incomplete suggesting that a catalyst is needed for its disassembly, such as Rad54. Rad51 is associated to an increased mobility of the broken chromosome revealing its native role for homology search (Hilario et al., 2009).

Resolution

The resultant double Holliday junction is a substrate either for resolution into crossover products or dissolution to non-crossover products (Matos and West, 2014).

Checkpoint proteins regulating initiation of DSB repair

Microscopy on *S. cerevisiae* live cells expressing tagged damage checkpoint and DNA repair proteins gave a comprehensive view of the order of arrival of the different proteins at DSB site and confirmed that MRX complex is the earliest sensor of DSB. Also, the recruitment of the MRX complex is associated to Tel1 (ATM in mammals) which is a regulatory checkpoint for HR (Lisby et al., 2004) (**Figure 7.A**). Cells detect DNA damage during mitosis in different ways by activating either a G1-to-S or a G2-to-M checkpoint. Checkpoint-mediated arrest of cell cycle prevents division which would segregate broken chromosomes, giving more time for DNA repair processes to occur.

The role of resection in initiating HR over NHEJ

Main determinant of repair pathway choice is the initiation of 5′-3′ resection of DNA ends, which commits cells to homology-dependent repair (Symington and Gautier, 2011). In eukaryotes this commitment is linked to the cell cycle and CDK (Cyclin-dependent kinase) activity (Hustedt and Durocher, 2016). In *S. cerevisiae,* Cdc28p phosphorylates Sae2p (Ira et al., 2004), further stimulating end-resection (Huertas et al., 2008). Cdc28p also phosphorylates Dna2p (Chen et al., 2011), and the chromatin remodeler Fun30p (Ferrari et al., 2015). As a result, end-resection and HR-mediated DSB repair are prevalent in late S and G2/M phases of the cell cycle, when CDK activity is high. Similar CDK-dependent mechanisms promote end-resection in mammalian cells (Huertas, 2010). In addition, RIF1 plays an important role in blocking end-resection and promoting NHEJ in mammalian cells. RIF1 colocalizes to the ssDNA/dsDNA end-resection junction in an ATM and 53BP1 dependent manner (Silverman et al., 2004). RIF1 was also shown to reduce the accumulation of RPA, impeding the recruitment

and activation of the apical checkpoint kinase Mec1 (Mitosis entry checkpoint 1; ATR in human) (Xue et al., 2011). 53BP1 and RIF1 both repress end-resection at DSBs (Zimmermann et al., 2013), however it is not yet understood how they cooperate to inhibit the end-resection machinery. In budding yeast, Rif1p is part of a protective cap of telomeres preventing end resection at telomeres (Shi et al., 2013). By analogy with budding yeast, it was speculated that 53BP1 and RIF1 might form structures at DSBs making it less amenable for resection (Panier and Boulton, 2014).

Single-Strand Annealing

Single Strand Annealing (SSA) is a mutagenic Rad51-independent repair mechanism that operates between long direct repeats flanking a DSB and results in loss of one of the repeats and of the intervening sequence. Although resection initiation is well regulated, its ending is not and resection processes at a constant rate of 4 kb/hr for more than 24 hours in *S. cerevisiae*. SSA may be a byproduct of this long resection. In yeast few dispersed elements except in rDNA are present in the genome whereas in mammals many dispersed elements exist (transposons and retrotransposon) and SSA may have more dramatic consequences in higher eukaryotes (Bhargava et al., 2016).

SSA can be studied in yeast using a linearized plasmid containing a region homologous to the yeast genome and an auxotrophy marker gene (Sugawara et al., 2000). By designing sequences with increasing homology length, it was shown that SSA requires at least 30 bp identical on each side of the DSB and efficiency increased linearly up to 400 bp. Above this threshold, SSA efficiency was reduced. Another assay in which two homologous sequences were separated by an increasing long sequence, showed that repair products appeared 30 minutes after the DSB when separated by a few bp and 2 hours when separated by 5kb, which is consistent with a rate of resection of 4 kb/hr (Jain et al., 2009). SSA was shown to be Rad52-dependent. FRET experiments were used to follow the interaction between fluorescent probes located on two different ssDNA strands sharing homologies. Upon addition of RAD52, the two ssDNA strands are brought together revealing another fluorescence. No loss of signal fluorescence was observed after reannealing of ssDNA suggesting that further search for homology occurs without dissociation of RAD52 to ssDNA (Rothenberg et al., 2008). SSA can occur between mismatch sequences but will require MMR proteins Msh2 and Msh3 to remove mismatches (Sugawara et al., 1997).

Yeast cells deficient for *RAD10* are unable to complete SSA when a HO-induced DSB is induced into a plasmid containing two uncomplete LacZ genes (Fishman-Lobell and Haber, 1992). SSA is restored when provided with a substrate in which the lacZ donor sequences are homologous except for one single base pair mutation at the HO cleavage site. Triplication sequence of LacZ was also observed alongside SSA products. These results indicate that *RAD1* is necessary to remove non-homologous ends after annealing. Since RAD1 and RAD10 were shown to form a protein complex (Bailly et al., 1992), RAD10 role may also play a similar role; indeed, *RAD10* was also shown to be necessary for non-homologous end clipping (Ivanov and Haber, 1995) (**Figure 7.A**).

Reporter assays to study DSB repair efficiency

Reporters of spontaneous homologous recombination exist *in vitro* and *in vivo.* Homologous recombination assays based on auxotrophic markers (Schiestl and Prakash, 1988) or antibiotic resistance in mammalian cell lines (Moynahan et al., 1999) were used to study DSB repair efficacy. Maria Jasin was the first to set up a GFP reporter assay to measure the efficacy of repair by homologous recombination after a single DSB in the genome, as fluorescence assays are easier to use and more quantitative. In the first assay the irs1SF cell line was studied. It is a hamster cell line which is defective in XRCC3, a protein involved in DSB repair by homologous recombination. Since this cell line was isolated from mutagenized cultures, it was possible that DSB repair impairments observed may not be solely due to XRCC3 defect (Pierce et al., 1999). A GFP reporter assay containing an I-*Sce*I cutting site was introduced into this cell line (**Figure 15**). When I-*Sce*I was co-expressed alongside XRCC3, GFP-levels were restored to wild type levels, showing that XRCC3 was directly involved in the DSB repair defect of this cell line, as no other protein involved in DSB repair was able to restore WT GFP levels. The same assay was introduced in murine cells lacking BRCA2 and showed that BRCA2 was required for homologous recombination (Moynahan et al., 2001). The same reporter gene was integrated into C57BL/6 mice and primary cell culture from different tissues were transfected with I-*Sce*I (Kass et al., 2013). This mice model was used to interrogate the effect of BRCA1 loss and ATM loss. The same mice were crossed to generate Dox-I-SceI inducible DR-GFP mice (Kass et al., 2016). A low % of GFP+ cells was measured upon I-*Sce*I induction, due in part to repair through NHEJ which does not reconstitute any GFP and that is favored over HR in most tissues. In regard of breast cancer development linked to BRCA2 mutations, authors observed that highest GFP+% was achieved in luminal population of the mammary gland as compared to basal

population which contains less proliferative cells. Furthermore, they showed that HR efficiency in pubertal epithelial tissue varies across cell subpopulation, with stem cells having the highest GFP-positive percent. The authors postulated that it is due to their proliferative state during puberty. BRCA2 deficiency was also demonstrated to lower HR efficiency in all tissues tested and not only in mammary glands.

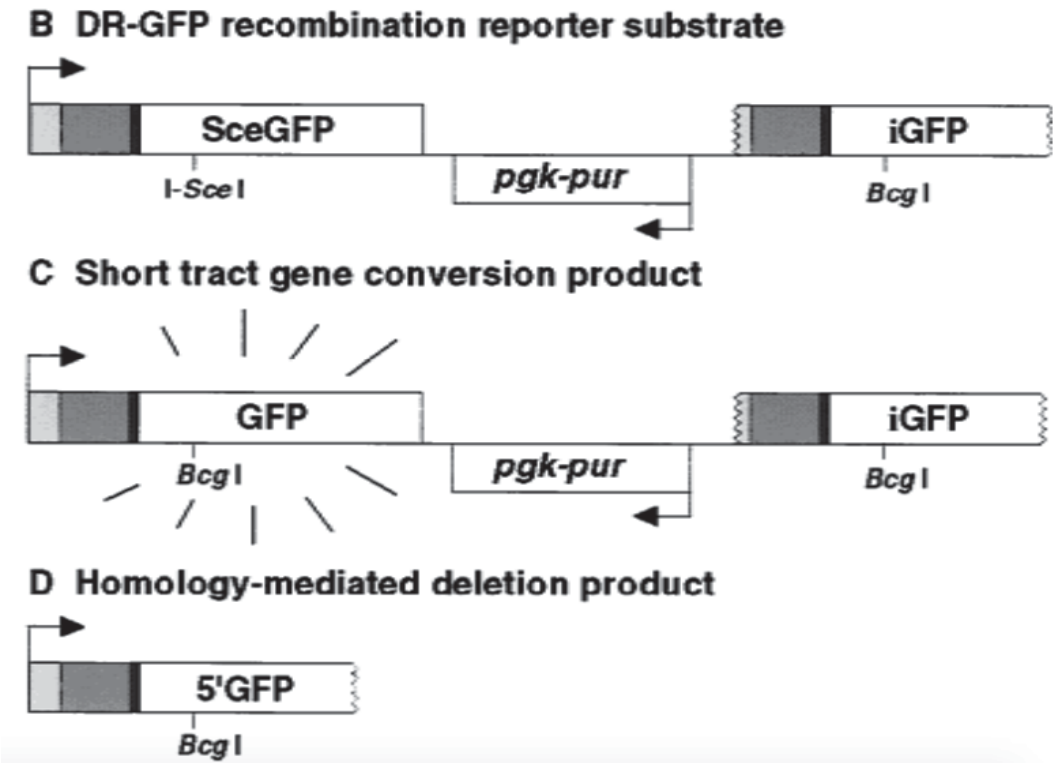


Figure 15 : Homologous recombination assay. DR–GFP recombination substrate: GFP is modified to SceGFP to contain an I-SceI site and in-frame termination codons (underline). Downstream of the SceGFP gene is iGFP, an internal GFP gene. Two homologous recombination products are possible: C: The Short Tract Gene Conversion product: a DSB at the I-Sce I site is repaired from the iGFP gene on the same chromatid or sister chromatid, to result in a functional GFP gene. D: Homology-mediated deletion product. From (Pierce et al., 1999)

Similar reporter assays were tested in mice harboring a cassette with I-SceI site flanked by two direct repeats of *LagoZ*- an engineered derivative of *LacZ* gene that lacks CpG sequences (Gouble et al., 2006). Upon DSB induction after injection of AAV particles containing I-SceI, 1% of liver cells exhibited a ß-galactosidase positive phenotype that are measured by histological staining demonstrating that homologous recombination following nuclease induction can occur in living animals. The p^{un} mice harbor a duplication of the pink-eyed unstable allele, which upon recombination reconstitutes a functional protein responsible for the assembly of black color melanin complex. Expression of this gene can be observed as black dots either on the fur or on the eye (Reliene and Schiestl, 2003). Similarly, a bipartite GFP gene was integrated at the ubiquitously expressed Rosa26 locus in mice to follow spontaneous recombination (Sukup-Jackson et al., 2014).

A common way to assess nuclease efficiency is to target a highly specific nuclease against a GFP gene in a cell line constitutively expressing GFP gene. Efficient DSB induction will result in the creation of indels into the sequence and GFP fluorescence signal loss can be quantified by flow cytometry. This is commonly called EGFP disruption assay, and was first used to test TALENs (Reyon et al., 2012). It later served to validate Cas9-HF1 activity (Kleinstiver et al., 2016b) or to design better guide RNAs (Doench et al., 2014). Additionally, in order to discriminate events resulting from NHEJ or HR, an assay relying on GFP conversion to blue fluorescent protein (BFP) was designed. A single point mutation in the chromophore of GFP could shift its fluorescence absorption and emission toward the blue spectrum, thus creating BFP (Heim et al., 1994); using a template carrying this mutation, cells constitutively expressing GFP repairing by HR can be quantified as BFP+ cells while cells repairing by NHEJ lose their fluorescence (Glaser et al., 2016).

Reporter assays studying DSB repair into repeated sequences are rare. Previous assays in yeast relied on the insertion of an I-SceI cleavage site in between repeated sequences. This was done for CTG repeats (Richard et al., 1999). A recent assay was used to quantify both contractions and expansions at a CTG/CAG repeat tract. Repeats were inserted into an intron of a miniEGFP gene. This cassette was stably integrated into T-REx HEK293 cells. Long repeats will interfere with the expression of the GFP and decrease fluorescence levels. A ZFN designed to cut CAG repeats into a cell line carrying 89 repeats resulted in a 3.5 fold increase in GFP fluorescence – called GFP+ cells (Santillan et al., 2014). Quantification of GFP- cells enables to quantify expansion events; this assay was used to determine which nuclease may be used to trigger contractions in a repeated sequence while not inducing expansions. D10A nickase was found to be the best (Cinesi et al., 2016). The limitation of this technique is that fluorescence intensity

of cells is equally distributed and setting a threshold for GFP+ and GFP- cells has to be carefully carried out. No reporter assay exists for other repeated sequences.

Cell cycle stage and cell type influence DSB repair pathway choice

Cell cycle stage

HR is supposed to be active in actively dividing cells, while NHEJ is active in non-dividing cells. Hence, one highly specific nuclease targeting the same locus in different cell types is most likely to lead to different outcomes as the repair mechanism will be different, especially between dividing and non-dividing cells. DNA end resection seems to be the critical element for regulation of DSB repair since end resected DNA is not amenable to NHEJ repair anymore (Symington and Gautier, 2011).

Cell type

Also depending on cell type, DNA repair efficiency and pathway of choice may differ. Using HR assay, ear fibroblasts, adult ovary, neonatal brain and virgin mammary epithelium showed different HR rates (Kass et al., 2013). Hematopoietic stem cells (HSC) were shown to have a different activation of repair pathways (Mohrin et al., 2010). Similarly, terminally differentiated astrocytes showed different DNA repair responses (Schneider et al., 2012). One hypothesis explaining changes in DNA repair response across cell types is that terminally differentiated cells such as neurons do not need to divide and do not need extensive DNA repair, except at transcriptionally active sites (Nouspikel and Hanawalt, 2002). Muscle cells also show a different DNA damage response (Narciso et al., 2007). Finally quiescent HSCs in G0 have an attenuated DNA repair and response pathway leading to DNA damage accumulation over age (Beerman et al., 2014). Upon reentry into cell cycle, the expression of DSB repair proteins such as Rad51, Brca1, Exo1 was found to be upregulated.

Successful gene edition is achieved when the DSB repair following highly specific nuclease induction is effective. DSB repair pathway choice will differ from one cell type to another and will also be influenced by cell cycle stage. *In vivo* genome editing was achieved in many model organisms and will be discussed in the next chapter with an emphasis on muscular dystrophies, including DM1. However, many issues remain to be solved in order to obtain reliable, reproducible and efficient *in vivo* gene corrections that could be later used in human patients.

Gene editing of muscular dystrophies, successes and hurdles

Success in genome edition

Editing the genome for therapeutic purposes was started 30 years ago with gene therapy. First assays were based on lentiviral vectors which would integrate into patient cells of interest to bring back a functional mutation of a deficient gene. Successes were obtained to treat Severe Combined Immunodeficiency (SCID), by modifying *ex vivo* CD34+ hematopoietic progenitor cells with a retrovirus carrying the γc subunit receptor of T-cells to restore immune functions. Seventeen of 20 treated patients had their immunodeficiency corrected and did not show any adverse effect up to 20 years after the treatment (Fischer et al., 2010). However, five patients developed leukemia due to random integration of the vector activating the expression of oncogenes. Leukemia was fatal for one patient and was cured in the four others. Sickle cell disease was also successfully treated by *ex vivo* modification of autologous HSC to express a functional ß-globin gene (*HBB)* (Ribeil et al., 2017). Non-integrative vectors such as AAV vectors were also used to bring a missing gene. A therapy was developed for Leber congenital amaurosis which is due to a mutation in the *RPE65* gene inducing dysfunction and death photoreceptor cells, eventually leading to blindness. Spark Therapeutics is currently commercializing a gene therapy treatment which provides a functional copy of *RPE65* gene (the product is called Luxturna). Targeted gene modifications are also starting to be tested in the clinic. For example to treat cancer, T-cells are modified *ex vivo* using TALEN to introduce a modified T-cell receptor called CAR for chimeric antigen receptors that recognizes CD19, commonly expressed by tumor cells (Park et al., 2016).
In this chapter, I will focus on gene editing approaches for muscular disorders, including DM1.

Advances in Duchenne Muscular Dystrophy gene editing

Duchenne muscular dystrophy is a severe, progressive muscle-wasting disease leading to disability and premature death. It is due to a lack of dystrophin in muscles, a protein essential for the connection of the muscle fiber cytoskeleton to the surrounding extracellular matrix (Nowak and Davies, 2004). Exon skipping in the dystrophin gene in order to restore the reading frame, and exclude exon 23 which bears a stop codon, would result in the expression of a shorter, but still functional dystrophin (**Figure 16**). In 2016, three papers came out reporting

exon skipping correction in mdx mice model (Long et al., 2016; Nelson et al., 2016; Tabebordbar et al., 2016). All of them use recombinant Adeno Associated Virus (rAAV) in order to deliver SaCas9 and gRNAs. rAAVs are derived from AAV which were modified to be non-replicative and to infect target cells and deliver recombinant DNA (Daya and Berns, 2008). These vectors have a limited cargo capacity of 4kb, hence the use of dual vectors to deliver bigger transgenes. These *in vivo* gene edition studies are summarized in **Table 2**. Either route of administration resulted in production of dystrophin. This promising Cas9-mediated exon skipping strategy was later confirmed to be efficient in 4 DMD dog models where dystrophin expression was restored from 3% to 90% of normal level, depending on the AAV dose administered systematically and depending on the muscle type, the heart being the most refractory to gene edition. The higher dose (1.10^{14} vg/kg, with vg corresponding to vector genome) was administered alongside immune suppression drugs to avoid immune response (Amoasii et al., 2018).

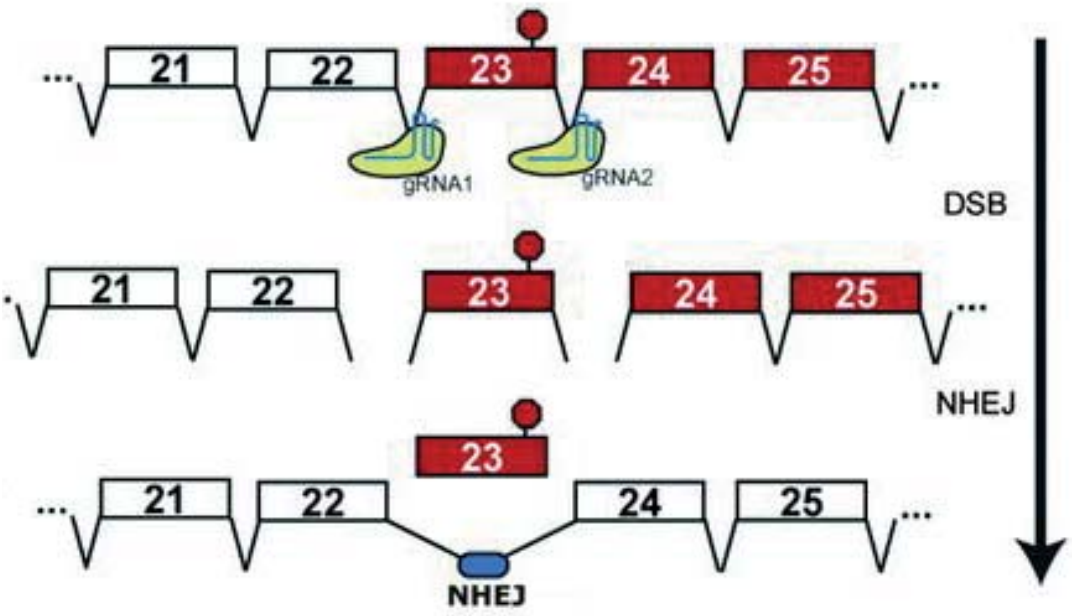


Figure 16: Exon skipping strategy for DMD. The Cas9 nuclease is targeted to introns 22 and 23 by two gRNAs. Two DSBs made by Cas9 leads to excision of the region surrounding the mutated exon 23, containing a strop codon. The ends are rejoined by NHEJ, reconstituting a shorter but functional dystrophin gene. Figure from (Nelson et al., 2016).

Advances in Myotonic Dystrophy Type I gene editing

A CTG repeat expansion from the 3' UTR of the DMPK locus was integrated into a yeast strain. A TALEN called $TALEN_{CTG}$ was designed to recognize and cut this CTG triplet repeat and was very efficient at shortening it (>99% cells showed contraction) and highly specific as no other mutation was detected in yeast cells (Richard et al., 2014). Similar approaches to delete CTG repeats were attempted: SaCas9 and gRNAs cutting upstream and downstream the repeats were expressed *in vitro* in DM1 patient cells. Clones exhibiting correctly excised CTG repeats showed reversed phenotype (Provenzano et al., 2017) (van Agtmaal et al., 2017). Those two studies used different cell types, respectively, myogenic DM1 myoblasts and DM1 fibroblasts and different target loci. They achieved, respectively, 46% and 14% of successfully edited cells. Indels were found in both cases at cut sites and few loci were tested for off-target effects. These approaches are summarized in a review which I coauthored, see Annex 1 (Mosbach et al., 2019a). More recently, dual AAV9 vectors encoding SaCas9 and two gRNAs targeting downstream and upstream CTG repeats were injected in tibialis anterior (TA) muscle of DMSXL homozygous mice. While the efficacy of edition was very low, 24% of muscle cells showed a decreased number of RNA foci, suggesting that this approach can be applied to curing DM1 (Lo Scrudato et al., 2019) **(Table 2).**

Mice model	Age at injection	AAV serotype	AAV quantity	Mode of injection	Cas9 type	Cell % expressing Cas9	Cell% edited	Phenotypic rescue?	Off targets?	Reference
mdx	P1	AAV9	6.10^{12}vg (Dual AAVs)	Intraperitoneal	SpCas9 (two gRNAs for exon 23 deletion strategy)	NA	1.8% in muscle expressing dystrophin, 3.2% in heart expressing dystrophin (8 weeks post-injection)	Yes, forelimb grip strength (4 weeks post injection)	None *in vitro* (over the top 10 predicted off-target sites)	(Long et al., 2016)
	P12		1.10^{13}vg (Dual AAVs)	Intramuscular (TA)			25% expressing dystrophin (6 weeks post-injection)	NA		
	P18		$1.8.10^{13}$vg (Dual AAVs)	Retro-orbital			6.1% expressing dystrophin in muscle, 5% in heart expressing dystrophin (4 weeks post-injection)	NA		
mdx	adult	AAV9	$1.5.10^{12}$vg (Dual AAVs)	Intramuscular (TA)	SaCas9 (two gRNAs for exon 23 deletion strategy)	NA	39% exon deletion	Muscle-specific force and decrease in force after eccentric damage (4 weeks post injection)	None *in vivo* (over the top 8 predicted off-target	(Tabebordbar et al., 2016)
	P3		3.10^{12}vg (Dual AAVs)	Intraperitoneal		NA	6% exon deletion in muscle, 5% exon deletion in heart (3 weeks post-injection)	NA	NA	
mdx	8 to 6-week old adult	AAV8	$5.6.10^{11}$vg to $7.46.10^{11}$vg (Dual AAVs)	Intramuscular (TA)	SaCas9 (two gRNAs for exon 23 deletion strategy)	NA	2% exon deletion 67% of myofibers expressing dystrophin (8 weeks post-injection)	Yes, twitch and tetanic force improved (8 weeks post-injection)	None to low (1% in one case) *in vivo* (over the top 10 predicted off-target	(Nelson et al., 2016)
	P2		$2.8.10^{11}$vg (Dual AAVs)	Intraperitoneal		NA	Dystrophin expression in muscle and heart (7 weeks post injection)	NA	NA	
	6-week old adult		$2.7.10^{12}$vg (Dual AAVs)	Intravenous		NA	Dystrophin expression in heart	NA	NA	
Homozygous DMSXL	5 to 9-week old adult	AAV9	1.10^{11}vg (Dual AAVs)	Intramuscular (TA)	SaCas9 (two gRNAs for CTG repeat expansion deletion strategy)	21% of myofibers expressing SaCas9 (HA-tagged), 76% of myofibers expressing gRNAs and 18% both (4 weeks post injection) (+Small inflammatory foci in 1 of 10 mice)	6% to 9% deletion	Yes, nuclei containing foci decreased by 24% No, DMPK mRNA level, and muscle strength unchanged (4 weeks post injection)	*In vivo*, no indels were found in 20 potential off-target sites	(Lo Scrudato et al., 2019)

Table 2: Summary of *in vivo* gene edition using CRISPR-Cas9 in disease models of muscular dystrophies.

Main issues raised by Cas9-mediated gene editing

Immune response

Immune response to viral vectors

One of the first gene therapy clinical trial was stopped because one person died from acute immune response to the vector leading to sepsis. This young man was receiving this treatment for a non-life threatening disease: ornithine transcarbomylase. This dramatic outcome was linked to preexisting immunity to the viral vector, an adenovirus Ad5 carrying the gene of interest, probably due to an infection to adenovirus before the clinical trial (Raper et al., 2003). rAAvs elicit immune response although no infection associated to AAV was ever reported (Blacklow et al., 1968). Immune ignorance in neonatal mice can be exploited to readminister AAV at adult age to ensure a longer term expression. This was successfully carried out using rAAV encoding β-Gal gene and administered in mice airways 3 and 6 months after initial perinatal gene transfer (Carlon et al., 2014).

Immune response to Cas9

In addition to immune response to the vector, Cas9 itself may be immunogenic. More than half of humans have antibodies against both SpCas9 and SaCas9, 78% and 58% respectively among 48 donors. Antigen-reactive T-cells against SpCas9 and SaCas9 were also detected, 78% and 68% respectively among 18 donors. It demonstrates the existence of both a humoral and cell-mediated preexisting adaptive immunity against Cas9 in human due to prior exposure to *Streptococcus pyogenes* or *Staphylococcus aureus* (Charlesworth et al., 2019). However, the small sample size might have bias; it would be worth extend this analysis to a bigger population. It may be possible to circumvent this problem by using Cas9 from bacteria that are less wide spread in our environment (Burstein et al., 2017). Nevertheless, it still indicates that when exposed to Cas9, the human immune system will develop antibodies against it. Mice also develop antibodies against Cas9. Immune response elicited by AAV encoding in mice was extensively characterized by the analysis of transcriptome by total RNA sequencing and retrieving expression signature of immune cell populations (Chew et al., 2016). It showed an enrichment in activated T cells and lymphoid lineages corresponding to adaptive immune response specific to Cas9 expression. Cas9 immune response was avoided by injecting neonatal mice. Later re-exposure to Cas9 antigen did not trigger any acquired immune response (Nelson et al., 2019).

Efficacy and toxicity

Which edition efficacy is required to fix mutations in vivo?

As highlighted in **Table 2**, edition efficiency is rather low, not every cell expressing the nuclease will bear the desired modification. Maybe correcting a small fraction of cells is enough to achieve therapeutic effect and improving edition rate may not be required. For example, in mosaic DMD mice generated by germline editing of a dystrophin mutation, as little as 15% genetic correction was sufficient to restore dystrophin expression to normal levels in nearly all myofibers (Long et al., 2014). This low gene edition % may be an exception and the reflect of the edition of a particular population of cells such as stem cells which will produce other identical and corrected cells. Furthermore, the muscle is a peculiar tissue, one syncitia of cells is made of many nuclei and one corrected nucleus may be enough to restore dystrophin levels to therapeutic level.

Controlling DNA repair outcome by favoring NHEJ or HR

Controlling DNA repair outcome may be a good way to enhance gene edition. It is pointless to enhance cleavage efficiency if the cells are going to religate the breaks by NHEJ without integrating the desired modifications. Downregulating NHEJ protein can enhance edition by HR. For example, in cells defective for DNA-PKc, HR was enhanced (Allen et al., 2002). Inhibiting DNA ligase IV expression was enough to increase homology directed recombination in human cell lines (Maruyama et al., 2015). Similarly, ku70 and DNA ligase IV suppression yielded the same results (Chew et al., 2016). Alternatively, promoting HR can be achieved by upregulating proteins involved in HR. Overexpression of hRad51 in human fibrosarcoma HT1080 cells enhances gene recombination with a donor template by 2-fold (Yáñez and Porter, 1999). However, overexpression of Rad52p in the same experimental conditions results in a 2-fold decrease in gene targeting as well as prolonged G1 phase, and cell survival impairment (Yáñez and Porter, 2002). On the contrary, expression of *S. cerevisiae* Rad52 in Hela cells increased targeted homologous recombination with a donor template up to 37-fold. Random integration, supposedly through the NHEJ pathway, was also significantly decreased. Number of Rad51 foci was also decreased indicating that pathway choice may be independent of Rad51 in this assay (Di Primio et al., 2005). Fusion proteins to Cas9, first CtiP (Charpentier et al., 2018) and then extended to Rad52 and Mre11 (Tran et al., 2019) increased homology directed recombination. A screen of 204 ORF involved in DNA damage repair identified RAD18 as an aid for CRISPR-Cas9- mediated homology directed repair. The assay performed in HEK293T

relied on the conversion of a BFP gene into GFP by HR with a template donor. Various domains of RAD18 were suppressed to generate variants that were tested using the same assay, and enhanced 18 (e18) was isolated. Expression of e18 promoted HR by inhibiting 53BP1 localization to DSBs and thereby inhibiting NHEJ (Nambiar et al., 2019). Finally, fusion of Cas9 to gemini protein, which first 110 amino acids were shown to confer nuclear localization and cell-cycle dependent expression, led to increased efficacy of HDR in cultured and dividing cells: HEK-293T (Gutschner et al., 2016) and hPSC cell lines (Howden et al., 2016). Thus, enhancing DSB repair by overexpressing proteins involved in HR and downregulating proteins involved in NHEJ may be a good strategy to induce more efficient targeted modifications in cells.

Edited cells must retain a functional repair pathway

However, two studies have shown that cells edited in CRISPR wide screenings are often mutated for p53. TP53 controls the cellular response to double-strand breaks. Cells with a functional p53 pathway were counter selected due to cell arrest triggered by p53 upon DSB formation. TP53 inhibition could alleviate toxicity of CRISPR edition for the cells but has the potential to increase off-target mutations and poses a risk for cancer. Therefore, checkpoint activity integrity should be controlled when developing cell-based therapies utilizing CRISPR–Cas9 (Ihry et al., 2018) (Haapaniemi et al., 2018).

Undesired modifications: off-target and on-target edition

Off and on-target mutations

Off-target edition can result in uncontrollable changes in cells and to cell death. Such events were reported in many studies (Fu et al., 2013) (Zhang et al., 2015). It is of big concern for clinical applications as unexpected outcomes may arise from off-target genome modifications. Another concern is the edition at on-target locus, as highlighted by the TALEN modified cattle bearing plasmid sequences at the intended target locus (Norris et al., 2019), rearrangements in mES cells, with deletions of up to 2kb (Kosicki et al., 2018), asymmetric deletions and large deletions into mouse zygotes injected with Cas9 (Shin et al., 2017). Study of long-term effect of AAV expressing cells *in vivo* revealed unexpected outcomes. mdx mice treated with a dual AAV vector encoding SaCas9 and two guides designed to excise exon 23 were analyzed by unbiased sequencing 8 weeks or 1-year post-injection (Nelson et al., 2019). Heterogenous genome-editing events were found at the on-target locus, including deletions, inversions, indels

and AAV integrations in all treated mice in all of the analyzed organs: TA, heart, diaphragm, liver. AAV integrations were also found at predicted off-target sites and at previously described integration sites; AAV integrations had already been reported in the past (Miller et al., 2004), however the induction of a double-strand break by Cas9 may change the integration landscape and genotoxicity profile of AAV. In this study, the frequency of AAV integration was higher than the intended edition event.

Limiting undesired editions

To reduce unwanted modifications, nuclease expression could be restricted in time. A self-limiting plasmid encoding the region targeted by the gRNA was designed to limit in time the expression of Cas9. It resulted in sustained genome editing events while limiting the expression of Cas9 (Ruan et al., 2017). Assessing Cas9 specificity at a particular locus should be assessed. *In silico* simulations can be used to design gRNAs with a limited potential off-targets in the genome (Concordet and Haeussler, 2018). Techniques to measure off-targets *in vitro* were developed such as GUIDE-Seq (Tsai et al., 2015), or VIVO (Akcakaya et al., 2018). The use of nucleases that were engineered to be more specific as described in IIB, may help limiting off target concerns. Using two paired nickases instead of one Cas9 increases the sequence length to be recognized and can result in higher specificity (Gopalappa et al., 2018).

Methods of delivery

Viral gene delivery

I will present the two main viral delivery vectors: rAAVs and lentiviruses. AAV is a protein shell surrounding a single-stranded DNA genome of around 4.8kb (Rose et al., 1966). AAV belongs to the parvovirus family and requires co-infection with other viruses, mainly adenoviruses, in order to replicate. Its genome contains three genes, Rep (Replication), Cap (Capsid), and aap (Assembly), giving rise to at least nine gene products. These coding sequences are flanked by inverted terminal repeats (ITRs) that are required for genome replication and packaging (Samulski and Muzyczka, 2014). To generate gene delivery vectors, recombinant AAV (rAAV), which lacks viral DNA, was designed and consists of a protein shell engineered to traverse the cell membrane, where it can ultimately traffic and deliver its DNA into the nucleus of a cell. In the absence of Rep proteins, ITR-flanked transgenes encoded within rAAV can form circular concatemers that persist as episomes in the nucleus of transduced cells. Because recombinant episomal DNA does not integrate into host genomes, it

is diluted over time and over division cycles, leading to the loss of the transgene (Choi et al., 2006). AAVs carry single stranded DNA; upon entry in the cell, ssDNA is converted into dsDNA which delays the expression of the transgene. This issue can be circumvented by producing AAVs containing self-complementary DNA. But this approach requires a transgene of less than 2.2kb (McCarty, 2008). AAV exhibit different tropisms, which makes them suitable to target specific tissues, although they elicit immune response which has to be carefully monitored (Zincarelli et al., 2008). AAVs are currently the delivery method for genes in various tissues (liver, muscle, central nervous system) for more than 100 clinical trials with no reported adverse effect caused by the vector. Production of rAAVs rely on the co-transfection of three vectors in HEK293T cell line; one vector carrying the gene of interest, one carrying AAV *gag* and *pol* genes and one carrying Adenovirus helper sequences. Production efficiency of rAAV particles can still be improved; one possibility is the establishment of stable cell lines, overcoming the need for a triple transfection (Clément and Grieger, 2016).

Lentiviral vectors were developed based on lentiviruses, such as HIV-1, which are retroviruses. The basic encoded genes are *gag*, encoding structural proteins, *pol* encoding genes required for reverse transcription and *env* encoding viral envelope glycoproteins (Escors and Breckpot, 2010). The two main steps of the retroviral life cycle are reverse transcription, converting viral RNA into double stranded viral DNA, followed by the integration into the host genome. The process of integration is not random and each class of retroviruses has its own preferences. For example, HIV-1 preferentially inserts within transcriptional units (Lewinski et al., 2006). First-generation lentiviral vectors contained the whole HIV genome except genes not essential for growth. In second-generation lentiviral vectors, further genes were removed (Vannucci et al., 2013). Finally, third-generation vectors were further improved as they were made virtually non-replicative: the viral genome was split into separate plasmids. One packaging plasmid carried *gag*, *pol* and *rev* genes, a second plasmid carried *env* gene and the third plasmid carried gene of interest flanked by LTR sequences used for insertion (Dull et al., 1998). Recombination between the three is unlikely but may still be a concern hence long-term follow up of patients is required by the FDA (Approval letter-ucm574106). Production of lentiviral vectors remains a challenge and relies on triple transfection; packaging cell lines are still in development (Sanber et al., 2015). Lentiviral vectors were extensively used to transduce hematopoeitic stem cells, for the treatment of HIV infection (McGarrity et al., 2013) and ß-thalassemia for example (Cavazzana-Calvo et al., 2010).

Non-viral delivery

Another alternative method is to use cationic lipidic vectors to deliver any nuclease coupled with a polyanionic molecule. Conjugation of Cas9 to cell-penetrating peptides was used to efficiently edit various cell types in culture (Ramakrishna et al., 2014). Effective delivery *in vivo* of functional Cre recombinase and functional Cas9:sgRNA complexes to hair cells in the inner ear of live mice was achieved. GFP disruption assay in hair cells revealed successful edition by Cas9 (Zuris et al., 2015). *In vivo* delivery studies are still needed to confirm the efficacy of these delivery methods. A very promising method is the use of gold nanoparticles complexed with Cas9 ribonucleoprotein and PAsp(DET), a coating polymer. This system was called CRISPR-Gold. Successful delivery of Cas9 ribonucleoprotein and donor DNA was achieved in *mdx* mice by intramuscular injection at 6 mg/kg in TA muscle. 5.4% of the dystrophin gene was corrected at the wild type gene after CRISPR-Gold treatment, Reduced fibrosis was also observed on TA sections (Lee et al., 2017a).

Cell population targeted

Which cell population to edit?

The goal of *in vivo* genome editing is to correct affected cells in order to cure genetic disorders. In the case of DM1, modifying muscle cells would alleviate myotonia symptoms and serve as a treatment for DM1. However, it is unclear how many cells need to be edited to have a positive outcome on the disease symptoms. Muscle cells are continuously replaced throughout life and satellite cells form new muscle fibers. It would then be of great advantage to be able to edit stem cells. First evidence of CRISPR-Cas mediated genome modification in stem cells was brought in mdx mice on muscle satellite cells (Tabebordbar et al., 2016). The same team subsequently showed that other population of stem cells are amenable to Cre-Lox edition: Cre-containing AAVs were administered to Lox mice which resulted in the expression of a fluorescent Tdt tomato reporter gene. In total, three lineages of stem and progenitor cells (myogenic, mesenchymal, hematopoietic) present in three distinct niches (skeletal muscle, bone marrow and skin) were efficiently targeted by AAV encoding Cre gene and retained their differentiation potential *in vitro*. Next question is to know whether these cell populations can be edited by Cas9 *in vivo* (Goldstein et al., 2019). Stem cells correction is envisioned as a therapeutic strategy to cure many rare disorders. *In vivo* genome modification of stem cells would avoid the isolation and transplantation of these cells and so maintain key regulatory interactions present in endogenous niches and preserve stem cells integrity. Deleterious effects

of *ex vivo* manipulation often result in failed engraftment when reinjected in patients (Wagers, 2012).

Are all the cells amenable to gene editing?

Cells in human body are not equal in terms of DNA repair potency. Edition outcome may greatly vary *in vivo* depending on the cell type targeted. Mouse embryonic stem cells preferentially use HR to repair DSBs. Even when overexpressing DNA Ligase IV, and thereby elevating NHEJ in ES cells, HR is still the preferred pathway. When cells are differentiated, the predominant pathway is NHEJ (Tichy et al., 2010). In muscle cells, genome editing for DM1 and DMD showed that muscle cells can repair by NHEJ resulting in large deletions of CTG expansions (Lo Scrudato et al., 2019) or in exon skipping (Tabebordbar et al., 2016). Muscle cells can also repair by homologous recombination to remove a deleterious stop codon (Lee et al., 2017b). More work needs to be done in order to characterize genome modification outcomes in therapeutically relevant cell populations.

Many issues still need to be adressed before using Cas9 or other highly specific nucleases to induce *in vivo* genome editing. In the present thesis, the goal is to investigate approaches to edit expanded repeats involved in microsatellite disorders, with an emphasis on DM1. DM1 CTG repeat expansion was integrated into a yeast strain. A TALEN was designed to recognize and cut the CTG triplet repeat and was very efficient at shortening it in yeast cells (>99% cells showed contraction) and highly specific as no other mutation was detected (Richard et al., 2014). In order to induce a double-strand break (DSB) into microsatellites, different types of nucleases can be used: meganucleases, Zinc Finger Nucleases (ZFN), Transcription activator-like effector nucleases (TALEN) and CRISPR-Cas9. Previous experiments using the I-SceI meganuclease to induce a DSB into CTG repeat tract in which an I-SceI recognition site was integrated, showed that repair occurred by annealing between the CTG repeats (Richard et al., 2000). ZFNs were used to induce DSBs into CAG/CTG repeats which mostly led to contractions in CHO cells (Mittelman et al., 2009) and in a HEK293 cells GFP reporter assay (Santillan et al., 2014). As only one arm was enough to induce DSBs into repeat tracts and since CAG fingers can recognize CTG triplets and vice versa, authors concluded that the specificity was not good enough for future applications. In our hands, TALENs are specific, and were used in yeast to cut long CTG repeat tracts and were shown to induce repeat contractions through single-strand annealing (SSA) between the repeats by a *RAD52*, *RAD50* and *SAE2* dependent mechanism (Mosbach et al., 2018)(annex 3).

<u>First problematic: Finding active nucleases to cut microsatellites</u>

Cutting repeated sequences like microsatellites may be difficult due to stable secondary structures that may form either on target DNA or on the guide RNA -when using the CRISPR-system- making some repeats more or less permissive to nuclease recognition and cleavage. In addition, secondary structure formation could impede double-strand break resection or later steps of the repair mechanism. Eukaryotic genomes contain thousands of identical microsatellites; therefore, the specificity issue may become a real problem when targeting one single locus. Here, we developed an *in vivo* biochemistry assay in order to test different nucleases belonging to the CRISPR-Cas system on microsatellites associated to human disorders.

<u>Second problematic: Gene editing for DM1</u>

$TALEN_{CTG}$ was tested in yeast cells on expanded microsatellites. Proof of concept in patient cells and DMSXL mice would be a great step toward applying CTG repeat shortening as a

therapeutic approach in humans. I studied the effect of $TALEN_{CTG}$ expression in DM1 patient cells and DM1 mouse model.

Third problematic: A human cell reporter assay to test nucleases on DM1 CTG expansions

Finally, I developed a reporter cell line for CTG repeats in HEK293FS cell line, to test nuclease activity in human cells on CTG repeat tracts.

Results

Efficacy of Cas nucleases on microsatellites involved in human disorders

I have set up a GFP reporter assay to test nucleases from the CRISPR system on various microsatellites involved on human disorders. I did all the experiments except Illumina libraries preparation which was performed by Lisa. Sequencing analysis was performed by Stéphane. I wrote the manuscript. This work is currently submitted and deposited in bioarchive (Poggi et al., 2019).

Differential efficacies of Cas nucleases on microsatellites involved in human disorders and associated off-target mutations

Lucie Poggi [1,2,3], Lisa Emmenegger [1,4], Stéphane Descorps-Declère [1,5], Bruno Dumas[3], Guy-Franck Richard[1,2]

1 Institut Pasteur, CNRS, UMR3525, 25 rue du Dr Roux, F-75015 Paris, France

2 Sorbonne Université, Collège Doctoral, 4 Place Jussieu, F-75005 Paris, France

3 Biologics Research, Sanofi R&D, 13 Quai Jules Guesde, 94403 Vitry sur Seine, France

4 Present address: Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Strasse 10, 13125 Berlin, Germany.

5 Institut Pasteur, Bioinformatics and Biostatistics Hub, Department of Computational Biology, USR3756 CNRS, F-75015 Paris, France

Keywords: Microsatellites, resection, CRISPR-Cas9, homologous recombination, off-targets

Abstract

Microsatellite expansions are the cause of more than 20 neurological or developmental human disorders. Shortening expanded repeats using specific DNA endonucleases may be envisioned as a gene editing approach. Here, a new assay was developed to test several CRISPR-Cas nucleases on microsatellites involved in human diseases, by measuring at the same time double-strand break rates, DNA end resection and homologous recombination efficacy. Broad variations in nuclease performances were detected on all repeat tracts. *Streptococcus pyogenes* Cas9 was the most efficient of all. All repeat tracts did inhibit double-strand break resection. We demonstrate that secondary structure formation on the guide RNA was a major determinant of nuclease efficacy. Using deep sequencing, off-target mutations were assessed genomewide. Out of 221 CAG/CTG or GAA/TTC trinucleotide repeats of the yeast genome, three were identified as carrying statistically significant low frequency mutations, corresponding to off-target effects.

Introduction

A growing number of neurological disorders were identified to be linked to microsatellite expansions (Orr and Zoghbi, 2007). Each disease is associated to a repeat expansion at a specific locus (Table 1). No cure exists for any of these dramatic disorders. Shortening the expanded array to non-pathological length could suppress symptoms of the pathology and could be used as a new gene therapy approach (Richard, 2015). Indeed, when a trinucleotide repeat contraction occurred during transmission from father to daughter of an expanded myotonic dystrophy type 1 allele, clinical examination of the daughter showed no sign of the disease (O'Hoy et al., 1993) (Shelbourne et al., 1992).

Table 1. Summary of the main microsatellite disorders and associated repeat expansions.

Sequence	**Disease**	**Locus**	**Expansion length (bp)**
$(CAG)_n$	Huntington Disease	*HTT* exon	30-180
$(GCN)_n$	Synpolydactyly, type 1	*HOXD13* exon	15
$(CTG)_n$	Myotonic dystrophy type 1 (DM1)	*DMPK* 3'UTR	50-10,000
$(CGG)_n$	Fragile X syndrome	*FRAXA* 5'UTR	60-200
$(GAA)_n$	Friedreich ataxia	*FRDA* exon	200-1,700
$(CCTG)_n$	Myotonic dystrophy (DM2)	*ZNF9* intron	75-11,000
$(ATTCT)_n$	Spinocerebellar ataxia, type 10	*ATXN10* intron	500-4500
$(TGGAA)_n$	Spinocerebellar ataxia, type 31	*TK2* / *BEAN* intron	500-760
$(GGCCTG)_n$	Spinocerebellar ataxia, type 36	*NOP56* intron	>650
$(GGGGCC)_n$	Amyotrophic lateral sclerosis	C9orf72 intron	700-1,600

In order to induce a double-strand break (DSB) into a microsatellite, different types of nucleases can be used: meganucleases, Zinc Finger Nucleases (ZFN), Transcription activator-like effector nucleases (TALEN) and CRISPR-Cas9. Previous experiments using the I-*Sce*I meganuclease

to induce a DSB into a CTG repeat tract showed that repair occurred by annealing between the flanking CTG repeats (Richard et al., 1999). Later on, ZFNs were used to induce DSBs into CAG or CTG repeats, which mostly led to contractions in CHO cells (Mittelman et al., 2009) and in a HEK293 cell GFP reporter assay (Santillan et al., 2014). As only one arm was enough to induce a DSB into the repeat tract and since CAG zinc fingers can recognize CTG triplets and *vice versa*, the authors concluded that the specificity was too low for further medical applications. As a proof of concept of the approach, a myotonic dystrophy type 1 CTG repeat expansion was integrated into a yeast strain. A TALEN was designed to recognize and cut the CTG triplet repeat and was very efficient at shortening it in yeast cells (>99% cells showed contraction) and highly specific as no other mutation was detected (Richard et al., 2014). The TALEN was shown to induce specific repeat contractions through single-strand annealing (SSA) by a *RAD52*, *RAD50* and *SAE2* dependent mechanism (Mosbach et al., 2018). As a proof of concept of the approach, a myotonic dystrophy type 1 CTG repeat expansion was integrated into a yeast strain. A TALEN was designed to recognize and cut the CTG triplet repeat and was very efficient at shortening it in yeast cells (>99% cells showed contraction) and highly specific as no other mutation was detected (Richard et al., 2014).

The CRISPR-Cas system is the easiest to manipulate and to target any locus, as sequence recognition is based on the complementarity to a guide RNA (gRNA). To recognize its sequence, Cas9 requires a specific protospacer adjacent motif (PAM) that varies depending on the bacterial species of the Cas9 gene. The most widely used Cas9 is wild-type *Streptococcus pyogenes* Cas9 (SpCas9) (Cong et al., 2013). Its Protospacer Adjacent Motif (PAM) is NGG and induces a blunt cut 3-4 nucleotides away from it, through concerted activation of two catalytical domains, RuvC and HNH, each catalyzing one single-strand break (SSB). Issues were recently raised about the specificity of SpCas9, leading to the engineering of more specific variants. In eSpCas9, three positively charged residues interacting with the phosphate backbone

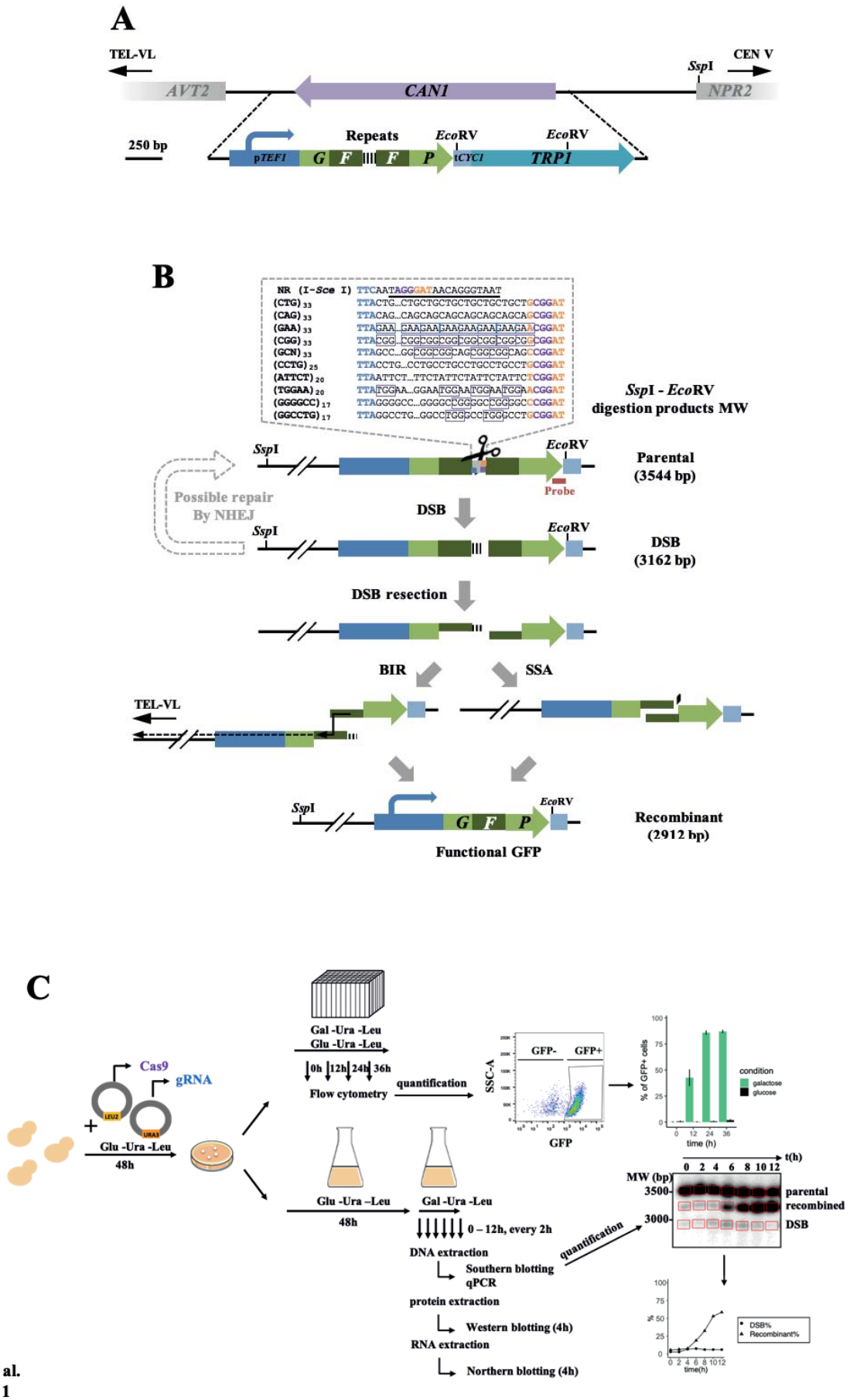
of the non-target strand were neutralized, conferring an increased specificity (Kleinstiver et al., 2016a). Similarly, Cas9-HF1 was mutated on 4 residues interacting through hydrogen bonds with the target strand (Slaymaker et al., 2016). *Staphylococcus aureus* is a smaller Cas9, its PAM is NNGRRT, having a similar structure to SpCas9 with two catalytic sites. Finally, type V CRISPR-Cas, Cpf1 nucleases, exhibit very different features including a T-rich PAM located 3' of target DNA and making staggered cuts leaving five-nucleotide overhangs by iterative activation of a single RuvC catalytic site (Zetsche et al., 2015).

Cutting repeated sequences like microsatellites may be difficult due to stable secondary structures that may form either on target DNA or on the guide RNA, making some repeats more or less permissive to nuclease recognition and cleavage. In addition, secondary structure formation could impede DSB resection or later repair steps. Eukaryotic genomes contain thousands of identical microsatellites, therefore the specificity issue may become a real problem when targeting one single locus. Here we developed an *in vivo* assay in the yeast *Saccharomyces cerevisiae* in order to test different nucleases belonging to the CRISPR-Cas family on synthetic microsatellites associated to human disorders. Our experiments revealed that these sequences may be cut, with surprisingly different efficacies between nucleases and between microsatellites. SpCas9 was the most efficient and nuclease efficacy relied mainly on gRNA stability, strongly suggesting that secondary structures are the limiting factor in inducing a DSB *in vivo*. DSB resection was decreased to different levels in all repeated tracts. In addition, we analyzed off-target mutations genomewide and found that three microsatellites with similar sequences were also edited by the nuclease. The mutation pattern was different depending on the microsatellite targeted.

Results

A GFP reporter assay integrated in the *Saccharomyces cerevisiae* genome enables the quantification of nuclease activity

The goal of the present experiments was to design and build a reporter system in the yeast *S. cerevisiae* to determine efficacy and specificity of different Cas nucleases on various microsatellites. In order to accurately compare experiments, we decided to use synthetic microsatellites integrated at the same position in the yeast genome. The advantage of this approach -as compared to using the human repeat tract sequences- was that all nucleases could be tested on the same genomic and chromatinian environment. In addition, we made the synthetic constructs in such a way that PAM sequences were available to each nuclease, which was not possible with human sequences. We therefore built a set of 11 isogenic yeast strains, differing only by the repeat sequence cloned in a cassette containing two synthetic GFP halves flanking 100 bp-repeats, integrated at the same genomic locus and replacing the *CAN1* gene on yeast chromosome V (Figure 1A). Note that given the repeated nature of the target DNA, some of them also harbor internal PAM sequences (Figure 1B).



Poggi et al.
Figure 1

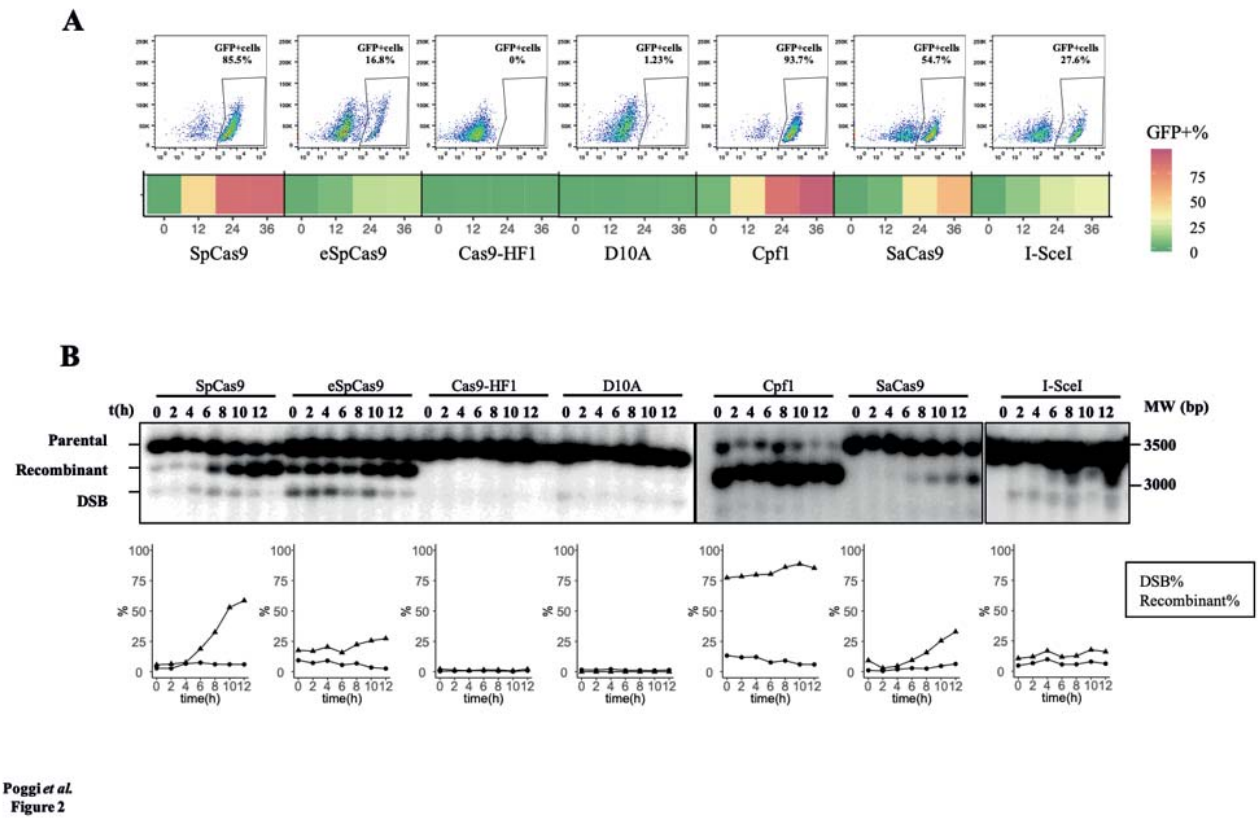
Figure 1: GFR reporter assay. A: The *CAN1* locus was replaced by recombinant GFP cassettes. Each synthetic GFP contains the constitutive *TEF1* promoter, followed by the bipartite eGFP interrupted by one of the ten microsatellites or by the I-*Sce* I recognition sequence and the *CYC1* terminator. The *TRP1* gene was used as an auxotrophic marker. **B**: Upon double-strand break induction, cells may repair by single-strand annealing, break-induced replication to reconstitute a functional GFP. Alternatively, repair by NHEJ may occur but will never lead to functional GFP. PAM sequences common to all constructs are colored (orange: SaCas9, purple: SpCas9, blue: FnCpf1) and additional PAM are boxed (same color code). *Ssp* I and *Eco* RV restriction site positions are indicated. Predicted molecular weights of each molecular species are indicated on the right. **C:** Cartoon depicting the experimental protocol (see text). Dot-plot axes are FSC-A/SSC-A.

Upon DSB induction, haploid yeast cells may fix the break by three different pathways. Homology regions flanking the DSB site may be used to repair the DSB either by single-strand annealing (SSA) between the two GFP halves, or by break-induced replication (BIR) to the end of the chromosome. In both cases, a fully functional GFP gene will be reconstituted. Note that in our experimental system, we cannot distinguish between BIR and SSA events. Alternatively, the DSB may be repaired by end-joining (NHEJ) between the two DNA ends. However, this is very unlikely, Homologous recombination is the preferred pathway in *S. cerevisiae*. In any case, perfectly religated DSB ends could be recut by the nuclease, until a functional GFP could be reconstituted by homologous recombination.

All experiments were performed as follows: independent yeast colonies expressing each Cas nuclease and its cognate gRNA were picked from glucose plates and seeded either in 96-deep well plates for flow cytometry measurements over a 36-h time period. Simultaneously, a colony from the same strain was expanded in a 500 mL flask to recover sufficient cells for further

molecular analyses (Figure 1C). As a control in all experiments, we used a non-repeated sequence containing the I-*Sce*I recognition site.

By flow cytometry, two distinct populations separated by one or two fluorescence intensity logarithms, corresponding to GFP-negative and GFP-positive cells, were observed upon nuclease induction (Figure 2A). Tested nucleases showed very different efficacies, SpCas9, FnCpf1 and SaCas9 were all more efficient than I-*Sce*I itself, as indicated by a higher number of GFP-positive cells. In order to know whether GFP-positive cells were a good readout of DSB efficacy, Southern blots were performed to detect and quantify parental and recombinant products as well as the DSB. A time course was run over a 12-hour period of time for each strain and each nuclease (except for the N863A Cas9 nickase). Parental, recombinant and DSB signals were quantified using phosphorimaging technology. In all cases, the DSB and recombinant products were detected, although in variable amounts (Figure 2B). The only exceptions were Cas9-HF1 in which no DSB nor recombinant band were detected, and Cas9-D10A in which a faint DSB signal was recorded but no recombinant molecules could be seen (see later). In subsequent experiments, the I-*Sce* I sequence will be used as our reference and called 'NR' (for Non Repeated).



Poggi *et al.*
Figure 2

Figure 2: CRISPR-Cas nuclease induction on non-repeated sequence containing an I-*Sce* I recognition site. A: Top: Percentage of GFP+ cells was measured throughout a time course of 36 hours. Dot plots indicate final populations at 36 hours. X-axis: FITC, Y-axis: SSC. Bottom: GFP+ cells are represented by a color code: from low recombination rates in dark green to high recombination in dark red. **B:** Top: Repair time courses were carried out during 12 hours. Parental (3500 bp), recombinant (3100 bp) and DSB (2900 bp) products were quantified (see Materials & Methods). Bottom: DSB (circles) and recombinant (triangles) products are represented as a percentage of the total signal in each lane.

Streptococcus pyogenes Cas9 variants exhibit a wide range of efficacies

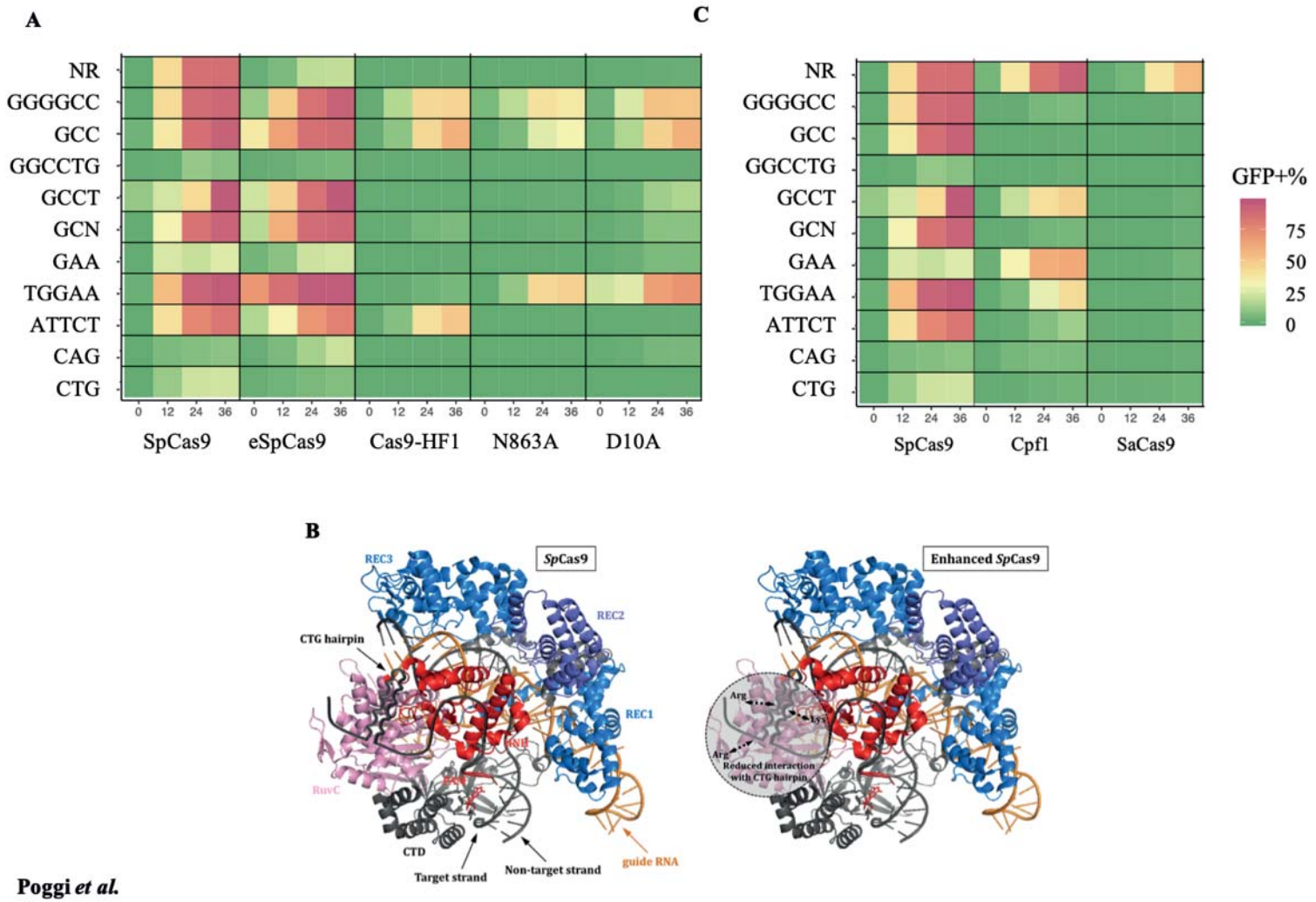
Once the experimental setup was optimized with I-*Sce* I (NR), the same exact assay was performed using seven different nucleases on ten microsatellites, including tri-, tetra-, penta- and hexanucleotide repeats (Supplemental Figures S1 & S2 and Supplemental Table S4). SpCas9 was able to cut every repeated sequence although $(GGCCTG)_{15}$, $(GAA)_{33}$, $(CAG)_{33}$

and $(CTG)_{33}$ were less efficiently cut (Figure 3A). Surprisingly, the G-quadruplex forming sequence GGGGCC was the most efficiently cut, although it is supposed to form stable secondary structures *in vitro* (Parkinson et al., 2002). This suggests that, despite possible secondary structures, this sequence is accessible to the nuclease *in vivo*. Alternatively, the presence of multiple PAMs at this locus may increase the chance that the nuclease would bind and make a DSB (Figure 1B). Two engineered variants of SpCas9 were then assayed. SpCas9 was more efficient that eSpCas9, itself consistently 2-10 times more efficient than Cas9-HF1 (Figure 3A). The NR sequence was also less efficiently cut, showing a general trend for these two variant nucleases. CTG repeats and CAG repeats were not cut the same way, eSpCas9 being more efficient on $(CAG)_{33}$ than SpCas9, although the contrary was found for $(CTG)_{33}$ (Figure 3A). It is known that CTG hairpins are more stable than CAG hairpins. Amrane *et al.*, (2005) showed that the Tm of a $(CTG)_{25}$ repeat was 58°C-61°C, depending on the method used for the measurement. In the same article, it was also shown that the Tm of a $(CAG)_{25}$ repeat was only 54°C, proving that it was less stable. However, to the best of our knowledge, there is no evidence at the present time for formation of such secondary structures in living cells. However, given that we found opposite results with CAG and CTG repeat tracts, it is possible some kind of secondary structures may occur *in vivo*. Given that eSpCas9 shows a reduced interaction with the non-target strand, it may be inferred that a CTG hairpin on this strand should not affect eSpCas9 as much as its wild-type counterpart (Figure 3B). Therefore, CAG repeats on the target strand (CTG on the non-target strand) should be cut more efficiently by eSpCas9, as it was observed in the present experiments.

FnCpf1 was then tested on the same repeats (Figure 3C). $(GAA)_{33}$ was the only one that was more efficiently cut by FnCpf1 than by SpCas9. This may be due to particular folding of the repeated sequence that makes it easier to cut by this nuclease. Alternatively, it may be due to

the presence of several PAM on the complementary strand which may more easily attract this nuclease at this specific locus (Figure 1B).

None to very low level of recombinant cells were observed when SaCas9 was induced, although it efficiently cut the NR sequence (Figure 3C). This may be due to particular conformation issues of the DNA and/or of the guide or to their expression levels.



Poggi *et al.*
Figure 3

Figure 3: GFP-positive cells after DSB repair. A: SpCas9 and variants. NR: I-*Sce* I recognition site. Each microsatellite is shown on an horizontal line and called by its sequence motif. Recombination efficacies are indicated by the same color code as in Figure 2A. **B:** Reconstructed models of SpCas9 (left) and eSpCas9 (right) interacting with a structured CAG/CTG repeat, according to the SpCas9 crystal structure (PDB: 4UN3). In this model the CAG sequence is on the target strand whereas the CTG hairpin is on the non-target one. The three recognition domains are indicated in different shades of blue. The RuvC and the HNH nuclease domains are shown in pink and red, respectively. The three mutated amino acids in eSpCas9 (two arginine and one lysine residues) are also indicated. C: GFP-positive cells after SpCas9, SaCas9 or FnCpf1 inductions.

gRNA and protein levels do not explain differences observed between nucleases

In order to determine whether DSB efficacies could be due to differences in protein levels or gRNA expression, we performed Western and Northern blots. For each guide, the signal corresponding to the expected RNA was quantified and compared to the signal of a control *SNR44* probe, corresponding to a snoRNA gene (Supplemental Figure S3A). For SpCas9, in one strain, $(GCN)_{33}$, smaller species were detected, around 75 nt, that may correspond to degradation or abortive transcription. Using the classical phenol-glass beads protocol and despite numerous attempts, the FnCpf1 gRNA could not be detected. We hypothesized that it may be so tightly associated to its nuclease that phenol could not extract it, or that its amount was too low to be detected by Northern blot. Therefore, an alternative protocol used to extract very low levels of small RNAs was performed (see Materials & Methods), but did not allow to detect FnCpF1 gRNA. For SpCas9 and SaCas9, guide RNA levels were different among the ten strains. No correlation was found between gRNA quantification and GFP-positive cells,

showing that gRNA steady state level was not the limiting factor in this reaction (Supplemental Figure S3B, left panel).

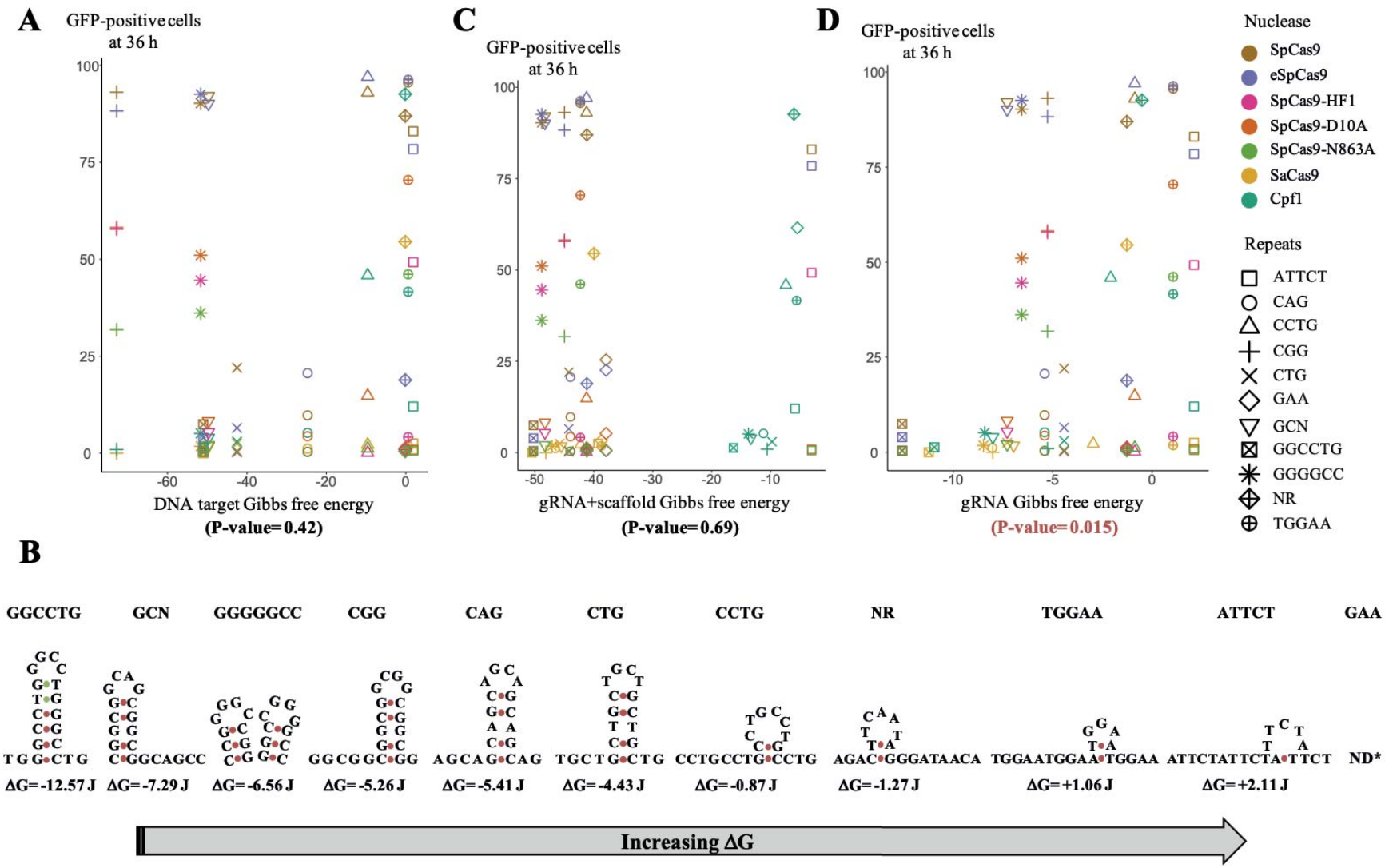
To assess the level of protein, total extracts were performed from yeast cells containing the NR sequence and the seven different nucleases. Note that different antibodies were used since proteins were not tagged. SpCas9 and its derivative mutant forms were detected with the same antibody, whereas SaCas9 and FnCpf1 were each detected with a specific monoclonal antibody (Supplemental Figure S3C). The same membranes were then stripped and rehybridized with an antibody directed against the product of *ZWF1*, encoding the ubiquitous glucose-6-phosphate dehydrogenase protein. Nuclease levels over control protein levels did not correlate with GFP-positive cells (Supplemental Figure S3B, right panel). Interestingly, the steady state level of eSpCas9 was found to be six times higher than SpCas9. This may be due to a higher stability of the protein, which could explain the high background of GFP-positive cells observed in repressed conditions (Supplemental Figure S2).

Overall, we concluded from these experiments that DSB efficacies were not obviously correlated to gRNA levels (at least for Sp- and SaCas9), nor to nuclease levels. This conclusion must be tempered by the fact that different antibodies with different affinities were used to detect nucleases. Therefore, we cannot totally rule out that SpCas9 was much more abundant than SaCas9 and/or FnCpf1 in our experiments.

Secondary structure stability partly explains DSB efficacy

Trinucleotide repeats involved in human disorders are known to form stable secondary structures *in vitro*. This has been extensively studied and reviewed over the last 25 years (Gacy et al., 1995; Lenzmeier and Freudenreich, 2003; McMurray, 2010; Mirkin, 2006; Pearson et al., 2005; Richard et al., 2008; Usdin et al., 2015). Secondary structures are known to form both at DNA and at RNA levels (Kiliszek and Rypniewski, 2014; Kiliszek et al., 2010). It is however

unclear if such structures actually exist in living cells, although genetic data strongly suggest that some kind of secondary DNA structures may be transiently encountered during replication and/or DNA repair. In our present experiments, secondary structures may possibly form on target DNA and on guide RNA. We therefore calculated theoretical Gibbs free energy for each target DNA and did not find any obvious correlation between structure stability and GFP-positive cells (Figure 4A, ANOVA test p-value=0.42) (see Materials & Methods). We subsequently performed the same calculation for the 20 nt guide RNA with or without their cognate scaffolds. Predicted structures of gRNA are shown in Figure 4B. When RNA scaffolds were taken into account, theoretical Gibbs energies were very low and comparable to each other, except for FnCpf1 guide RNA, which is much smaller than the others and for ATTCT gRNA that do not form secondary structure. This indicates that scaffold stability most frequently outweighs the 20 nt guide sequence stability. There was no correlation between scaffold stability and GFP-positive cells (Figure 4C, ANOVA test p-value=0.69). Finally, a statistically significant inverse correlation was found between the 20 nt guide RNA stability and GFP-positive cells (Figure 4D, ANOVA test p-value=0.015). We concluded that the 20 nt guide RNA stability was negatively correlated to GFP-positive cell formation, although it was not the sole determinant.



Poggi *et al.*
Figure 4

Figure 4: Secondary structures and Gibbs free energy. A: GFP-positive cells as a function of Gibbs energy calculated for each DNA sequence. Each nuclease is represented by a different color code and each microsatellite by a different shape. **B**: Predicted secondary structure of each SpCas9 gRNA. The *mfold* algorithm was used to model each structure. Only the most stable one is shown here. No structure could be calculated for GAA repeats. **C**: GFP-positive cells as a function of Gibbs energy calculated for each gRNA and its containing scaffold. **D**: GFP-positive cells as a function of Gibbs energy calculated for each gRNA alone. P-values are given below each graph.

DSB-resection of microsatellites

Resection rate at *Sty* I restriction sites (Supplemental Figure S4) was measured by qPCR as previously described (Zierhut and Diffley, 2008) (Chen et al., 2013). Resected single-stranded DNA will not be digested by *Sty* I and will generate a PCR product whereas double-stranded DNA will be digested and will not be amplified. Resection ratios at 12h were calculated as resection at the repeat-containing end over resection at the non-repeated DSB end. They were normalized to the NR sequence whose ratio was set to 1. Resection values were only determined when the DSB was detected unambiguously at 12 hours. When SpCas9 was induced, resection rates were reduced at the repeated end as compared to the non-repeated end. When FnCpf1 was induced, resection rates were also lower on the repeated end for $(GAA)_{33}$, $(CTG)_{33}$ and $(ATTCT)_{20}$, $(CCTG)_{25}$ but not for $(TGGAA)_{20}$. In conclusion, almost all repeats tested here inhibited resection to some level.

Correlation between nuclease efficacy measured by flow cytometry and double strand break rate.

To determine whether the flow cytometry assay recapitulates nuclease efficacy at molecular level, time courses were performed over 12-hour time periods for each nuclease-repeat couple. GFP-positive cell percentage at 12, 24 or 36 hours was plotted as a function of cumulative DSB over 12 hours (Figure 5). Given the number of different strains and nucleases tested, only one time course was performed in each condition (Supplemental Figure S5). However, data were very consistent between time points, showing that experimental variability was low. A linear correlation between the number of GFP-positive cells at 12 hours and the total signal of DSB accumulated during the same time period was found (linear regression test p-value=$1.1x10^{-9}$, R^2= 0.62) (Figure 5A). A good linear correlation was also found at later time points, 24 hours (p-value=$2.6x10^{-8}$, R^2=0.56) and 36 hours (p-value=$4.8x10^{-8}$, R^2=0.48) (Figures 5B, 5C). In conclusion, this GFP reporter assay is a good readout of double-strand break efficacy, and could be used in future experiments with other repeated sequences and different nucleases.

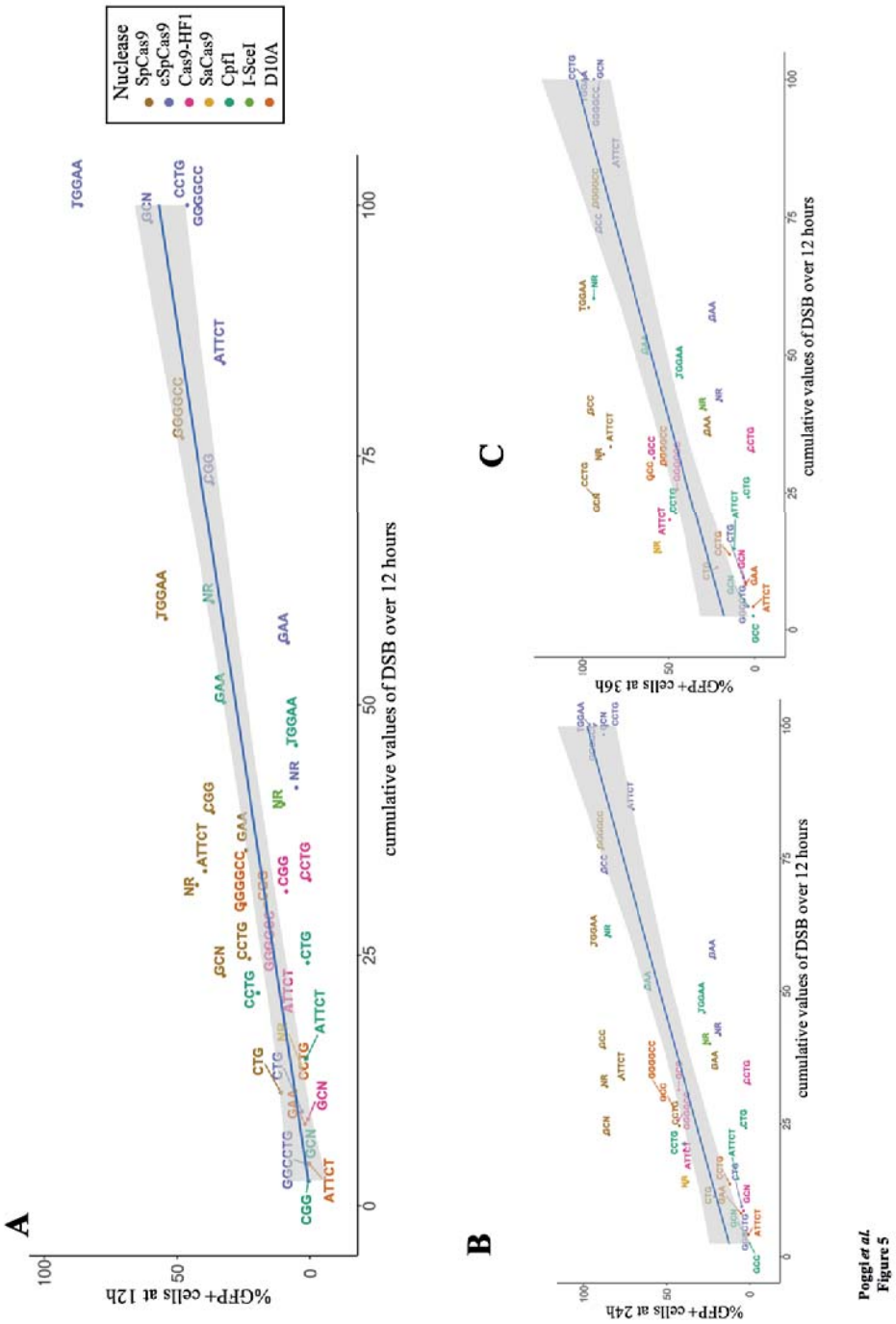


Figure 5: GFP-positive cell percentages as a function of DSB. Correlation between cumulative DSB level and GFP-positive cell percentage. Each color represents a nuclease. The blue line corresponds to the linear regression model. In grey: 95% confidence interval of the model. **A:** After 12 hours. **B**: After 24 hours. **C:** After 36 hours.

Cas9-D10A nicks are converted to DSB *in vivo*

The Cas9 mutant D10A was more efficient than N863A on all repeats (Figure 3A). Surprisingly, both nickases were able to induce recombinogenic events on $(GGGGCC)_{15}$, $(GCC)_{33}$, $(CCTG)_{25}$, $(GCN)_{33}$, $(GAA)_{33}$ and $(TGGAA)_{20}$ repeats. SSBs are repaired by a specific machinery in yeast involving Base Excision Repair (BER) (Krokan and Bjørås, 2013). Nicks do not trigger homologous recombination unless they are converted to DSB. However, in our experiments, nicks trigger homologous recombination on some repeats. By Southern blot, DSBs were visible when Cas9-D10A was induced on those repeats (Supplemental Figure S5). These may be due to mechanical breakage during DNA preparation procedure, which converts SSB into DSB. Therefore, genomic DNA was prepared in agarose plugs to check this hypothesis. DNA extraction was carried out on Cas9-D10A time courses for $(GAA)_{33}$ and $(CGG)_{33}$ and NR sequences. DSBs were visible, suggesting that nicks were indeed converted into DSBs *in vivo* (Supplemental Figure S6).

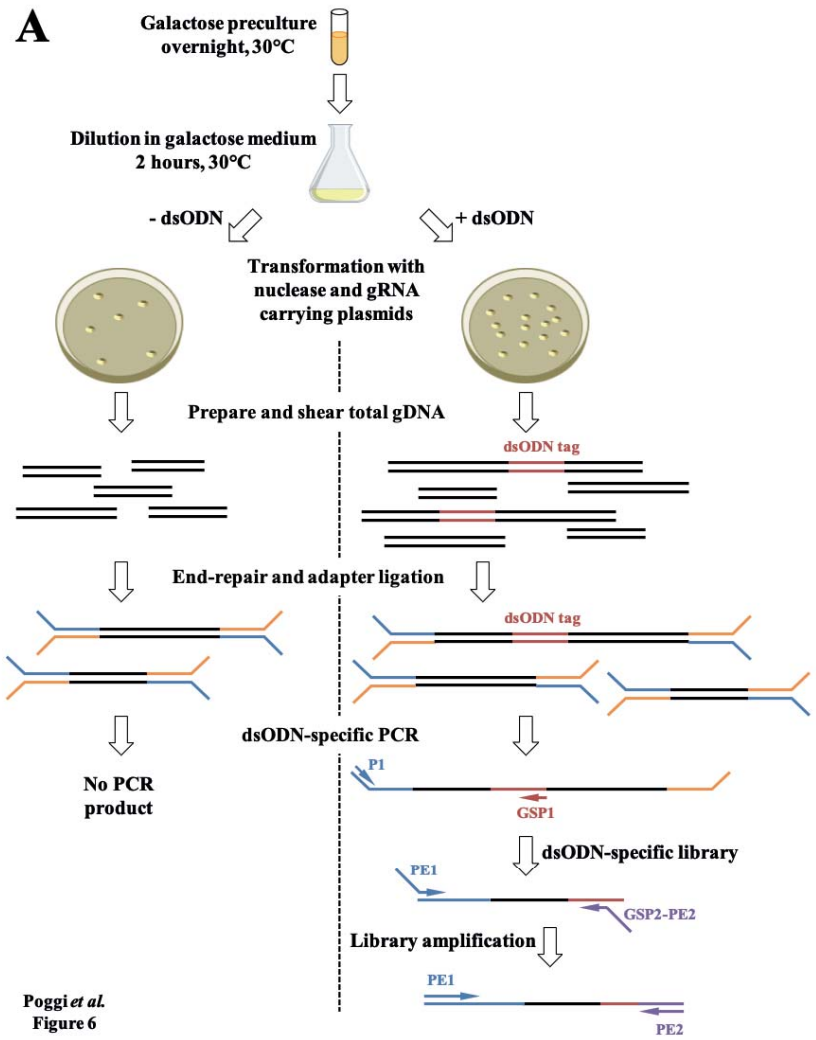
Genome-wide determination of off-target mutations

Microsatellites are very common elements of all eukaryotic genomes and the yeast genome contains 1,818 di-, tri- and tetranucleotide repeats (Malpertuy et al., 2003). In our present experiments, it was possible that other microsatellites of the yeast genome could also be mutated. We therefore decided to use an unbiased approach to determine all possible off-target sequences. The GUIDE-seq method was described in 2015 as a global approach to detect genome-wide DSB, and it was decided to adapt this method to budding yeast (see Materials & Methods). Shortly, cells were transformed with SpCas9 or FnCpf1, a gRNA and a modified double-stranded oligodeoxynucleotide (dsODN) to serve as a tag for targeted amplification. We chose the NR sequence as a control, as well as 6 out of 10 microsatellites cut by SpCas9 and the three most efficiently cut by FnCpf1. Colonies were collected, pooled and total genomic

DNA extracted. Following random shearing and repair of DNA ends, two successive rounds of PCR were performed, using a primer complementary to the dsODN. DNA yield was unfortunately too low to be directly sequenced, and an additional round of PCR was performed (Figure 6A). The resulting libraries were loaded on an Illumina sequencer. Out of 78 millions reads, only 1.5 millions (1.9%) contained the dsODN. These reads were mapped to the yeast genome and found to be specially enriched at the rDNA locus and mitochondrial DNA (Supplemental Figure S7). In addition, from 7 – 2103 gene loci whose coverage was above twice the median coverage were identified in each library. These positions were compared to predicted off-targets using the CRISPOR web tool (Haeussler et al., 2016). Out of 68 genes in the CAG library, only one was predicted as a possible off-target and out of 2103 genes in the GAA library, only six were predicted as possible off-targets. In the libraries, there was an overlap between CRISPOR predictions and dsODN-containing gene loci. We therefore concluded that this approach was not efficient to identify real off-targets in the yeast genome.

We decided to use a different approach to try to identify off-target sites, since that for each library, millions of reads homogeneously covered the whole genome. Classical SNP and indel calling algorithms aim at identifying frequent variants. However, off-targets are rare events, therefore the following pipeline of analysis was developed. In each library, variant reads were identified at each position predicted by CRISPOR. This ended up in 56 positions containing variant reads within microsatellites (Supplemental Table S5). Next, among these 56 positions, all positions containing only one mutant read were discarded. This left us with 14 genes containing at least one mutant read at a predicted off-target position. In order to determine whether these mutant reads were statistically significant, they were compared to the number of mutant reads at the same positions in the NR library used as a control. Given that colony number differed from one transformation to another one, the mean coverage per colony was used to normalize read number in each library (Mean coverage/CFU= genome coverage/CFU, Figure

6B). Once normalized, mutant reads in each library were compared to the NR control, using the Fisher exact test. Out of 14 possible off-target genes, only three exhibited read numbers significantly different from the NR control (Figure 6C). In the end, one gene (*YMR124w*) was identified to be a valid off-target for SpCas9 targeting CTG repeats, and two genes (*QCR6* and *LEO1*) were validated as off-targets for GAA repeats respectively targeted by SpCas9 and FnCpf1. Interestingly, all mutations in *YMR124w* were deletions of one or more triplets, but the two validated off-targets in the GAA library were all point mutations (Supplemental Figure S8). In conclusion, in the present experiments only nucleases targeted to CTG and GAA repeats exhibited some off-target effects.



Poggi *et al.*
Figure 6

B

Repeats	Nuclease	CFU	Nbr reads	Genome coverage	Mean coverage/ CFU
NR (I-Sce I)	SpCas9	90	4,618,586	55	0.6
CTG	SpCas9	150	3,429,892	41	0.3
CAG	SpCas9	70	7,407,743	89	1.3
GAA	SpCas9	70	14,029,331	168	2.4
CGG	SpCas9	150	3,237,063	39	0.3
TGGAA	SpCas9	150	9,068,696	109	0.7
GGGGCC	SpCas9	200	8,695,246	104	0.5
GAA	Cpf1	200	8,180,824	98	0.5
CCTG	Cpf1	200	11,028,944	132	0.7
TGGAA	Cpf1	200	8,591,751	103	0.5

C

Repeats	Nuc.	Locus	Mutant	Total	NR mutant	NR total	Fisher exact test
CTG	Cas9	*YMR124w*	27	63	2	95	7.5×10^{-11}
CAG	Cas9	*UBP7*	2	91	0	118	NS
CAG	Cas9	*SLM1*	4	81	2	90	NS
CAG	Cas9	*MGA1*	2	100	2	82	NS
CAG	Cas9	*YMR124w*	2	90	2	95	NS
GAA	Cas9	*QCR6*	20	74	10	108	2.3×10^{-3}
GAA	Cas9	*SCW11*	1	62	0	92	NS
GAA	Cas9	*YNL058c*	2	106	0	87	NS
GAA	Cas9	*Intergene*	4	79	8	63	NS
GAA	Cas9	*LEO1*	9	58	2	72	2.3×10^{-2}
GAA	Cas9	*Intergene*	8	60	8	77	NS
GAA	Cas9	*CLB5*	10	144	3	102	NS
GAA	Cas9	*Intergene*	4	60	0	45	NS
GAA	Cpf1	*FIG2*	4	218	0	45	NS

Figure 6: Off-target analysis. A: Cartoon depicting the experimental protocol (see text). **B**: Deep-sequencing results. For each library, the number of yeast colonies (CFU) after transformation and read numbers are given. Genome coverage was calculated by dividing (read number x 150 nucleotides) by 12.5 x 10^6 nucleotides (haploid yeast genome). Mean coverage was found by dividing genome coverage by CFU. **C**: Statistical analysis. For each of the 14 putative off-targets, the Fisher exact test was used to compare mutants reads in each library to mutant reads in the NR library.

Discussion

Here, we successfully designed an assay for determining Cas9 variant efficacy on various microsatellites. Type II CRISPR-Cas nucleases were classified according to decreasing efficacies in the following order: SpCas9, eSpCas9, Cas9-HF1, SaCas9. FnCpf1, the only type V nuclease tested, was shown to exhibit substrate preferences different from type II nucleases. We also demonstrated that gRNA and protein levels did not generally correlate to nuclease activity and thus are not limiting factors in our experimental assay, ensuring that we are measuring nuclease activity and DSB repair *per se*.

In vivo nuclease activities correlate to activities observed *in vitro*

Previous biophysical analyses showed that Cas9-HF1 and eSpCas9 bound to DNA similarly to SpCas9, but variants were trapped in an inactive state when bound to off-target sequences (Chen et al., 2017). Cas9-HF1 was more efficiently trapped in this inactive state than eSpCas9, showing more drastic impairment of cleavage. In our experiments, SpCas9 was more efficient than the two variants, confirming these biochemical data. Single molecule analyses enabled the precise determination of Cas9 binding and cleavage: first, the nuclease scrolls the genome for a PAM, then sequentially unwinds DNA starting from it (Sternberg et al., 2014). This explains

why SpCas9 is not tolerant to mutations in the region proximal to the PAM. In our experiments, it may also explain why PAM-rich repeats were more easily cleaved, more protein could be recruited at the locus. However, Malina *et al* (2015) observed the opposite, decreased DSB repair when additional PAMs were present within the target sequence.

A very good correlation was generally observed between DSB efficacy and recombination (Figure 5). However, for some repeat/nuclease couples this was not the case (TGGAA/SpCas9 and GAA/eSpCas9 for example). We hypothesized that resection defects may lead to the observed phenotype, as we previously showed that a $(CTG)_{80}$ repeat tract reduced resection efficacy in yeast in a *SAE2*-dependent manner (Mosbach et al., 2018). Comparison of resection values between repeated and non-repeated ends demonstrated that all repeats inhibit resection (Supplemental Figure S4). Therefore, differences in recombination are not due to resection defects alone. Note that in the present work, much shorter repeats (33 CTGs) were used as compared to our previous experiments with 80 CTGs.

Nickases trigger homologous recombination on some repeat tracts

We confirm earlier findings that the RuvC Cas9-D10A mutant was more efficient than the HNH N863A variant (Gopalappa et al., 2018). Surprisingly, both nickases induced homologous recombination into GGGGCC, GCC, TGGAA repeat tracts and to a lower extent into CCTG, GCN and GAA repeat tracts (Supplemental Figure S2). Nicks are usually formed in the course of the BER pathway and trigger specific protein recruitment (Krokan and Bjørås, 2013). Nicks are therefore normally not processed by double-strand break repair machineries. However, there is some evidence supporting the hypothesis that nicks may be recombinogenic (Maizels and Davis, 2018; Strathern et al., 1991) which is in agreement with our data. For example, in *S. pombe,* mating type switching occurs by homologous recombination after the conversion of a nick into a DSB during replication (Arcangioli, 1998; Dalgaard and Klar, 2001). In our assay,

replication may also convert a nick into a DSB, triggering homologous recombination in repeated sequences as suggested by the presence of a DSB observed throughout repair time course (Supplemental Figure S5). In a former work in human cells, Cas9-D10A was found to induce CTG/CAG repeat contractions, which may be due to the fact that many nicks were created into the target strand due to the repeated nature of the sequence, which then led to gap repair (Cinesi et al., 2016). In our experiment, gaps due to multiple nicks may arise in GGGGCC, CGG and TGGAA repeat tracts (Figure 1B), and GFP-positive cells were indeed observed in these three strains when Cas9-D10A was expressed. These multiple nicks may be either converted into DSB or form gaps that will then be converted into DSB.

Correlation between secondary structure formation and nuclease efficacy

The sgRNA plays a crucial role in orchestrating conformational rearrangements of Cas9 (Wright et al., 2015). Stable secondary structure of the guide RNA as well as close state of the chromatin negatively affect Cas9 efficiency (Chari et al., 2015; Jensen et al., 2017). Possible secondary structures formed by the guide RNA are important to determine nuclease activity although there is no clear rule that can be sorted out and it is still challenging to know which hairpins will be detrimental (Thyme et al., 2016). We found that more stable gRNAs were correlated to less efficient DSBs (Figure 4D). This is consistent with former studies on non-repeated gRNAs showing that stable structured gRNAs (<-4 kcal/mol) were not efficient at inducing cleavage (Jensen et al., 2017). In addition, improperly folded inactive gRNAs could be competing with active and properly folded gRNAs within the same cell, to form inactive or poorly active complexes with Cas9, that will inefficiently induce a DSB (Thyme et al., 2016). Differential folding of CAG and CTG gRNA may explain the difference of efficacy observed with SpCas9 and eSpCas9. *In vitro* assays revealed that CTG hairpins were more stable than CAG hairpins because purines occupy more space than pyrimidines and are most likely to

interfere with hairpin stacking forces (Amrane et al., 2005). This is probably also true *in vivo*, since CAG/CTG trinucleotide repeats are more unstable when the CTG triplets are located on the lagging strand template, supposedly more prone to form single-stranded secondary structures, than the leading strand template (Freudenreich et al., 1997; Viterbo et al., 2016). This difference in hairpin stability may impede Cas9/gRNA complex formation and/or impede recognition of target DNA by the complex. In our assay, SpCas9 cuts CTG more efficiently than CAG, whereas this is the other way around for eSpCas9. This may be due to reduced interaction between SpCas9 and the non-target strand, that is probably more prone to form secondary structures (Figure 3B).

Finally, a lower preference for T and a higher preference for G next to the PAM was previously reported (Chari et al., 2015). Other nucleotide preferences were found (Doench et al., 2014) but the preference for a G at position 20 of the guide is consistent across studies. This may explain why SpCas9 may be more efficient on CGG, CCTG and less on ATTCT repeats (Figure 3A).

Defining the best nuclease to be used in gene therapy

It was previously shown that a DSB made into CTG repeat tracts by a TALEN was very efficient to trigger its shortening (Mosbach et al., 2018; Richard, 2015). Other approaches may be envisioned to specifically target toxic repeats in human using the CRISPR toolkit: i) Cas9-D10A induced CAG contractions (Cinesi et al., 2016), ii) dCas9 targeting microsatellites was able to partially block transcription, reversing partly phenotype in DM1, DM2 and ALS cell models (Pinto et al., 2017), iii) efficient elimination of microsatellite-containing toxic RNA using RNA-targeting Cas9 was also reported (Batra et al., 2017). Finally, if using CRISPR endonucleases to shorten toxic repeats involved in microsatellite disorders was envisioned, our study will help finding the best nuclease. For example, Fragile X syndrome CGG repeats could be efficiently targeted with SpCas9. It must be noted that all human microsatellites may not be

targeted by all nucleases tested here, for some of them lacking a required PAM. However, our results allow to discard inefficient nucleases for further human studies.

However, specificity must also be taken into consideration. Previous analyses showed that the yeast genome contained 88 CAG/CTG, 133 GAA/CTT and no CGG/CCG trinucleotide repeats (Malpertuy et al., 2003). The dsODN tag was preferentially found at the rDNA locus and in mitochondrial DNA. This suggests that random breakage occurs frequently within these repeated sequences. This is compatible with the high recombination rate observed at the rDNA locus following replication stalling (Mirkin and Mirkin, 2007; Rothstein et al., 2000). However, using an alternative method to detect rare variants we were able to identify three real off-targets in the yeast genome. Off-target mutations were found in one CTG repeat out of 88 and two GAA repeats out of 133, for SpCas9 and FnCpf1. This shows that although very frequent sequences like microsatellites were predicted to be off-targets, few real mutations were indeed retrieved. By comparison, the human genome contains 900 or 1356 CAG/CTG repeats, depending on authors (Kozlowski et al., 2010)(Lander et al., 2001). Given our results, we can predict that ca. 1% of these would be real off-targets for a SpCas9 directed to a specific CTG microsatellite. However, in our experiments, the nuclease was continuously expressed, which is not envisioned in human genome editing approaches. Reducing the expression period of the nuclease should also help reducing off-target mutations, but this has now to be thoughtfully investigated.

The GUIDE-seq method was very successful at identifying off-target site sin the human genome, following Cas9 expression. In *S. cerevisiae*, we showed here that this approach was not efficient, most probably because NHEJ is not as active as in human cells, particularly in haploid yeast in which it is downregulated (Frank-Vaillant and Marcand, 2001; Valencia et al., 2001). Altogether, our results give a new insight into which nuclease could be efficiently used to induce a DSB into a microsatellite in other eukaryotes.

Methods

Yeast plasmids

A synthetic cassette (synYEGFP) was ordered from ThermoFisher (GeneArt). It is a pUC57 vector containing upstream and downstream *CAN1* homology sequences flanking a bipartite eGFP gene interrupted by the I-*Sce* I recognition sequence (18 bp) under the control of the *TEF1* promoter and followed by the *CYC1* terminator. The *TRP1* selection marker along with its own promoter and terminator regions was added downstream the eGFP sequences (Figure 1A). The I-*Sce* I site was flanked by *Sap* I recognition sequences, in order to clone the different repeat tracts. Nine out the 10 repeat tracts were ordered from ThermoFisher (GeneArt) as 151 bp DNA fragments containing 100 bp of repeated sequence flanked by *Sap* I sites. The last repeat (GGGGCC) was ordered from Proteogenix. All these repeat tracts were cloned at the *Sap* I site of synYEGFP by standard procedures, to give plasmids pLPX101 to pLPX110 (Supplemental Table S1). All nucleases were cloned in a centromeric yeast plasmid derived from pRS415 (Sikorski and Hieter, 1989), carrying a *LEU2* selection marker. Each open reading frame was placed under the control of the GalL promoter, derived from *GAL10*, followed by the *CYC1* terminator (DiCarlo et al., 2013).These plasmids were cloned directly into yeast cells by homology-driven recombination (Muller et al., 2012) using 34-bp homology on one side and 40-bp homology on the other and were called pLPX10 to pLPX16. Primers used to amplify each nuclease are indicated in Supplemental Table S2. Nucleases were amplified from Addgene plasmids indicated in Supplemental Table S1. The I-*Sce* I gene was amplified from pTRi103 (Richard et al., 2003).Guide RNAs for SpCas9 (and variants) were ordered from ThermoFisher (GeneArt), flanked by *Eco* RI sites for subsequent cloning into pRS416 (Sikorski and Hieter, 1989). SaCas9 and FnCpf1 guide RNAs were ordered at Twist

Biosciences, directly cloned into pRS416 (see Supplemental Table S1 for plasmid names). Each guide RNA was synthetized under the control of the *SNR52* promoter.

Yeast strains

Each synYEGFP cassette containing repeat tracts was digested by *Bam* HI in order to linearize it and transformed into the FYBL1-4D strain (Gietz et al., 1995). Correct integrations at the *CAN1* locus were first screened as [CanR, Trp+] transformants, on SC -ARG -TRP +Canavanine (60 µ/ml) plates. Repeats were amplified by PCR using LP30b-LP33b primers and sequenced (Eurofins/GATC). As a final confirmation, all transformants were also analyzed by Southern blot and all the [CanR, Trp+] clones showed the expected profile at the *CAN1* locus. Derived strains were called LPY101 to LPY111 (Supplemental Table S3).

Flow cytometry assay

Cells were transformed using standard lithium-acetate protocol (Gietz et al., 1995) with both guide and nuclease and selected on 2% glucose SC -URA -LEU plates and grown for 36 hours. Each colony was then picked and seeded into a 96-well plate containing 300 µL of either 2% glucose SC -URA -LEU or 2% galactose SC -URA -LEU. At each time point (0h, 12h, 24h, 36h) cells were diluted in PBS and quantified by flow cytometry after gating on homogenous population, single cells and GFP-positive cells. The complete protocol was extensively described in (Poggi et al., 2020)(annex 2).

Time courses of DSB inductions

Cells were transformed using standard lithium-acetate protocol (Gietz et al., 1995) with both guide and nuclease and selected on 2% glucose SC -URA -LEU plates and grown for 36 hours. Each colony was seeded into 2 mL of 2% glucose SC -URA -LEU for 24 hours and then diluted

into 10 mL of 2% glucose SC -URA -LEU for 24 hours as a pre-culture step. Cells were washed twice in water and diluted at ca. 7 x 10^6 cells/mL in 2% galactose SC-URA -LEU, before being harvested at each time point (0h, 2h, 4h, 6h, 8h, 10h, 12h) for subsequent DNA extractions.

Southern blot analyses

For each Southern blot, 3-5 µg of genomic DNA digested with *Eco* RV and *Ssp* I were loaded on a 1% agarose gel and electrophoresis was performed overnight at 1V/cm. The gel was manually transferred overnight in 20X SSC, on a Hybond-XL nylon membrane (GE Healthcare), according to manufacturer recommendations. Hybridization was performed with a ^{32}P-randomly labeled *CAN1* probe amplified from primers CAN133 and CAN135 (Supplemental Table S2) (Viterbo et al., 2018). The membrane was exposed 3 days on a phosphor screen and quantifications were performed on a FujiFilm FLA-9000 phosphorimager, using the Multi Gauge (v. 3.0) software. Percentages of DSB and recombinant molecules were calculated as the amount of each corresponding band divided by the total amount of signal in the lane.

Agarose plug DNA preparation

During time courses of DSB induction (see above), 2 x 10^9 cells were collected at each time point and centrifuged. Each pellet was resuspended in 330 µL 50 mM EDTA (pH 9.0), taking into account the pellet volume. Under a chemical hood, 110 µL of Solution I (1 M sorbitol, 10 mM EDTA (pH 9.0), 100 mM sodium citrate (pH 5.8), 2.5% β-mercaptoethanol and 10 µL of 100 mg/mL Zymolyase 100T-Seikagaku) were added to the cells, before 560 mL of 1% InCert agarose (Lonza) were delicately added and mixed. This mix was rapidly poured into plug molds and left in the cold room for at least 10 minutes. When solidified, agarose plugs were removed from the molds and incubated overnight at 37°C in Solution II (450 mM EDTA (pH 9.0), 10

mM Tris-HCl (pH 8.0), 7.5% β-mercaptoethanol). In the morning, tubes were cooled down on ice before Solution II was delicately removed with a pipette and replaced by Solution III (450 mM EDTA (pH 9.0), 10 mM Tris-HCl (pH 8.0), 1% N-lauryl sarcosyl, 1 mg/mL Proteinase K). Tubes were incubated overnight at 65°C, before being cooled down on ice in the morning. Solution III was removed and replaced by TE (10 mM Tris pH 8.0, 1 mM EDTA). Blocks were incubated in 1 mL TE in 2 mL eppendorf tubes for one hour at 4°C, repeated four times. TE was replaced by 1 mL restriction enzyme buffer (Invitrogen REACT 2) for one hour, then replaced by 100 µl buffer containing 100 units of each enzyme (*Eco* RV and *Ssp* I) and left overnight at 37°C. Agarose was melted at 70°C for 10 minutes without removing the buffer, 100 units of each enzyme *Eco* RV and *Ssp* I was added and left at 37°C for one hour. Then, 2 µL of β-agarase (NEB M0392S) and 2 µL of RNAse A (Roche 1 119 915) were added and left for one hour at 37°C. Microtubes were centrifuged at maximum speed in a tabletop centrifuge for one minute to pellet undigested agarose. The liquid phase was collected with wide bore 200 µL filter tips (Fisher Scientific #2069G) and loaded on a 1% agarose gel, subsequently processed as for a regular Southern blot (see above).

Northern blot analyses

Each repeat-containing strain transformed with its cognate gRNA and nucleases was grown for 4 hours in 2% galactose SC-URA-LEU. Total RNAs were extracted using standard phenol-chloroform procedure (Richard et al., 1997) or the miRVANA kit, used to extract very low levels of small RNAs with high efficacy (ThermoFisher). Total RNA samples were loaded on 50% urea 10% polyacrylamide gels and run at 20 W for one hour. Gels were electroblotted on N^+ nylon membranes (GE Healthcare), hybridized at 42°C using a SpCas9, SaCas9, FnCpf1 or *SNR44* oligonucleotidic probe. Each probe was terminally labeled with γ-^{32}P ATP in the presence of polynucleotide kinase.

Western blot analyses

Total proteins were extracted in 2X Laemmli buffer and denatured at 95°C before being loaded on a 12% polyacrylamide gel. After migration, the gel was electroblotted (0.22 A, constant voltage) on a Nytran membrane (Whatman), blocked for one hour in 3% NFDM/TBS-T and hybridized using either anti-SpCas9 (ab202580, dilution 1/1000), anti-SaCas9 (ab203936, dilution 1/1000), anti-HA (ab9110, dilution 1/1000) or anti-ZWF1 (A9521, dilution 1/100 000) overnight. Membranes were washed in TBS-T for 10 minutes twice. Following anti-SaCas9 and anti-HA hybridization, a secondary hybridization using secondary antibody Goat anti-Rabbit 31460 (dilution 1/5000). Membranes were read and quantified on a Bio-Rad ChemiDoc apparatus.

Analysis of DSB end resection

A real-time PCR assay using primer pairs flanking *Sty* I sites 282 bp away from 5' end of the repeat sequence and 478 bp away from the 3′ end of the repeat tract (LP001/LP002 and LP003/LP004, respectively) was used to quantify end resection. Another pair of primers was used to amplify a region of chromosome X to serve as an internal control of the DNA amount (JEM1f-JEM1r). Genomic DNA of cells collected at t = 12h was split in two fractions; one was used for *Sty* I digestion and the other one for a mock digestion in a final volume of 15 µL. Samples were incubated for 5h at 37°C and then the enzyme was inactivated for 20 min at 65°C. DNA was subsequently diluted by adding 55 µL of ice-cold water, and 4 µL was used for each real-time PCR reaction in a final volume of 25 µL. PCRs were performed with EurobioProbe qPCR Mix Lo-ROX in a CFX96 Real time machine (Bio-Rad) using the following program: 95°C for 15 min, 95°C for 15 s, 55°C for 30 s, and 72°C for 30 s repeated 40 times, followed by a 20-min melting curve. Reactions were performed in triplicate, and the mean value was

used to determine the amount of resected DNA using the following formula: raw resection = $2/(1+2\Delta Ct)$ with $\Delta C_t = C_{t,StyI} - C_{t,mock}$. Relative resection values were calculated by dividing raw resection values by the percentage of DSB quantified at the corresponding time point (Chen et al., 2013). Ratios of relative resection rates from both sides of the repeated sequence were calculated and compared to a non-repeated control sequence.

Determination of off-target mutations

Cell were grown overnight in YPGal medium and diluted for 2 more hours. Cells were incubated in 20ml 0.1M of Lithium Acetate/TE buffer for 45 minutes at 30°C. 500 µl of 1M DTT was added and cells were incubated for a further 15 minutes at the same temperature. Cells were washed in water, then in 1M Sorbitol and resuspended in 120 µl ice-cold 1M Sorbitol. Then, 40 µl of competent cells were mixed with 150 ng of guide RNA-expressing plasmid, 300 ng of nuclease-expressing plasmid and 100 µM of dsODN (5′-P G*T*TAATTGAGTTGTCATATGTTAATAACGGT*A*T-3’; where P represents a 5′ phosphorylation and * indicates a phosphorothioate linkage). Cells were electroporated at 1.5 kV, 25 µF, 200 Ω. Right after electroporation (BioRad Micropulser), 1 ml 1M Sorbitol was added to the mixture. Cells were centrifuged and supernatant was removed to plate a volume of 200 µl on 2% galactose SC -URA -LEU plates and grown for 72 hours. Negative control consisted of the same procedure without dsODN. Genomic DNA was extracted and approximately 10 µg of total genomic DNA was extracted and sonicated to an average size of 500 bp, on a Covaris S220 (LGC Genomics) in microtubes AFA (6x16 mm) using the following setup: Peak Incident Power: 105 Watts, Duty Factor: 5%, 200 cycles, 80 seconds. DNA ends were subsequently repaired with T4 DNA polymerase (15 units, NEBiolabs) and Klenow DNA polymerase (5 units, NEBiolabs) and phosphorylated with T4 DNA kinase (50 units, NEBiolabs). Repaired DNA was purified on two MinElute columns (Qiagen) and eluted in 16

µl (32 µl final for each library). Addition of a 3' dATP was performed with Klenow DNA polymerase (exo-) (15 units, NEBiolabs). Home-made adapters containing a 4-bp unique tag used for multiplexing, were ligated with 2 µl T4 DNA ligase (NEBiolabs, 400,000 units/ml). DNA was size fractionated on 1% agarose gels and 500-750 bp DNA fragments were gel extracted with the Qiaquick gel extraction kit (Qiagen). A first round of PCR was performed using primers GSP1 and P1 (First denaturation step: 98°C for 30s; 98°C for 30s, 50°C for 30s, 72°C for 30s repeated 30 times; followed by 72°C for 7min). A second round of PCR was performed using primers PE1 and GSP2-PE2 (First denaturation step: 98°C for 30s; 98°C for 30s, 65°C for 30s, 72°C for 30s repeated 30 times; followed by 72°C for 7min) A final round of PCR was performed using primers PE1 and PE2 (First denaturation step: 98°C for 30s; 98°C for 30s, 65°C for 30s, 72°C for 30s repeated 15 times; followed by 72°C for 7min). Libraries were purified on agarose gel and quantified on a Bioanalyzer. Equimolar amounts of each library were loaded on a Next-Seq Mid output flow cell cartridge (Illumina NextSeq 500/550 #20022409).

Computer analysis of off-target mutations

In a first step, all fastq originating from the different libraries were scanned in order to identify reads coming from the dsODN specific amplification. The test was carried out with the standard unix command grep and the result was used to split each former fastq file in two: with or without the dsODN tag. In a second step, all the resulting fastq files were mapped against the S288C reference genome obtained from the SGD database (release R64-2-1_20150113, https://www.yeastgenome.org/). Mapping was carried out by minimap2 (Li, 2018) using "-ax sr --secondary=no" parameters. Sam files resulting from mappings were then all sorted and indexed by the samtools software suite (Li et al., 2009). Subsequently, for dsODN-containing sequences, double strand break positions were identified by searching coverage peaks. Peaks

were defined as region showing a coverage at least equal to twice the median coverage. Regarding reads that did not contain the dsODN tag, mutations within predicted off-target sites were detected by the mean of samtools pileup applied to all regions of interest identified by crispor (Haeussler et al., 2016). Each of the 56 positions exhibiting mutations was manually examined using the IGV visualization software and validated or not, as explained in the text.

Analysis of Gibbs free energy for gRNA guide sequences

The Gibbs free energy formation for gRNA secondary structures were determined using the MFOLD RNA 2.3 (http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form2.3) with temperature parameter set to 30°C.

Statistical Analysis

All statistical tests were performed with R3.5.1. Linear regression model was performed to test the correlation between DSB value and the percentage of GFP-positive cells at different time points. Linear regression was performed to determine statistical significance of proteins levels and gRNA levels over the percentage of GFP-positive cells. For each linear regression, R^2 and p-value were calculated. One-way analysis of variance (ANOVA) was used to determine the impact of gRNA free energy over the percentage of GFP-positive cells at 36 h. P-values less than 0.05 were considered significant. Figures were plotted using the package ggplot2.

Data Access

Accession number of Illumina sequencing is PRJEB35597.

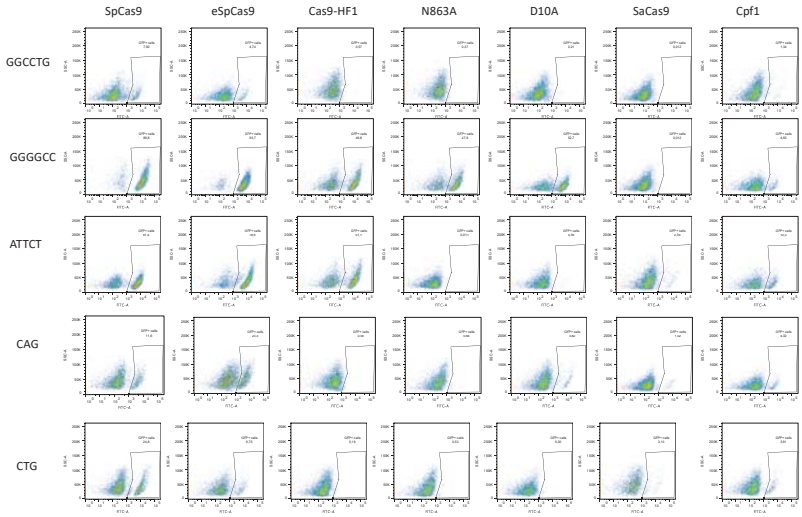
Acknowledgments

L. P. was supported by a CIFRE PhD fellowship from Sanofi. Off-target studies were supported by the AFM-Telethon. We thank Heloïse Muller for sharing her unpublished protocol for yeast transformation by electroporation, and Carine Giovannangeli for the generous gift of CRISPR-Cas plasmids. This work was supported by Sanofi, the Institut Pasteur and the Centre National de la Recherche Scientifique (CNRS).

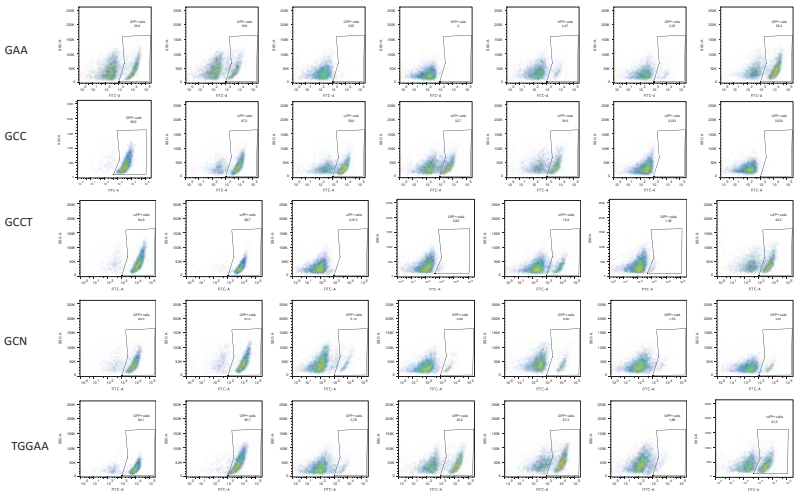
Disclosure declaration

None declared.

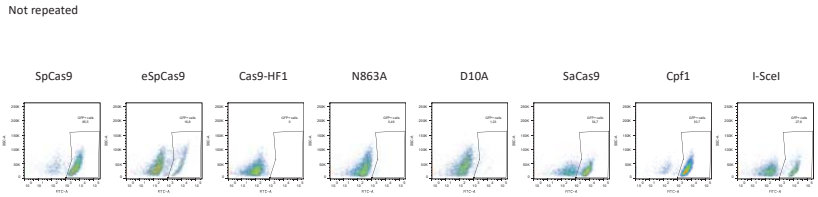
Supplemental figures



Poggi *et al.*
Figure S1

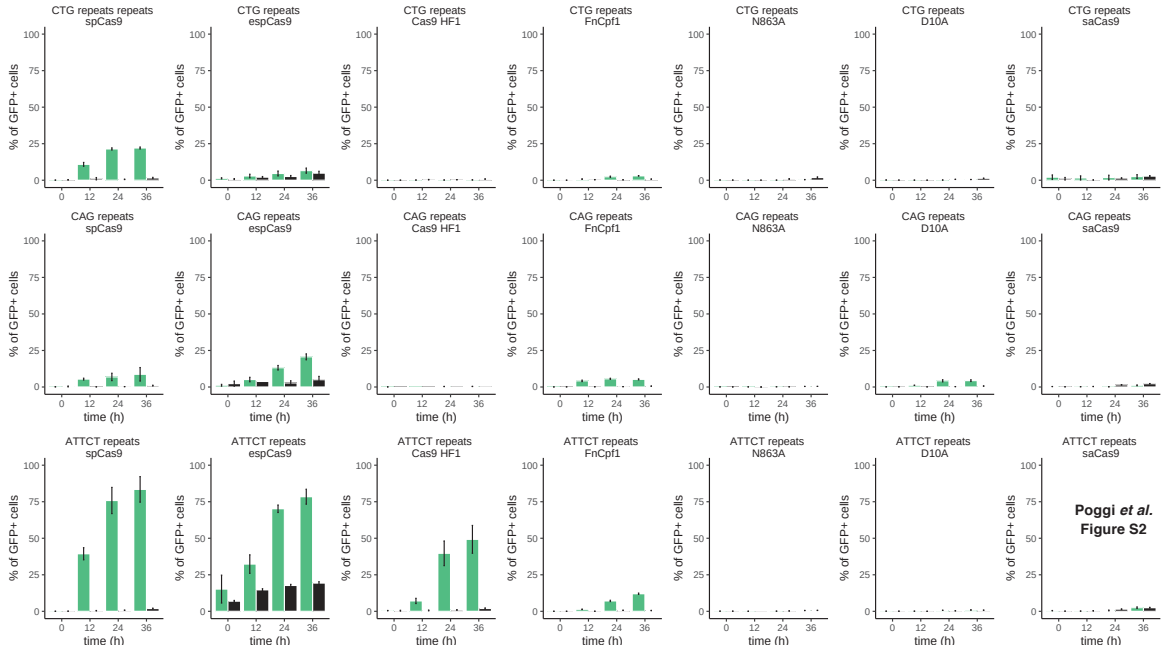


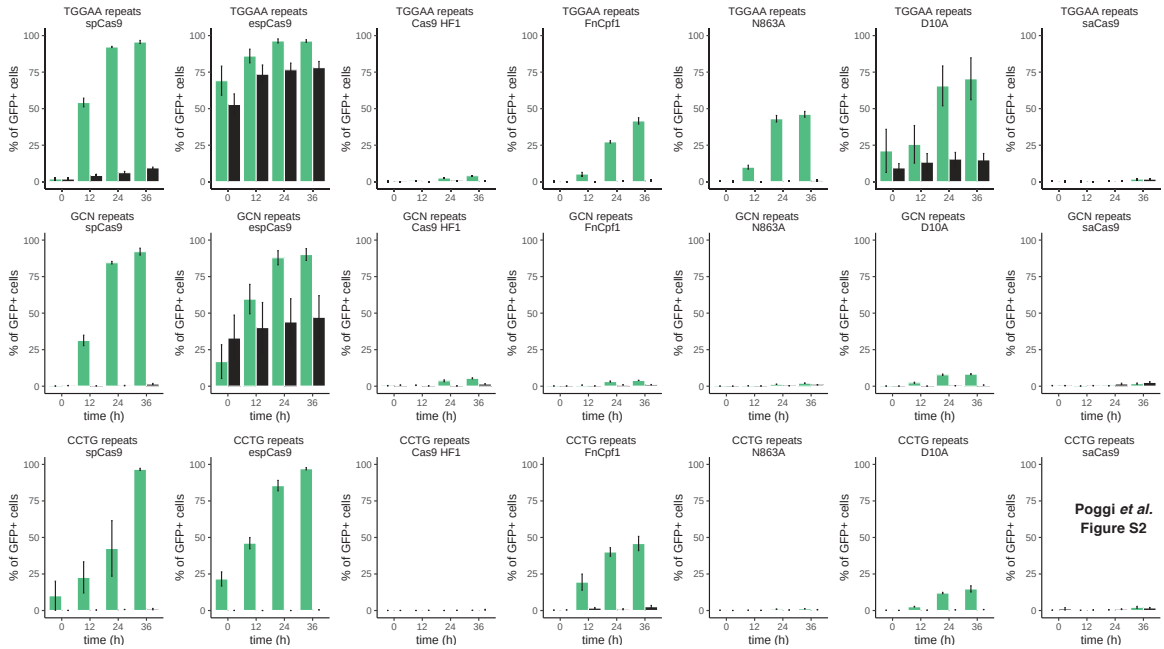
Poggi *et al.*
Figure S1

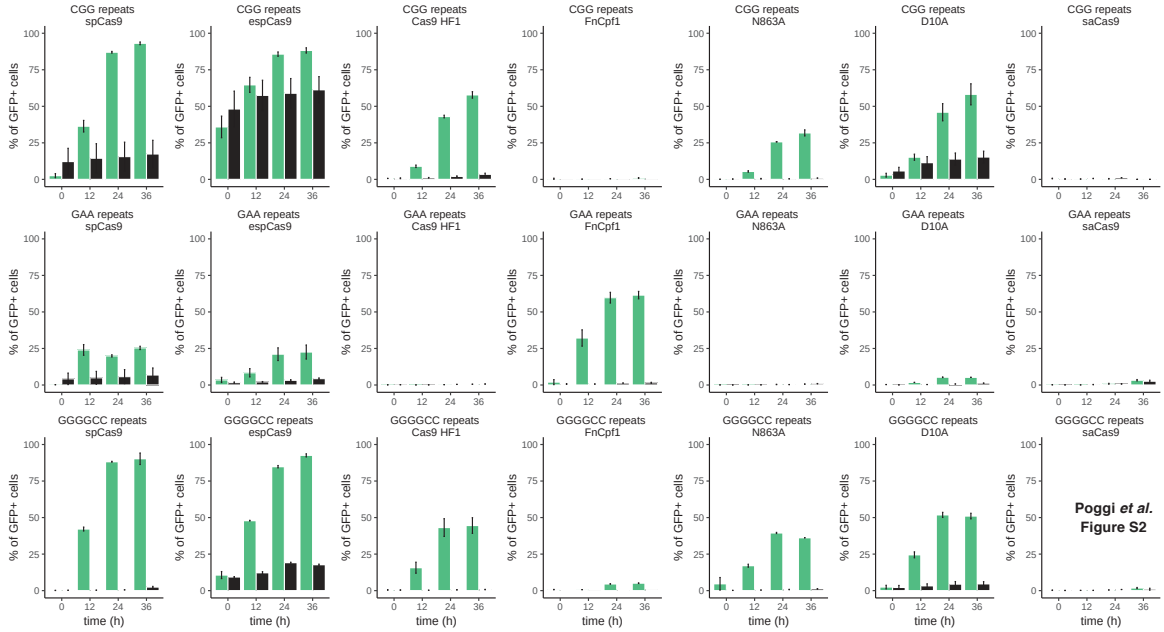


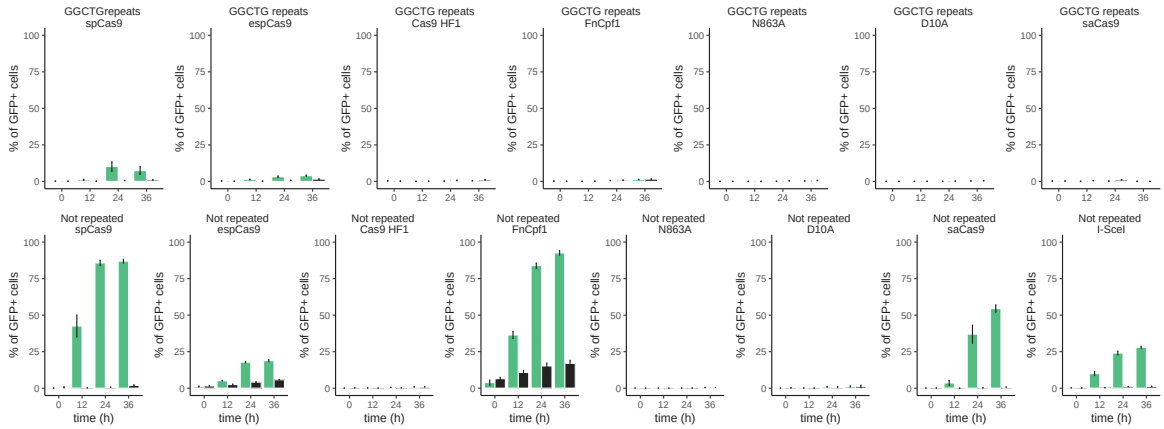
Poggi *et al.*
Figure S1

Supplemental Figure S1: Flow cytometry at 36 hours for each repeat. For each repeat (horizontal) the corresponding dot plot is shown for each nuclease (vertical). Gates are drawn to separate recombined from non-recombined populations.


CTG repeats repeats spCas9
CTG repeats espCas9
CTG repeats Cas9 HF1
CTG repeats FnCpf1
CTG repeats N863A
CTG repeats D10A
CTG repeats saCas9
CAG repeats spCas9
CAG repeats espCas9
CAG repeats Cas9 HF1
CAG repeats FnCpf1
CAG repeats N863A
CAG repeats D10A
CAG repeats saCas9
ATTCT repeats spCas9
ATTCT repeats espCas9
ATTCT repeats Cas9 HF1
ATTCT repeats FnCpf1
ATTCT repeats N863A
ATTCT repeats D10A
ATTCT repeats saCas9
% of GFP+ cells
time (h)
Poggi et al.
Figure S2

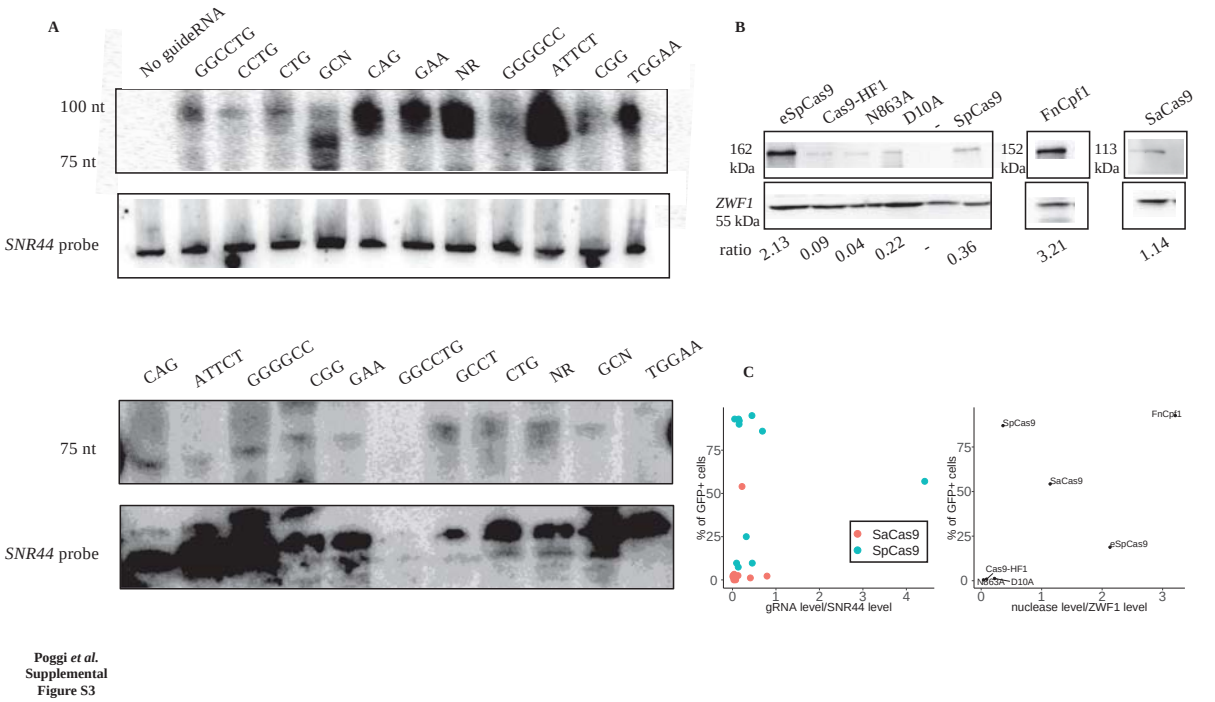

TGGAA repeats spCas9
TGGAA repeats espCas9
TGGAA repeats Cas9 HF1
TGGAA repeats FnCpf1
TGGAA repeats N863A
TGGAA repeats D10A
TGGAA repeats saCas9
GCN repeats spCas9
GCN repeats espCas9
GCN repeats Cas9 HF1
GCN repeats FnCpf1
GCN repeats N863A
GCN repeats D10A
GCN repeats saCas9
CCTG repeats spCas9
CCTG repeats espCas9
CCTG repeats Cas9 HF1
CCTG repeats FnCpf1
CCTG repeats N863A
CCTG repeats D10A
CCTG repeats saCas9
% of GFP+ cells
time (h)
Poggi et al.
Figure S2





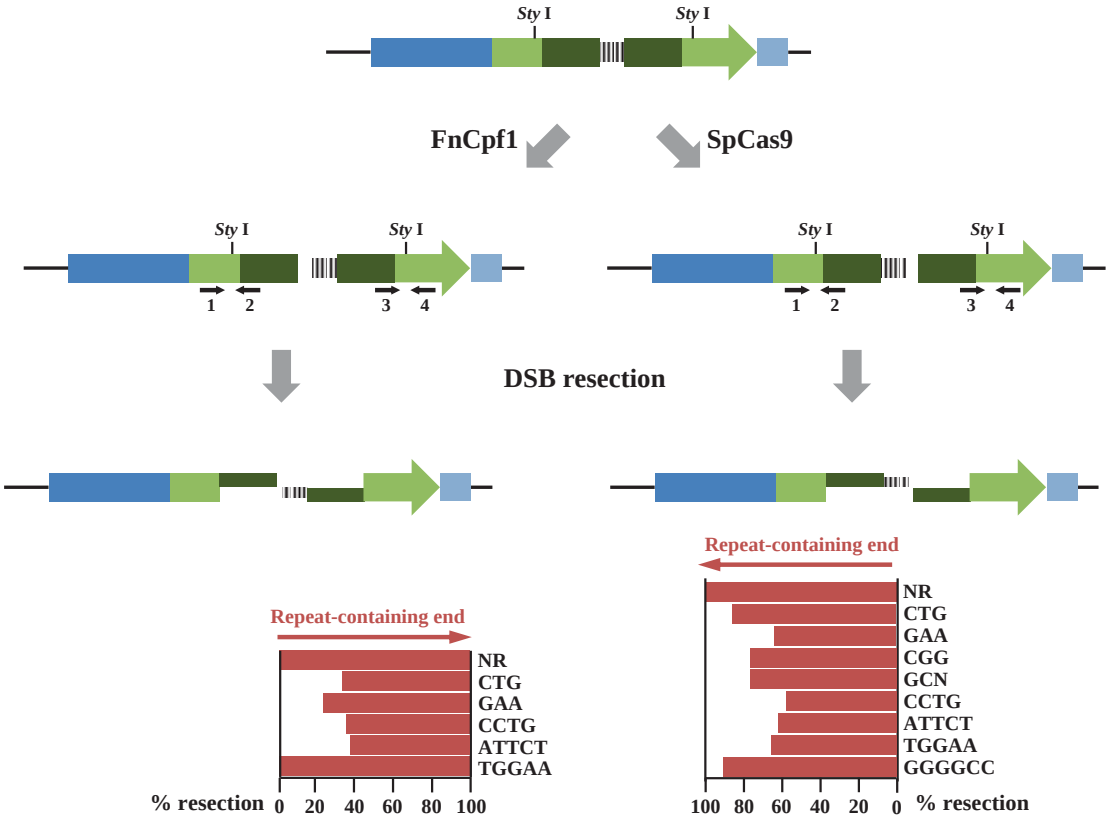
Poggi *et al.*
Figure S2

Supplemental Figure S2: Histogram of GFP-positive cell percentages for each repeat and each nuclease. Galactose condition is highlighted in green and glucose condition in black for each of the four time points. Each experiment was performed 3-8 times, depending on the strain. Error bars are standard errors.



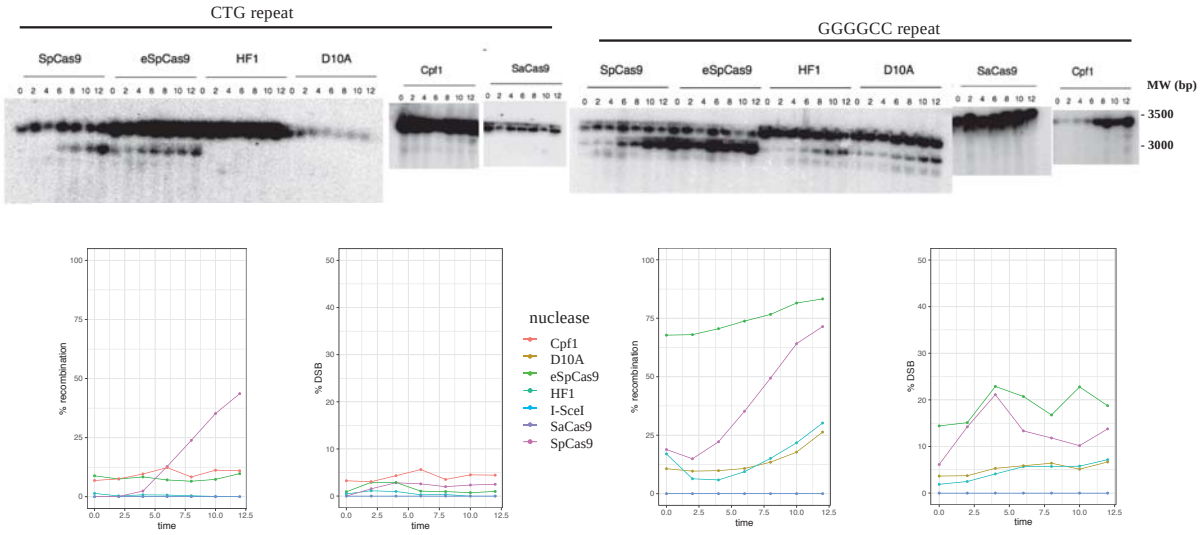
Poggi *et al.* Supplemental Figure S3

Supplemental Figure S3: gRNA and protein levels. A: gRNA expression levels measured by Northern blot. Top: Northern blot was hybridized with a SpCas9 gRNA scaffold probe. Bottom: Same for SaCas9. The same Northern blots were rehybridized with a control *SNR44* probe, corresponding to a snoRNA gene. **B**: Signals of both gRNA and *SNR44* were quantified and their ratios compared to nuclease efficacy measured by the percentage of GFP-positive cells at 36 hours. **C:** Nuclease expression levels measured by Western blot. Blots were successively hybridized with Cas-specific antibodies and Zwf1p antibody. Ratios of Cas/Zwf1 signals are shown below the blots. Left: GFP+ cell percentage as a function of gRNA levels. Right: GFP+ cell percentage as a function of protein levels.

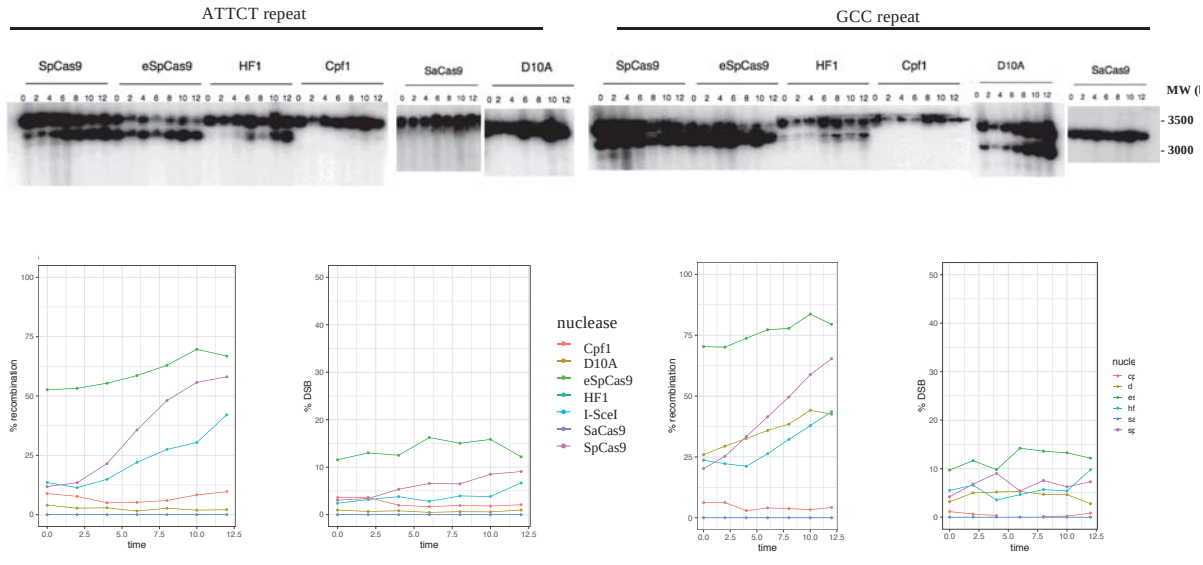


Poggi et al.
Supplemental Figure S4

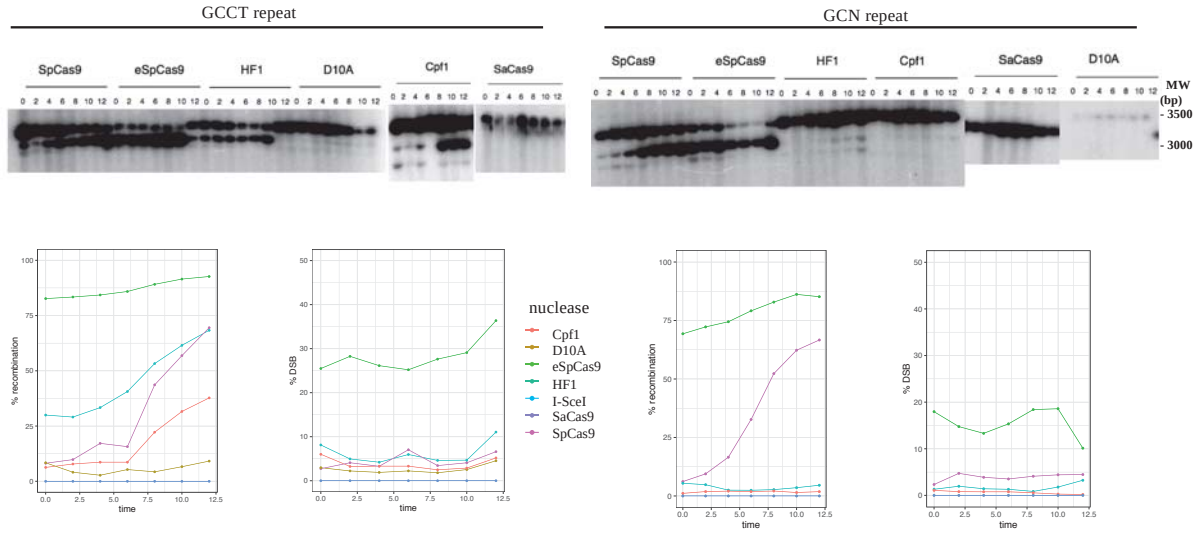
Supplemental Figure S4: Resection analysis of SpCas9 and FnCpf1 induced double-strand breaks. Sty I restriction sites located on each side of the DSB site are indicated. Only yeast strains for which a DSB was detectable at 12 hours post-nuclease induction were analyzed. Resection at NR was set at 100% to normalize the data. Lower resection values indicate resection inhibition.



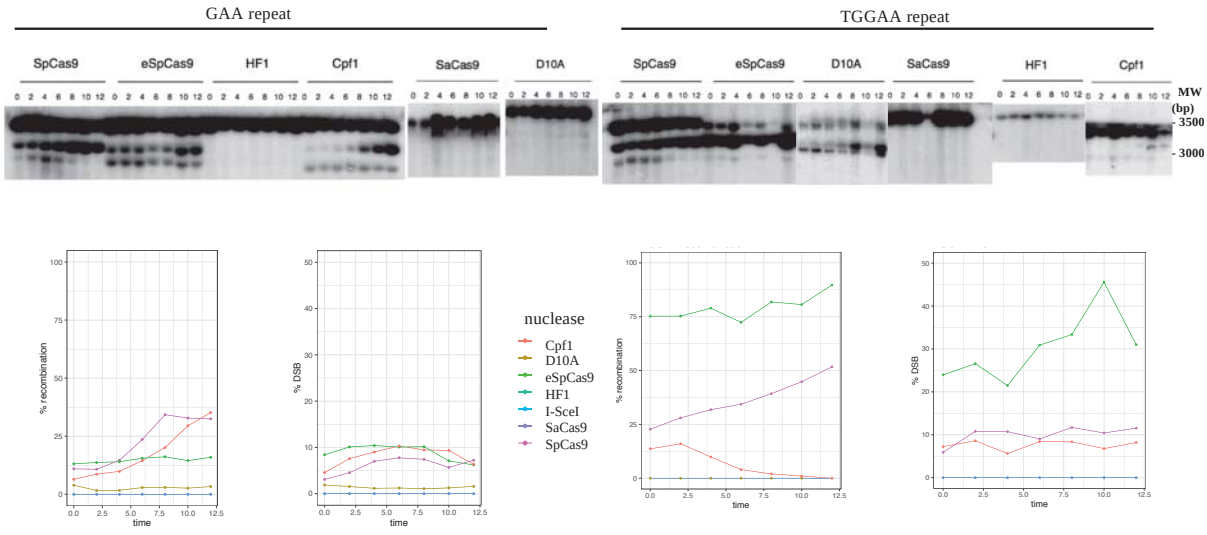
Poggi *et al.*
Supplemental
Figure S5



Poggi *et al.*
Supplemental
Figure S5



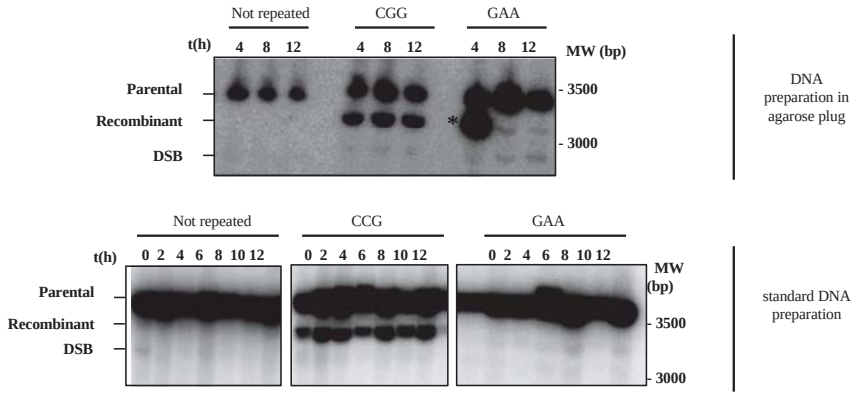
Poggi *et al.*
Supplemental Figure S5



Poggi *et al.*
Supplemental Figure S5

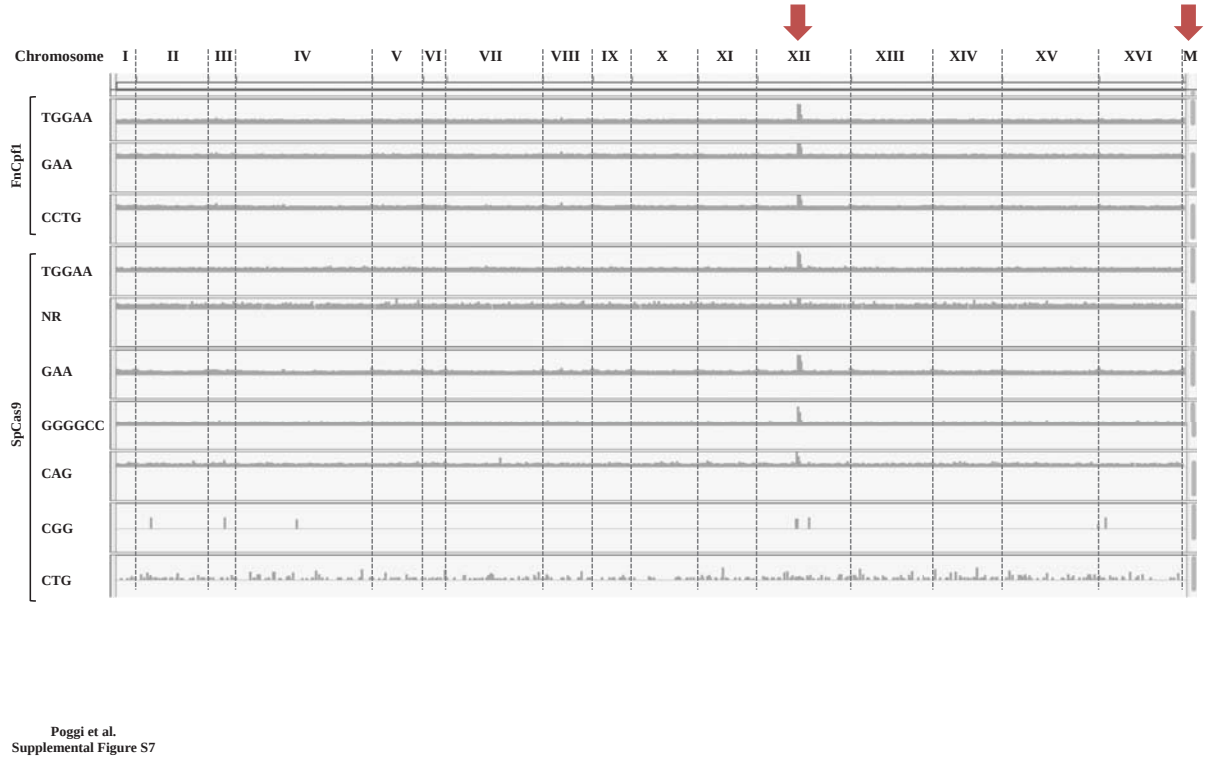
Supplemental Figure S5: Southern blots and quantifications. For each repeat, a Southern blot of an induction time course over 12 hours is shown above quantification graphs.

Recombination and DSB percentages were calculated as fractions of the total signal in each lane.



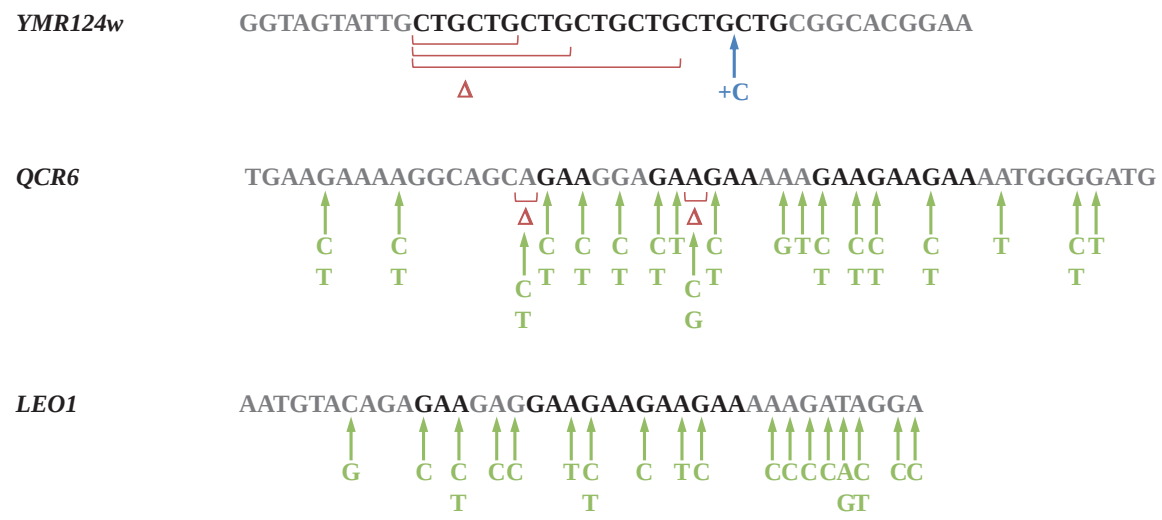
Poggi *et al.*
Supplemental Figure S6

Supplemental Figure S6: Evidence for *in vivo* DSBs generated by nickases. Top: Southern blot of DNA prepared in agarose plugs for three strains (Non-repeated, CGG and GAA repeats) after 4, 8 and 12 hours of nuclease induction. Parental, recombinant and DSB bands are indicated. DSBs are visible for CGG and GAA strains, increasing with time. The asterisk indicates a migration artefact commonly observed when overloading DNA in a lane. Bottom: The same experiment was performed but DNA was prepared by the standard protocol. The DSB was faintly detected at similar time points. Note that for the control non-repeated strain, some DSB signal is detected at the T0 time point only, suggesting that it could correspond to mechanically broken DNA molecules.



Poggi et al.
Supplemental Figure S7

Supplemental Figure S7: dsODN genome coverage. Each horizontal line represents a whole genome, in which each chromosome is separated by a dashed line and identified by a roman number. Red arrows point to regions enriched for the dsODN tag (rDNA and mitochondrial DNA).



Poggi et al.
Supplemental Figure S8

Supplemental Figure S8: Mutations identified in each of the three off-target sequences. Deletions are indicated by a red Δ, insertion in blue, and base substitutions in green. Triplets different from the microsatellite consensus are grey.

Supplemental Table S1: list of plasmids used in this study

Plasmid no.	Description	Reference
pRS416	URA3 selection marker	(Sikorski and Hieter, 1989)
pLPX210	Sp-Cas9 guide CTG cloned at SpeI-XhoI in pRS416	This study
pLPX209	Sp-Cas9 guide CAG cloned at SpeI-XhoI in pRS416	This study
pLPX206	Sp-Cas9 guide TGGAA cloned at SpeI-XhoI in pRS416	This study
pLPX203	Sp-Cas9 guide GCN cloned at SpeI-XhoI in pRS416	This study
pLPX208	Sp-Cas9 guide CCTG cloned at SpeI-XhoI in pRS416	This study
pLPX204	Sp-Cas9 guide CGG cloned at SpeI-XhoI in pRS416	This study

pLPX202	Sp-Cas9 guide GAA cloned at SpeI-XhoI in pRS416	This study
pLPX207	Sp-Cas9 guide ATTCT cloned at SpeI-XhoI in pRS416	This study
pLPX201	Sp-Cas9 guide GGGGCC cloned at SpeI-XhoI in pRS416	This study
pLPX205	Sp-Cas9 guide GGCCTG cloned at SpeI-XhoI in pRS416	This study
pLPX211	Sp-Cas9 guide I-SceI cloned at SpeI-XhoI in pRS416	This study
pLPX310	Cpf1 guide CTG cloned at SpeI-XhoI in pRS416	This study
pLPX309	Cpf1 guide CAG cloned at SpeI-XhoI in pRS416	This study
pLPX306	Cpf1guide TGGAA cloned at SpeI-XhoI in pRS416	This study
pLPX303	Cpf1guide GCN cloned at SpeI-XhoI in pRS416	This study
pLPX308	Cpf1guide CCTG cloned at SpeI-XhoI in pRS416	This study
pLPX304	Cpf1guide CGG cloned at SpeI-XhoI in pRS416	This study
pLPX302	Cpf1guide GAA cloned at SpeI-XhoI in pRS416	This study
pLPX307	Cpf1guide ATTCT cloned at SpeI-XhoI in pRS416	This study
pLPX301	Cpf1guide GGGGCC cloned at SpeI-XhoI in pRS416	This study
pLPX305	Cpf1guide GGCCTG cloned at SpeI-XhoI in pRS416	This study
pLPX311	Cpf1guide I-SceI cloned at SpeI-XhoI in pRS416	This study
pLPX410	Sa-Cas9 guide CTG cloned at SpeI-XhoI in pRS416	This study
pLPX409	Sa-Cas9 guide CAG cloned at SpeI-XhoI in pRS416	This study
pLPX406	Sa-Cas9 guide TGGAA cloned at SpeI-XhoI in pRS416	This study
pLPX403	Sa-Cas9 guide GCN cloned at SpeI-XhoI in pRS416	This study
pLPX408	Sa-Cas9 guide CCTG cloned at SpeI-XhoI in pRS416	This study
pLPX404	Sa-Cas9 guide CGG cloned at SpeI-XhoI in pRS416	This study
pLPX402	Sa-Cas9 guide GAA cloned at SpeI-XhoI in pRS416	This study
pLPX407	Sa-Cas9 guide ATTCT cloned at SpeI-XhoI in pRS416	This study
pLPX401	Sa-Cas9 guide GGGGCC cloned at SpeI-XhoI in pRS416	This study
pLPX405	Sa-Cas9 guide GGCCTG cloned at SpeI-XhoI in pRS416	This study
pLPX411	Sa-Cas9 guide I-SceI cloned at SpeI-XhoI in pRS416	This study
Addgene #43804	p415-GalL-Cas9-CYC1t LEU2 selection marker	(DiCarlo et al., 2013)
Addgene #72247	Cas9-HF1	(Kleinstiver et al., 2016b)

Addgene #71814	eSpCas9	(Slaymaker et al., 2016)
Addgene #68706	Cas9-N863A	(Nishimasu et al., 2014)
Addgene #48873	Cas9-D10A	(Ran et al., 2013)
Addgene #84039	SaCas9	(Ran et al., 2015b)
Addgene #69976	FnCpf1	(Zetsche et al., 2015)
pTRI103	I-SceI - NLS	(Richard et al., 2003)
pLPX10	pGAL-Cas9-HF1	This study
pLPX11	pGAL-eSpCas9	This study
pLPX12	pGAL-N863A	This study
pLPX13	pGAL-D10A	This study
pLPX14	pGAL-SaCas9	This study
pLPX15	pGAL-FnCpf1	This study
pLPX16	pGAL-I-SceI	This study
synYEGFP	pUC57 backone containing bipartite EGFP gene	This study
pLPX110	bipartite EGFP gene + CTG repeat (100bp)	This study
pLPX109	bipartite EGFP gene + CAG repeat (100bp)	This study
pLPX106	bipartite EGFP gene + TGGGAA repeat (100bp)	This study
pLPX103	bipartite EGFP gene + GCN repeat (100bp)	This study
pLPX108	bipartite EGFP gene + CGG repeat (100bp)	This study
pLPX104	bipartite EGFP gene + CCTG repeat (100bp)	This study
pLPX102	bipartite EGFP gene + ATTCT repeat (100bp)	This study
pLPX107	bipartite EGFP gene + GGGGCC repeat (100bp)	This study
pLPX101	bipartite EGFP gene + GGCCTG repeat (100bp)	This study
pLPX105	bipartite EGFP gene + GAA repeat (100bp)	This study

Supplemental Table S2: list and sequence of primers used in this study

Name	Oligonucleotide sequence (5' – 3')	Use
JEM1f	TGTGATTTGGCTGAGTTACAACG	qPCR assay
JEM1r	AACTGCCCAGCGATCCATT	qPCR assay
LP001	AGTAGTGACTAAGGTTGGCC	qPCR assay
LP002	GGTGAAGGTGATGCTACTTAC	qPCR assay
LP003	GGGGCCTGTTTATTTGTACAAT	qPCR assay
LP004	GATCCAAACGAAAAGAGAGACC	qPCR assay
LP400	CCCCGGATTCTAGAACTAGTGGATCCCCCGGGaaaaaaATGGACTATAAGGACCACGACG	eSpCas9 cloning into pGAL plasmid
LP401	TAAGCGTGACATAACTAATTACATGACTCGAGAAGAGATTACTTTTTCTTTTTTGCCTGGCC	
LP402	ACCCCGGATTCTAGAACTAGTGGATCCCCCGGGaaaaaaGCCGCCACCATGGATAAAAAG	Cas9-HF1 cloning into pGAL plasmid
LP403	TGTAAGCGTGACATAACTAATTACATGACTCGAGAAGAGAGTCATCCTGCAGCCTTGTCA	
LP408	AACCCCGGATTCTAGAACTAGTGGATCCCCCGGGaaaaaaatgagcatctaccag	FnCpf1 cloning into pGAL plasmid
LP407	TGTAAGCGTGACATAACTAATTACATGACTCGAGAAGAGAatgagcatctaccag	
LP410	AACCCCGGATTCTAGAACTAGTGGATCCCCCGGGaaaaaacaccatggccccaaagaagaa	N863A cloning into pGAL plasmid
LP411	TGTAAGCGTGACATAACTAATTACATGACTCGAGAAGAGATTActtgtcatcgtcatccttttttctttttgcctggcc	
LP412	AACCCCGGATTCTAGAACTAGTGGATCCCCCGGGaaaaaaCcatggactataaggaccacg	D10A cloning into pGAL plasmid

LP413	TGTAAGCGTGACATAACTAATTACATGACTCGAGAAGAGAttcttacttttttcttttttgcc	
LP417	AACCCCGGATTCTAGAACTAGTGGATCCCCCGGGaaaaaaGGCTGCAGATGCCTCCAAAAAAG	I-SceI cloning into pGAL plasmid
LP418	TGTAAGCGTGACATAACTAATTACATGACTCGAGAAGAGACCCCTCGACTTATTATTTCA	
LP419	AACCCCGGATTCTAGAACTAGTGGATCCCCCGGGATGGCCCCAAAGAAGAAGCG	SaCas9 cloning into pGAL plasmid
LP422	TGTAAGCGTGACATAACTAATTACATGACTCGAGAAGAGAttacttttttcttttttgc	
CAN133	ACATTTCCACGCCATTTCGC	CAN1 probe
CAN135	GGTTCTAGGTTCGGGTGACG	
SNR44p	GATAACGGACTAGCCTTATTTT	SNR44 probe
Spgrnap	GATAACGGACTAGCCTTATTTT	SpCas9 gRNA probe
Sagrnap	Tgccttgttttagtagattctg	SaCas9 gRNA probe
LP30b	TTCCATGGCCAAC	Verification of cassette integration
LP33b	ATGGCTGACAAACAAAA	
P1	ACACTCTTTCCCTACACGA	Illumina libraries amplification
GSP1	ATACCGTTATTAACATATGACA	
PE1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT	

GSP2-PE2	CAAGCAGAAGACGGCATACTACCGTTATTAACATATGACAACTCAA	
PE2	CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT	

Supplemental Table S3: list of strains used in this study

Name	**Genotype**	**Origin**
FYBL1-4D	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202	FYBL1 spore
LPY110	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(CTG)_{33}$-*TRP1*	This study
LPY109	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(CAG)_{33}$-*TRP1*	This study
LPY106	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(TGGGAA)_{20}$-*TRP1*	This study
LPY111	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP(I-SceI site)-*TRP1*	This study
LPY103	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(GCN)_{33}$-*TRP1*	This study
LPY108	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(CCTG)_{25}$-*TRP1*	This study
LPY104	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(CGG)_{33}$-*TRP1*	This study
LPY102	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(GAA)_{33}$-*TRP1*	This study
LPY107	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(ATTCT)_{20}$-*TRP1*	This study
LPY101	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(GGGGCC)_{17}$-*TRP1*	This study
LPY105	*MAT*a *ura3*Δ851 *trp*Δ63 *leu2*Δ1 *his3*Δ200 *lys2*Δ202 *can1*Δ:: yEGFP$(GGCCTG)_{17}$-*TRP1*	This study

Supplemental Table S4. Summary of flow cytometry experiments quantifications.

Repeat	Nuclease	Number of replicas	Condition	time (h)	GFP+% cells	standard error	standard deviation	confidence interval at 95%
GGCCTG	Cpf1	3	Gal	0	0.29	0.19	0.11	0.47
GGCCTG	D10A	3	Gal	0	0.08	0.04	0.02	0.09
GGCCTG	eSpCas9	3	Gal	0	0.33	0.13	0.08	0.32
GGCCTG	HF1	3	Gal	0	0.38	0.28	0.16	0.70
GGCCTG	N863A	3	Gal	0	0.07	0.01	0.00	0.01
GGCCTG	SaCas9	3	Gal	0	0.29	0.24	0.14	0.59
GGCCTG	SpCas9	4	Gal	0	0.19	0.12	0.06	0.20
GGCCTG	Cpf1	3	Gal	12	0.09	0.01	0.00	0.02
GGCCTG	D10A	3	Gal	12	0.05	0.04	0.02	0.10
GGCCTG	eSpCas9	3	Gal	12	1.33	0.47	0.27	1.16
GGCCTG	HF1	3	Gal	12	0.09	0.03	0.02	0.07
GGCCTG	N863A	3	Gal	12	0.03	0.02	0.01	0.04
GGCCTG	SaCas9	3	Gal	12	0.10	0.04	0.02	0.10
GGCCTG	SpCas9	4	Gal	12	1.13	0.73	0.37	1.17
GGCCTG	Cpf1	3	Gal	24	0.60	0.17	0.10	0.42
GGCCTG	D10A	3	Gal	24	0.08	0.03	0.02	0.07
GGCCTG	eSpCas9	3	Gal	24	3.25	0.84	0.48	2.08
GGCCTG	HF1	3	Gal	24	0.20	0.08	0.05	0.21
GGCCTG	N863A	3	Gal	24	0.09	0.05	0.03	0.13
GGCCTG	SaCas9	3	Gal	24	0.25	0.18	0.11	0.46
GGCCTG	SpCas9	4	Gal	24	10.19	6.34	3.17	10.09
GGCCTG	Cpf1	3	Gal	36	1.30	0.22	0.12	0.54
GGCCTG	D10A	3	Gal	36	0.33	0.18	0.10	0.44
GGCCTG	eSpCas9	3	Gal	36	3.95	0.61	0.35	1.52
GGCCTG	HF1	3	Gal	36	0.45	0.22	0.13	0.55
GGCCTG	N863A	3	Gal	36	0.34	0.12	0.07	0.30
GGCCTG	SaCas9	3	Gal	36	0.04	0.06	0.04	0.16
GGCCTG	SpCas9	4	Gal	36	7.44	5.23	2.61	8.32
GGCCTG	Cpf1	3	Glu	0	0.12	0.03	0.02	0.07
GGCCTG	D10A	3	Glu	0	0.08	0.01	0.01	0.03
GGCCTG	eSpCas9	3	Glu	0	0.14	0.05	0.03	0.12
GGCCTG	HF1	3	Glu	0	0.23	0.15	0.09	0.37

GGCCTG	N863A	3	Glu	0	0.02	0.01	0.01	0.03
GGCCTG	SaCas9	3	Glu	0	0.26	0.18	0.10	0.44
GGCCTG	SpCas9	3	Glu	0	0.17	0.06	0.03	0.14
GGCCTG	Cpf1	3	Glu	12	0.16	0.03	0.02	0.08
GGCCTG	D10A	3	Glu	12	0.05	0.05	0.03	0.13
GGCCTG	eSpCas9	3	Glu	12	0.19	0.01	0.01	0.03
GGCCTG	HF1	3	Glu	12	0.15	0.01	0.01	0.03
GGCCTG	N863A	3	Glu	12	0.04	0.03	0.02	0.06
GGCCTG	SaCas9	3	Glu	12	0.56	0.19	0.11	0.47
GGCCTG	SpCas9	3	Glu	12	0.17	0.05	0.03	0.12
GGCCTG	Cpf1	3	Glu	24	0.88	0.06	0.03	0.14
GGCCTG	D10A	3	Glu	24	0.23	0.09	0.05	0.22
GGCCTG	eSpCas9	3	Glu	24	0.76	0.13	0.07	0.31
GGCCTG	HF1	3	Glu	24	0.69	0.30	0.17	0.74
GGCCTG	N863A	3	Glu	24	0.39	0.14	0.08	0.36
GGCCTG	SaCas9	3	Glu	24	1.05	0.31	0.18	0.76
GGCCTG	SpCas9	3	Glu	24	0.64	0.23	0.13	0.57
GGCCTG	Cpf1	3	Glu	36	1.57	0.11	0.06	0.27
GGCCTG	D10A	3	Glu	36	0.46	0.09	0.05	0.21
GGCCTG	eSpCas9	3	Glu	36	1.59	0.19	0.11	0.48
GGCCTG	HF1	3	Glu	36	1.01	0.18	0.10	0.44
GGCCTG	N863A	3	Glu	36	0.69	0.18	0.10	0.43
GGCCTG	SaCas9	3	Glu	36	0.00	0.00	0.00	0.00
GGCCTG	SpCas9	3	Glu	36	1.04	0.12	0.07	0.30
GGGGCC	Cpf1	2	Gal	0	0.70	0.18	0.13	1.65
GGGGCC	D10A	3	Gal	0	2.43	2.13	1.23	5.30
GGGGCC	eSpCas9	3	Gal	0	10.65	4.18	2.41	10.38
GGGGCC	HF1	4	Gal	0	0.47	0.38	0.19	0.61
GGGGCC	N863A	3	Gal	0	4.67	7.48	4.32	18.57
GGGGCC	SaCas9	3	Gal	0	0.27	0.26	0.15	0.64
GGGGCC	SpCas9	2	Gal	0	0.14	0.11	0.08	0.99
GGGGCC	Cpf1	2	Gal	12	0.69	0.03	0.02	0.25
GGGGCC	D10A	3	Gal	12	24.50	3.56	2.05	8.83
GGGGCC	eSpCas9	3	Gal	12	47.87	0.42	0.24	1.03
GGGGCC	HF1	3	Gal	12	15.68	6.57	3.80	16.33
GGGGCC	N863A	3	Gal	12	17.07	1.75	1.01	4.34
GGGGCC	SaCas9	3	Gal	12	0.14	0.03	0.02	0.08
GGGGCC	SpCas9	2	Gal	12	42.15	1.91	1.35	17.15
GGGGCC	Cpf1	2	Gal	24	4.57	0.25	0.18	2.22

GGGGCC	D10A	3	Gal	24	51.83	3.02	1.75	7.51
GGGGCC	eSpCas9	3	Gal	24	84.80	1.41	0.81	3.50
GGGGCC	HF1	4	Gal	24	43.18	12.16	6.08	19.35
GGGGCC	N863A	3	Gal	24	39.47	0.68	0.39	1.69
GGGGCC	SaCas9	3	Gal	24	0.45	0.10	0.05	0.24
GGGGCC	SpCas9	2	Gal	24	88.20	0.42	0.30	3.81
GGGGCC	Cpf1	2	Gal	36	5.10	0.32	0.23	2.86
GGGGCC	D10A	3	Gal	36	51.03	3.35	1.93	8.32
GGGGCC	eSpCas9	3	Gal	36	92.57	1.86	1.07	4.62
GGGGCC	HF1	4	Gal	36	44.55	10.74	5.37	17.09
GGGGCC	N863A	3	Gal	36	36.17	0.23	0.13	0.57
GGGGCC	SaCas9	3	Gal	36	1.78	0.66	0.38	1.63
GGGGCC	SpCas9	2	Gal	36	90.25	5.59	3.95	50.19
GGGGCC	Cpf1	1	Glu	0	0.21	NA	NA	NA
GGGGCC	D10A	3	Glu	0	2.19	2.37	1.37	5.88
GGGGCC	eSpCas9	3	Glu	0	9.33	0.44	0.25	1.09
GGGGCC	HF1	4	Glu	0	0.39	0.29	0.15	0.46
GGGGCC	N863A	3	Glu	0	0.15	0.07	0.04	0.16
GGGGCC	SaCas9	3	Glu	0	0.13	0.14	0.08	0.35
GGGGCC	SpCas9	2	Glu	0	0.24	0.05	0.04	0.44
GGGGCC	Cpf1	1	Glu	12	0.01	NA	NA	NA
GGGGCC	D10A	3	Glu	12	3.28	2.42	1.40	6.01
GGGGCC	eSpCas9	3	Glu	12	12.20	1.35	0.78	3.36
GGGGCC	HF1	3	Glu	12	0.26	0.05	0.03	0.13
GGGGCC	N863A	3	Glu	12	0.38	0.04	0.02	0.09
GGGGCC	SaCas9	3	Glu	12	0.15	0.07	0.04	0.17
GGGGCC	SpCas9	2	Glu	12	0.30	0.04	0.03	0.32
GGGGCC	Cpf1	1	Glu	24	0.50	NA	NA	NA
GGGGCC	D10A	3	Glu	24	4.45	3.05	1.76	7.58
GGGGCC	eSpCas9	3	Glu	24	19.13	0.55	0.32	1.37
GGGGCC	HF1	4	Glu	24	0.55	0.33	0.16	0.52
GGGGCC	N863A	3	Glu	24	0.76	0.23	0.13	0.57
GGGGCC	SaCas9	3	Glu	24	0.77	0.07	0.04	0.16
GGGGCC	SpCas9	2	Glu	24	0.55	0.05	0.04	0.44
GGGGCC	Cpf1	1	Glu	36	0.71	NA	NA	NA
GGGGCC	D10A	3	Glu	36	4.63	2.63	1.52	6.53
GGGGCC	eSpCas9	3	Glu	36	17.73	0.74	0.43	1.83
GGGGCC	HF1	4	Glu	36	0.74	0.27	0.14	0.43
GGGGCC	N863A	3	Glu	36	1.21	0.29	0.17	0.72

GGGGCC	SaCas9	3	Glu	36	0.97	1.16	0.67	2.87
GGGGCC	SpCas9	2	Glu	36	2.43	0.70	0.50	6.29
ATTCT	Cpf1	3	Gal	0	0.04	0.05	0.03	0.12
ATTCT	D10A	3	Gal	0	0.10	0.04	0.02	0.09
ATTCT	eSpCas9	3	Gal	0	15.21	16.55	9.55	41.10
ATTCT	HF1	4	Gal	0	0.46	0.26	0.13	0.42
ATTCT	N863A	3	Gal	0	0.11	0.08	0.04	0.19
ATTCT	SaCas9	3	Gal	0	0.45	0.19	0.11	0.47
ATTCT	SpCas9	3	Gal	0	0.05	0.03	0.02	0.07
ATTCT	Cpf1	3	Gal	12	1.38	0.31	0.18	0.77
ATTCT	D10A	3	Gal	12	0.10	0.05	0.03	0.12
ATTCT	eSpCas9	3	Gal	12	32.33	11.00	6.35	27.32
ATTCT	HF1	4	Gal	12	7.11	3.59	1.79	5.70
ATTCT	N863A	3	Gal	12	0.05	0.05	0.03	0.11
ATTCT	SaCas9	3	Gal	12	0.15	0.06	0.03	0.14
ATTCT	SpCas9	3	Gal	12	39.40	7.20	4.16	17.89
ATTCT	Cpf1	3	Gal	24	7.12	0.77	0.44	1.91
ATTCT	D10A	3	Gal	24	0.69	0.08	0.04	0.19
ATTCT	eSpCas9	3	Gal	24	70.20	4.45	2.57	11.06
ATTCT	HF1	4	Gal	24	39.70	16.90	8.45	26.90
ATTCT	N863A	3	Gal	24	0.12	0.09	0.05	0.21
ATTCT	SaCas9	3	Gal	24	0.61	0.47	0.27	1.18
ATTCT	SpCas9	3	Gal	24	75.83	15.55	8.98	38.62
ATTCT	Cpf1	3	Gal	36	12.03	0.81	0.47	2.01
ATTCT	D10A	3	Gal	36	0.91	0.31	0.18	0.77
ATTCT	eSpCas9	3	Gal	36	78.43	8.87	5.12	22.03
ATTCT	HF1	4	Gal	36	49.25	19.05	9.52	30.31
ATTCT	N863A	3	Gal	36	0.56	0.15	0.08	0.36
ATTCT	SaCas9	3	Gal	36	2.50	0.93	0.54	2.31
ATTCT	SpCas9	3	Gal	36	83.43	15.17	8.76	37.68
ATTCT	Cpf1	3	Glu	0	0.03	0.04	0.02	0.09
ATTCT	D10A	3	Glu	0	0.09	0.11	0.06	0.27
ATTCT	eSpCas9	3	Glu	0	6.93	0.93	0.54	2.32
ATTCT	HF1	4	Glu	0	0.44	0.56	0.28	0.89
ATTCT	N863A	3	Glu	0	0.10	0.05	0.03	0.12
ATTCT	SaCas9	3	Glu	0	0.12	0.03	0.02	0.08
ATTCT	SpCas9	7	Glu	0	0.10	0.04	0.02	0.04
ATTCT	Cpf1	3	Glu	12	0.06	0.08	0.05	0.20
ATTCT	D10A	3	Glu	12	0.03	0.02	0.01	0.05

ATTCT	eSpCas9	3	Glu	12	14.70	1.35	0.78	3.34
ATTCT	HF1	4	Glu	12	0.61	0.56	0.28	0.89
ATTCT	N863A	3	Glu	12	0.02	0.04	0.02	0.10
ATTCT	SaCas9	3	Glu	12	0.34	0.26	0.15	0.64
ATTCT	SpCas9	7	Glu	12	0.43	0.38	0.15	0.36
ATTCT	Cpf1	3	Glu	24	0.70	0.12	0.07	0.29
ATTCT	D10A	3	Glu	24	0.36	0.05	0.03	0.11
ATTCT	eSpCas9	3	Glu	24	17.70	1.28	0.74	3.17
ATTCT	HF1	4	Glu	24	0.93	0.44	0.22	0.70
ATTCT	N863A	3	Glu	24	0.26	0.12	0.07	0.29
ATTCT	SaCas9	3	Glu	24	1.47	0.36	0.21	0.89
ATTCT	SpCas9	7	Glu	24	0.75	0.11	0.04	0.10
ATTCT	Cpf1	3	Glu	36	0.53	0.20	0.12	0.50
ATTCT	D10A	3	Glu	36	0.78	0.14	0.08	0.35
ATTCT	eSpCas9	3	Glu	36	19.33	1.66	0.96	4.13
ATTCT	HF1	4	Glu	36	1.93	0.84	0.42	1.34
ATTCT	N863A	3	Glu	36	0.65	0.13	0.07	0.32
ATTCT	SaCas9	3	Glu	36	2.47	0.46	0.27	1.15
ATTCT	SpCas9	7	Glu	36	1.91	0.75	0.28	0.69
CAG	Cpf1	5	Gal	0	0.07	0.02	0.01	0.03
CAG	D10A	3	Gal	0	0.09	0.04	0.02	0.10
CAG	eSpCas9	7	Gal	0	1.20	1.61	0.61	1.49
CAG	HF1	3	Gal	0	0.07	0.04	0.03	0.11
CAG	N863A	3	Gal	0	0.09	0.04	0.02	0.09
CAG	SaCas9	3	Gal	0	0.32	0.06	0.03	0.14
CAG	SpCas9	4	Gal	0	0.13	0.13	0.07	0.21
CAG	Cpf1	5	Gal	12	4.22	1.11	0.50	1.38
CAG	D10A	3	Gal	12	1.17	0.03	0.02	0.08
CAG	eSpCas9	4	Gal	12	5.11	3.07	1.53	4.88
CAG	HF1	1	Gal	12	0.06	NA	NA	NA
CAG	N863A	3	Gal	12	0.10	0.03	0.02	0.08
CAG	SaCas9	3	Gal	12	0.20	0.08	0.05	0.20
CAG	SpCas9	3	Gal	12	5.41	1.09	0.63	2.71
CAG	Cpf1	5	Gal	24	5.61	0.85	0.38	1.05
CAG	D10A	3	Gal	24	4.25	1.28	0.74	3.19
CAG	eSpCas9	7	Gal	24	13.17	4.15	1.57	3.83
CAG	HF1	3	Gal	24	0.34	0.15	0.08	0.36
CAG	N863A	3	Gal	24	0.16	0.06	0.03	0.14
CAG	SaCas9	3	Gal	24	0.17	0.03	0.02	0.08

CAG	SpCas9	4	Gal	24	6.89	5.10	2.55	8.11
CAG	Cpf1	5	Gal	36	5.23	0.75	0.34	0.93
CAG	D10A	3	Gal	36	4.37	1.03	0.60	2.57
CAG	eSpCas9	7	Gal	36	20.64	5.40	2.04	4.99
CAG	HF1	3	Gal	36	0.39	0.14	0.08	0.35
CAG	N863A	3	Gal	36	0.38	0.04	0.02	0.10
CAG	SaCas9	3	Gal	36	1.15	0.49	0.28	1.22
CAG	SpCas9	4	Gal	36	8.68	9.28	4.64	14.76
CAG	Cpf1	5	Glu	0	0.10	0.06	0.03	0.08
CAG	D10A	3	Glu	0	0.14	0.07	0.04	0.18
CAG	eSpCas9	2	Glu	0	2.33	2.47	1.75	22.24
CAG	HF1	1	Glu	0	0.04	NA	NA	NA
CAG	N863A	2	Glu	0	0.06	0.06	0.04	0.51
CAG	SaCas9	3	Glu	0	0.16	0.01	0.00	0.01
CAG	SpCas9	4	Glu	0	0.37	0.61	0.31	0.97
CAG	Cpf1	5	Glu	12	0.17	0.11	0.05	0.14
CAG	D10A	3	Glu	12	0.11	0.04	0.02	0.10
CAG	eSpCas9	1	Glu	12	3.81	NA	NA	NA
CAG	HF1	1	Glu	12	0.07	NA	NA	NA
CAG	N863A	2	Glu	12	0.00	0.00	0.00	0.00
CAG	SaCas9	3	Glu	12	0.34	0.11	0.06	0.28
CAG	SpCas9	3	Glu	12	0.11	0.03	0.01	0.06
CAG	Cpf1	5	Glu	24	0.25	0.10	0.04	0.12
CAG	D10A	3	Glu	24	0.27	0.05	0.03	0.13
CAG	eSpCas9	2	Glu	24	2.77	2.14	1.52	19.25
CAG	HF1	1	Glu	24	0.15	NA	NA	NA
CAG	N863A	2	Glu	24	0.15	0.02	0.02	0.19
CAG	SaCas9	3	Glu	24	1.47	0.29	0.16	0.71
CAG	SpCas9	4	Glu	24	0.27	0.07	0.03	0.11
CAG	Cpf1	5	Glu	36	0.59	0.23	0.10	0.29
CAG	D10A	3	Glu	36	0.65	0.23	0.13	0.57
CAG	eSpCas9	2	Glu	36	4.99	3.25	2.30	29.16
CAG	HF1	1	Glu	36	0.21	NA	NA	NA
CAG	N863A	2	Glu	36	0.41	0.01	0.00	0.06
CAG	SaCas9	3	Glu	36	2.26	0.19	0.11	0.47
CAG	SpCas9	4	Glu	36	0.96	0.55	0.28	0.88
CTG	Cpf1	3	Gal	0	0.03	0.01	0.01	0.03
CTG	D10A	6	Gal	0	0.15	0.10	0.04	0.10
CTG	eSpCas9	3	Gal	0	1.31	0.67	0.38	1.65

CTG	HF1	3	Gal	0	0.03	0.03	0.02	0.07
CTG	N863A	6	Gal	0	0.15	0.07	0.03	0.07
CTG	SaCas9	6	Gal	0	2.07	3.82	1.56	4.01
CTG	SpCas9	8	Gal	0	0.07	0.06	0.02	0.05
CTG	Cpf1	3	Gal	12	0.77	0.32	0.18	0.79
CTG	D10A	6	Gal	12	0.06	0.02	0.01	0.02
CTG	eSpCas9	3	Gal	12	2.90	2.10	1.21	5.22
CTG	HF1	3	Gal	12	0.12	0.07	0.04	0.17
CTG	N863A	6	Gal	12	0.05	0.02	0.01	0.02
CTG	SaCas9	6	Gal	12	1.57	3.47	1.42	3.64
CTG	SpCas9	8	Gal	12	10.86	3.73	1.32	3.12
CTG	Cpf1	3	Gal	24	2.38	0.71	0.41	1.77
CTG	D10A	6	Gal	24	0.14	0.07	0.03	0.07
CTG	eSpCas9	3	Gal	24	4.62	2.93	1.69	7.27
CTG	HF1	3	Gal	24	0.15	0.07	0.04	0.19
CTG	N863A	6	Gal	24	0.06	0.03	0.01	0.03
CTG	SaCas9	6	Gal	24	1.80	4.02	1.64	4.22
CTG	SpCas9	8	Gal	24	21.41	2.11	0.75	1.77
CTG	Cpf1	3	Gal	36	3.00	0.30	0.17	0.74
CTG	D10A	6	Gal	36	0.49	0.14	0.06	0.15
CTG	eSpCas9	3	Gal	36	6.57	3.06	1.77	7.60
CTG	HF1	3	Gal	36	0.22	0.11	0.07	0.28
CTG	N863A	6	Gal	36	0.41	0.20	0.08	0.21
CTG	SaCas9	6	Gal	36	2.49	3.26	1.33	3.42
CTG	SpCas9	8	Gal	36	22.06	1.97	0.70	1.64
CTG	Cpf1	3	Glu	0	0.01	0.01	0.01	0.03
CTG	D10A	6	Glu	0	0.10	0.06	0.03	0.06
CTG	eSpCas9	3	Glu	0	0.81	0.56	0.32	1.38
CTG	HF1	3	Glu	0	0.03	0.02	0.01	0.05
CTG	N863A	6	Glu	0	0.10	0.02	0.01	0.02
CTG	SaCas9	6	Glu	0	1.01	2.17	0.89	2.28
CTG	SpCas9	8	Glu	0	0.26	0.53	0.19	0.44
CTG	Cpf1	3	Glu	12	0.37	0.16	0.09	0.40
CTG	D10A	6	Glu	12	0.04	0.03	0.01	0.03
CTG	eSpCas9	3	Glu	12	2.09	0.78	0.45	1.94
CTG	HF1	3	Glu	12	0.37	0.29	0.17	0.73
CTG	N863A	6	Glu	12	0.06	0.06	0.02	0.06
CTG	SaCas9	6	Glu	12	0.18	0.13	0.05	0.14
CTG	SpCas9	8	Glu	12	1.04	2.28	0.80	1.90

CTG	Cpf1	3	Glu	24	0.29	0.10	0.06	0.25
CTG	D10A	6	Glu	24	0.60	0.16	0.07	0.17
CTG	eSpCas9	3	Glu	24	2.56	1.16	0.67	2.87
CTG	HF1	3	Glu	24	0.34	0.16	0.09	0.40
CTG	N863A	6	Glu	24	0.85	0.64	0.26	0.68
CTG	SaCas9	6	Glu	24	1.45	0.51	0.21	0.54
CTG	SpCas9	8	Glu	24	0.61	0.45	0.16	0.38
CTG	Cpf1	3	Glu	36	0.76	0.16	0.09	0.39
CTG	D10A	6	Glu	36	1.32	0.28	0.12	0.30
CTG	eSpCas9	3	Glu	36	4.84	2.25	1.30	5.59
CTG	HF1	3	Glu	36	0.81	0.26	0.15	0.64
CTG	N863A	6	Glu	36	1.83	1.13	0.46	1.18
CTG	SaCas9	6	Glu	36	2.79	0.98	0.40	1.02
CTG	SpCas9	8	Glu	36	1.57	1.14	0.40	0.96
GAA	Cpf1	6	Gal	0	1.97	4.23	1.73	4.44
GAA	D10A	3	Gal	0	0.31	0.07	0.04	0.17
GAA	eSpCas9	3	Gal	0	3.48	3.08	1.78	7.66
GAA	HF1	3	Gal	0	0.12	0.09	0.05	0.22
GAA	N863A	3	Gal	0	0.08	0.01	0.01	0.03
GAA	SaCas9	3	Gal	0	0.16	0.11	0.06	0.27
GAA	SpCas9	6	Gal	0	0.21	0.09	0.04	0.10
GAA	Cpf1	3	Gal	12	32.17	9.72	5.61	24.15
GAA	D10A	3	Gal	12	1.77	0.10	0.06	0.26
GAA	eSpCas9	3	Gal	12	8.36	4.94	2.85	12.28
GAA	HF1	3	Gal	12	0.06	0.03	0.02	0.09
GAA	N863A	3	Gal	12	0.09	0.06	0.04	0.15
GAA	SaCas9	3	Gal	12	0.08	0.01	0.00	0.02
GAA	SpCas9	6	Gal	12	24.00	9.00	3.67	9.44
GAA	Cpf1	6	Gal	24	59.75	9.01	3.68	9.45
GAA	D10A	3	Gal	24	5.31	0.36	0.21	0.91
GAA	eSpCas9	3	Gal	24	21.10	7.57	4.37	18.81
GAA	HF1	3	Gal	24	0.25	0.03	0.02	0.07
GAA	N863A	3	Gal	24	0.37	0.23	0.13	0.56
GAA	SaCas9	3	Gal	24	0.85	0.51	0.29	1.27
GAA	SpCas9	6	Gal	24	19.92	1.89	0.77	1.98
GAA	Cpf1	6	Gal	36	61.52	6.25	2.55	6.56
GAA	D10A	3	Gal	36	5.24	0.06	0.04	0.16
GAA	eSpCas9	3	Gal	36	22.57	8.36	4.83	20.76
GAA	HF1	3	Gal	36	0.53	0.07	0.04	0.18

GAA	N863A	3	Gal	36	0.62	0.19	0.11	0.46
GAA	SaCas9	3	Gal	36	3.26	0.85	0.49	2.10
GAA	SpCas9	6	Gal	36	25.38	2.80	1.14	2.93
GAA	Cpf1	4	Glu	0	0.68	0.56	0.28	0.89
GAA	D10A	3	Glu	0	0.09	0.04	0.02	0.09
GAA	eSpCas9	3	Glu	0	1.63	0.77	0.44	1.90
GAA	HF1	3	Glu	0	0.20	0.13	0.07	0.32
GAA	N863A	3	Glu	0	0.03	0.02	0.01	0.05
GAA	SaCas9	3	Glu	0	0.09	0.04	0.02	0.09
GAA	SpCas9	6	Glu	0	4.24	9.68	3.95	10.16
GAA	Cpf1	3	Glu	12	0.66	0.20	0.11	0.48
GAA	D10A	3	Glu	12	0.33	0.07	0.04	0.17
GAA	eSpCas9	3	Glu	12	2.14	0.41	0.23	1.01
GAA	HF1	3	Glu	12	0.11	0.03	0.02	0.08
GAA	N863A	3	Glu	12	0.11	0.08	0.04	0.19
GAA	SaCas9	3	Glu	12	0.28	0.19	0.11	0.47
GAA	SpCas9	6	Glu	12	4.92	10.82	4.42	11.35
GAA	Cpf1	3	Glu	24	1.32	0.09	0.05	0.22
GAA	D10A	3	Glu	24	0.69	0.07	0.04	0.17
GAA	eSpCas9	3	Glu	24	3.20	0.81	0.47	2.02
GAA	HF1	3	Glu	24	0.42	0.07	0.04	0.18
GAA	N863A	3	Glu	24	0.30	0.09	0.05	0.23
GAA	SaCas9	3	Glu	24	0.74	0.29	0.17	0.72
GAA	SpCas9	6	Glu	24	5.73	11.80	4.82	12.38
GAA	Cpf1	6	Glu	36	1.49	0.86	0.35	0.90
GAA	D10A	3	Glu	36	1.26	0.16	0.09	0.39
GAA	eSpCas9	3	Glu	36	4.36	0.84	0.49	2.09
GAA	HF1	3	Glu	36	0.68	0.18	0.10	0.45
GAA	N863A	3	Glu	36	0.92	0.18	0.10	0.44
GAA	SaCas9	3	Glu	36	2.59	1.24	0.72	3.08
GAA	SpCas9	6	Glu	36	6.84	11.84	4.83	12.43
CGG	Cpf1	2	Gal	0	0.43	0.55	0.39	4.93
CGG	D10A	4	Gal	0	2.88	2.61	1.30	4.15
CGG	eSpCas9	6	Gal	0	35.90	18.16	7.41	19.05
CGG	HF1	3	Gal	0	0.65	0.11	0.06	0.27
CGG	N863A	2	Gal	0	0.08	0.02	0.02	0.22
CGG	SaCas9	3	Gal	0	0.59	0.34	0.20	0.85
CGG	SpCas9	5	Gal	0	2.55	3.15	1.41	3.91
CGG	Cpf1	2	Gal	12	0.17	0.05	0.04	0.44

CGG	D10A	4	Gal	12	15.18	4.19	2.10	6.67
CGG	eSpCas9	6	Gal	12	64.68	12.66	5.17	13.28
CGG	HF1	3	Gal	12	8.96	1.50	0.86	3.71
CGG	N863A	2	Gal	12	5.44	0.69	0.49	6.23
CGG	SaCas9	3	Gal	12	0.23	0.15	0.08	0.36
CGG	SpCas9	5	Gal	12	36.38	8.90	3.98	11.05
CGG	Cpf1	2	Gal	24	0.45	0.07	0.05	0.64
CGG	D10A	4	Gal	24	45.95	11.78	5.89	18.75
CGG	eSpCas9	6	Gal	24	85.73	3.66	1.50	3.84
CGG	HF1	3	Gal	24	43.00	1.59	0.92	3.94
CGG	N863A	2	Gal	24	25.70	0.00	0.00	0.00
CGG	SaCas9	3	Gal	24	0.40	0.31	0.18	0.78
CGG	SpCas9	5	Gal	24	87.00	1.30	0.58	1.62
CGG	Cpf1	2	Gal	36	0.93	0.42	0.30	3.81
CGG	D10A	4	Gal	36	58.20	14.50	7.25	23.07
CGG	eSpCas9	6	Gal	36	88.25	4.45	1.82	4.67
CGG	HF1	3	Gal	36	57.80	3.85	2.22	9.57
CGG	N863A	2	Gal	36	31.80	2.97	2.10	26.68
CGG	SaCas9	3	Gal	36	0.00	0.00	0.00	0.00
CGG	SpCas9	5	Gal	36	93.10	2.13	0.95	2.64
CGG	Cpf1	1	Glu	0	0.01	NA	NA	NA
CGG	D10A	3	Glu	0	5.78	4.31	2.49	10.70
CGG	eSpCas9	6	Glu	0	48.17	30.02	12.25	31.50
CGG	HF1	3	Glu	0	0.76	0.48	0.28	1.19
CGG	N863A	2	Glu	0	0.23	0.07	0.05	0.64
CGG	SaCas9	3	Glu	0	0.26	0.38	0.22	0.94
CGG	SpCas9	5	Glu	0	12.20	20.29	9.07	25.19
CGG	Cpf1	1	Glu	12	0.03	NA	NA	NA
CGG	D10A	3	Glu	12	11.43	7.21	4.16	17.91
CGG	eSpCas9	6	Glu	12	57.48	25.43	10.38	26.69
CGG	HF1	3	Glu	12	1.10	0.33	0.19	0.82
CGG	N863A	2	Glu	12	0.43	0.32	0.23	2.86
CGG	SaCas9	3	Glu	12	0.60	0.19	0.11	0.48
CGG	SpCas9	5	Glu	12	14.46	22.30	9.97	27.69
CGG	Cpf1	1	Glu	24	0.38	NA	NA	NA
CGG	D10A	3	Glu	24	13.91	7.19	4.15	17.87
CGG	eSpCas9	6	Glu	24	58.93	24.69	10.08	25.91
CGG	HF1	3	Glu	24	2.03	0.64	0.37	1.58
CGG	N863A	2	Glu	24	0.39	0.06	0.05	0.57

CGG	SaCas9	3	Glu	24	1.15	0.11	0.06	0.27
CGG	SpCas9	5	Glu	24	15.59	22.11	9.89	27.46
CGG	Cpf1	1	Glu	36	0.69	NA	NA	NA
CGG	D10A	3	Glu	36	15.29	6.99	4.04	17.37
CGG	eSpCas9	6	Glu	36	61.30	22.13	9.03	23.22
CGG	HF1	3	Glu	36	3.45	1.38	0.80	3.44
CGG	N863A	2	Glu	36	0.88	0.02	0.02	0.19
CGG	SaCas9	3	Glu	36	0.00	0.00	0.00	0.00
CGG	SpCas9	5	Glu	36	17.38	20.97	9.38	26.04
CCTG	Cpf1	6	Gal	0	0.24	0.13	0.05	0.13
CCTG	D10A	3	Gal	0	0.17	0.08	0.05	0.20
CCTG	eSpCas9	4	Gal	0	21.60	9.70	4.85	15.44
CCTG	HF1	3	Gal	0	0.03	0.01	0.00	0.02
CCTG	N863A	3	Gal	0	0.16	0.11	0.06	0.26
CCTG	SaCas9	3	Gal	0	0.26	0.16	0.09	0.39
CCTG	SpCas9	3	Gal	0	10.11	17.23	9.95	42.79
CCTG	Cpf1	3	Gal	12	19.44	9.55	5.52	23.73
CCTG	D10A	3	Gal	12	2.60	0.37	0.22	0.93
CCTG	eSpCas9	4	Gal	12	46.10	7.68	3.84	12.23
CCTG	HF1	3	Gal	12	0.02	0.03	0.02	0.06
CCTG	N863A	3	Gal	12	0.17	0.10	0.06	0.26
CCTG	SaCas9	3	Gal	12	0.16	0.03	0.02	0.08
CCTG	SpCas9	6	Gal	12	22.69	26.14	10.67	27.43
CCTG	Cpf1	6	Gal	24	40.08	7.16	2.92	7.51
CCTG	D10A	3	Gal	24	12.00	0.72	0.42	1.79
CCTG	eSpCas9	4	Gal	24	85.48	7.15	3.57	11.37
CCTG	HF1	3	Gal	24	0.02	0.02	0.01	0.05
CCTG	N863A	3	Gal	24	0.90	0.13	0.07	0.31
CCTG	SaCas9	3	Gal	24	0.49	0.27	0.15	0.66
CCTG	SpCas9	6	Gal	24	42.49	46.59	19.02	48.89
CCTG	Cpf1	6	Gal	36	45.87	11.92	4.86	12.50
CCTG	D10A	3	Gal	36	14.80	3.70	2.14	9.19
CCTG	eSpCas9	4	Gal	36	97.05	1.53	0.76	2.43
CCTG	HF1	3	Gal	36	0.12	0.05	0.03	0.14
CCTG	N863A	3	Gal	36	1.20	0.04	0.02	0.10
CCTG	SaCas9	3	Gal	36	2.19	1.10	0.63	2.73
CCTG	SpCas9	3	Gal	36	96.63	1.16	0.67	2.88
CCTG	Cpf1	6	Glu	0	0.42	0.14	0.06	0.15
CCTG	D10A	3	Glu	0	0.09	0.03	0.01	0.06

CCTG	eSpCas9	3	Glu	0	0.04	0.05	0.03	0.12
CCTG	HF1	3	Glu	0	0.03	0.01	0.00	0.02
CCTG	N863A	3	Glu	0	0.02	0.01	0.01	0.03
CCTG	SaCas9	3	Glu	0	1.15	1.45	0.84	3.61
CCTG	SpCas9	5	Glu	0	0.06	0.06	0.03	0.08
CCTG	Cpf1	3	Glu	12	1.68	0.67	0.38	1.65
CCTG	D10A	3	Glu	12	0.24	0.08	0.05	0.21
CCTG	eSpCas9	3	Glu	12	0.00	0.01	0.00	0.02
CCTG	HF1	3	Glu	12	0.11	0.03	0.02	0.07
CCTG	N863A	3	Glu	12	0.10	0.09	0.05	0.24
CCTG	SaCas9	3	Glu	12	0.36	0.26	0.15	0.64
CCTG	SpCas9	2	Glu	12	0.35	0.16	0.11	1.40
CCTG	Cpf1	6	Glu	24	0.89	0.93	0.38	0.98
CCTG	D10A	3	Glu	24	0.46	0.14	0.08	0.34
CCTG	eSpCas9	3	Glu	24	0.01	0.02	0.01	0.04
CCTG	HF1	3	Glu	24	0.11	0.02	0.01	0.05
CCTG	N863A	3	Glu	24	0.36	0.17	0.10	0.42
CCTG	SaCas9	3	Glu	24	0.91	0.41	0.24	1.02
CCTG	SpCas9	2	Glu	24	0.70	0.03	0.02	0.25
CCTG	Cpf1	6	Glu	36	2.63	2.05	0.84	2.16
CCTG	D10A	3	Glu	36	0.65	0.11	0.07	0.28
CCTG	eSpCas9	3	Glu	36	0.47	0.15	0.09	0.38
CCTG	HF1	3	Glu	36	0.50	0.50	0.29	1.24
CCTG	N863A	3	Glu	36	0.55	0.24	0.14	0.60
CCTG	SaCas9	3	Glu	36	1.79	0.50	0.29	1.25
CCTG	SpCas9	5	Glu	36	0.95	1.01	0.45	1.26
GCN	Cpf1	3	Gal	0	0.16	0.02	0.01	0.04
GCN	D10A	3	Gal	0	0.12	0.08	0.05	0.20
GCN	eSpCas9	5	Gal	0	16.85	25.83	11.55	32.07
GCN	HF1	4	Gal	0	0.30	0.33	0.16	0.52
GCN	N863A	3	Gal	0	0.07	0.03	0.02	0.08
GCN	SaCas9	3	Gal	0	0.38	0.22	0.13	0.54
GCN	SpCas9	3	Gal	0	0.13	0.12	0.07	0.29
GCN	Cpf1	3	Gal	12	0.74	0.08	0.04	0.19
GCN	D10A	3	Gal	12	2.34	0.62	0.36	1.55
GCN	eSpCas9	4	Gal	12	59.63	20.16	10.08	32.08
GCN	HF1	4	Gal	12	0.68	0.44	0.22	0.69
GCN	N863A	3	Gal	12	0.27	0.09	0.05	0.23
GCN	SaCas9	3	Gal	12	0.15	0.16	0.09	0.39

GCN	SpCas9	3	Gal	12	31.33	6.21	3.59	15.43
GCN	Cpf1	3	Gal	24	3.20	0.58	0.33	1.43
GCN	D10A	3	Gal	24	7.84	0.83	0.48	2.07
GCN	eSpCas9	5	Gal	24	87.92	10.86	4.86	13.49
GCN	HF1	4	Gal	24	3.71	1.27	0.63	2.02
GCN	N863A	3	Gal	24	1.32	0.17	0.10	0.43
GCN	SaCas9	3	Gal	24	0.29	0.15	0.09	0.37
GCN	SpCas9	3	Gal	24	84.53	1.46	0.85	3.64
GCN	Cpf1	3	Gal	36	3.96	0.05	0.03	0.12
GCN	D10A	3	Gal	36	8.26	0.62	0.36	1.54
GCN	eSpCas9	5	Gal	36	90.12	9.01	4.03	11.19
GCN	HF1	4	Gal	36	5.42	0.68	0.34	1.08
GCN	N863A	3	Gal	36	2.06	0.29	0.16	0.71
GCN	SaCas9	3	Gal	36	1.77	0.71	0.41	1.77
GCN	SpCas9	3	Gal	36	92.10	4.09	2.36	10.15
GCN	Cpf1	3	Glu	0	0.15	0.08	0.04	0.19
GCN	D10A	3	Glu	0	0.11	0.03	0.02	0.07
GCN	eSpCas9	4	Glu	0	33.00	31.34	15.67	49.87
GCN	HF1	4	Glu	0	0.74	0.33	0.17	0.53
GCN	N863A	3	Glu	0	0.03	0.03	0.02	0.08
GCN	SaCas9	3	Glu	0	0.36	0.23	0.13	0.56
GCN	SpCas9	3	Glu	0	0.39	0.23	0.13	0.56
GCN	Cpf1	3	Glu	12	0.22	0.04	0.02	0.10
GCN	D10A	3	Glu	12	0.09	0.05	0.03	0.12
GCN	eSpCas9	4	Glu	12	40.14	34.29	17.14	54.56
GCN	HF1	4	Glu	12	0.23	0.16	0.08	0.25
GCN	N863A	3	Glu	12	0.16	0.21	0.12	0.51
GCN	SaCas9	3	Glu	12	0.36	0.23	0.13	0.56
GCN	SpCas9	3	Glu	12	0.23	0.09	0.05	0.23
GCN	Cpf1	3	Glu	24	0.91	0.16	0.09	0.39
GCN	D10A	3	Glu	24	0.33	0.13	0.07	0.32
GCN	eSpCas9	4	Glu	24	44.00	31.77	15.89	50.56
GCN	HF1	4	Glu	24	0.21	0.33	0.16	0.52
GCN	N863A	3	Glu	24	0.31	0.08	0.05	0.20
GCN	SaCas9	3	Glu	24	1.57	0.81	0.47	2.01
GCN	SpCas9	3	Glu	24	0.40	0.30	0.17	0.73
GCN	Cpf1	3	Glu	36	1.08	0.13	0.08	0.32
GCN	D10A	3	Glu	36	0.78	0.08	0.05	0.20
GCN	eSpCas9	4	Glu	36	47.18	29.63	14.81	47.15

GCN	HF1	4	Glu	36	1.54	0.28	0.14	0.44
GCN	N863A	3	Glu	36	0.96	0.24	0.14	0.59
GCN	SaCas9	3	Glu	36	2.62	0.90	0.52	2.24
GCN	SpCas9	3	Glu	36	1.51	0.66	0.38	1.63
NR	Cpf1	4	Gal	0	3.81	3.52	1.76	5.60
NR	D10A	3	Gal	0	0.16	0.04	0.03	0.11
NR	eSpCas9	3	Gal	0	1.17	0.51	0.29	1.26
NR	HF1	3	Gal	0	0.06	0.01	0.00	0.02
NR	ISCEI	3	Gal	0	0.21	0.01	0.01	0.03
NR	N863A	3	Gal	0	0.10	0.01	0.01	0.03
NR	SaCas9	4	Gal	0	0.09	0.03	0.02	0.05
NR	SpCas9	6	Gal	0	0.18	0.15	0.06	0.15
NR	Cpf1	4	Gal	12	36.48	4.76	2.38	7.58
NR	D10A	3	Gal	12	0.17	0.15	0.09	0.37
NR	eSpCas9	3	Gal	12	5.16	0.45	0.26	1.12
NR	HF1	3	Gal	12	0.22	0.12	0.07	0.31
NR	ISCEI	3	Gal	12	9.96	2.28	1.32	5.67
NR	N863A	3	Gal	12	0.06	0.02	0.01	0.04
NR	SaCas9	4	Gal	12	3.60	3.39	1.70	5.40
NR	SpCas9	6	Gal	12	42.57	18.43	7.52	19.34
NR	Cpf1	4	Gal	24	83.90	3.48	1.74	5.54
NR	D10A	3	Gal	24	0.95	0.54	0.31	1.34
NR	eSpCas9	3	Gal	24	17.80	0.72	0.42	1.79
NR	HF1	3	Gal	24	0.50	0.02	0.01	0.04
NR	ISCEI	3	Gal	24	24.07	2.20	1.27	5.47
NR	N863A	3	Gal	24	0.10	0.08	0.05	0.20
NR	SaCas9	4	Gal	24	36.98	12.37	6.18	19.68
NR	SpCas9	6	Gal	24	85.78	4.26	1.74	4.47
NR	Cpf1	4	Gal	36	92.60	2.94	1.47	4.68
NR	D10A	3	Gal	36	1.27	0.55	0.32	1.36
NR	eSpCas9	3	Gal	36	18.87	1.10	0.64	2.74
NR	HF1	3	Gal	36	0.90	0.28	0.16	0.69
NR	ISCEI	3	Gal	36	27.97	1.16	0.67	2.88
NR	N863A	3	Gal	36	0.46	0.18	0.10	0.45
NR	SaCas9	4	Gal	36	54.53	4.94	2.47	7.86
NR	SpCas9	6	Gal	36	86.95	3.04	1.24	3.19
NR	Cpf1	4	Glu	0	6.54	1.71	0.85	2.72
NR	D10A	3	Glu	0	0.24	0.28	0.16	0.69
NR	eSpCas9	3	Glu	0	1.58	0.38	0.22	0.93

NR	HF1	3	Glu	0	0.20	0.19	0.11	0.48
NR	ISCEI	3	Glu	0	0.17	0.13	0.08	0.33
NR	N863A	3	Glu	0	0.05	0.04	0.02	0.10
NR	SaCas9	4	Glu	0	0.14	0.06	0.03	0.09
NR	SpCas9	6	Glu	0	0.62	1.06	0.43	1.11
NR	Cpf1	4	Glu	12	10.79	2.81	1.40	4.47
NR	D10A	3	Glu	12	0.10	0.07	0.04	0.17
NR	eSpCas9	3	Glu	12	2.50	0.59	0.34	1.47
NR	HF1	3	Glu	12	0.09	0.09	0.05	0.21
NR	ISCEI	3	Glu	12	0.30	0.12	0.07	0.30
NR	N863A	3	Glu	12	0.01	0.01	0.01	0.02
NR	SaCas9	4	Glu	12	0.03	0.02	0.01	0.03
NR	SpCas9	6	Glu	12	0.23	0.23	0.10	0.25
NR	Cpf1	4	Glu	24	15.25	4.00	2.00	6.36
NR	D10A	3	Glu	24	0.77	0.66	0.38	1.63
NR	eSpCas9	3	Glu	24	4.16	0.55	0.32	1.36
NR	HF1	3	Glu	24	0.28	0.07	0.04	0.18
NR	ISCEI	3	Glu	24	1.09	0.06	0.03	0.14
NR	N863A	3	Glu	24	0.15	0.03	0.02	0.07
NR	SaCas9	4	Glu	24	0.22	0.08	0.04	0.13
NR	SpCas9	6	Glu	24	0.70	0.38	0.16	0.40
NR	Cpf1	4	Glu	36	17.00	4.29	2.14	6.82
NR	D10A	3	Glu	36	1.22	1.39	0.80	3.44
NR	eSpCas9	3	Glu	36	5.79	0.30	0.17	0.75
NR	HF1	3	Glu	36	0.93	0.17	0.10	0.43
NR	ISCEI	3	Glu	36	1.26	0.11	0.06	0.27
NR	N863A	3	Glu	36	0.47	0.10	0.06	0.26
NR	SaCas9	4	Glu	36	0.80	0.13	0.06	0.20
NR	SpCas9	6	Glu	36	1.89	1.15	0.47	1.21
TGGAA	Cpf1	3	Gal	0	0.15	0.06	0.03	0.15
TGGAA	D10A	6	Gal	0	21.12	36.12	14.74	37.90
TGGAA	eSpCas9	3	Gal	0	69.17	17.24	9.95	42.83
TGGAA	HF1	3	Gal	0	0.12	0.11	0.07	0.28
TGGAA	N863A	3	Gal	0	0.18	0.04	0.02	0.10
TGGAA	SaCas9	3	Gal	0	0.26	0.17	0.10	0.41
TGGAA	SpCas9	3	Gal	0	1.93	0.81	0.47	2.02
TGGAA	Cpf1	3	Gal	12	5.30	1.82	1.05	4.53
TGGAA	D10A	3	Gal	12	25.59	22.27	12.86	55.33
TGGAA	eSpCas9	3	Gal	12	86.03	8.13	4.69	20.19

TGGAA	HF1	3	Gal	12	0.69	0.13	0.07	0.31
TGGAA	N863A	3	Gal	12	9.94	2.29	1.32	5.68
TGGAA	SaCas9	3	Gal	12	0.21	0.18	0.10	0.45
TGGAA	SpCas9	3	Gal	12	54.23	5.15	2.97	12.79
TGGAA	Cpf1	3	Gal	24	27.30	1.39	0.80	3.45
TGGAA	D10A	6	Gal	24	65.58	33.45	13.66	35.11
TGGAA	eSpCas9	3	Gal	24	96.37	2.40	1.39	5.96
TGGAA	HF1	3	Gal	24	2.58	0.19	0.11	0.48
TGGAA	N863A	3	Gal	24	43.10	3.87	2.24	9.62
TGGAA	SaCas9	3	Gal	24	0.43	0.07	0.04	0.17
TGGAA	SpCas9	3	Gal	24	92.23	0.68	0.39	1.69
TGGAA	Cpf1	3	Gal	36	41.63	3.76	2.17	9.34
TGGAA	D10A	6	Gal	36	70.44	35.17	14.36	36.91
TGGAA	eSpCas9	3	Gal	36	96.30	1.56	0.90	3.88
TGGAA	HF1	3	Gal	36	4.15	0.05	0.03	0.13
TGGAA	N863A	3	Gal	36	46.13	3.39	1.96	8.43
TGGAA	SaCas9	3	Gal	36	1.84	0.57	0.33	1.42
TGGAA	SpCas9	3	Gal	36	95.63	1.78	1.03	4.42
TGGAA	Cpf1	3	Glu	0	0.13	0.03	0.02	0.08
TGGAA	D10A	6	Glu	0	9.48	7.16	2.92	7.52
TGGAA	eSpCas9	2	Glu	0	52.90	10.18	7.20	91.48
TGGAA	HF1	3	Glu	0	0.05	0.02	0.01	0.04
TGGAA	N863A	3	Glu	0	0.17	0.09	0.05	0.22
TGGAA	SaCas9	3	Glu	0	0.26	0.19	0.11	0.46
TGGAA	SpCas9	3	Glu	0	1.88	0.83	0.48	2.06
TGGAA	Cpf1	3	Glu	12	0.10	0.02	0.01	0.05
TGGAA	D10A	3	Glu	12	13.42	10.12	5.84	25.14
TGGAA	eSpCas9	2	Glu	12	73.60	8.91	6.30	80.05
TGGAA	HF1	3	Glu	12	0.12	0.03	0.02	0.08
TGGAA	N863A	3	Glu	12	0.13	0.03	0.02	0.07
TGGAA	SaCas9	3	Glu	12	0.07	0.02	0.01	0.06
TGGAA	SpCas9	3	Glu	12	4.32	1.37	0.79	3.40
TGGAA	Cpf1	3	Glu	24	0.61	0.23	0.14	0.58
TGGAA	D10A	6	Glu	24	15.54	11.06	4.52	11.61
TGGAA	eSpCas9	2	Glu	24	76.65	6.43	4.55	57.81
TGGAA	HF1	3	Glu	24	0.68	0.08	0.04	0.19
TGGAA	N863A	3	Glu	24	0.40	0.17	0.10	0.42
TGGAA	SaCas9	3	Glu	24	0.67	0.16	0.09	0.39
TGGAA	SpCas9	3	Glu	24	6.22	1.44	0.83	3.59

TGGAA	Cpf1	3	Glu	36	0.78	0.20	0.12	0.49
TGGAA	D10A	6	Glu	36	15.00	10.58	4.32	11.10
TGGAA	eSpCas9	2	Glu	36	78.10	5.94	4.20	53.37
TGGAA	HF1	3	Glu	36	0.64	0.06	0.03	0.15
TGGAA	N863A	3	Glu	36	0.91	0.29	0.17	0.73
TGGAA	SaCas9	3	Glu	36	1.80	0.41	0.24	1.02
TGGAA	SpCas9	3	Glu	36	9.50	0.89	0.51	2.20

Supplemental Table S5: Off-target mutations detected by deep sequencing

Library	Guide sequence	Off-target sequence	Mismatches	Mismatch position	MIT specificity score	CFD specificity score	Chromosome	Start	End	Position	Mutation	Locus	Mutant reads	Total reads	NR mutant reads	NR total reads
CTG_Cas9	TGCTGCTGCTGCTGCTGCTGCGG	TGCTGCTGCTGCTGTTTCTGTGG	2	.....*.*...	1	0,0	Chromosome XII	511510	511532	511518	-CTG	*YLR177W*	1	12	0	53
		TGCTGCTGTTGTTGTTGATGAGG	4	.....*..*..*..*..	0	0,0	Chromosome XII	556330	556352	556344	-CTG	*ENT2*	1	22	0	42
		TGCTGCTGCTGCTGCTGCTGCGG	0	.....	100	1,0	Chromosome XIII	514627	514649	514628	-CTG	*YMR124W*	6	19	1	57
		TGCTGCTGCTGCTGCTGCTGCGG	0	.....	100	1,0	Chromosome XIII	514627	514649	514632	+C	*YMR124W*	2	19	0	59
CAG_Cas9	AGCAGCAGCAGCAGCAGCAGCGG	ATCAGCAGCAGCAGCAGCAGCAG	1	.*.....	20	0,2	Chromosome II	649336	649358	649337	-CAG	*NGR1*	1	121	0	46
		AGCAACAACAACAGCAGCAGCAG	3	...*..*..*.....	0	0,2	Chromosome II	780334	780356	780344	-CAG	*SNF5*	1	80	0	32
		AACAACAGCAGCAGCAGCATCAG	3	.*..*.....*	0	0,1	Chromosome II	780340	780362	780362	+CAA	*SNF5*	1	80	1	36
		AACAACAGCAGCAGCAGCAGCAG	2	.*..*.....	1	0,2	Chromosome IV	923048	923070	923063	-CAG	*PCF11*	1	74	0	44
		AACAACAACAGCAGCAGCAGCAG	3	.*..*.*.....	1	0,2	Chromosome IV	923051	923073	923069	+TG	*PCF11*	1	76	0	41
		AGGAGCAGCAGCAGCAGCAGCAG	1	..*.....	20	0,1	Chromosome IX	49505	49527	49507	-CAG	*UBP7*	2	118	0	71
		AGCAGCAGCAGCAGCAACAGCAG	1	.....*...	8	0,2	Chromosome IX	169456	169478	169477	-CAA	*SLM1*	5	105	1	54
		TTCAGCAGCAGCAGCAGCAGCAG	2	**.....	1	0,2	Chromosome VII	379161	379183	379162	-CAG	*SGF73*	1	133	0	38
		ATTTACAGCAGCAGCAGCAGCAG	4	.****.....	0	0.06	Chromosome VII	728883	728905	728904	-T	*NUP57*	1	117	0	44
		CCCAGCAGCAGCAGCAGCAGCAG	2	**.....	1	0.15	Chromosome VII	988893	988915	988894	-CAG	*MGA1*	3	130	1	49
		AGTGGCAGCAGCAGCAGCAGTAG	2	..**.....	1	0.11	Chromosome VIII	341081	341103	341084	-CAG	*DMA1*	1	135	0	68
		AGTGGCAGCAGCAGCAGCAGTAG	2	..**.....	1	0.11	Chromosome VIII	341081	341103	341085	-C	*DMA1*	1	135	0	68
		TGCAGCAGCAGCAGCAGCAGCAG	1	*.....	20	0.25	Chromosome XI	567100	567122	567100	-CAG	*CCP1*	1	127	0	48
		AGCAGCAGCAGCAGCAGCAGTGA	0	.....	20	0.06	Chromosome XI	613689	613711	613691	-CAG	*SRP40*	1	92	0	53
		AACAGCAGCAGCAGCAGCAGCAG	1	.*.....	20	0.21	Chromosome XII	511517	511539	511518	-CAG	*YLR177W*	1	111	0	53
		AACAGGAGCAGCAGCAGCAGCAG	2	.*..*.....	1	0.10	Chromosome XII	1011934	1011956	1011952	-CAA	*DIF1*	1	86	0	43
		CGCAGCAGCAGCAGCAGCAGCAG	1	*.....	20	0.22	Chromosome XIII	514628	514650	514628	-CAG	*YMR124W*	3	117	1	57
		AGTTGTAGCAGCAGCAGCAGCGG	3	..**.*.....	1	0.4	Chromosome XIV	34056	34078	34061	-CAG	*KRE1*	1	140	0	49
		GCCAGCAGCAGCAGCAGCAACAG	3	**.....*	0	0.16	Chromosome XIV	527488	527510	527489	-CAG	*VAC7*	1	108	0	52
		ACCGGCAGCAGCAGCAGCAGCGA	2	.*.*.....	1	0.03	Chromosome XV	111519	111541	111520	-CAG	*INO4*	1	102	0	56
		CTCTACAGCAGCAGCAGCAGCAG	4	**.**.....	0	0.07	Chromosome XVI	522050	522072	522053	-AC	*SWI1*	1	131	0	48
GAA_Cas9	AAGAAGAAGAAGAAGAAGAACGG	CAGAAGAAGAAGAAGAAGAGGAG	2	*.....*	2	0.16	Chromosome IX	128729	128751	128732	-CAA	*SIM1*	1	158	1	50
		AAGAAGACAAAGAAGGAAAAGGG	4	.....**.....*.*..	0	0.03	Chromosome IX	218236	218258	218247	+A	*RPN2*	1	109	0	46
		AAGAAGACAAAGAAGGAAAAGGG	4	.....**.....*.*..	0	0.03	Chromosome IX	218236	218258	218247	+T	*RPN2*	1	109	0	47
		AAGAAAAAGAAGAAGAAAATGGG	3	.....*.....*.*	0	0.41	Chromosome VI	224626	224648	224646	G/A->C/T	*QCR6*	47	178	6	65
		AAGAAGTAGAAGAAGAAGTAGAG	2	.....*.....*..	0	0.06	Chromosome VII	442210	442232	442220	+GAC	*SCW11*	1	149	0	55
		AAGAAGTAGAAGAAGAAGTAGAG	2	.....*.....*..	0	0.06	Chromosome VII	442210	442232	442220	A->G/T	*SCW11*	2	149	0	55
		AGGAAGGATAAGAAGAAGAAAAG	3	.*...*.*.....	0	0.07	Chromosome VIII	69001	69023	69021	+A	*YHL019W-A*	1	143	0	58
		AAGAAGAAGAAGAAGAAGAGGAC	1	.....*	8	0.19	Chromosome X	189027	189049	189030	-AGA	*MTC1*	1	142	2	57
		AGAAAGAAGAAGAAGAAGAAGAC	2	**.....	1	0.15	Chromosome XI	38635	38657	38637	-GAA	*intergenic:TR*	1	170	0	53
		AGGAGGAGGAAGAAGAAGAAGAC	3	.*..*.*.....	1	0.10	Chromosome XI	386884	386906	386891	-GAA	*TFA1*	1	113	0	49
		AAGAGGATGAGGAAGAAGAAGAG	3	...*..*..*.....	0	0.05	Chromosome XI	386902	386924	386909	+CACA	*TFA1*	1	118	0	47
		ACGAGGAAGATGAAGAAGAAGAG	3	.*..*.....*.....	0	0.04	Chromosome XI	613750	613772	613754	+T, G->C	*SRP40*	1	140	0	47
		AGGAAGATGAAGAAGAAGAGGAG	3	.*.....*.....*	0	0.06	Chromosome XI	613753	613775	613760	-GAA	*SRP40*	1	135	0	45
		CAGCAGAAGAAGAAGAAGAAGGA	2	*..*.....	1	0.02	Chromosome XII	287199	287221	287202	-AGA	*SIC1*	1	144	0	50
		AAAAAAAAGAAGAAATAGAAAGG	4	..*..*.....**....	0	0.00	Chromosome XII	1002226	1002248	1002231	+C	*intergenic:AT*	1	82	0	26
		AAGAAGAAGAAGAAGAAGAAGAC	0	.....	20	0.25	Chromosome XII	1031403	1031430	1031403	-AGA	*FPR4*	1	124	1	67
		GAGAAGAAAAAGAAGAAAAAAGC	3	*..*.....*.....*..	0	0.44	Chromosome XIII	331570	331592	331584	G->T, G->C, -A	*RSF1*	1	123	0	39
		AGGAGGAAGAAGAAGAGGAAGAC	3	.*..*.....*...	0	0.02	Chromosome XIV	51605	51627	51611	-AGA	*SKP2*	1	171	0	55
		AAGTGGAAGAGAAAGAAGAAAGG	4	...**....**.....	0	0.27	Chromosome XIV	60081	60103	60098	-A	*intergenic:BA*	1	197	0	63
		AAGAAGAAGAAGAAGAAGAGAGA	1	.....*	8	0.05	Chromosome XIV	515982	516004	515985	-AGA	*YNL058C/YN*	2	106	0	52
		AACGAGAAGAAGAAGAAGAGGAG	3	..**.....	0	0.04	Chromosome XV	315028	315050	315031	-AGA	*TOP1*	1	109	0	49
		AAAAAAAAGAAAAAGAAAAAGGC	4	..*..*.....*.....*..	0	0.48	Chromosome XV	430180	430202	430186	-AAG, -AG, -G	*intergenic:YC*	4	79	5	38
		AGGAAGAAGAAGAAAAAGATAGG	3	.*.....*....	0	0.45	Chromosome XV	553233	553255	553255	T->G, T->A, C->G	*LEO1*	21	140	1	43
		AAAAAAAAAAAAGAAAAAGAAAGC	4	..*..*..*.....*.....	0	0.45	Chromosome XVI	528948	528970	528959	-T, -TT, -TTT, +T	*intergenic:YF*	8	60	5	46
		AAGAAGAAGAAGAAGAAGAAGGA	0	.....	20	0.06	Chromosome XVI	774826	774848	774827	-CTT	*CLB5*	10	144	2	61
		AAGGAGACGTAGAAGAAGAAAGA	3	...*..*.*.....	0	0.01	Chromosome XVI	829679	829701	829699	+T	*intergenic:YF*	4	60	0	27
TGGAA_Cas9	TGGAATGGAATGGAATGGAACGG	TGATATAGAATGGAATGGAATGA	3	..**..*.....	0	0.03	Chromosome XV	256276	256298	256282	-TGGAA	*intergenic:PF*	1	100	0	35
GAA_Cpf1	TTTAGAAGAAGAAGAAGAAGAAGAAGA	TTTTGAAGAAGGAGAAGGAGAATA	3	.....*.....*....*.....	0	0.00	Chromosome III	268091	268117	268113	-T	*FIG2*	2	109	0	27
		TTTGGAAGAAGATGAAGAGAGAGA	4	.....*.....***.....	0	0.00	Chromosome XIII	414048	414074	414055	-G	*RCO1/YMR0*	1	98	0	62
CGG_Cas9	GGCGGCGGCGGCGGCGGCGGCGG	GGCGGCGGCGGTGCCGCTGGAGG	4	.....*..*..**..	0	0.03	Chromosome VI	105827	105849	105829	-G	*MDJ1*	1	149	0	59
		AGCGGCGGGGAGCGGCGGCTGGGG	4	*.....**.....*.	0	0.34	Chromosome XVI	465077	465099	465090	-G	*CAM1*	1	122	0	43

Mismatches: number of base mispairing between the guide and the off-target sequence

Mismatch position: asterisks show mismatches positions

MIT specificity score: the highest score (100) indicates a very specific guide

CFD specificity score: the highest score (1.0) indicates a very specific guide

Start and End: coordinates of the off-target site

Position: coordinate at which the highest rate of mutations was detected

Mutant reads: number of mutants reads within the off-target locus (between Start and End)

Total reads: total number of reads covering the Position

NR mutant reads: number of mutant reads in the NR library, within the off-target locus (between Start and End)

NR total reads: total number of reads in the NR library, covering the Position

Locus found mutated in only one read

Locus found mutated in more than one read

Efficacy of the TALEN in relevant myotonic dystrophy type I models

Introduction

Figure 17: Design of the TALEN$_{CTG}$. The TALEN induces a DSB into CTG repeats from the DM1 locus, located in the 3'UTR of the *DMPK* gene. In green: FokI catalytic domain. RVDs are shown in different colors depending on the recognized DNA base: A in pink, C in dark brown, T in light brown and G in yellow.

The TALEN$_{CTG}$ was shown to be active on DM1 CTG expansions of 70 to 90 repeats integrated into the *S. cerevisiae* genome. The nuclease was designed to induce a double strand break into the CTG tract (**Figure 17**). Upon induction, TALEN$_{CTG}$ triggers repeat shortening in 99% cases down to 4 to 10 repeats. The minimal spacer sequence required for the TALEN to dimerize is around 10 bp. It means that when the number of repeats is below 6 (recognized by left arm) + 2 (recognized by the right arm) + 3 (approximate spacer length) equal 11 repeats, the TALEN$_{CTG}$ can no longer induce a DSB, preventing shorter contractions. I have subsequently tested the TALEN$_{CTG}$ in DM1 patient cells (1) and in DM1 mice model (2). I have also set up a reporter assay in human cells to test other nucleases active on DM1 CTG expansions, and test their efficacy compared to the TALEN$_{CTG}$ (3).

Results

$TALEN_{CTG}$ effect on patient cells

Expression of the $TALEN_{CTG}$ in DM1 ASA cells

Preliminary experiments conducted in ASA cells showed that transfection efficiency using chemical reagents was very low, around 3-5% of cells expressing a plasmid containing the two $TALEN_{CTG}$ arms and a GFP reporter gene (Valentine Mosbach PhD thesis). Hence, in subsequent experiments, the $TALEN_{CTG}$ was expressed using lentiviral expression vectors, each arm expressed on a different vector encoding an antibiotic resistance gene. Antibiotic kill curve was carried out to determine the minimum concentration of an antibiotic that can kill all the cells in one week. Geneticin was used at 750µg/ml and hygromycin at 75µg/ml (**Figure 18.A**). Successful expression of each protein arm was determined by western blot after one week of selection following infection with the right arm of the $TALEN_{CTG}$. These cells were then infected with the left arm of the $TALEN_{CTG}$ (**Figure 18.B**). Protein expression was followed by Western blot since both arms of the $TALEN_{CTG}$ carries an HA tag. Different ratio of virus:cells (Multiplicity Of Infection, MOI) were tested and MOI=1 was found to result in higher expression and lesser mortality. Cells that are first transduced with higher MOI, upon second transduction with MOI as low as MOI=1 stopped growing.

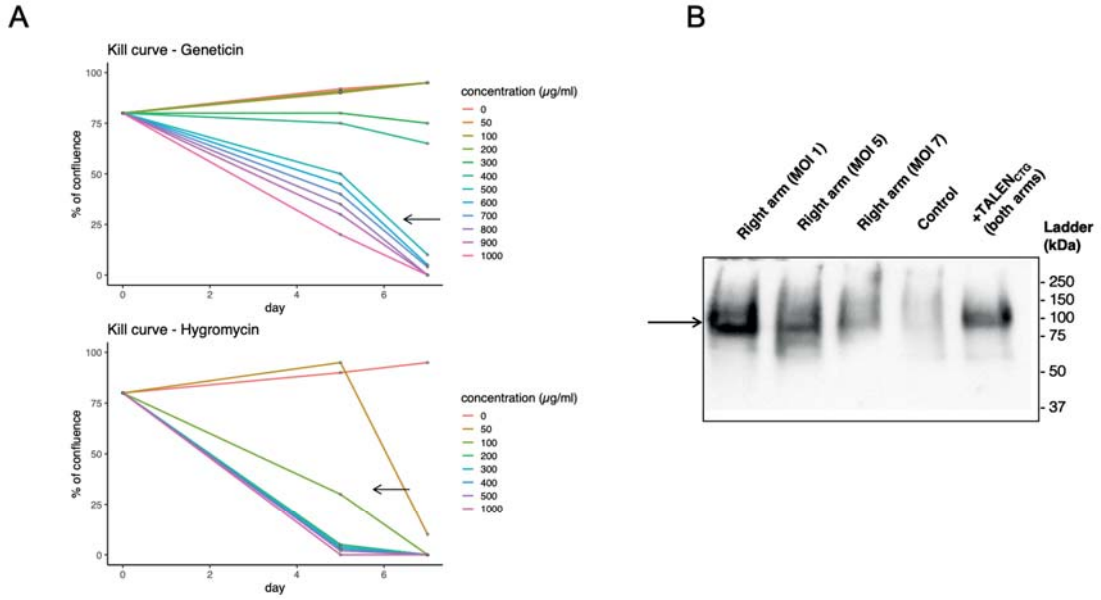
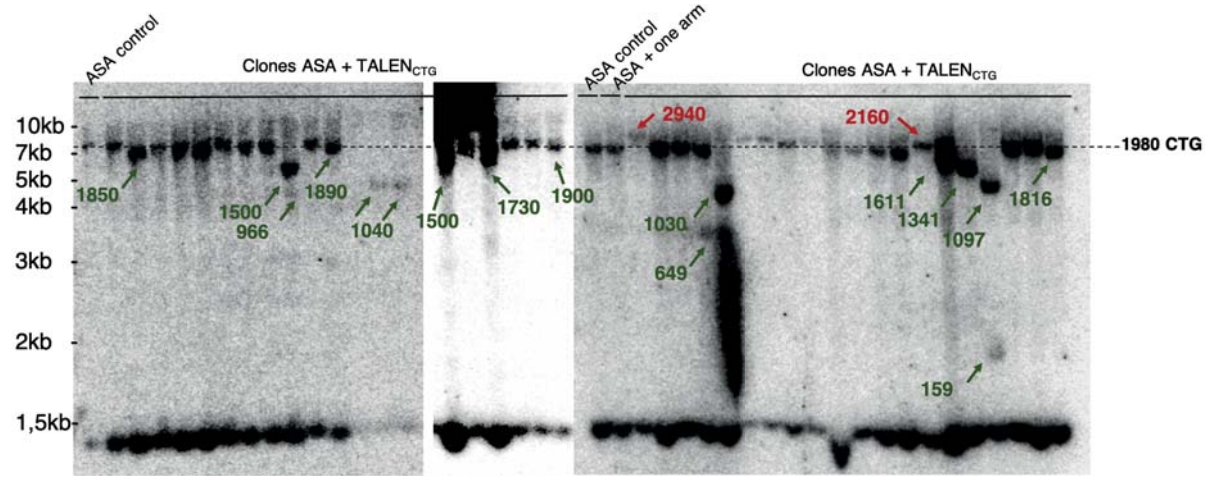


Figure 18.A: Kill curves on ASA cells of Geneticin and Hygromycin. Confluency is measured every 2 days for 7 days. Geneticin was later on used at a concentration of 750 µg/ml and hygromycin was used at 75 µg/ml. **1.B: Western blot of cell extracts after infection of viruses at different MOI.** Later on, MOI=1 was used as resulting in higher expression and lower mortality rate. The arrow indicates full arm length.

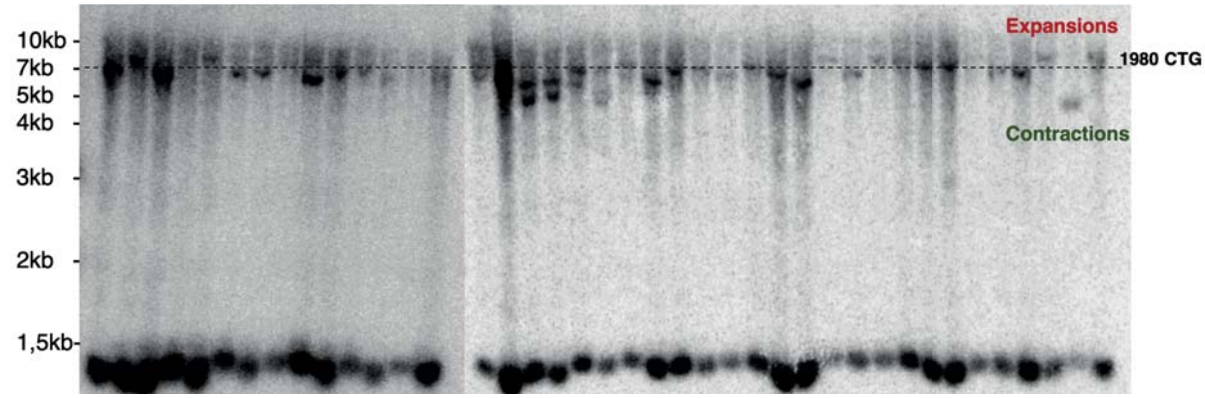
Effect of TALEN$_{CTG}$ on CTG repeat expansion length

ASA cells expressing both arms of the TALEN$_{CTG}$ were cloned. Each clonal population was expanded, DNA was extracted and CTG length was assessed by Southern blot. Fourty clones expressing both arms (**Figure 19.A**) and 41 clones expressing only the left arm, targeting the CTG repeat, were analyzed (**Figure 19.B**). ASA cells contain a short allele of around 30 CTG, migrating as a 1400bp band after *Bam*HI digestion and a large allele which size was estimated to be around 1980 CTGs and migrating at 7kb. Size range was determined for each clone in both conditions and was compared. In most cases, no size change was detected. Contractions and expansions were observed when only one arm or when two arms of the TALEN$_{CTG}$ were expressed. However, given the high instability in control cells, no statistical difference was found between both sets of data (**Figure 19.C**).

A



B



C

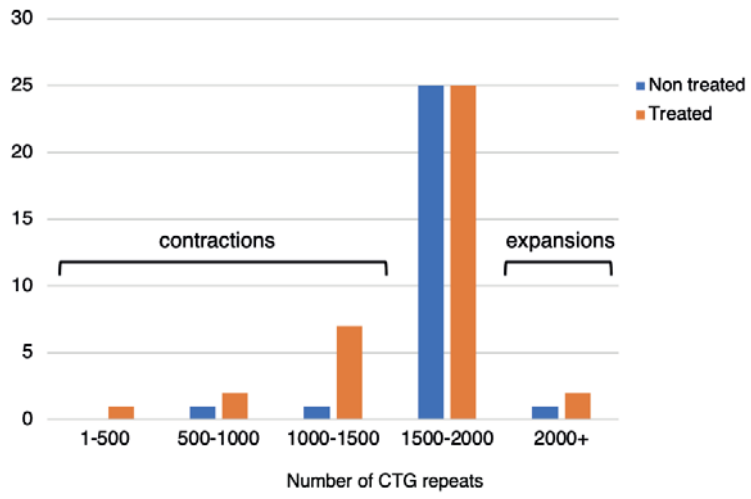


Figure 19: Southern blot of ASA clones expressing one or two TALEN$_{CTG}$ arms. A. Clones expressing two arms. 1980 CTGs is the starting allele length and was used as a baseline to quantify expansion (in red) and contraction (in green) events. **B: Clones expressing one arm.** Contraction and expansion events are visible. **C: Quantification of contraction and expansion events.** p-value >0.05.

Effect of TALEN$_{CTG}$ on foci number

The number of RNA foci in non-clonal populations expressing the TALEN$_{CTG}$ was compared to control untransduced ASA cells. An increase of the number of cells carrying one or zero foci was observed when expressing the TALEN$_{CTG}$ in replicate experiment 2 (**Figure 20**). However, no effect was observed in replicate 1. This may suggest that counting foci is not a reliable way to assess TALEN efficacy. Indeed, it was shown that foci tend to accumulate in the cells with age regardless of the number of triplets (Pettersson et al., 2015).

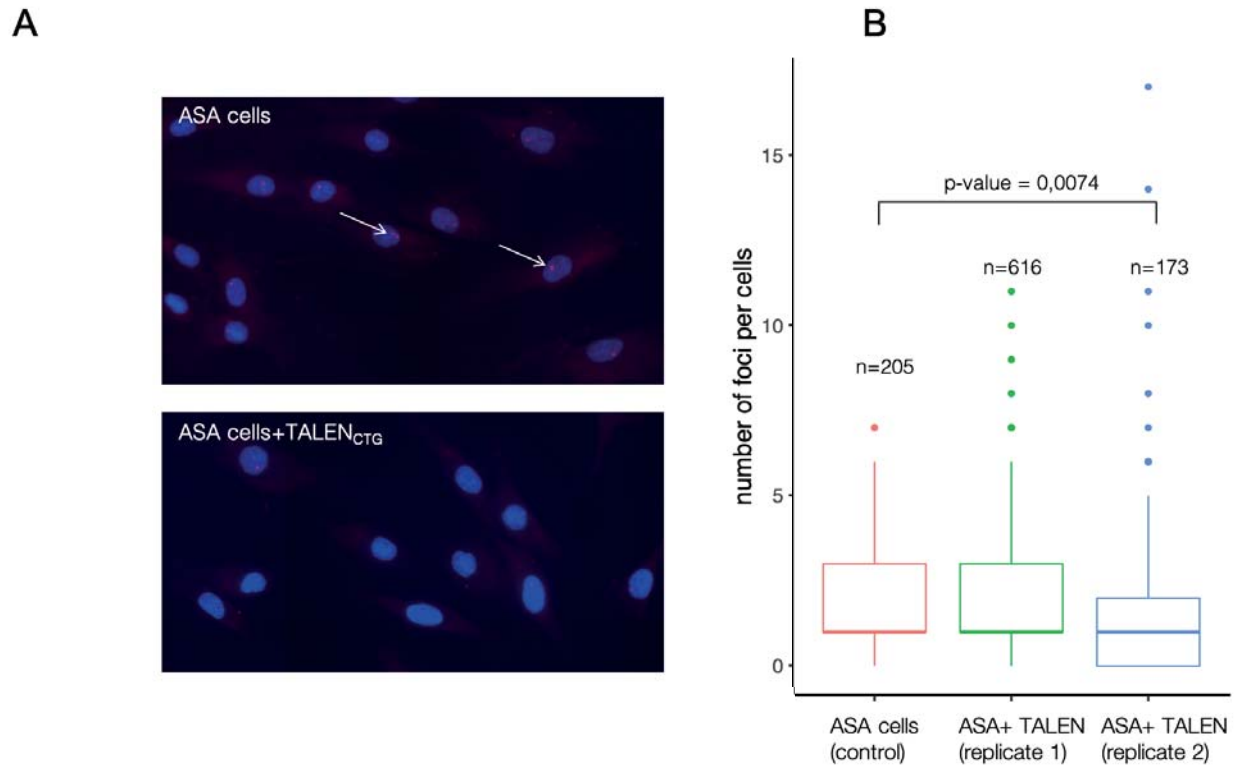


Figure 20: Foci number after TALEN$_{CTG}$ expression. A. Cells observed under fluorescence microscope (40x). White arrow: foci. **B: Quantification of foci number per cells.** Number of foci was quantified using imageJ. An enrichment in cells exhibiting 0 foci is observed.

Stability of CTG repeat over passaging in cultured cells expressing the TALEN$_{CTG}$

Five clones from Figure 19 were maintained for additional generations over one month and their size was assessed by Southern blot (data not shown). No further change in length was observed, suggesting that the lengthening induced by the TALEN$_{CTG}$ arose during the first days following TALEN$_{CTG}$ expression and that the length does not decreased afterwards.

TALEN$_{CTG}$ effect in DMSXL mice

All experiments with mice were supervised by Aline Huguet (Geneviève Gourdon lab). All mice experiments, from injections to sample processing were carried out with Olivia Frenoy, engineer in our lab. I obtained animal experimentation certification in order to be able to carry out these experiments.

Each TALEN$_{CTG}$ arm was packaged into recombinant AAV particles of serotypes 6 and 9 (hereafter called AAV6 and AAV9). Serotype 6 preferentially transduces muscles and serotype is routinely used for systemic injections (Zincarelli et al., 2008).

Expression of the TALEN in Tibialis anterior muscle of DMSXL mice

TA from mice after TALEN HA-tagged arm or PBS injection were dissected at different time points. Proteins were extracted and analyzed by Western blot to detect the TALEN arm (**Figure 21.A**). The TALEN arm is visible only one week after injection, the expression is lost at 2- and 3-weeks post injection. The HA tag is located in 5' of the TALEN arm. Bands below the full-length TALEN arm may be the result of protein degradation upon muscle collection and sample processing. However, when the same membrane was hybridized with an antibody against actin, no such bands were visible (data not shown). Hence, the degradation of the protein probably arose in muscle cells. This band pattern was also observed in yeast (control + in Western blot)) and was interpreted as a side-effect of overexpression of the protein (Mosbach et al., 2018). Loss of transgene expression has already been reported in muscles, probably due to promoter methylation resulting in silencing of the transgene (Brooks et al., 2004) (Nelson et al., 2019). In the present experiments, alternative hypothesis may be envisioned: toxicity and/or immunogenicity of the protein may lead to cell death resulting in loss of expression. Muscle histology after injection may give first answers regarding cell death following TALEN expression.

Assessment of muscle integrity after TALEN arm injection

TA from mice after either TALEN arm injection and PBS injection at different time points were stained by hematoxylin and eosin (H&E) to visualize muscle integrity (**Figure 21.B**). Many central nuclei were observed at 2 and 3 weeks when either TALEN arm is expressed. No central nuclei were observed when PBS was injected. Central nuclei are indicative of newly formed fibers. The nuclei slowly migrate to the periphery of the fiber. Hence, TALEN arm expression leads to the degradation of the muscle cells which die and are replaced by new cells that do not

express the TALEN arms. Two hypotheses can explain this result: the protein is toxic and/or the protein is immunogenic, in such way that it is the immune system that kill nuclease-expressing cells.

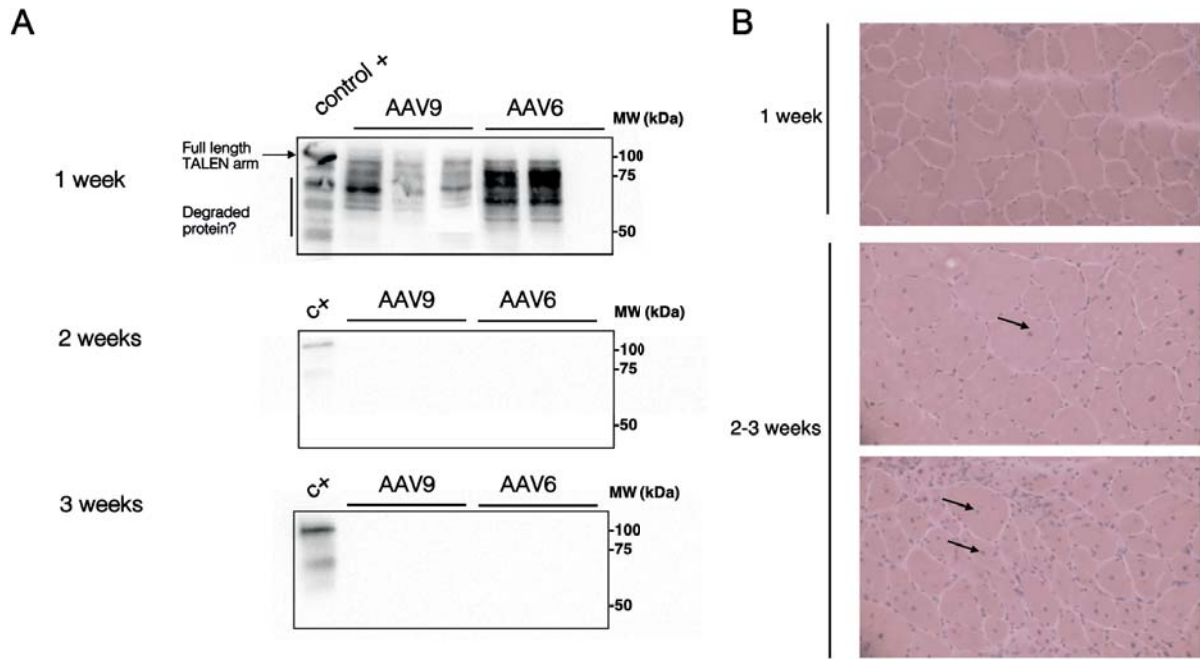


Figure 21: Expression of $TALEN_{CTG}$ in one-month heterozygous DMSXL mice. 3.A: Expression of HA-tagged TALEN arm followed by Western blot. AAV9 serotype was injected for a total amount of 4.10^{11} vg and AAV6 at 8.10^{10} vg. Positive control is a yeast extract expressing the TALEN. **3.B: H&E stained transversed muscle section.** At 2- and 3-weeks post injection, central nuclei (black arrows) are visible in 100% of the cells.

Dose response effect of AAV6 and AAV9 serotypes encoding TALEN arm

The HA-tagged TALEN arm carried by either serotype 6 or 9 was injected in the TA at lower doses 10^{10}, 10^{9} or 10^{8} vg. Expression of the TALEN arm in each condition in two mice was measured by Western blot at one and three weeks (**Figure 22**). The same pattern additional bands at lower molecular weight was observed. For AAV6 serotype, 10^{8} vg was not enough to induce any protein expression, 10^{9} and 10^{10} vg resulted in protein expression, lost at 3 weeks, probably due to cell death as observed in **Figure 21**. For AAV9, 10^{8},10^{9} and 10^{10} vg were not enough to induce any protein expression in TA muscle. No lower dose was selected that allows for sustained expression for 3 weeks.

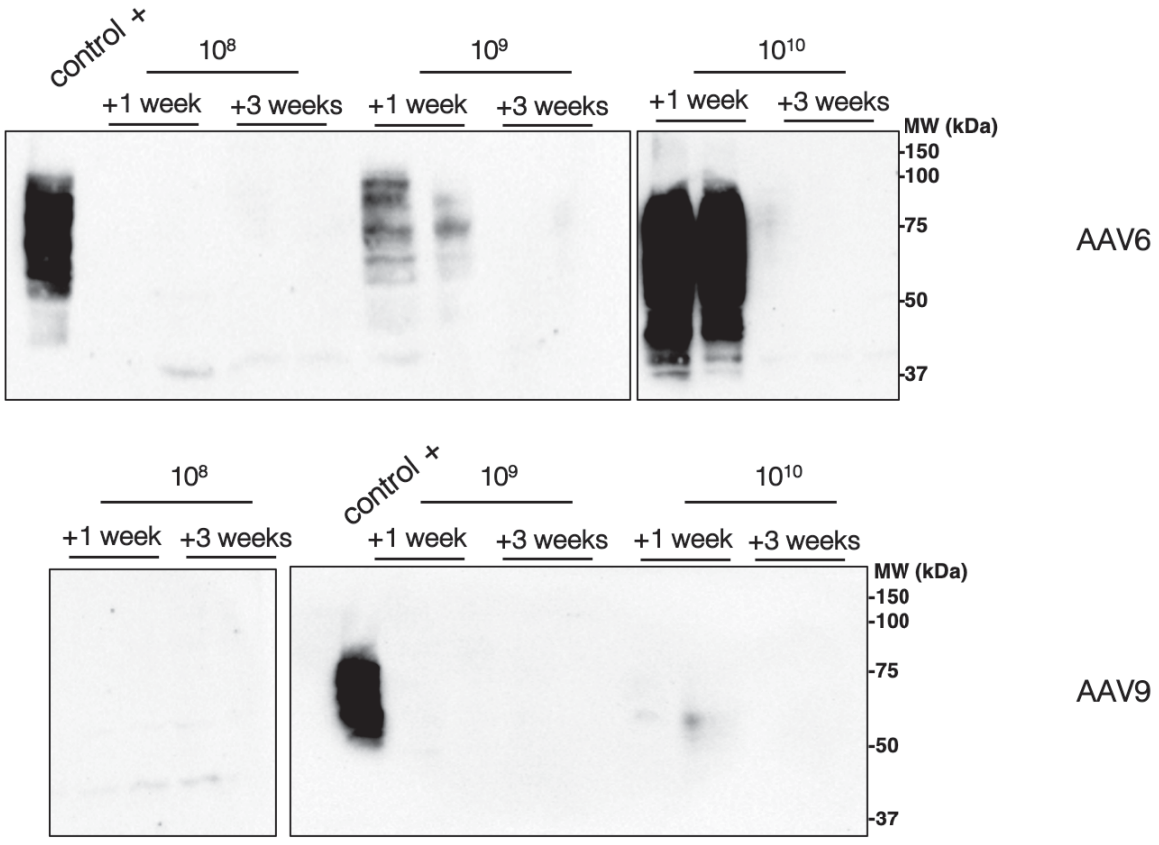


Figure 22: Expression of $TALEN_{CTG}$ in one-month heterozygous DMSXL mice after 1 and 3 weeks. Lower doses were injected: 10^{8}, 10^{9},10^{10} vg in one TA for each serotype. AAV6 is too low at 10^{8} to be detected. Higher concentrations are toxic after 3 weeks. AAV9: tested doses are all too low to be detected in the muscle. Positive control is a yeast extract expressing the TALEN.

Effect of TALEN expression in neonatal mice

TALEN was injected in neonatal mice to circumvent immune response since mice at this age do not have a developed immune system yet. This method has been used many times to the same purpose (Carlon et al., 2014)(Nelson et al., 2019). After systemic injection, expression of the HA-tagged TALEN arm was assessed by Western blot in the heart, muscle and liver of mice. The arm was expressed in the heart and at different levels after one week (**Figure 23**). With AAV6, no mice died after 1 or 3 week. The protein is clearly visible in both cases and mice are still alive. With AAV9, mice died between 1 and 3 weeks. The protein is expressed at higher levels than with AAV6 and at 1 week mice are alive. This suggests that protein level is responsible for cell death, either due to toxicity or immunogenicity. AAV9 work better for systemic injection which explains higher proteins levels observed in the heart (Zincarelli et al., 2008).

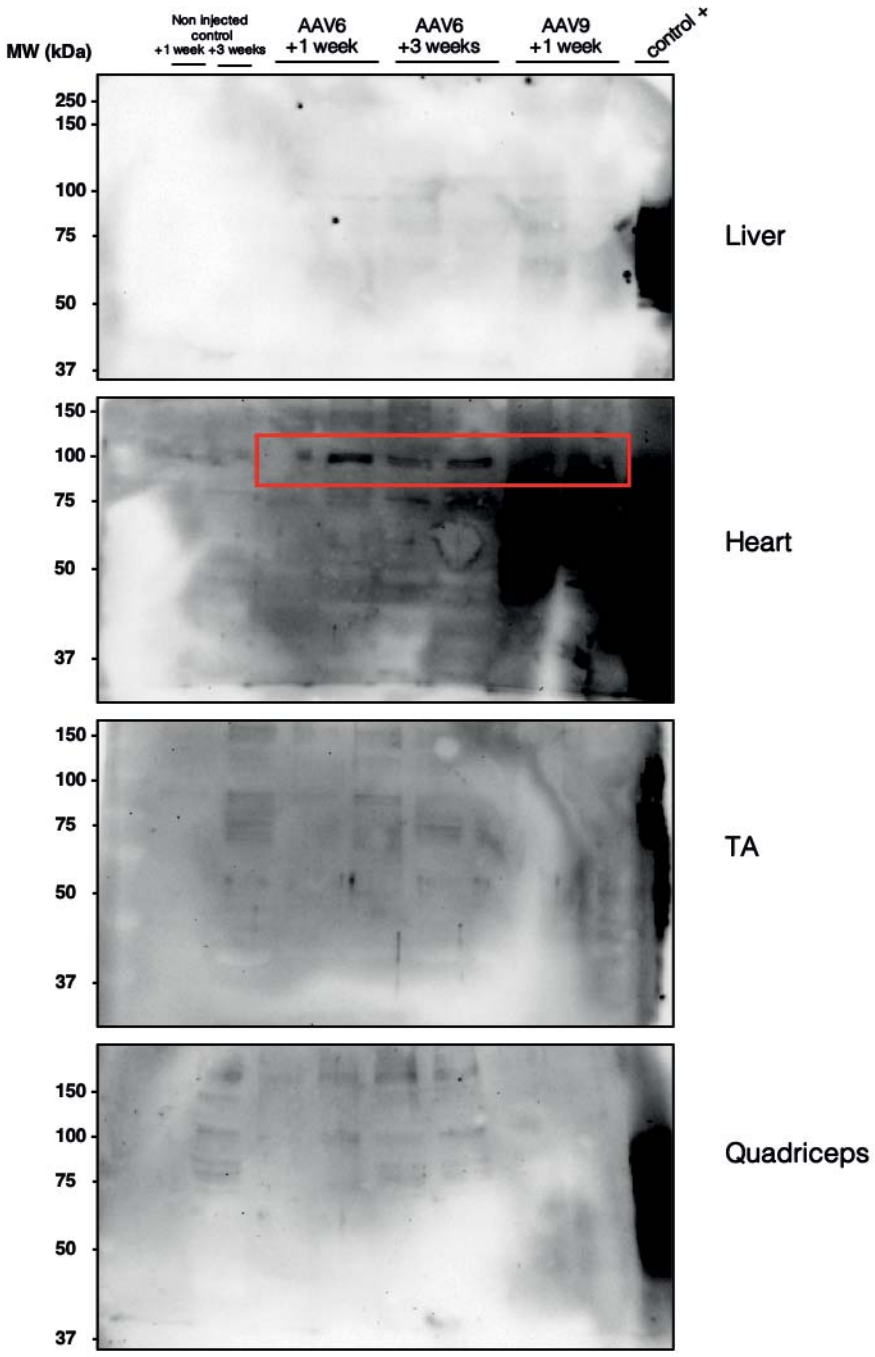


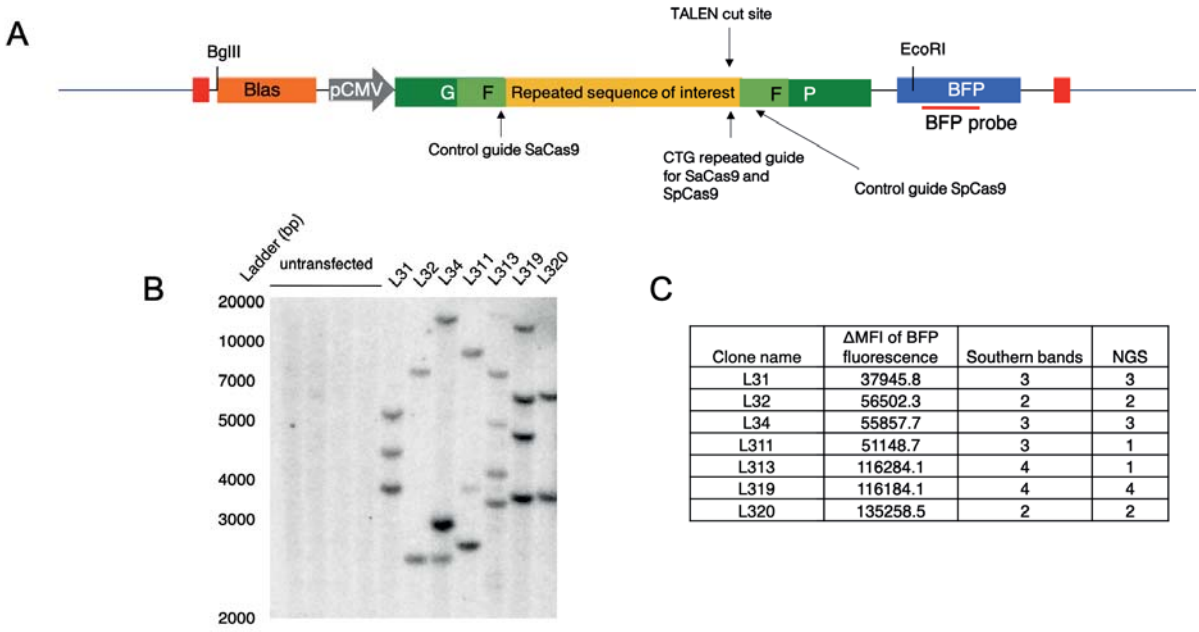
Figure 23: Expression of TALEN$_{CTG}$ in neonatal heterozygous DMSXL or WT mice 1 and 3 weeks post-injection in four tissues. Doses injected were: 4.1010 vg for AAV6 and AAV9 for 1.8.1011. Expression of the protein is found in the heart after 1 week in one AAV6 injected mouse and sustained till 3 weeks. Higher expression was found for AAV9 at 1 week. The two other injected mice died before reaching 3 weeks post-injection.

A human cell reporter assay to test nucleases on DM1 CTG expansions

Following yeast experiments, I wondered whether the same system could be transposed in human cells and to determine to what extent results in yeast are transposable to human assays. In this part, I constructed a reporter cell line for CTG repeats in HEK293FS cell line. This work was done at Sanofi R&D, Vitry-sur-Seine.

Establishment of clonal cell lines expressing the GFP cassette and 200 CTG repeats

A GFP cassette flanking 200 CTG from a DM1 patient was cloned into a PiggyBac transposon plasmid (**Figure 24.A**). HEK293FS cells were electroporated with two plasmids: one containing a transposase gene and one carrying the sequence of interest flanked by two inverted terminal repeat sequences (ITR). When the two plasmids are expressed in the same cell, the transposase recognizes ITRs and catalyzes the integration of the cassette flanked by ITR into TTAA chromosomal sites. Pool of cells was cloned and expanded as clonal populations. DNA was extracted and analyzed by southern blot to measure the number of integration sites (**Figure 24.B**). For each clone, BFP fluorescence intensity, and integration site was determined (**Figure 24.C**). Integration sites were determined by transposon insertion site sequencing (TIS-Seq) by Veeranagouda Y. at Sanofi (Veeranagouda and Didier, 2017). This method relies on the targeted amplification of regions of the genome bearing repeated ITR sequences used by the transposon to integrate into the genome.



Clone name	ΔMFI of BFP fluorescence	Southern bands	NGS
L31	37945.8	3	3
L32	56502.3	2	2
L34	55857.7	3	3
L311	51148.7	3	1
L313	116284.1	4	1
L319	116184.1	4	4
L320	135258.5	2	2

Figure 24: Establishment of a stable cell line with a reporter cassette for DSB repair into CTG repeats. A. Cartoon of the cassette integrated. In red: sequences for transposon integration; Blas is a resistance gene to blasticidin; pCMV is the promoter of Cytomegalovirus; two halves of GFP flank 200 CTGs from a DM1 patient. BFP is the gene encoding the Blue Fluorescent Protein which expression can be quantified by flow cytometry. Sites targeted by CRISPR-Cas9 in subsequent experiments are indicated by black arrows and labelled. **B. Southern blot for the verification of the number of integration sites of the transposon cassette.** The probe used was the BFP probe which location is indicated in A. Genomic DNA was digested by BglII and EcoRI as indicated in A. **C. Table summarizing the characterization of the established cell lines.** MFI: Mean fluorescent intensity. NGS: Next Generation Sequencing, performed at Sanofi using a TIS-Seq method.

B

Nuclease	Number of off-target sites in the human genome
$TALEN_{CTG}$	0
CRISPR-SpCas9	178 (58 exonic, 80 intergenic, 40 intronic)
CRISPR-SaCas9	1 (exonic)

Figure 25: Gene edition of the CTG repeats at the DMPK 3' UTR locus. A. Sequence targeted and gRNA design, for SpCas9 (up, -NGG PAM) and SaCas9 (bottom, -NNGRRT PAM). PAMs are underlined. **B. Number of off-target sites**, compared to $TALEN_{CTG}$ in the human genome. Calculated with CRISPOR online tool (Concordet and Haeussler, 2018).

Induction of SaCas9 on clones carrying different transgene copy number

In silico simulations of potential guides and Cas9 at the DMPK 3'UTR locus pointed out to SaCas9 as having the less potential off-target sites (**Figure 25**). Hence, SaCas9 was expressed alongside a guide cutting into CTG repeats or a guide cutting the non-repeated sequence between GFP halves containing homology regions. CTG repeats are from a DM1 patient sequence containing ≈200 CTGs. Integration events with PiggyBac transposon system are random and copy number of the integrated transgene can vary. SaCas9 + guides were expressed into each clone. Various levels of GFP+ cells % were observed (**Figure 26**). DSB induction into CTG repeats repaired by Single Strand annealing resulted in functional GFP reconstitution.

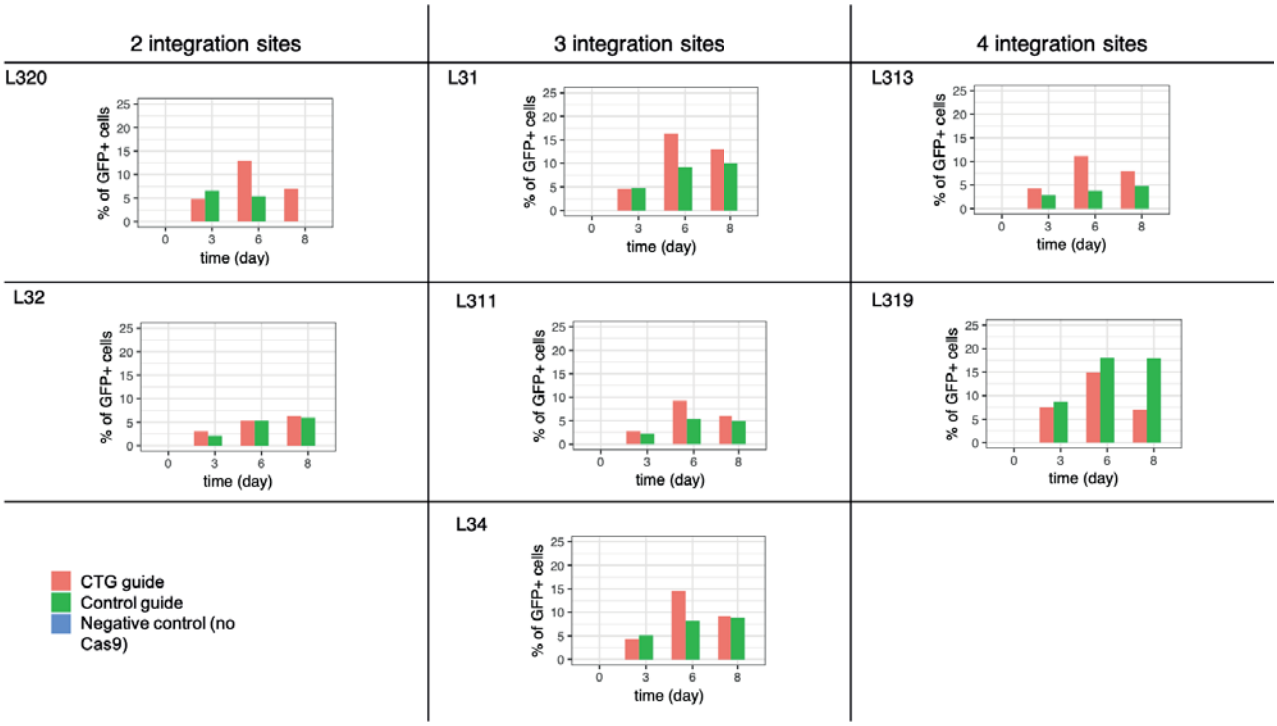


Figure 26: Comparison of SaCas9 efficacy with control or CTG guide in seven clones. Fluorescence was measured on the day of the transfection and +3, +6, +8 days after transfection.

Comparison of nuclease efficacy in the L320 clone

SaCas9, SpCas9 and the TALEN were expressed in the L320 clone carrying two copies of the transgene. SpCas9 was expressed alongside two different guides: one CTG guide, to induce an DSB in CTG repeats, and one control guide, inducing a DSB next to the CTG repeat. SaCas9 was expressed alongside three different guides: one CTG guide, to induce an DSB in CTG repeats, one control guide, inducing a DSB next to the CTG repeat, and one additional control,

inducing a DSB into the AAVS1 locus (**Figure 24.A**). Various efficacies were observed, with SaCas9 being more efficient than SpCas9 on CTG repeats (**Figure 27**). It is different from what was observed in yeast (First section in Results). Additionally, both TALEN arms were able to cut the repeat and trigger recombination, although at lower levels than Cas9. It is in agreement with previous experiments (Valentine Mosbach PhD thesis) where SpCas9 was more efficient than the TALEN at inducing a DSB, as measured by Southern blot. Detailed flow cytometry dot plots are in annex 5.

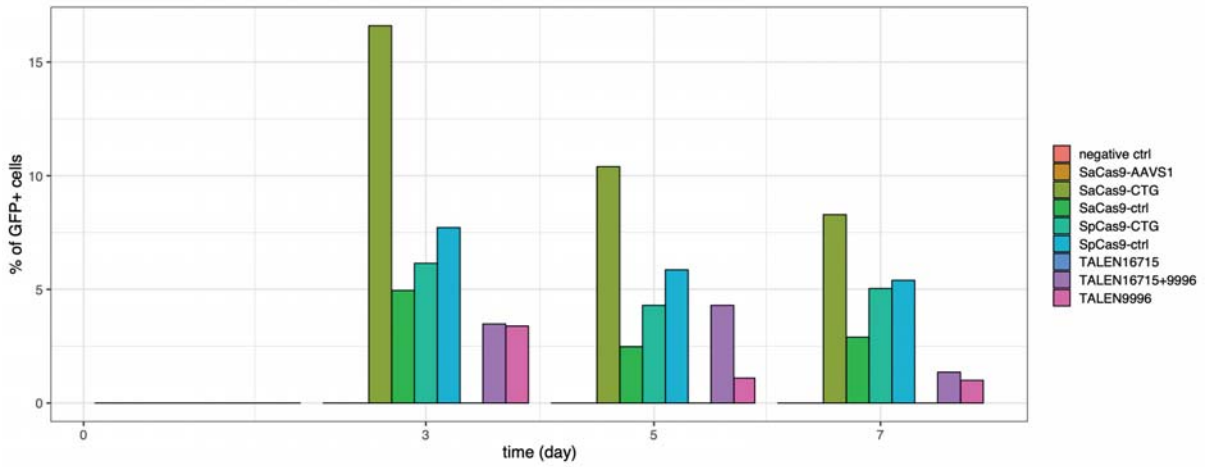


Figure 27: Comparison of TALEN, SaCas9, SpCas9 in L320 clone. Fluorescence was measured on the day of the transfection and +3 , +5, +7 days after transfection.

Discussion

During the course of my thesis, I worked on three different but interconnected projects.

(1) I constructed a yeast GFP reporter assay to quantify the efficacy of Cas nucleases on repeated sequences. Ten microsatellites involved in human disorders were inserted between two overlapping GFP halves. Upon nuclease induction into the repeated sequences, cells repaired by homologous recombination reconstituting a functional GFP that could be quantified. This work highlighted the fact that GFP positive cell count is a good readout of DSB efficiency measured by Southern blot and so can be used as a screening method to assess nuclease efficacy on various substrates. The analysis of efficacy differences of nucleases on structured microsatellites revealed that the main modifier was the structure formed by the gRNA. The more stable the gRNA is, the less likely the nuclease will induce a DSB into the repeat.

(2) I also tested the $TALEN_{CTG}$ effect on DM1 models in patient cells and in mice. It was found to induce contractions in patient cells, up to -2000 CTG, but the basal instability level of expanded CTG in cultured cells over time led to a non-significant difference from control cells. In mice, issues with the expression and potential toxicity of the $TALEN_{CTG}$ still need to be solved before assessing the nuclease effect on CTG repeats.

(3) Finally, I constructed a human cell model in HEK cells to quantify nuclease efficacy on DM1 CTG expansions based on the same reporter assay used in yeast. Using this model, I showed that SaCas9 was more efficient that SpCas9, itself more efficient than the $TALEN_{CTG}$ to induce a DSB in CTG repeats.

Secondary structure effect on Cas9 efficacy

Secondary structures may impede the access to the genome and hinder recognition. Additionally, since the gRNA is complementary to the sequence targeted, the gRNA secondary structure itself can destabilize Cas9 interaction to its gRNA. The sgRNA plays a crucial role in orchestrating conformational rearrangements of Cas9 (Wright et al., 2015). Stable secondary structure of the guide RNA as well as close state of the chromatin negatively affect Cas9 efficiency (Chari et al., 2015; Jensen et al., 2017). Possible secondary structures formed by the guide RNA are important to determine nuclease activity although there is no clear rule that can be sorted out and it is still challenging to know which hairpins will be detrimental (Thyme et al., 2016). In addition, improperly folded inactive gRNAs could be competing with active and properly folded gRNAs within the same cell, to form inactive or poorly active complexes with Cas9, that will inefficiently induce a DSB (Thyme et al., 2016).

Is it possible to treat other microsatellite disorders by this approach?

The most efficient Cas9 on microsatellite disorders were determined using a GFP reporter assay. Hence it appears that these structured microsatellites can be cut, except GGCCTG repeated sequence which appears to be cut at very low levels. Therefore, targeting human pathological expansions will depend on the surrounding sequences: a PAM must be present and *in silico* simulations will help to determine whether the guide RNA has a unique target in the genome. At this point, different Cas9 and gRNA pairs may be candidates. The yeast model assay will help determine which nucleases are more promising.

For DM1, I constructed an HEK cell model to test the efficacy of any nuclease on the DM1 endogenous locus, carrying 200 CTGs. This sequence was integrated between two GFP halves and nuclease efficacy is given by the percentage of GFP+ cells. The results of the nuclease testing, although still preliminary in HEK cells does not confirm results from yeast pre-screening. Indeed, on CTG repeats, SpCas9 is more efficient than SaCas9 in yeast while in HEK cells it is the opposite. The difference might be due to the sequence that is different, 33 repeats in yeast versus 200 repeats in HEK cells would indicate that SpCas9 is more sensitive to an increased repeat number than SaCas9. Using the HEK cells assay, we could screen for the best nuclease on the DM1 locus, the same approach could be applied to other microsatellites.

The best candidates, TALEN and SaCas9 may then be tested in a relevant mice model. For DM1, this model would be DMSXL as these mice carry the endogenous DM1 locus and 1000 CTGs and exhibit disease phenotype which would be monitored when the nuclease is expressed

to give insights into the nuclease potential as a therapeutic treatment. A complete screening pipeline now exists for DM1.

What is the mechanism underlying DSB repair into expanded microsatellites?

In yeast, a DSB into CTG repeats is repaired by SSA indicated by the essential role played by *RAD50, SAE2,* and *RAD52* at repairing the break. The repair between annealed stretches of repeated sequence leads to its contraction. Secondary structure formation also impedes resection (Mosbach et al., 2018)(Mosbach et al., 2019b). The model in HEK cells could be used to see whether these results hold true in human cells. By using RNA interference to silence genes involved in resection and DNA repair, such as CtiP, BRCA1, BRCA2, we would be able to dissect the role of resection in DSB repair at CTG expansions. Alternatively, the use of a CRISPR-based genome wide deletion library (Sanson et al., 2018) could be used to find which genes enhance or impede DNA repair at CTG repeat tract.

Cell to cell variations in DSB repair

Until now, all experiments designed to study DNA repair in eukaryotic cells were considering large cell populations on the order of millions of mammalian cells, or billions of yeast cells. However, there may be cell to cell differences in the way a DSB is processed and repaired. Preliminary experiments using droplets conducted with Antoine Barizien in Charles Baroud lab were carried out. The experiments were performed on a microfluidic chip which consists of a chamber containing 1495 cubic holes. The loading protocol consisted of priming the chip with oil, then injecting the yeast cell suspension. Finally, oil was pushed again in the chip to form the aqueous droplets trapped in the holes, where they remained for the duration of the experiment. Using this assay, we could: (i) visualize cell-cycle arrest following DSB induction, (ii) see whether repaired cells invade the culture, and (iii) by comparing yeast cells with various microsatellite see whether repair kinetic is different depending on the repeat. Additionally, it would be of interest to more precisely characterize the DSB repair not only at CTG repeats but also in other microsatellites. I transformed the cells with Cas9 plasmids as described in Section 1 of Results, and brought them to Antoine. Together we loaded the chip. He analyzed the results. Yeast cells containing GFP reporter gene can be studied at single cell level to dissect DSB repair. First experiments were carried out using the CTG repeat containing strain. The growth of each population in each droplet was followed by time-lapse microscopy (Barizien et al., 2019). Cell growth was inferred from the area occupied by the cell in one droplet at different

time points. Strikingly a wide variation from one population to another was observed, from droplets where the cell did not divide to saturated droplets (**Figure 28)**.

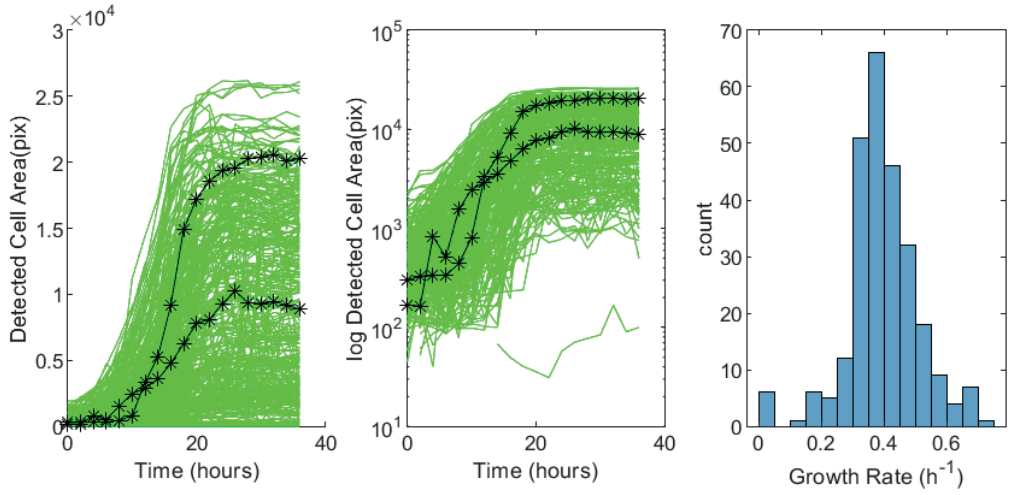


Figure 28: Growth curves of yeast cells from the LPY110 strain transformed with SpCas9 and gRNA targeting CTG repeats. Two days after the transformation, one colony was picked and loaded into a microfluidic chip. Detection of the population of yeast cells in the droplets for 240 droplets with example curves highlighted in black. The first graph shows the area of the cells, the second graph is the conversion in log scale and the last graph is a histogram of the growth rates measured in the droplets. Growth rate corresponds to the doubling time. The growth rate mean is 0.4h−1. Antoine Barizien analyzed microfluidics data and did these graphs.

Is the TALEN$_{CTG}$ a good candidate as a genome editing treatment for DM1?

In yeast the mechanism and the outcome of TALEN$_{CTG}$ is clear and reproducible: single-strand annealing among repeats, probably through multiple rounds, leads in 99.9% of the cases to a complete contraction of the repeat tract.

Results in patient cells carrying 2000 CTGs where the induction of the TALEN leads to contraction events up to a reduction of 5kb are encouraging. However, the TALEN$_{CTG}$ is not as efficient as in yeast on shorter repeats (up to 80 repeats). This may indicate that long repeats are more difficult to target. This can be due to secondary structures impeding the recognition and cutting of the nuclease. Our former data (Mosbach et al., 2018) suggest that secondary structures formed by various microsatellites including CTG, can hinder Cas9 activity. Additionally, it was shown that long CTG repeats can induce heterochromatin formation (Otten and Tapscott, 1995), which would further impede nuclease DSB induction. It is also possible that the delivery method, by lentiviral transduction may not be the most appropriate way to express the TALEN. It was shown in HeLa cells that when a TALEN was expressed from lentiviruses, the expression was quickly silenced and rearrangements between recombinogenic

sequence formed by consecutive RVDs led to its complete silencing (Holkers et al., 2013). It is possible to envision a delivery as purified proteins, or testing different constructs using different or inducible promoters may facilitate its expression. I was able to check by Western blot the good expression of the HA-tagged TALEN arm after a long time but I cannot rule out that the protein may be incomplete, lacking a few repeats, not visible on a Western blot. Since the stability for several weeks of the repeat tract remains unchanged in clone expressing the TALEN, it may be possible that the expressed nuclease is no longer able to cut and that contractions or expansions events occurred earlier, soon after TALEN expression. There may also be a selection against full-length TALEN in cell population, since these cells would not undergo DSBs and would have a proliferative advantage.

We would like to perform these experiments in cells carrying less CTG triplets, as the TALEN was found to be efficient on shorter repeats in yeast. However, availability of such models is limited and the cell line tested so far by Olivia Frenoy – DM300 primary skin fibroblasts, patient immortalized lymphoblasts- were found difficult to use due to their fragility and difficulty to be transduced or transfected.

Results in mice revealed technical issues about the expression of the TALEN *in vivo*. We were not able to obtain a sustained expression of the nuclease to be able to assess its effect on CTG repeats. Cells expressing the TALEN died after a short expression time (1 week) and are replaced by newly formed cells not expressing the TALEN (visible in the muscle by cells exhibiting central nuclei, **Figure 21**). Whether it is due to intrinsic toxicity of the protein or due to its immunogenicity is unknown. In any case it is not due to its ability to induce DSBs as cell mortality is the same whether expressing one or two arms. To circumvent any immune response and have the TALEN expressed for a longer time, it was injected in neonatal mice, previously described to tolerate immunogenic proteins due to the immaturity of their immune system which will ignore immunogenic proteins. Mice either died when the TALEN was expressed at very high levels (AAV9-1week, **Figure 23**) or survived to 3 weeks when the TALEN was expressed at lower levels (AAV6-1 and 3 weeks, **Figure 23**). It is thus difficult to conclude, both toxicity and immunogenicity can explain this result. To rule out the involvement of the immune system, immune cell staining to reveal markers of inflammatory cell infiltration (macrophages, neutrophils, interleukin 1ß and 12ß) located near the cells expressing the TALEN would have to be performed. As for the likely toxicity of the TALEN, dose lowering was not enough to lower toxicity. AAV6 and AAV9 are both inducing toxicity. It is possible that the promoter is too strong using a milder or an inducible promoter may help reducing the mortality. Testing alternative ways of expression might improve TALEN expression.

An alternative would be to use SaCas9, since it has only one off-target when targeting DM1 CTG expansion (**Figure 25**). SaCas9 has already been used to do genome editing in muscle and its toxicity and immunogenicity are still tolerable. Ongoing experiments in ASA cells expressing SaCas9 will indicate whether Cas9 is more efficient than the $TALEN_{CTG}$. In the HEK cell model, SaCas9 is more efficient than the $TALEN_{CTG}$ as GFP+ cells number is higher (**Figure 27**). We are also currently producing rAAVs containing SaCas9 and the CTG guide for expression in DMSXL mice.

The efficacy and outcome of $TALEN_{CTG}$ (Mosbach et al., 2018)(annex 3) and SpCas9 (Mosbach et al., 2019b)(annex 4) were compared in the same yeast model containing DM1 CTG repeats integrated at the SUP4 locus. Both nucleases were targeted to cut at the same position, inside the CTG repeat tract. Induction of the $TALEN_{CTG}$ lead to contractions in 99% of cases while SpCas9 lead to large deletions and rearrangements. Since most recombination events involved recombination between repeated Ty1 elements, it may be hypothesized that the high recombination rate observed with SpCas9 is particular to the SUP4 locus, which is surrounded by Ty1 retrotransposon LTRs. The difference in the repair repair outcome between the two nucleases is striking. Single-molecule experiments fluorescently labeled DNA and SpCas9 revealed that SpCas9 remained bound to the two DNA ends after cleavage. This was confirmed by biochemical gel shift assays (Sternberg et al., 2014). Although no evidence exist so far to confirm this finding *in vivo*, it may explain the high recombination observed of the SUP4 locus: the break is not repaired due to Cas9 binding and entering into mitosis results in dramatic rearrangements. Even though no such rearrangements using the yeast GFP reporter assay was ever observed with any repeated sequence (Poggi et al., 2019), special care should be taken while envisioning using Cas9 for therapeutic purposes, notably by extensively characterizing repair outcomes.

Finally, off-target effects need to be carefully assessed before envisioning a therapeutic assay. In yeast the $TALEN_{CTG}$ is very specific and sequencing of yeast cells expressing the $TALEN_{CTG}$ did not reveal any mutation (Mosbach et al., 2018). Additionally, *in silico* simulations in the human genome did not reveal any potential off target. In order to test for off-targets, techniques such as GUIDE-Seq should be applied to patient cells expressing the $TALEN_{CTG}$.

Sequencing DM1 patient cells expressing the $TALEN_{CTG}$

There is an ongoing project in the lab: A PacBio sequencing of different clones expressing the $TALEN_{CTG}$ after a long time and showing either contractions, expansions or no change (**Figure 19.C**). It would give several insights, and this would be the first DM1 patient genome to be ever

sequenced. (i) we will assess with more precision the length of the CTG repeats and quantify the exact number of repeats that were shortened. The sequencing technology was chosen because it enables the sequencing of long reads which is the only way to accurately retrieve a repeated sequence. Repeated sequences have always been difficult to analyze in any genome, particularly with short reads like those generated by Illumina. In a preliminary run, we were able to retrieve one read of more than 10 kb indicating that untransduced ASA cells carry 3060 CTGs. (ii) We would detect additional instability in other repeats elsewhere in the DM1 genome. (iii) We will check the sequence integrity of the integrated $TALEN_{CTG}$ arms, and control for possible recombination events between RVDs. (iv) We would be able to detect off-target effects by revealing single nucleotide polymorphism in the genome, compared to the untreated reference clone, if unfaithful repair after off-target DSB was made by the $TALEN_{CTG}$. (v) Finally, sequence modifiers in the genome of this DM1 patient may be linked to the CTG instability; however, to be able to find significant genetic linkage, the sequencing of many different DM1 families will be needed.

What is the future of in vivo genome editing for DM1?

Many clinical trials involving ZFN, TALEN or CRISPR-Cas9 are ongoing. Regarding gene editing approach for DM1, many obstacles remain before going into clinical trials. Two main approaches exist to remove pathological expansions of CTGs: (i) cutting upstream and downstream the repeat (van Agtmaal et al., 2017; Lo Scrudato et al., 2019; Provenzano et al., 2017) or (ii) cutting inside the expansion and promote contraction of the repeat by homologous recombination (SSA or gene conversion) (Mosbach et al., 2018). The first approach was proven to work in patient cell models, alleviating pathological phenotype including RNA foci decrease, restored splicing of affected genes. In DMSXL mice, successful deletion of the CTG was achieved, resulting in a reduced number of RNA foci. Both *in vitro* and *in vivo* assays revealed that the genome editing efficiency was rather low: 14% (Provenzano et al., 2017), 46% (van Agtmaal et al., 2017) and 6 to 9% in mice (Lo Scrudato et al., 2019). The low efficacy may be an issue for treatments in human, especially if sustained expression of the nuclease is required, increasing the risk for off-target mutations. Regarding undesired modifications, the three studies looked at the top *in silico* predicted off-targets and did not find any mutation. On-target mutations were heterogenous with indels frequently observed. More extensive assessments of the nuclease off-targets using deep-sequencing methods will be required for future applications. A recent paper showed that long-term expression of Cas9 in mice led to many unwanted modifications, in a bigger proportion than on-target mutations (Nelson et al., 2019).

Nevertheless, mice were alive and did not show any sign of premature deaths. The second approach was tested during the course of my doctoral research using a TALEN. The TALEN was moderately active in patient cells inducing contraction events. In mice, the delivery and the expression of the nuclease were challenging and did not enable to confirm that the shortening observed in yeast and in patient cells was reproducible *in vivo.* Experiments in DMSXL mice revealed the difficulty to express a nuclease and characterize its effect. Toxicity and immunogenicity of exogenous proteins expressed *in vivo* have been known since the beginning of gene therapy. Cas9 elicits immune response (Nelson et al., 2019) and human being, contrary to mice bred in a confined environment have neutralizing antibodies toward SpCas9 and SaCas9 (Charlesworth et al., 2019) and probably towards other CRISPR-Cas nucleases. An ongoing clinical trial aims at assessing the safety of CRISPR-Cas9 in human modified cells into a patient with aggressive lung cancer (clinical trial number NCT02793856, (Cyranoski, 2016)). Other clinical trials followed such as one testing CD19-directed T-cell immunotherapy via allogeneic T-cells genetically modified ex vivo using CRISPR-Cas9 gene editing components safety and efficacy (clinical trial number NCT04035434). These trials rely on ex vivo modifications of the cells, which is not possible for DM1 which is a multisystemic disorder and because muscles cannot be extracted. More work is needed to characterize nuclease expression, efficacy and repair outcome on targeted cells. Cell to cell variations from one cell type to another and among a same cell type will also probably be a concern; dividing and non-dividing cells are both present in the muscle tissue and may not process DSB in the same way nor express the nuclease at the same level. $TALEN_{CTG}$ systemic expression profile was different from organ to organ with the heart expressing the highest amount of nuclease in DMSXL mice. Expression vector of the nuclease and method of delivery will also influence edition efficacy.

The use of reporter systems such as those developed during my doctoral work might be of great help to choose the best nuclease and the best construct to target CTG expansion. Both yeast and HEK reporter assays can easily be modified by cloning another sequence of interest instead of the CTG repeat. Difference nucleases could be screened for their activity on other sequences in order to which one will be the most efficient, before future testing in relevant disease models.

Materials and Methods

ASA cells

Vectors

pLVbcNEOPGK-IRES plasmid hereafter called pNEO, an empty lentiviral backbone containing IRES-Neomycin cassette was a kind gift from Denis Furling. Plasmid #23138 hereafter called pHYG, a lentiviral backbone containing IRES-Hygromycin cassette was ordered from Addgene. TALEN arm was cloned into pIRES and PNEO. Packaging into lentiviral particles was done by the ICM platform. TALEN9996 arm was cloned into pHYG and TALEN arm 16716 in pNEO. pHYG was digested with BamHi, XhoI and pNEO with SalI and AgeI. TALEN arm 9996 was retrieved by digesting pCMha182KN9996 with PmeI and PstI and TALEN arm 16715 by digesting pCMha183KN16715 with PmeI and PstI. Adaptor 1 was the result of the annealing of LP100 (GATCCAGCTTCATCTACTGA) and LP101 (TCAGTAGATGAAGCTG) oligos, adaptor 2 of LP104 (AGCTAGTTCATAAGC)- LP105 (TCGAGCTTATGAACTAGCTTGCA), adaptor 3 of LP102(CCGGTCAGTCGGTACTAAGC)- LP103 (GCTTAGTACCGACTGA) and adaptor 4 of LP106 (AGCTAGTTCATAAGG)- LP107 (TCGACCTTATGAACTAGCTTGCA). Ligation reaction of digested 9996 arm+adaptors 1 and 2 +digested pHYG and ligation reaction of digested 16715 arm+adaptors 3 and 4 +digested pNEO were transformed into XL10 gold and grown at RT. Transformants were picked and plasmid were extracted and verified by restriction digestion and sequencing. pTRI204 (9996-NEO) and pTRI205 (16715-HYG) were retrieved. ICM platform produced viral particles for these two plasmids.

Cell culture

Cells were cultivated in DMEM (Gibco #61965) supplemented with 15% Fœtal Bovine Serum (FBS) at 37°C, 5% CO_2 and in humidified atmosphere. Cells were passaged by washing with PBS before using trypsin for 5 minutes at 37°C to dissociate cells. Cells are split into new flasks. Cells were frozen in 90% FBS+10% DMSO in liquid nitrogen after an overnight slow freezing into a Mr. Frosty at -80°C.

Lentiviral transductions

Cells were seeded one day before at 55 000 cells/ well in 12-well plates. In the morning, media was change for 500ul of: DMEM+5% FBS+ 4 µg/ml of polybrene (TR-1003-G)+ virus diluted

to the desired MOI. Cells were incubated at 37°C, 5% CO_2 and in humidified atmosphere for the day. 500 µl of DMEM+25%FBS was added. The following day, media is changed for DMEM + 15%FBS. The following day, selection pressure is added and maintained for 1 week.

Ring cloning

Cells were diluted to reach between 10 to 100 cells per plate. Media was changed every week until small colonies of cells were detectable under the microscope (usually around 1 month). Media was removed, plastic rings were glued around each colony using silicone. After washing with PB, Trypsin was added inside the ring to dissociate the cells. Colonies were seeded into 96-well plates. Cells were expanded reaching the desired number of cells, by passaging into 24-well plates, 6-well plates, T25 and T75 flask (2-3 months of culture).

RNA-FISH

The day before staining, cells were plated onto a coated coverslip (coated with an autoclaved 0.5% gelatin solution). After overnight incubation, cells coated on the coverslip were rinsed with PBS and incubated with 4%PFA for 15 minutes. Cells were then washed twice for ten minutes in PBS+5mM $MgCl_2$. Cells were washed in 70% EtOH and then washed twice for 10 minutes in PBS+5mM $MgCl_2$. Coverslips were incubated on 20ul of hybridization buffer (40% formamide, 2X SSC, 0.2% BSA, 1/1000 2'OMe CAG7-Cy3 probe) in a humidified chamber at 37°C for 90 minutes. Coverslip were then washed 5 minutes in PBS-Tween 0.1% at RT and 30 minutes in PBS-Tween 0.1% at 45°C. Coverslip were washed in PBS and mounted onto a slide with 20µl of Prolong Gold antifade mounting medium containing DAPI. Slides were left to dry overnight before imaging.

ImageJ was used to count foci per cells, DAPI channel was used to define Regions of Interest and the function Find Maxima was used to count the number of foci was applied on the Cy3 channel. For each image, the number of foci per cells in each Region of Interest was calculated.

DNA extraction

Around 1.10^6 cells were harvested, washed in PBS and resuspended in 300µl lysis buffer (100mM Tris-Hcl pH8, 5mM EDTA pH8, 0.2% SDS and 200mM NaCl) supplemented with 600µg of proteinase K and incubated overnight in a water bath at 55°C. 300µl of phenol chloroform was added. After vortexing and centrifugating at maximal speed, aqueous phase was retrieved and washed using Chloroform. DNA was precipitated by adding 1/10 3M NaCl

and 3V of absolute ethanol. Centrifugation at maximal speed was carried out at 4°C for 1 hour. DNA pellet was washed in 70% ethanol and resuspended in water.

Western blot

Around 100 000 cells were washed into PBS and resuspended into RIPA buffer supplemented with Compete tablet. Samples were incubated for 1 hour on ice. Protein extracts were loaded on a 8 – 12% acrylamide gel alongside a positive control (yeast extract expressing the $TALEN_{CTG}$). After migration was complete, proteins were transferred on a nytran membrane by electroblotting. Membrane was incubated in 5% NFDM/PBS for 1 hour (blocking) and overnight in 5% NFDM/PBS + ab9110 (1/1,000). The following day, membrane was incubated with secondary antibody Goat anti-Rabbit 31460 (dilution 1/5000), and washed in 5% NFDM/PBS. Membrane was revealed using Amersham ECL detection kit and quantified on a Bio-Rad ChemiDoc apparatus.

Southern blotting

10µg of DNA was digested with BamHI overnight. DNA was loaded on a 1% agarose/TBE gel +BET. Run was carried out overnight at 50V, 4°C. A picture of the gel was taken under a BioRad imaging system. Gel was cleaned in distilled water and treated for one hour in 1M NaOH, followed by 2 hours in Tris 1M-NaCl 3M pH 8,5. Tranfer onto a positively charged nylon membrane (#RPN 203 S) was carried out manually, overnight in 6X SSC. The membrane was crosslinked under a stratalinker. Prehybridation was carried tout into Sigma PerfectHybridPlus solution at 68°C for 30 minutes. DNA probe was prepared by digesting B1.4 plasmid containing a cloned 3'UTR of DMPK gene with BamHI and PvuII and gel-purifying the 690-bp band. The probe was labeled with alpha-32P CTP using the High Prime DNA labeling kit (Roche, #11 585 584 001). The labelled probe was purified on a G50 sephadex column and denatured for 5 minutes at 95°C. Hybridation was carried out at 68°C overnight. Three 20-min washes were carried out (Two in High Stringency (0,5% SSC + 0,1%SDS) followed by one in Ultra high Stringency (0,1% SSC + 0,1% SDS)). Membrane was wrapped in saran and exposed to phosphorimager screen. The following day, the membrane is revealed under a phosphorimager.

DMSXL mice

Molecular cloning

Each TALEN arm was cloned into the pAAV-MCS vector (Clontech) at EcoRI/HindII, to give pAAV16715 and pAAV9996. rAAV production and packaging into serotypes 6 and 9 was carried out by Sanofi Genzyme following GMP procedures.

Animal breeding

Animal breeding and care was performed by the animal facility at Institut Imagine. Genotyping was performed by PCR300 on DNA extracted from tail tip. Animals were identified by toe tattooing.

Intramuscular injections and dissections

Heterozygous DMSXL mice aged of one month were anesthetized with isoflurane 5% delivered with an anesthetic vaporizer. Once anesthetized, mice were weighed and placed on their back under a sterile hood with a nose cone supplied with isoflurane gas to maintain anesthetic depth. Legs were shaved. 50µl of either rAAV or PBS (negative control) were injected using a syringe into Tibialis anterior muscles.

After either 1, 2 or 3 weeks, mice were sacrificed by cervical dislocation, weighed and dissected under a hood. TAs were either snap-frozen into liquid nitrogen or embedded in OCT, mounted onto a wood cork piece, precooled in isopentane before freezing in liquid nitrogen. Tissues were stored at -80°C.

Intraperitoneal injection and dissections

Neonatal mice at P2 were injected with 50µl of rAAVs into the right side of the peritoneal cavity. After 1 week, mice were sacrificed by decapitation, weighed and dissected under a hood to retrieve organs of interest (heart, liver, TA and quadriceps). Tissues were either snap-frozen into liquid nitrogen or embedded in OCT, mounted onto a wood cork piece, precooled in isopentane before freezing in liquid nitrogen. Tissues were stored at -80°C. After 3 weeks, mice were sacrificed by cervical dislocation, weighed and placed under a hood for dissection. Same organs were collected.

Tissue processing and Hematoxylin/eosin staining

Tissues were cut at -23°C in a cryostat into small sections of 10µm, and placed onto charged slides. Slides were stored at -80°C. For hematoxylin/eosin staining, slides were thawed at RT and then washed into successive baths containing: 80% EtOH, 70% EtOH, Hematoxylin, water, Eosin, twice in 70% EtOH, 95% EtOH, 100% EtOH. Slides were left to dry after washing with Xylene. Finally, slides were mounted in DPX and left to dry overnight.

DNA extraction

Tissues were incubated in 400µl lysis buffer (100mM Tris-Hcl pH8, 5mM EDTA pH8, 0.2% SDS and 200mM NaCl) supplemented with 800µg of proteinase K and incubated overnight in a water bath at 55°C. 1 volume of phenol chloroform was added. After vortexing and centrifugating at maximal speed, aqueous phase was retrieved and washed using Chloroform. DNA was precipitated by adding 1/10 volume 3M NaCl and 3 volume of absolute ethanol. Centrifugation at maximal speed was carried out at 4°C for 1 hour. DNA pellet was washed in 70% ethanol and resuspended in water.

PCR300

SM005 buffer was supplemented with □-mercaptoethanol to a final concentration of 69 µM. 15 ng of DNA, 1X of buffer SM005, 0.4µM final of each primer ST300-F and ST300-R, 0.04U of Taq were mixed. The following program was used: 96°C at 60°C for 5 minutes, a 30-times cycling of: 96°C for 45s, 60°C for 30 seconds, 72°C for 3 minutes, and a final 60°C for 1 minute and 72°C for 10 minutes.

Protein extraction and Western blot

Tissues were mixed with 300µl of homogenization buffer (RIPA 1X, CHAPS, Complete tablet, PhosphoSTOP tablet, 200mM orthovanadate) into a Precellys tube (#P000912-LYSK0). The tubes were processed into a precellys homogenizator (50Hz for 20 seconds, break of 30 seconds, 50Hz for 20 seconds, repeated twice). Samples were sonicated at 50Hz 15 times for 2 seconds. Tubes were placed at 4°C on a rotator for 2 hours. After centrifugating at 13,000 rpm at 4°C for 15 minutes, supernatant was stored on ice and remaining pellet was resuspended into 100µl of homogenization buffer, sonicated and placed at 4°C on a rotator for 2 hours. After centrifugation at 13,000 rpm at 4°C for 15 minutes, supernatant was added to stored supernatant. Protein concentration was quantified using a Qubit. Proteins were loaded on a 8 – 12% acrylamide gel alongside a positive control (yeast extract expressing the $TALEN_{CTG}$).

After migration was complete, proteins were transferred on a nytran membrane by electroblotting. Membrane was incubated in 5% NFDM/PBS for 1 hour (blocking) and overnight in 5% NFDM/PBS + ab9110 (1/1,000). The following day, membrane was incubated with secondary antibody Goat anti-Rabbit 31460 (dilution 1/5000), and washed in 5% NFDM/PBS. Membrane was revealed using Amersham ECL detection kit and quantified on a Bio-Rad ChemiDoc apparatus.

HEK293FS cell model

Vectors

200 CTG repeats from pRW3332 were cloned into synmammalian-EGFP synthetic sequence ordered from GenArt. This sequence contains a CMV promoter, two GFP halves and cloning sites. Both plasmids were digested with BamHI and PstI. The cassette EGFP::200CTGs was cloned PiggyBac ITR sequences into a proprietary vector of Sanofi at HindIII / EcoRI.

SaCas9 fused to mcherry was expressed using pX601-mcherry plasmid (addgene #84039). Guide RNAs were cloned at BsaI site. CTG guide was the result of reannealing of oligos px601guidefwd (CACCTGCTGCTGCTGCTGCTGCTGG) and px601guiderev (AAACCCAGCAGCAGCAGCAGCAGCA), control guide of px601ctrlfwd (CACCTCTAGAGTCGTCCTTGTAGCC) and px601ctrlrev (AAACGGCTACAAGGACGACTCTAGA), AAVS1 guide of px601AAVS1fwr (CACCGTGTGTGTAGCACCGCGTAAA) and px601AAVS1rev (AAACTTTACGCGGTGCTACACACAC). Spas9 fused to mcherry was expressed using pU6-(BbsI)_CBh-Cas9-T2A-mCherry plasmid (addgene #64324). Guide RNAs were cloned at BbsI site. CTG guide was the result of reannealing of oligos LP1000 (CACCCTGCTGCTGCTGCTGCTGGG) and LP1001 (AAACCCCAGCAGCAGCAGCAGCAG), control guide of LP1004 (CACCTCTTTCTTTCGGCCAGGCTG) and LP1005 (AAACCAGCCTGGCCGAAAGAAAGA).

Cell culture

HEK293FS cells and derivatives are cultivated at 37°C, 5% CO_2 under humidified atmosphere, at 120 rpm. Media is FreeStyle media + Glutamax (Gibco # 12338018) changed twice a week and cells were passaged at a density of $0.3.10^6$ cells/ml.

Stable cell line generation

Cells were electroporated using a MaxCyte electroporator with 3µg of the plasmid carrying transposase (proprietary vector of Sanofi) and 27µg of plasmid carrying the cassette flanked by Piggybac ITR sequences. After 48h of recovery, selection pressure was maintained for 1 week. Cells were then cloned by limiting dilution into 96-well plates at a density of 0.5 cells/well. Media was changed twice a week. After 3 weeks, visible colonies were expanded into 24-well plates then 6-well plates and in flasks. Clones were frozen into FreeStyle medium containing 7.5% DMSO in liquid nitrogen after an overnight slow freezing into a Mr. Frosty at -80°C.

Expression of Cas9

HEK293FS cells were transfected using fectin293 (gibco #12347019) according to manufacturer's instructions, with 30µg of plasmid.

Flow cytometry

Cells were rinsed in PBS before being resuspended in PBS for flow cytometry processing. On a MACSQuant cytometer, the following parameters were used: FSC channel hlog, 197V; SSC channel hlog, 139V; Y2 channel hlog, 400V; V1 channel hlog, 300V; B1 channel hlog, 300V. Homogenous population was gated on FSC-A/SSC-A, single cells were gated on FSC-A/FSC-H.

References

van Agtmaal, E.L., André, L.M., Willemse, M., Cumming, S.A., van Kessel, I.D.G., van den Broek, W.J.A.A., Gourdon, G., Furling, D., Mouly, V., Monckton, D.G., et al. (2017). CRISPR/Cas9-Induced (CTG·CAG)n Repeat Instability in the Myotonic Dystrophy Type 1 Locus: Implications for Therapeutic Genome Editing. Molecular Therapy *25*, 24–43.

Akarsu, A.N., Stoilov, I., Yilmaz, E., Sayli, B.S., and Sarfarazi, M. (1996). Genomic structure of HOXD13 gene: a nine polyalanine duplication causes synpolydactyly in two unrelated families. Hum. Mol. Genet. *5*, 945–952.

Akcakaya, P., Bobbin, M.L., Guo, J.A., Lopez, J.M., Clement, M.K., Garcia, S.P., Fellows, M.D., Porritt, M.J., Firth, M.A., Carreras, A., et al. (2018). In vivo CRISPR-Cas gene editing with no detectable genome-wide off-target mutations. BioRxiv 272724.

Allen, C., Kurimasa, A., Brenneman, M.A., Chen, D.J., and Nickoloff, J.A. (2002). DNA-dependent protein kinase suppresses double-strand break-induced and spontaneous homologous recombination. PNAS *99*, 3758–3763.

Amiel, J., Trochet, D., Clément-Ziza, M., Munnich, A., and Lyonnet, S. (2004). Polyalanine expansions in human. Hum Mol Genet *13*, R235–R243.

Amoasii, L., Hildyard, J.C.W., Li, H., Sanchez-Ortiz, E., Mireault, A., Caballero, D., Harron, R., Stathopoulou, T.-R., Massey, C., Shelton, J.M., et al. (2018). Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. Science *362*, 86–91.

Amrane, S., Saccà, B., Mills, M., Chauhan, M., Klump, H.H., and Mergny, J.-L. (2005). Length-dependent energetics of (CTG)n and (CAG)n trinucleotide repeats. Nucleic Acids Res. *33*, 4065–4077.

Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., et al. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature 1–1.

Arandel, L., Polay Espinoza, M., Matloka, M., Bazinet, A., De Dea Diniz, D., Naouar, N., Rau, F., Jollet, A., Edom-Vovard, F., Mamchaoui, K., et al. (2017). Immortalized human myotonic dystrophy muscle cell lines to assess therapeutic compounds. Dis Model Mech *10*, 487–497.

Arbizu, T., Santamaría, J., Gomez, J.M., Quílez, A., and Serra, J.P. (1983). A family with adult spinal and bulbar muscular atrophy, X-linked inheritance and associated testicular failure. J. Neurol. Sci. *59*, 371–382.

Arcangioli, B. (1998). A site- and strand-specific DNA break confers asymmetric switching

potential in fission yeast. EMBO J *17*, 4503–4510.

Arnott, S., Fuller, W., Hodgson, A., and Prutton, I. (1968). Molecular conformations and structure transitions of RNA complementary helices and their possible biological significance. Nature *220*, 561–564.

Arnould, S., Chames, P., Perez, C., Lacroix, E., Duclert, A., Epinat, J.-C., Stricher, F., Petit, A.-S., Patin, A., Guillier, S., et al. (2006). Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. J. Mol. Biol. *355*, 443–458.

Ashizawa, T., Gonzales, I., Ohsawa, N., Singer, R.H., Devillers, M., Ashizawa, T., Balasubramanyam, A., Cooper, T.A., Khajavi, M., Lia-Baldini, A.-S., et al. (2000). New nomenclature and DNA testing guidelines for myotonic dystrophy type 1 (DM1). Neurology *54*, 1218–1221.

Aumiller, V., Graebsch, A., Kremmer, E., Niessing, D., and Förstemann, K. (2012). Drosophila Pur-α binds to trinucleotide-repeat containing cellular RNAs and translocates to the early oocyte. RNA Biol *9*, 633–643.

Bailly, V., Sommers, C.H., Sung, P., Prakash, L., and Prakash, S. (1992). Specific complex formation between proteins encoded by the yeast DNA repair and recombination genes RAD1 and RAD10. Proc. Natl. Acad. Sci. U.S.A. *89*, 8273–8277.

Barizien, A., Suryateja Jammalamadaka, M.S., Amselem, G., and Baroud, C.N. (2019). Growing from a few cells: combined effects of initial stochasticity and cell-to-cell variability. J R Soc Interface *16*, 20180935.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Batra, R., Nelles, D.A., Pirie, E., Blue, S.M., Marina, R.J., Wang, H., Chaim, I.A., Thomas, J.D., Zhang, N., Nguyen, V., et al. (2017). Elimination of Toxic Microsatellite Repeat Expansion RNA by RNA-Targeting Cas9. Cell *170*, 899-912.e10.

Beerman, I., Seita, J., Inlay, M.A., Weissman, I.L., and Rossi, D.J. (2014). Quiescent Hematopoietic Stem Cells Accumulate DNA Damage during Aging that Is Repaired upon Entry into Cell Cycle. Cell Stem Cell *15*, 37–50.

Berg, I.L., Neumann, R., Lam, K.-W.G., Sarbajna, S., Odenthal-Hesse, L., May, C.A., and Jeffreys, A.J. (2010). PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nat Genet *42*, 859–863.

Bhargava, R., Onyango, D.O., and Stark, J.M. (2016). Regulation of Single-Strand Annealing

and its Role in Genome Maintenance. Trends Genet. *32*, 566–575.

Bibikova, M., Carroll, D., Segal, D.J., Trautman, J.K., Smith, J., Kim, Y.G., and Chandrasegaran, S. (2001). Stimulation of homologous recombination through targeted cleavage by chimeric nucleases. Mol. Cell. Biol. *21*, 289–297.

Bibikova, M., Beumer, K., Trautman, J.K., and Carroll, D. (2003). Enhancing Gene Targeting with Designed Zinc Finger Nucleases. Science *300*, 764–764.

Bitinaite, J., Wah, D.A., Aggarwal, A.K., and Schildkraut, I. (1998). FokI dimerization is required for DNA cleavage. Proc. Natl. Acad. Sci. U.S.A. *95*, 10570–10575.

Blacklow, N.R., Hoggan, M.D., Kapikian, A.Z., Austin, J.B., and Rowe, W.P. (1968). Epidemiology of adenovirus-associated virus infection in a nursery population. Am. J. Epidemiol. *88*, 368–378.

Blier, P.R., Griffith, A.J., Craft, J., and Hardin, J.A. (1993). Binding of Ku protein to DNA. Measurement of affinity for ends and demonstration of binding to nicks. J. Biol. Chem. *268*, 7594–7601.

Boch, J., Scholze, H., Schornack, S., Landgraf, A., Hahn, S., Kay, S., Lahaye, T., Nickstadt, A., and Bonas, U. (2009). Breaking the Code of DNA Binding Specificity of TAL-Type III Effectors. Science *326*, 1509–1512.

Boulton, S.J., and Jackson, S.P. (1996). Saccharomyces cerevisiae Ku70 potentiates illegitimate DNA double-strand break repair and serves as a barrier to error-prone DNA repair pathways. EMBO J. *15*, 5093–5103.

Bourn, R.L., De Biase, I., Pinto, R.M., Sandi, C., Al-Mahdawi, S., Pook, M.A., and Bidichandani, S.I. (2012). Pms2 Suppresses Large Expansions of the (GAA·TTC)n Sequence in Neuronal Tissues. PLoS One *7*.

van den Broek, W.J.A.A., Nelen, M.R., van der Heijden, G.W., Wansink, D.G., and Wieringa, B. (2006). Fen1 does not control somatic hypermutability of the (CTG)(n)*(CAG)(n) repeat in a knock-in mouse model for DM1. FEBS Lett. *580*, 5208–5214.

Brook, J.D., McCurrach, M.E., Harley, H.G., Buckler, A.J., Church, D., Aburatani, H., Hunter, K., Stanton, V.P., Thirion, J.-P., Hudson, T., et al. (1992). Molecular basis of myotonic dystrophy: Expansion of a trinucleotide (CTG) repeat at the 3′ end of a transcript encoding a protein kinase family member. Cell *68*, 799–808.

Brooks, A.R., Harkins, R.N., Wang, P., Qian, H.S., Liu, P., and Rubanyi, G.M. (2004). Transcriptional silencing is associated with extensive methylation of the CMV promoter following adenoviral gene delivery to muscle. J Gene Med *6*, 395–404.

Brouns, S.J.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J.H., Snijders, A.P.L.,

Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. Science *321*, 960–964.

Burrow, A.A., Marullo, A., Holder, L.R., and Wang, Y.-H. (2010). Secondary structure formation and DNA instability at fragile site FRA16B. Nucleic Acids Res *38*, 2865–2877.

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR-Cas systems from uncultivated microbes. Nature *542*, 237–241.

Calado, A., Tomé, F.M., Brais, B., Rouleau, G.A., Kühn, U., Wahle, E., and Carmo-Fonseca, M. (2000). Nuclear inclusions in oculopharyngeal muscular dystrophy consist of poly(A) binding protein 2 aggregates which sequester poly(A) RNA. Hum. Mol. Genet. *9*, 2321–2328.

Callahan, J.L., Andrews, K.J., Zakian, V.A., and Freudenreich, C.H. (2003). Mutations in yeast replication proteins that increase CAG/CTG expansions also increase repeat fragility. Mol. Cell. Biol. *23*, 7849–7860.

Carlon, M.S., Vidović, D., Dooley, J., da Cunha, M.M., Maris, M., Lampi, Y., Toelen, J., Van den Haute, C., Baekelandt, V., Deprest, J., et al. (2014). Immunological Ignorance Allows Long-Term Gene Expression After Perinatal Recombinant Adeno-Associated Virus-Mediated Gene Transfer to Murine Airways. Hum Gene Ther *25*, 517–528.

Casper, A.M., Nghiem, P., Arlt, M.F., and Glover, T.W. (2002). ATR regulates fragile site stability. Cell *111*, 779–789.

Cavazzana-Calvo, M., Payen, E., Negre, O., Wang, G., Hehir, K., Fusil, F., Down, J., Denaro, M., Brady, T., Westerman, K., et al. (2010). Transfusion independence and HMGA2 activation after gene therapy of human β-thalassaemia. Nature *467*, 318–322.

Cejka, P. (2015). DNA End Resection: Nucleases Team Up with the Right Partners to Initiate Homologous Recombination. J. Biol. Chem. *290*, 22931–22938.

Cermak, T., Doyle, E.L., Christian, M., Wang, L., Zhang, Y., Schmidt, C., Baller, J.A., Somia, N.V., Bogdanove, A.J., and Voytas, D.F. (2011). Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res. *39*, e82.

Chargaff, E., Magasanik, B., Vischer, E., Green, C., Doniger, R., and Elson, D. (1950). Nucleotide composition of pentose nucleic acids from yeast and mammalian tissues. J. Biol. Chem. *186*, 51–67.

Chari, R., Mali, P., Moosburner, M., and Church, G.M. (2015). Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. Nat. Methods *12*, 823–826.

Charlesworth, C.T., Deshpande, P.S., Dever, D.P., Camarena, J., Lemgart, V.T., Cromer, M.K.,

Vakulskas, C.A., Collingwood, M.A., Zhang, L., Bode, N.M., et al. (2019). Identification of preexisting adaptive immunity to Cas9 proteins in humans. Nat. Med. *25*, 249–254.

Charpentier, M., Khedher, A.H.Y., Menoret, S., Brion, A., Lamribet, K., Dardillac, E., Boix, C., Perrouault, L., Tesson, L., Geny, S., et al. (2018). CtIP fusion to Cas9 enhances transgene integration by homology-dependent repair. Nat Commun *9*, 1–11.

Chen, H., Lisby, M., and Symington, L.S. (2013). RPA coordinates DNA end resection and prevents formation of DNA hairpins. Mol Cell *50*.

Chen, J.S., Dagdas, Y.S., Kleinstiver, B.P., Welch, M.M., Sousa, A.A., Harrington, L.B., Sternberg, S.H., Joung, J.K., Yildiz, A., and Doudna, J.A. (2017). Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. Nature *550*, 407–410.

Chen, X., Niu, H., Chung, W.-H., Zhu, Z., Papusha, A., Shim, E.Y., Lee, S.E., Sung, P., and Ira, G. (2011). Cell cycle regulation of DNA double-strand break end resection by Cdk1-dependent Dna2 phosphorylation. Nat Struct Mol Biol *18*, 1015–1019.

Cherng, N., Shishkin, A.A., Schlager, L.I., Tuck, R.H., Sloan, L., Matera, R., Sarkar, P.S., Ashizawa, T., Freudenreich, C.H., and Mirkin, S.M. (2011). Expansions, contractions, and fragility of the spinocerebellar ataxia type 10 pentanucleotide repeat in yeast. PNAS *108*, 2843–2848.

Cheung, I., Schertzer, M., Rose, A., and Lansdorp, P.M. (2002). Disruption of dog-1 in Caenorhabditis elegans triggers deletions upstream of guanine-rich DNA. Nat. Genet. *31*, 405–409.

Chew, W.L., Tabebordbar, M., Cheng, J.K.W., Mali, P., Wu, E.Y., Ng, A.H.M., Zhu, K., Wagers, A.J., and Church, G.M. (2016). A multi-functional AAV-CRISPR-Cas9 and its host response. Nat Methods *13*, 868–874.

Choi, V.W., McCarty, D.M., and Samulski, R.J. (2006). Host Cell DNA Repair Pathways in Adeno-Associated Viral Genome Processing. J Virol *80*, 10346–10356.

Choulika, A., Perrin, A., Dujon, B., and Nicolas, J.F. (1995). Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. Mol. Cell. Biol. *15*, 1968–1973.

Christian, M., Cermak, T., Doyle, E.L., Schmidt, C., Zhang, F., Hummel, A., Bogdanove, A.J., and Voytas, D.F. (2010). Targeting DNA double-strand breaks with TAL effector nucleases. Genetics *186*, 757–761.

Cinesi, C., Aeschbach, L., Yang, B., and Dion, V. (2016). Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. Nature Communications *7*, 13272.

Cleary, J.D., Tomé, S., López Castel, A., Panigrahi, G.B., Foiry, L., Hagerman, K.A., Sroka,

H., Chitayat, D., Gourdon, G., and Pearson, C.E. (2010). Tissue- and age-specific DNA replication patterns at the CTG/CAG-expanded human myotonic dystrophy type 1 locus. Nature Structural & Molecular Biology *17*, 1079–1087.

Clément, N., and Grieger, J.C. (2016). Manufacturing of recombinant adeno-associated viral vectors for clinical trials. Molecular Therapy - Methods & Clinical Development *3*, 16002.

Colleaux, L., d'Auriol, L., Betermier, M., Cottarel, G., Jacquier, A., Galibert, F., and Dujon, B. (1986). Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into E. coli as a specific double strand endonuclease. Cell *44*, 521–533.

Colleaux, L., D'Auriol, L., Galibert, F., and Dujon, B. (1988). Recognition and cleavage site of the intron-encoded omega transposase. PNAS *85*, 6022–6026.

Concordet, J.-P., and Haeussler, M. (2018). CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res *46*, W242–W245.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., et al. (2013). Multiplex genome engineering using CRISPR/Cas systems. Science *339*, 819–823.

Conlon, E.G., Lu, L., Sharma, A., Yamazaki, T., Tang, T., Shneider, N.A., and Manley, J.L. (2016). The C9ORF72 GGGGCC expansion forms RNA G-quadruplex inclusions and sequesters hnRNP H to disrupt splicing in ALS brains. ELife *5*, e17820.

Costelloe, T., Louge, R., Tomimatsu, N., Mukherjee, B., Martini, E., Khadaroo, B., Dubois, K., Wiegant, W.W., Thierry, A., Burma, S., et al. (2012). The yeast Fun30 and human SMARCAD1 chromatin remodellers promote DNA end resection. Nature *489*, 581–584.

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. Science *358*, 1019–1027.

Crawford, D.C., Acuña, J.M., and Sherman, S.L. (2001). FMR1 and the Fragile X Syndrome: Human Genome Epidemiology Review. Genet Med *3*, 359–371.

Cumming, S.A., Hamilton, M.J., Robb, Y., Gregory, H., McWilliam, C., Cooper, A., Adam, B., McGhie, J., Hamilton, G., Herzyk, P., et al. (2018). De novo repeat interruptions are associated with reduced somatic instability and mild or absent clinical features in myotonic dystrophy type 1. Eur. J. Hum. Genet. *26*, 1635–1647.

Cyranoski, D. (2016). CRISPR gene-editing tested in a person for the first time. Nature News *539*, 479.

Dai, Y., Kysela, B., Hanakahi, L.A., Manolis, K., Riballo, E., Stumm, M., Harville, T.O., West, S.C., Oettinger, M.A., and Jeggo, P.A. (2003). Nonhomologous end joining and V(D)J recombination require an additional factor. Proc. Natl. Acad. Sci. U.S.A. *100*, 2462–2467.

Daley, J.M., Palmbos, P.L., Wu, D., and Wilson, T.E. (2005). Nonhomologous End Joining in Yeast. Annu. Rev. Genet. *39*, 431–451.

Dalgaard, J.Z., and Klar, A.J.S. (2001). A DNA replication-arrest site RTS1 regulates imprinting by determining the direction of replication at mat1 in S. pombe. Genes Dev *15*, 2060–2068.

Daya, S., and Berns, K.I. (2008). Gene Therapy Using Adeno-Associated Virus Vectors. Clin Microbiol Rev *21*, 583–593.

Deng, D., Yan, C., Pan, X., Mahfouz, M., Wang, J., Zhu, J.-K., Shi, Y., and Yan, N. (2012). Structural basis for sequence-specific recognition of DNA by TAL effectors. Science *335*, 720–723.

Dere, R., Napierala, M., Ranum, L.P.W., and Wells, R.D. (2004). Hairpin structure-forming propensity of the (CCTG.CAGG) tetranucleotide repeats contributes to the genetic instability associated with myotonic dystrophy type 2. J. Biol. Chem. *279*, 41715–41726.

Deveau, H., Barrangou, R., Garneau, J.E., Labonté, J., Fremaux, C., Boyaval, P., Romero, D.A., Horvath, P., and Moineau, S. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

Di Primio, C., Galli, A., Cervelli, T., Zoppè, M., and Rainaldi, G. (2005). Potentiation of gene targeting in human cells by expression of Saccharomyces cerevisiae Rad52. Nucleic Acids Res *33*, 4639–4648.

DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., and Church, G.M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res. *41*, 4336–4343.

Doench, J.G., Hartenian, E., Graham, D.B., Tothova, Z., Hegde, M., Smith, I., Sullender, M., Ebert, B.L., Xavier, R.J., and Root, D.E. (2014). Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. Nat. Biotechnol. *32*, 1262–1267.

Donaldson, K.M., Li, W., Ching, K.A., Batalov, S., Tsai, C.-C., and Joazeiro, C.A.P. (2003). Ubiquitin-mediated sequestration of normal cellular proteins into polyglutamine aggregates. Proc. Natl. Acad. Sci. U.S.A. *100*, 8892–8897.

Duan, J., Lu, G., Hong, Y., Hu, Q., Mai, X., Guo, J., Si, X., Wang, F., and Zhang, Y. (2018). Live imaging and tracking of genome regions in CRISPR/dCas9 knock-in mice. Genome Biol *19*.

Dull, T., Zufferey, R., Kelly, M., Mandel, R.J., Nguyen, M., Trono, D., and Naldini, L. (1998). A third-generation lentivirus vector with a conditional packaging system. J. Virol. *72*, 8463–8471.

Dunah, A.W., Jeong, H., Griffin, A., Kim, Y.-M., Standaert, D.G., Hersch, S.M., Mouradian, M.M., Young, A.B., Tanese, N., and Krainc, D. (2002). Sp1 and TAFII130 transcriptional activity disrupted in early Huntington's disease. Science *296*, 2238–2243.

Elrod-Erickson, M., Rould, M.A., Nekludova, L., and Pabo, C.O. (1996). Zif268 protein-DNA complex refined at 1.6 A: a model system for understanding zinc finger-DNA interactions. Structure *4*, 1171–1180.

Escors, D., and Breckpot, K. (2010). Lentiviral vectors in gene therapy: their current status and future potential. Arch. Immunol. Ther. Exp. (Warsz.) *58*, 107–119.

Felsenfeld, G., and Rich, A. (1957). Studies on the formation of two- and three-stranded polyribonucleotides. Biochimica et Biophysica Acta *26*, 457–468.

Ferrari, M., Dibitetto, D., De Gregorio, G., Eapen, V.V., Rawal, C.C., Lazzaro, F., Tsabar, M., Marini, F., Haber, J.E., and Pellicioli, A. (2015). Functional Interplay between the 53BP1-Ortholog Rad9 and the Mre11 Complex Regulates Resection, End-Tethering and Repair of a Double-Strand Break. PLoS Genet *11*.

Fischer, A., Hacein-Bey-Abina, S., and Cavazzana-Calvo, M. (2010). 20 years of gene therapy for SCID. Nature Immunology *11*, 457–460.

Fishman-Lobell, J., and Haber, J.E. (1992). Removal of nonhomologous DNA ends in double-strand break recombination: the role of the yeast ultraviolet repair gene RAD1. Science *258*, 480–484.

Franklin, R.E., and Gosling, R.G. (1953). Molecular Configuration in Sodium Thymonucleate. Nature *171*, 740–741.

Frank-Vaillant, M., and Marcand, S. (2001). NHEJ regulation by mating type is exercised through a novel protein, Lif2p, essential to the Ligase IV pathway. Genes Dev *15*, 3005–3012.

Freudenreich, C.H., Stavenhagen, J.B., and Zakian, V.A. (1997). Stability of a CTG/CAG trinucleotide repeat in yeast is dependent on its orientation in the genome. Mol Cell Biol *17*, 2090–2098.

Freudenreich, C.H., Kantrow, S.M., and Zakian, V.A. (1998). Expansion and length-dependent fragility of CTG repeats in yeast. Science *279*, 853–856.

Fry, M., and Loeb, L.A. (1994). The fragile X syndrome d(CGG)n nucleotide repeats form a stable tetrahelical structure. PNAS *91*, 4950–4954.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nature Biotechnology *31*, 822–826.

Furling, D., Lam, L.T., Agbulut, O., Butler-Browne, G.S., and Morris, G.E. (2003). Changes in

Myotonic Dystrophy Protein Kinase Levels and Muscle Development in Congenital Myotonic Dystrophy. Am J Pathol *162*, 1001–1009.

Gacy, A.M., Goellner, G., Juranić, N., Macura, S., and McMurray, C.T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell *81*, 533–540.

Gacy, A.M., Goellner, G.M., Spiro, C., Chen, X., Gupta, G., Bradbury, E.M., Dyer, R.B., Mikesell, M.J., Yao, J.Z., Johnson, A.J., et al. (1998). GAA Instability in Friedreich's Ataxia Shares a Common, DNA-Directed and Intraallelic Mechanism with Other Trinucleotide Diseases. Molecular Cell *1*, 583–593.

Gaj, T., Gersbach, C.A., and Barbas, C.F. (2013). ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol *31*, 397–405.

Garneau, J.E., Dupuis, M.-È., Villion, M., Romero, D.A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadán, A.H., and Moineau, S. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Gautron, A.-S., Juillerat, A., Guyot, V., Filhol, J.-M., Dessez, E., Duclert, A., Duchateau, P., and Poirot, L. (2017). Fine and Predictable Tuning of TALEN Gene Editing Targeting for Improved T Cell Adoptive Immunotherapy. Molecular Therapy - Nucleic Acids *9*, 312–321.

Gecz, J. (2000). The FMR2 gene, FRAXE and non-specific X-linked mental retardation: clinical and molecular aspects. Ann. Hum. Genet. *64*, 95–106.

Gellert, M., Lipsett, M.N., and Davies, D.R. (1962). Helix formation by guanylic acid. Proc. Natl. Acad. Sci. U.S.A. *48*, 2013–2018.

Gietz, R.D., Schiestl, R.H., Willems, A.R., and Woods, R.A. (1995). Studies on the transformation of intact yeast cells by the LiAc/SS-DNA/PEG procedure. Yeast *11*, 355–360.

Gilbert, L.A., Larson, M.H., Morsut, L., Liu, Z., Brar, G.A., Torres, S.E., Stern-Ginossar, N., Brandman, O., Whitehead, E.H., Doudna, J.A., et al. (2013). CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell *154*, 442–451.

Glaser, A., McColl, B., and Vadolas, J. (2016). GFP to BFP Conversion: A Versatile Assay for the Quantification of CRISPR/Cas9-mediated Genome Editing. Molecular Therapy - Nucleic Acids *5*.

Glover, T.W. (1981). FUdR induction of the X chromosome fragile site: evidence for the mechanism of folic acid and thymidine inhibition. Am J Hum Genet *33*, 234–242.

Glover, T.W., Arlt, M.F., Casper, A.M., and Durkin, S.G. (2005). Mechanisms of common fragile site instability. Hum. Mol. Genet. *14 Spec No. 2*, R197-205.

Goldstein, J.M., Tabebordbar, M., Zhu, K., Wang, L.D., Messemer, K.A., Peacker, B., Ashrafi Kakhki, S., Gonzalez-Celeiro, M., Shwartz, Y., Cheng, J.K.W., et al. (2019). In Situ

Modification of Tissue Stem and Progenitor Cell Genomes. Cell Rep *27*, 1254-1264.e7.

Gomes-Pereira, M., Foiry, L., Nicole, A., Huguet, A., Junien, C., Munnich, A., and Gourdon, G. (2007). CTG Trinucleotide Repeat "Big Jumps": Large Expansions, Small Mice. PLoS Genet *3*.

Gomes-Pereira, M., Cooper, T.A., and Gourdon, G. (2011). Myotonic dystrophy mouse models: towards rational therapy development. Trends Mol Med *17*.

Goodman, F.R., Mundlos, S., Muragaki, Y., Donnai, D., Giovannucci-Uzielli, M.L., Lapi, E., Majewski, F., McGaughran, J., McKeown, C., Reardon, W., et al. (1997). Synpolydactyly phenotypes correlate with size of expansions in HOXD13 polyalanine tract. Proc. Natl. Acad. Sci. U.S.A. *94*, 7458–7463.

Gopalappa, R., Suresh, B., Ramakrishna, S., and Kim, H. (Henry) (2018). Paired D10A Cas9 nickases are sometimes more efficient than individual nucleases for gene disruption. Nucleic Acids Res *46*, e71.

Gouble, A., Smith, J., Bruneau, S., Perez, C., Guyot, V., Cabaniols, J.-P., Leduc, S., Fiette, L., Avé, P., Micheau, B., et al. (2006). Efficient in toto targeted recombination in mouse liver by meganuclease-induced double-strand break. The Journal of Gene Medicine *8*, 616–622.

Gourdon, G., Radvanyi, F., Lia, A.S., Duros, C., Blanche, M., Abitbol, M., Junien, C., and Hofmann-Radvanyi, H. (1997). Moderate intergenerational and somatic instability of a 55-CTG repeat in transgenic mice. Nat. Genet. *15*, 190–192.

Grabczyk, E., and Usdin, K. (2000). The GAA*TTC triplet repeat expanded in Friedreich's ataxia impedes transcription elongation by T7 RNA polymerase in a length and supercoil dependent manner. Nucleic Acids Res. *28*, 2815–2822.

Guilinger, J.P., Thompson, D.B., and Liu, D.R. (2014). Fusion of catalytically inactive Cas9 to FokI nuclease improves the specificity of genome modification. Nat. Biotechnol. *32*, 577–582.

Gutschner, T., Haemmerle, M., Genovese, G., Draetta, G.F., and Chin, L. (2016). Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. Cell Reports *14*, 1555–1566.

Haapaniemi, E., Botla, S., Persson, J., Schmierer, B., and Taipale, J. (2018). CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. Nature Medicine *24*, 927.

Haeusler, A.R., Donnelly, C.J., Periz, G., Simko, E.A.J., Shaw, P.G., Kim, M.-S., Maragakis, N.J., Troncoso, J.C., Pandey, A., Sattler, R., et al. (2014). C9orf72 nucleotide repeat structures initiate molecular cascades of disease. Nature *507*, 195–200.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., et al. (2016). Evaluation of off-target and

on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biology *17*, 148.

Heath, P.J., Stephens, K.M., Monnat, R.J., and Stoddard, B.L. (1997). The structure of I-Crel, a group I intron-encoded homing endonuclease. Nat. Struct. Biol. *4*, 468–476.

Heim, R., Prasher, D.C., and Tsien, R.Y. (1994). Wavelength mutations and posttranslational autoxidation of green fluorescent protein. Proc. Natl. Acad. Sci. U.S.A. *91*, 12501–12504.

Herbers, K., Conrads-Strauch, J., and Bonas, U. (1992). Race-specificity of plant resistance to bacterial spot disease determined by repetitive motifs in a bacterial avirulence protein. Nature *356*, 172.

Higham, C.F., Morales, F., Cobbold, C.A., Haydon, D.T., and Monckton, D.G. (2012). High levels of somatic DNA diversity at the myotonic dystrophy type 1 locus are driven by ultra-frequent expansion and contraction mutations. Hum Mol Genet *21*, 2450–2463.

Hilario, J., Amitani, I., Baskin, R.J., and Kowalczykowski, S.C. (2009). Direct imaging of human Rad51 nucleoprotein dynamics on individual DNA molecules. PNAS *106*, 361–368.

Holkers, M., Maggio, I., Liu, J., Janssen, J.M., Miselli, F., Mussolino, C., Recchia, A., Cathomen, T., and Gonçalves, M.A.F.V. (2013). Differential integrity of TALE nuclease genes following adenoviral and lentiviral vector gene transfer into human cells. Nucleic Acids Res *41*, e63.

Hoogsteen, K. (1963). The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine. Acta Cryst *16*, 907–916.

Howden, S.E., McColl, B., Glaser, A., Vadolas, J., Petrou, S., Little, M.H., Elefanty, A.G., and Stanley, E.G. (2016). A Cas9 Variant for Efficient Generation of Indel-Free Knockin or Gene-Corrected Human Pluripotent Stem Cells. Stem Cell Reports *7*, 508–517.

Hsu, P.D., Scott, D.A., Weinstein, J.A., Ran, F.A., Konermann, S., Agarwala, V., Li, Y., Fine, E.J., Wu, X., Shalem, O., et al. (2013). DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. *31*, 827–832.

Huang, T.-Y., Chang, C.-K., Kao, Y.-F., Chin, C.-H., Ni, C.-W., Hsu, H.-Y., Hu, N.-J., Hsieh, L.-C., Chou, S.-H., Lee, I.-R., et al. (2017). Parity-dependent hairpin configurations of repetitive DNA sequence promote slippage associated with DNA expansion. Proc. Natl. Acad. Sci. U.S.A. *114*, 9535–9540.

Huertas, P. (2010). DNA resection in eukaryotes: deciding how to fix the break. Nat Struct Mol Biol *17*, 11–16.

Huertas, P., Cortés-Ledesma, F., Sartori, A.A., Aguilera, A., and Jackson, S.P. (2008). CDK targets Sae2 to control DNA-end resection and homologous recombination. Nature *455*, 689–

692.

Huguet, A., Medja, F., Nicole, A., Vignaud, A., Guiraud-Dogan, C., Ferry, A., Decostre, V., Hogrel, J.-Y., Metzger, F., Hoeflich, A., et al. (2012). Molecular, Physiological, and Motor Performance Defects in DMSXL Mice Carrying >1,000 CTG Repeats from the Human DM1 Locus. PLoS Genetics *8*.

Hustedt, N., and Durocher, D. (2016). The control of DNA repair by the cell cycle. Nat. Cell Biol. *19*, 1–9.

Ihry, R.J., Worringer, K.A., Salick, M.R., Frias, E., Ho, D., Theriault, K., Kommineni, S., Chen, J., Sondey, M., Ye, C., et al. (2018). p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. Nat. Med. *24*, 939–946.

Ira, G., Pellicioli, A., Balijja, A., Wang, X., Fiorani, S., Carotenuto, W., Liberi, G., Bressan, D., Wan, L., Hollingsworth, N.M., et al. (2004). DNA end resection, homologous recombination and DNA damage checkpoint activation require CDK1. Nature *431*, 1011–1017.

Isalan, M., Choo, Y., and Klug, A. (1997). Synergy between adjacent zinc fingers in sequence-specific DNA recognition. Proc. Natl. Acad. Sci. U.S.A. *94*, 5617–5621.

Ishino, Y., Shinagawa, H., Makino, K., Amemura, M., and Nakata, A. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J. Bacteriol. *169*, 5429–5433.

Ivanov, E.L., and Haber, J.E. (1995). RAD1 and RAD10, but not other excision repair genes, are required for double-strand break-induced recombination in Saccharomyces cerevisiae. Mol. Cell. Biol. *15*, 2245–2251.

Jacquemont, S., Hagerman, R.J., Hagerman, P.J., and Leehey, M.A. (2007). Fragile-X syndrome and fragile X-associated tremor/ataxia syndrome: two faces of FMR1. Lancet Neurol *6*, 45–55.

Jain, S., Sugawara, N., Lydeard, J., Vaze, M., Tanguy Le Gac, N., and Haber, J.E. (2009). A recombination execution checkpoint regulates the choice of homologous recombination pathway during DNA double-strand break repair. Genes Dev *23*, 291–303.

Jansen, G., Groenen, P.J.T.A., Bächner, D., Jap, P.H.K., Coerwinkel, M., Oerlemans, F., Broek, W. van den, Gohlsch, B., Pette, D., Plomp, J.J., et al. (1996). Abnormal myotonic dystrophy protein kinase levels produce only mild myopathy in mice. Nature Genetics *13*, 316.

Jansen, R., Embden, J.D.A. van, Gaastra, W., and Schouls, L.M. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. *43*, 1565–1575.

Jensen, K.T., Fløe, L., Petersen, T.S., Huang, J., Xu, F., Bolund, L., Luo, Y., and Lin, L. (2017). Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene

editing efficiency. FEBS Lett. *591*, 1892–1901.

Jensen, R.B., Carreira, A., and Kowalczykowski, S.C. (2010). Purified human BRCA2 stimulates RAD51-mediated recombination. Nature *467*, 678–683.

Jiang, H., Mankodi, A., Swanson, M.S., Moxley, R.T., and Thornton, C.A. (2004). Myotonic dystrophy type 1 is associated with nuclear foci of mutant RNA, sequestration of muscleblind proteins and deregulated alternative splicing in neurons. Hum. Mol. Genet. *13*, 3079–3088.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816–821.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). RNA-programmed genome editing in human cells. Elife *2*, e00471.

Johnson, K.R., Sweet, H.O., Donahue, L.R., Ward-Bailey, P., Bronson, R.T., and Davisson, M.T. (1998). A New Spontaneous Mouse Mutation of Hoxd13 with a Polyalanine Expansion and Phenotype Similar to Human Synpolydactyly. Hum Mol Genet *7*, 1033–1038.

Kadyk, L.C., and Hartwell, L.H. (1992). Sister chromatids are preferred over homologs as substrates for recombinational repair in Saccharomyces cerevisiae. Genetics *132*, 387–402.

Kanadia, R.N., Shin, J., Yuan, Y., Beattie, S.G., Wheeler, T.M., Thornton, C.A., and Swanson, M.S. (2006). Reversal of RNA missplicing and myotonia after muscleblind overexpression in a mouse poly(CUG) model for myotonic dystrophy. Proc. Natl. Acad. Sci. U.S.A. *103*, 11748–11753.

Karanjawala, Z.E., Adachi, N., Irvine, R.A., Oh, E.K., Shibata, D., Schwarz, K., Hsieh, C.-L., and Lieber, M.R. (2002). The embryonic lethality in DNA ligase IV-deficient mice is rescued by deletion of Ku: implications for unifying the heterogeneous phenotypes of NHEJ mutants. DNA Repair (Amst.) *1*, 1017–1026.

Karvelis, T., Gasiunas, G., Miksys, A., Barrangou, R., Horvath, P., and Siksnys, V. (2013). crRNA and tracrRNA guide Cas9-mediated DNA interference in Streptococcus thermophilus. RNA Biol *10*, 841–851.

Kass, E.M., Helgadottir, H.R., Chen, C.-C., Barbera, M., Wang, R., Westermark, U.K., Ludwig, T., Moynahan, M.E., and Jasin, M. (2013). Double-strand break repair by homologous recombination in primary mouse somatic cells requires BRCA1 but not the ATM kinase. Proc. Natl. Acad. Sci. U.S.A. *110*, 5564–5569.

Kass, E.M., Lim, P.X., Helgadottir, H.R., Moynahan, M.E., and Jasin, M. (2016). Robust homology-directed repair within mouse mammary tissue is not specifically affected by Brca2 mutation. Nat Commun *7*.

Kay, S., Hahn, S., Marois, E., Hause, G., and Bonas, U. (2007). A Bacterial Effector Acts as a Plant Transcription Factor and Induces a Cell Size Regulator. Science *318*, 648–651.

Kiliszek, A., and Rypniewski, W. (2014). Structural studies of CNG repeats. Nucleic Acids Res. *42*, 8189–8199.

Kiliszek, A., Kierzek, R., Krzyzosiak, W.J., and Rypniewski, W. (2010). Atomic resolution structure of CAG RNA repeats: structural insights and implications for the trinucleotide repeat expansion diseases. Nucleic Acids Res. *38*, 8370–8376.

Kim, H.-M., Narayanan, V., Mieczkowski, P.A., Petes, T.D., Krasilnikova, M.M., Mirkin, S.M., and Lobachev, K.S. (2008). Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. EMBO J *27*, 2896–2906.

Kim, J.C., Harris, S.T., Dinter, T., Shah, K.A., and Mirkin, S.M. (2017). The role of break-induced replication in large-scale expansions of (CAG)n•(CTG)n repeats. Nat Struct Mol Biol *24*, 55–60.

Kim, Y.G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc Natl Acad Sci U S A *93*, 1156–1160.

Kinniburgh, A.J. (1989). A cis-acting transcription element of the c-myc gene can assume an H-DNA conformation. Nucleic Acids Res. *17*, 7771–7778.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Topkar, V., Nguyen, N.T., Zheng, Z., Gonzales, A.P.W., Li, Z., Peterson, R.T., Yeh, J.-R.J., et al. (2015a). Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature *523*, 481–485.

Kleinstiver, B.P., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Topkar, V.V., Zheng, Z., and Joung, J.K. (2015b). Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat. Biotechnol. *33*, 1293–1298.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016a). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490–495.

Kleinstiver, B.P., Tsai, S.Q., Prew, M.S., Nguyen, N.T., Welch, M.M., Lopez, J.M., McCaw, Z.R., Aryee, M.J., and Joung, J.K. (2016b). Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. Nat Biotechnol *34*, 869–874.

Klesert, T.R., Cho, D.H., Clark, J.I., Maylie, J., Adelman, J., Snider, L., Yuen, E.C., Soriano, P., and Tapscott, S.J. (2000). Mice deficient in Six5 develop cataracts: implications for myotonic dystrophy. Nature Genetics *25*, 105.

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S., and Sternberg, S.H. (2019). Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. Nature *571*, 219–225.

Kobayashi, H., Abe, K., Matsuura, T., Ikeda, Y., Hitomi, T., Akechi, Y., Habu, T., Liu, W., Okuda, H., and Koizumi, A. (2011). Expansion of intronic GGCCTG hexanucleotide repeat in NOP56 causes SCA36, a type of spinocerebellar ataxia accompanied by motor neuron involvement. Am. J. Hum. Genet. *89*, 121–130.

Koike-Yusa, H., Li, Y., Tan, E.-P., Velasco-Herrera, M.D.C., and Yusa, K. (2014). Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. Nat. Biotechnol. *32*, 267–273.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.

Koob, M.D., Moseley, M.L., Schut, L.J., Benzow, K.A., Bird, T.D., Day, J.W., and Ranum, L.P. (1999). An untranslated CTG expansion causes a novel form of spinocerebellar ataxia (SCA8). Nat. Genet. *21*, 379–384.

Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat. Biotechnol. *36*, 765–771.

Kostriken, R., Strathern, J.N., Klar, A.J., Hicks, J.B., and Heffron, F. (1983). A site-specific endonuclease essential for mating-type switching in Saccharomyces cerevisiae. Cell *35*, 167–174.

Kovtun, I.V., and McMurray, C.T. (2001). Trinucleotide expansion in haploid germ cells by gap repair. Nat. Genet. *27*, 407–411.

Kovtun, I.V., Liu, Y., Bjoras, M., Klungland, A., Wilson, S.H., and McMurray, C.T. (2007). OGG1 initiates age-dependent CAG trinucleotide expansion in somatic cells. Nature *447*, 447–452.

Kozlowski, P., de Mezer, M., and Krzyzosiak, W.J. (2010). Trinucleotide repeats in human genome and exome. Nucleic Acids Res *38*, 4027–4039.

Krasilnikova, M.M., and Mirkin, S.M. (2004). Replication stalling at Friedreich's ataxia (GAA)n repeats in vivo. Mol. Cell. Biol. *24*, 2286–2295.

Krokan, H.E., and Bjørås, M. (2013). Base Excision Repair. Cold Spring Harb Perspect Biol *5*.

Kumari, D., Hayward, B., Nakamura, A.J., Bonner, W.M., and Usdin, K. (2015). Evidence for chromosome fragility at the frataxin locus in Friedreich ataxia. Mutat Res *781*, 14–21.

Kuyumcu-Martinez, N.M., Wang, G.-S., and Cooper, T.A. (2007). Increased Steady-State Levels of CUGBP1 in Myotonic Dystrophy 1 Are Due to PKC-Mediated Hyperphosphorylation. Molecular Cell *28*, 68–78.

Lam, S.L., Wu, F., Yang, H., and Chi, L.M. (2011). The origin of genetic instability in CCTG repeats. Nucleic Acids Res *39*, 6260–6268.

Lam, Y.C., Bowman, A.B., Jafar-Nejad, P., Lim, J., Richman, R., Fryer, J.D., Hyun, E.D., Duvick, L.A., Orr, H.T., Botas, J., et al. (2006). ATAXIN-1 interacts with the repressor Capicua in its native complex to cause SCA1 neuropathology. Cell *127*, 1335–1347.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Laoharawee, K., DeKelver, R.C., Podetz-Pedersen, K.M., Rohde, M., Sproul, S., Nguyen, H.-O., Nguyen, T., St Martin, S.J., Ou, L., Tom, S., et al. (2018). Dose-Dependent Prevention of Metabolic and Neurologic Disease in Murine MPS II by ZFN-Mediated In Vivo Genome Editing. Mol. Ther. *26*, 1127–1136.

Larrea, A.A., Lujan, S.A., and Kunkel, T.A. (2010). SnapShot: DNA mismatch repair. Cell *141*, 730.e1.

Lee, J.S., Burkholder, G.D., Latimer, L.J., Haug, B.L., and Braun, R.P. (1987). A monoclonal antibody to triplex DNA binds to eucaryotic chromosomes. Nucleic Acids Res *15*, 1047–1061.

Lee, K., Conboy, M., Park, H.M., Jiang, F., Kim, H.J., Dewitt, M.A., Mackley, V.A., Chang, K., Rao, A., Skinner, C., et al. (2017a). Nanoparticle delivery of Cas9 ribonucleoprotein and donor DNA in vivo induces homology-directed DNA repair. Nat Biomed Eng *1*, 889–901.

Lee, K., Conboy, M., Park, H.M., Jiang, F., Kim, H.J., Dewitt, M.A., Mackley, V.A., Chang, K., Rao, A., Skinner, C., et al. (2017b). Nanoparticle delivery of Cas9 ribonucleoprotein and donor DNA in vivo induces homology-directed DNA repair. Nature Biomedical Engineering *1*, 889.

Lemaire, P., Revelant, O., Bravo, R., and Charnay, P. (1988). Two mouse genes encoding potential transcription factors with identical DNA-binding domains are activated by growth factors in cultured cells. Proc. Natl. Acad. Sci. U.S.A. *85*, 4691–4695.

Lenzmeier, B.A., and Freudenreich, C.H. (2003). Trinucleotide repeat instability: a hairpin curve at the crossroads of replication, recombination, and repair. Cytogenet. Genome Res. *100*, 7–24.

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., Crawford, G., Collins, F., Shinn, P., Leipzig, J., Hannenhalli, S., et al. (2006). Retroviral DNA Integration: Viral and Cellular Determinants of Target-Site Selection. PLOS Pathogens *2*, e60.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, H., Haurigot, V., Doyon, Y., Li, T., Wong, S.Y., Bhagwat, A.S., Malani, N., Anguela, X.M., Sharma, R., Ivanciu, L., et al. (2011). In vivo genome editing restores haemostasis in a mouse model of haemophilia. Nature *475*, 217–221.

Lisby, M., Barlow, J.H., Burgess, R.C., and Rothstein, R. (2004). Choreography of the DNA damage response: spatiotemporal relationships among checkpoint and repair proteins. Cell *118*, 699–713.

Liu, L.F., and Wang, J.C. (1987). Supercoiling of the DNA template during transcription. Proc Natl Acad Sci U S A *84*, 7024–7027.

Liu, G., Chen, X., Bissler, J.J., Sinden, R.R., and Leffak, M. (2010a). Replication-dependent instability at (CTG)•(CAG) repeat hairpins in human cells. Nature Chemical Biology *6*, 652–659.

Liu, H., Chen, Y., Niu, Y., Zhang, K., Kang, Y., Ge, W., Liu, X., Zhao, E., Wang, C., Lin, S., et al. (2014). TALEN-Mediated Gene Mutagenesis in Rhesus and Cynomolgus Monkeys. Cell Stem Cell *14*, 323–328.

Liu, J., Doty, T., Gibson, B., and Heyer, W.-D. (2010b). Human BRCA2 protein promotes RAD51 filament formation on RPA-covered single-stranded DNA. Nat. Struct. Mol. Biol. *17*, 1260–1262.

Liu, Q., Segal, D.J., Ghiara, J.B., and Barbas, C.F. (1997). Design of polydactyl zinc-finger proteins for unique addressing within complex genomes. Proc. Natl. Acad. Sci. U.S.A. *94*, 5525–5530.

Lo Scrudato, M., Poulard, K., Sourd, C., Tomé, S., Klein, A.F., Corre, G., Huguet, A., Furling, D., Gourdon, G., and Buj-Bello, A. (2019). Genome editing of expanded CTG repeats within the human DMPK gene reduces nuclear RNA foci in muscle of DM1 mice. Molecular Therapy.

Lokanga, R.A., Zhao, X.-N., and Usdin, K. (2014). The mismatch repair protein MSH2 is rate limiting for repeat expansion in a fragile X premutation mouse model. Hum. Mutat. *35*, 129–136.

Long, C., McAnally, J.R., Shelton, J.M., Mireault, A.A., Bassel-Duby, R., and Olson, E.N. (2014). Prevention of muscular dystrophy in mice by CRISPR/Cas9–mediated editing of germline DNA. Science *345*, 1184–1188.

Long, C., Amoasii, L., Mireault, A.A., McAnally, J.R., Li, H., Sanchez-Ortiz, E., Bhattacharyya, S., Shelton, J.M., Bassel-Duby, R., and Olson, E.N. (2016). Postnatal genome

editing partially restores dystrophin expression in a mouse model of muscular dystrophy. Science *351*, 400–403.

Lubs, H.A. (1969). A marker X chromosome. Am J Hum Genet *21*, 231–244.

Ma, Y., Pannicke, U., Schwarz, K., and Lieber, M.R. (2002). Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. Cell *108*, 781–794.

Ma, Y., Lu, H., Tippin, B., Goodman, M.F., Shimazaki, N., Koiwai, O., Hsieh, C.-L., Schwarz, K., and Lieber, M.R. (2004). A biochemically defined system for mammalian nonhomologous DNA end joining. Mol. Cell *16*, 701–713.

Machuca-Tzili, L., Brook, D., and Hilton-Jones, D. (2005). Clinical and molecular aspects of the myotonic dystrophies: a review. Muscle Nerve *32*, 1–18.

Magenis, R.E., Hecht, F., and Lovrien, E.W. (1970). Heritable fragile site on chromosome 16: probable localization of haptoglobin locus in man. Science *170*, 85–87.

Maizels, N., and Davis, L. (2018). Initiation of homologous recombination at DNA nicks. Nucleic Acids Res *46*, 6962–6973.

Makarova, K.S., Wolf, Y.I., Alkhnbashi, O.S., Costa, F., Shah, S.A., Saunders, S.J., Barrangou, R., Brouns, S.J.J., Charpentier, E., Haft, D.H., et al. (2015). An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. *13*, 722–736.

Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E., and Church, G.M. (2013). RNA-guided human genome engineering via Cas9. Science *339*, 823–826.

Malina, A., Cameron, C.J.F., Robert, F., Blanchette, M., Dostie, J., and Pelletier, J. (2015). PAM multiplicity marks genomic target sites as inhibitory to CRISPR-Cas9 editing. Nat Commun *6*.

Malkov, V.A., Voloshin, O.N., Veselkov, A.G., Rostapshov, V.M., Jansen, I., Soyfer, V.N., and Frank-Kamenetskii, M.D. (1993). Protonated pyrimidine-purine-purine triplex. Nucleic Acids Res. *21*, 105–111.

Malpertuy, A., Dujon, B., and Richard, G.-F. (2003). Analysis of microsatellites in 13 hemiascomycetous yeast species: mechanisms involved in genome dynamics. J. Mol. Evol. *56*, 730–741.

Mankodi, A., Logigian, E., Callahan, L., McClain, C., White, R., Henderson, D., Krym, M., and Thornton, C.A. (2000). Myotonic dystrophy in transgenic mice expressing an expanded CUG repeat. Science *289*, 1769–1773.

Mankodi, A., Takahashi, M.P., Jiang, H., Beck, C.L., Bowers, W.J., Moxley, R.T., Cannon, S.C., and Thornton, C.A. (2002). Expanded CUG repeats trigger aberrant splicing of ClC-1

chloride channel pre-mRNA and hyperexcitability of skeletal muscle in myotonic dystrophy. Mol. Cell *10*, 35–44.

Martin, J.P., and Bell, J. (1943). A PEDIGREE OF MENTAL DEFECT SHOWING SEX-LINKAGE. J Neurol Psychiatry *6*, 154–157.

Maruyama, T., Dougan, S.K., Truttmann, M.C., Bilate, A.M., Ingram, J.R., and Ploegh, H.L. (2015). Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. Nat. Biotechnol. *33*, 538–542.

Masson, J.Y., Tarsounas, M.C., Stasiak, A.Z., Stasiak, A., Shah, R., McIlwraith, M.J., Benson, F.E., and West, S.C. (2001). Identification and purification of two distinct complexes containing the five RAD51 paralogs. Genes Dev. *15*, 3296–3307.

Matos, J., and West, S.C. (2014). Holliday junction resolution: Regulation in space and time. DNA Repair (Amst) *19*, 176–181.

Matsuura, T., Yamagata, T., Burgess, D.L., Rasmussen, A., Grewal, R.P., Watase, K., Khajavi, M., McCall, A.E., Davis, C.F., Zu, L., et al. (2000). Large expansion of the ATTCT pentanucleotide repeat in spinocerebellar ataxia type 10. Nat. Genet. *26*, 191–194.

McCarty, D.M. (2008). Self-complementary AAV Vectors; Advances and Applications. Molecular Therapy *16*, 1648–1656.

McGarrity, G.J., Hoyah, G., Winemiller, A., Andre, K., Stein, D., Blick, G., Greenberg, R.N., Kinder, C., Zolopa, A., Binder-Scholl, G., et al. (2013). Patient monitoring and follow-up in lentiviral clinical trials. J Gene Med *15*, 78–82.

McMurray, C.T. (2010). Mechanisms of trinucleotide repeat instability during human development. Nat. Rev. Genet. *11*, 786–799.

Miller, D.G., Petek, L.M., and Russell, D.W. (2004). Adeno-associated virus vectors integrate at chromosome breakage sites. Nat. Genet. *36*, 767–773.

Miller, J.C., Holmes, M.C., Wang, J., Guschin, D.Y., Lee, Y.-L., Rupniewski, I., Beausejour, C.M., Waite, A.J., Wang, N.S., Kim, K.A., et al. (2007). An improved zinc-finger nuclease architecture for highly specific genome editing. Nat. Biotechnol. *25*, 778–785.

Miller, J.C., Tan, S., Qiao, G., Barlow, K.A., Wang, J., Xia, D.F., Meng, X., Paschon, D.E., Leung, E., Hinkley, S.J., et al. (2011). A TALE nuclease architecture for efficient genome editing. Nat. Biotechnol. *29*, 143–148.

Miller, J.C., Patil, D.P., Xia, D.F., Paine, C.B., Fauser, F., Richards, H.W., Shivak, D.A., Bendaña, Y.R., Hinkley, S.J., Scarlott, N.A., et al. (2019). Enhancing gene editing specificity by attenuating DNA cleavage kinetics. Nature Biotechnology 1.

Miller, J.W., Urbinati, C.R., Teng-umnuay, P., Stenberg, M.G., Byrne, B.J., Thornton, C.A.,

and Swanson, M.S. (2000). Recruitment of human muscleblind proteins to (CUG)n expansions associated with myotonic dystrophy. The EMBO Journal *19*, 4439–4448.

Mimitou, E.P., and Symington, L.S. (2008). Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature *455*, 770–774.

Mirkin, S.M. (2006). DNA structures, repeat expansions and human hereditary disorders. Curr. Opin. Struct. Biol. *16*, 351–358.

Mirkin, E.V., and Mirkin, S.M. (2007). Replication Fork Stalling at Natural Impediments. Microbiol Mol Biol Rev *71*, 13–35.

Mittelman, D., Moye, C., Morton, J., Sykoudis, K., Lin, Y., Carroll, D., and Wilson, J.H. (2009). Zinc-finger directed double-strand breaks within CAG repeat tracts promote repeat instability in human cells. Proc. Natl. Acad. Sci. U.S.A. *106*, 9607–9612.

Mohrin, M., Bourke, E., Alexander, D., Warr, M.R., Barry-Holson, K., Le Beau, M.M., Morrison, C.G., and Passegué, E. (2010). Hematopoietic stem cell quiescence promotes error prone DNA repair and mutagenesis. Cell Stem Cell *7*, 174–185.

Mojica, F.J.M., Díez-Villaseñor, C., García-Martínez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. *60*, 174–182.

Mosbach, V., Poggi, L., Viterbo, D., Charpentier, M., and Richard, G.-F. (2018). TALEN-Induced Double-Strand Break Repair of CTG Trinucleotide Repeats. Cell Reports *22*, 2146–2159.

Mosbach, V., Poggi, L., and Richard, G.-F. (2019a). Trinucleotide repeat instability during double-strand break repair: from mechanisms to gene therapy. Curr Genet *65*, 17–28.

Mosbach, V., Viterbo, D., Descorps-Declère, S., Poggi, L., Vaysse-Zinkhöfer, W., and Richard, G.-F. (2019b). Resection and repair of a Cas9 double-strand break at CTG trinucleotide repeats induces local and extensive chromosomal rearrangements. BioRxiv 782268.

Moure, C.M., Gimble, F.S., and Quiocho, F.A. (2003). The crystal structure of the gene targeting homing endonuclease I-SceI reveals the origins of its target site specificity. J. Mol. Biol. *334*, 685–695.

Moynahan, M.E., Chiu, J.W., Koller, B.H., and Jasin, M. (1999). Brca1 Controls Homology-Directed DNA Repair. Molecular Cell *4*, 511–518.

Moynahan, M.E., Pierce, A.J., and Jasin, M. (2001). BRCA2 Is Required for Homology-Directed Repair of Chromosomal Breaks. Molecular Cell *7*, 263–272.

Muller, H., Annaluru, N., Schwerzmann, J.W., Richardson, S.M., Dymond, J.S., Cooper, E.M., Bader, J.S., Boeke, J.D., and Chandrasegaran, S. (2012). Assembling large DNA segments in

yeast. Methods Mol. Biol. *852*, 133–150.

Muragaki, Y., Mundlos, S., Upton, J., and Olsen, B.R. (1996). Altered Growth and Branching Patterns in Synpolydactyly Caused by Mutations in HOXD13. Science *272*, 548–551.

Nambiar, T.S., Billon, P., Diedenhofen, G., Hayward, S.B., Taglialatela, A., Cai, K., Huang, J.-W., Leuzzi, G., Cuella-Martin, R., Palacios, A., et al. (2019). Stimulation of CRISPR-mediated homology-directed repair by an engineered RAD18 variant. Nature Communications *10*, 3395.

Narciso, L., Fortini, P., Pajalunga, D., Franchitto, A., Liu, P., Degan, P., Frechet, M., Demple, B., Crescenzi, M., and Dogliotti, E. (2007). Terminally differentiated muscle cells are defective in base excision DNA repair and hypersensitive to oxygen injury. PNAS *104*, 17010–17015.

Nelson, C.E., Hakim, C.H., Ousterout, D.G., Thakore, P.I., Moreb, E.A., Castellanos Rivera, R.M., Madhavan, S., Pan, X., Ran, F.A., Yan, W.X., et al. (2016). In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. Science *351*, 403–407.

Nelson, C.E., Wu, Y., Gemberling, M.P., Oliver, M.L., Waller, M.A., Bohning, J.D., Robinson-Hamm, J.N., Bulaklak, K., Castellanos Rivera, R.M., Collier, J.H., et al. (2019). Long-term evaluation of AAV-CRISPR genome editing for Duchenne muscular dystrophy. Nat. Med. *25*, 427–432.

Nickoloff, J.A., Chen, E.Y., and Heffron, F. (1986). A 24-base-pair DNA sequence from the MAT locus stimulates intergenic recombination in yeast. Proc Natl Acad Sci U S A *83*, 7831–7835.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. Cell *156*, 935–949.

Niu, D., Wei, H.-J., Lin, L., George, H., Wang, T., Lee, I.-H., Zhao, H.-Y., Wang, Y., Kan, Y., Shrock, E., et al. (2017). Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. Science *357*, 1303–1307.

Norris, A.L., Lee, S.S., Greenlees, K.J., Tadesse, D.A., Miller, M.F., and Lombardi, H. (2019). Template plasmid integration in germline genome-edited cattle. BioRxiv 715482.

Nouspikel, T., and Hanawalt, P.C. (2002). DNA repair in terminally differentiated cells. DNA Repair *1*, 59–75.

Nowak, K.J., and Davies, K.E. (2004). Duchenne muscular dystrophy and dystrophin: pathogenesis and opportunities for treatment. EMBO Rep *5*, 872–876.

Oberlé, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boué, J., Bertheas, M.F., and Mandel, J.L. (1991). Instability of a 550-base pair DNA segment and abnormal methylation

in fragile X syndrome. Science *252*, 1097–1102.

Ohno, M., Fukagawa, T., Lee, J.S., and Ikemura, T. (2002). Triplex-forming DNAs in the human interphase nucleus visualized in situ by polypurine/polypyrimidine DNA probes and antitriplex antibodies. Chromosoma *111*, 201–213.

O'Hoy, K.L., Tsilfidis, C., Mahadevan, M.S., Neville, C.E., Barceló, J., Hunter, A.G., and Korneluk, R.G. (1993). Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission. Science *259*, 809–812.

Orr, H.T., and Zoghbi, H.Y. (2007). Trinucleotide repeat disorders. Annu. Rev. Neurosci. *30*, 575–621.

Otten, A.D., and Tapscott, S.J. (1995). Triplet repeat expansion in myotonic dystrophy alters the adjacent chromatin structure. PNAS *92*, 5465–5469.

Owen, B.A.L., Yang, Z., Lai, M., Gajec, M., Gajek, M., Badger, J.D., Hayes, J.J., Edelmann, W., Kucherlapati, R., Wilson, T.M., et al. (2005). (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. Nat. Struct. Mol. Biol. *12*, 663–670.

Panier, S., and Boulton, S.J. (2014). Double-strand break repair: 53BP1 comes into focus. Nat. Rev. Mol. Cell Biol. *15*, 7–18.

Park, J.H., Geyer, M.B., and Brentjens, R.J. (2016). CD19-targeted CAR T-cell therapeutics for hematologic malignancies: interpreting clinical outcomes to date. Blood *127*, 3312–3320.

Parkinson, G.N., Lee, M.P.H., and Neidle, S. (2002). Crystal structure of parallel quadruplexes from human telomeric DNA. Nature *417*, 876–880.

Parsons, R., Li, G.M., Longley, M.J., Fang, W.H., Papadopoulos, N., Jen, J., de la Chapelle, A., Kinzler, K.W., Vogelstein, B., and Modrich, P. (1993). Hypermutability and mismatch repair deficiency in RER+ tumor cells. Cell *75*, 1227–1236.

Paschon, D.E., Lussier, S., Wangzor, T., Xia, D.F., Li, P.W., Hinkley, S.J., Scarlott, N.A., Lam, S.C., Waite, A.J., Truong, L.N., et al. (2019). Diversifying the structure of zinc finger nucleases for high-precision genome editing. Nat Commun *10*, 1–12.

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

Pavletich, N.P., and Pabo, C.O. (1991). Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science *252*, 809–817.

Pearson, C.E., Ewel, A., Acharya, S., Fishel, R.A., and Sinden, R.R. (1997). Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. Hum. Mol. Genet. *6*, 1117–1123.

Pearson, C.E., Wang, Y.H., Griffith, J.D., and Sinden, R.R. (1998). Structural analysis of slipped-strand DNA (S-DNA) formed in (CTG)n. (CAG)n repeats from the myotonic dystrophy

locus. Nucleic Acids Res *26*, 816–823.

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

Pelletier, R., Krasilnikova, M.M., Samadashwily, G.M., Lahue, R., and Mirkin, S.M. (2003). Replication and Expansion of Trinucleotide Repeats in Yeast. Mol Cell Biol *23*, 1349–1357.

Petruska, J., Arnheim, N., and Goodman, M.F. (1996). Stability of intrastrand hairpin structures formed by the CAG/CTG class of DNA triplet repeats associated with neurological diseases. Nucleic Acids Res *24*, 1992–1998.

Pettersson, O.J., Aagaard, L., Jensen, T.G., and Damgaard, C.K. (2015). Molecular mechanisms in DM1 — a focus on foci. Nucleic Acids Res *43*, 2433–2441.

Philips, A.V., Timchenko, L.T., and Cooper, T.A. (1998). Disruption of Splicing Regulated by a CUG-Binding Protein in Myotonic Dystrophy. Science *280*, 737–741.

Pierce, A.J., Johnson, R.D., Thompson, L.H., and Jasin, M. (1999). XRCC3 promotes homology-directed repair of DNA damage in mammalian cells. Genes Dev. *13*, 2633–2638.

Pinto, B.S., Saxena, T., Oliveira, R., Méndez-Gómez, H.R., Cleary, J.D., Denes, L.T., McConnell, O., Arboleda, J., Xia, G., Swanson, M.S., et al. (2017). Impeding Transcription of Expanded Microsatellite Repeats by Deactivated Cas9. Molecular Cell *68*, 479-490.e5.

Poggi, L., Emmenegger, L., Descorps-Declère, S., Dumas, B., and Richard, G.-F. (2019). Differential efficacies of Cas nucleases on microsatellites involved in human disorders and associated off-target mutations. BioRxiv 857714.

Poggi, L., Dumas, B., and Richard, G.-F. (2020). Monitoring Double-Strand Break Repair of Trinucleotide Repeats Using a Yeast Fluorescent Reporter Assay. In Trinucleotide Repeats: Methods and Protocols, G.-F. Richard, ed. (New York, NY: Springer New York), pp. 113–120.

Porteus, M.H., and Baltimore, D. (2003). Chimeric nucleases stimulate gene targeting in human cells. Science *300*, 763.

Potaman, V.N., Bissler, J.J., Hashem, V.I., Oussatcheva, E.A., Lu, L., Shlyakhtenko, L.S., Lyubchenko, Y.L., Matsuura, T., Ashizawa, T., Leffak, M., et al. (2003). Unpaired structures in SCA10 (ATTCT)n.(AGAAT)n repeats. J. Mol. Biol. *326*, 1095–1111.

Provenzano, C., Cappella, M., Valaperta, R., Cardani, R., Meola, G., Martelli, F., Cardinali, B., and Falcone, G. (2017). CRISPR/Cas9-Mediated Deletion of CTG Expansions Recovers Normal Phenotype in Myogenic Cells Derived from Myotonic Dystrophy 1 Patients. Mol Ther Nucleic Acids *9*, 337–348.

Puccio, H., Simon, D., Cossée, M., Criqui-Filipe, P., Tiziano, F., Melki, J., Hindelang, C., Matyas, R., Rustin, P., and Koenig, M. (2001). Mouse models for Friedreich ataxia exhibit

cardiomyopathy, sensory nerve defect and Fe-S enzyme deficiency followed by intramitochondrial iron deposits. Nat. Genet. *27*, 181–186.

Raghavan, S.C., Chastain, P., Lee, J.S., Hegde, B.G., Houston, S., Langen, R., Hsieh, C.-L., Haworth, I.S., and Lieber, M.R. (2005). Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. J. Biol. Chem. *280*, 22749–22760.

Ramakrishna, S., Kwaku Dad, A.-B., Beloor, J., Gopalappa, R., Lee, S.-K., and Kim, H. (2014). Gene disruption by cell-penetrating peptide-mediated delivery of Cas9 protein and guide RNA. Genome Res. *24*, 1020–1027.

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., and Zhang, F. (2013). Genome engineering using the CRISPR-Cas9 system. Nat Protoc *8*, 2281–2308.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015a). In vivo genome editing using Staphylococcus aureus Cas9. Nature *520*, 186–191.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., et al. (2015b). In vivo genome editing using Staphylococcus aureus Cas9. Nature *520*, 186–191.

Ranum, L.P.W., and Cooper, T.A. (2006). RNA-mediated neuromuscular disorders. Annu. Rev. Neurosci. *29*, 259–277.

Raper, S.E., Chirmule, N., Lee, F.S., Wivel, N.A., Bagg, A., Gao, G., Wilson, J.M., and Batshaw, M.L. (2003). Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. Molecular Genetics and Metabolism *80*, 148–158.

Rau, F., Freyermuth, F., Fugier, C., Villemin, J.-P., Fischer, M.-C., Jost, B., Dembele, D., Gourdon, G., Nicole, A., Duboc, D., et al. (2011). Misregulation of miR-1 processing is associated with heart defects in myotonic dystrophy. Nat. Struct. Mol. Biol. *18*, 840–845.

Reddy, K., Schmidt, M.H.M., Geist, J.M., Thakkar, N.P., Panigrahi, G.B., Wang, Y.-H., and Pearson, C.E. (2014). Processing of double-R-loops in (CAG)·(CTG) and C9orf72 (GGGGCC)·(GGCCCC) repeats causes instability. Nucleic Acids Res *42*, 10473–10487.

Reliene, R., and Schiestl, R.H. (2003). Mouse models for induced genetic instability at endogenous loci. Oncogene *22*, 7000.

Reyon, D., Tsai, S.Q., Khayter, C., Foden, J.A., Sander, J.D., and Joung, J.K. (2012). FLASH assembly of TALENs for high-throughput genome editing. Nat. Biotechnol. *30*, 460–465.

Riballo, E., Woodbine, L., Stiff, T., Walker, S.A., Goodarzi, A.A., and Jeggo, P.A. (2009). XLF-Cernunnos promotes DNA ligase IV–XRCC4 re-adenylation following ligation. Nucleic

Acids Res *37*, 482–492.

Ribeil, J.-A., Hacein-Bey-Abina, S., Payen, E., Magnani, A., Semeraro, M., Magrin, E., Caccavelli, L., Neven, B., Bourget, P., El Nemer, W., et al. (2017). Gene Therapy in a Patient with Sickle Cell Disease. New England Journal of Medicine *376*, 848–855.

Richard, G.-F. (2015). Shortening trinucleotide repeats using highly specific endonucleases: a possible approach to gene therapy? Trends Genet. *31*, 177–186.

Richard, G.-F., Fairhead, C., and Dujon, B. (1997). Complete transcriptional map of yeast chromosome XI in different life conditions. J. Mol. Biol. *268*, 303–321.

Richard, G.-F., Dujon, B., and Haber, J.E. (1999). Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. Mol Gen Genet *261*, 871–882.

Richard, G.-F., Goellner, G.M., McMurray, C.T., and Haber, J.E. (2000). Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11–RAD50–XRS2 complex. The EMBO Journal *19*, 2381–2390.

Richard, G.-F., Cyncynatus, C., and Dujon, B. (2003). Contractions and Expansions of CAG/CTG Trinucleotide Repeats occur during Ectopic Gene Conversion in Yeast, by a MUS81-independent Mechanism. Journal of Molecular Biology *326*, 769–782.

Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev. *72*, 686–727.

Richard, G.-F., Viterbo, D., Khanna, V., Mosbach, V., Castelain, L., and Dujon, B. (2014). Highly Specific Contractions of a Single CAG/CTG Trinucleotide Repeat by TALEN in Yeast. PLOS ONE *9*, e95611.

Richards, R.I., and Sutherland, G.R. (1992). Heritable unstable DNA sequences. Nature Genetics *1*, 7.

Richards, B.W., Sylvester, P.E., and Brooker, C. (1981). Fragile X-linked mental retardation: the Martin-Bell syndrome. J Ment Defic Res *25 Pt 4*, 253–256.

Roessler, E., Lacbawan, F., Dubourg, C., Paulussen, A., Herbergs, J., Hehr, U., Bendavid, C., Zhou, N., Ouspenskaia, M., Bale, S., et al. (2009). The full spectrum of holoprosencephaly-associated mutations within the ZIC2 gene in humans predicts loss-of-function as the predominant disease mechanism. Hum. Mutat. *30*, E541-554.

Rose, J.A., Hoggan, M.D., and Shatkin, A.J. (1966). Nucleic acid from an adeno-associated virus: chemical and physical studies. Proc Natl Acad Sci U S A *56*, 86–92.

Ross, C.A., and Poirier, M.A. (2004). Protein aggregation and neurodegenerative disease. Nat. Med. *10 Suppl*, S10-17.

Rothenberg, E., Grimme, J.M., Spies, M., and Ha, T. (2008). Human Rad52-mediated homology search and annealing occurs by continuous interactions between overlapping nucleoprotein complexes. Proc. Natl. Acad. Sci. U.S.A. *105*, 20274–20279.

Rothstein, R., Michel, B., and Gangloff, S. (2000). Replication fork pausing and recombination or "gimme a break." Genes Dev. *14*, 1–10.

Rouet, P., Smih, F., and Jasin, M. (1994). Expression of a site-specific endonuclease stimulates homologous recombination in mammalian cells. Proc. Natl. Acad. Sci. U.S.A. *91*, 6064–6068.

Rowland, L.P., and Shneider, N.A. (2001). Amyotrophic lateral sclerosis. N. Engl. J. Med. *344*, 1688–1700.

Ruan, G.-X., Barry, E., Yu, D., Lukason, M., Cheng, S.H., and Scaria, A. (2017). CRISPR/Cas9-Mediated Genome Editing as a Therapeutic Approach for Leber Congenital Amaurosis 10. Mol. Ther. *25*, 331–341.

Ruff, P., Donnianni, R.A., Glancy, E., Oh, J., and Symington, L.S. (2016). RPA stabilization of single-stranded DNA is critical for break-induced replication. Cell Rep *17*, 3359–3368.

Rzuczek, S.G., Colgan, L.A., Nakai, Y., Cameron, M.D., Furling, D., Yasuda, R., and Disney, M.D. (2017). Precise small-molecule recognition of a toxic CUG RNA repeat expansion. Nature Chemical Biology *13*, 188–193.

Samadashwily, G.M., Raca, G., and Mirkin, S.M. (1997). Trinucleotide repeats affect DNA replication in vivo. Nat. Genet. *17*, 298–304.

Samulski, R.J., and Muzyczka, N. (2014). AAV-Mediated Gene Therapy for Research and Therapeutic Purposes. Annu Rev Virol *1*, 427–451.

Sanber, K.S., Knight, S.B., Stephen, S.L., Bailey, R., Escors, D., Minshull, J., Santilli, G., Thrasher, A.J., Collins, M.K., and Takeuchi, Y. (2015). Construction of stable packaging cell lines for clinical lentiviral vector production. Sci Rep *5*, 9021.

Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F., et al. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nat Commun *9*, 1–15.

Santillan, B.A., Moye, C., Mittelman, D., and Wilson, J.H. (2014). GFP-Based Fluorescence Assay for CAG Repeat Instability in Cultured Human Cells. PLOS ONE *9*, e113952.

Sato, N., Amino, T., Kobayashi, K., Asakawa, S., Ishiguro, T., Tsunemi, T., Takahashi, M., Matsuura, T., Flanigan, K.M., Iwasaki, S., et al. (2009). Spinocerebellar ataxia type 31 is associated with "inserted" penta-nucleotide repeats containing (TGGAA)n. Am. J. Hum. Genet. *85*, 544–557.

Savkur, R.S., Philips, A.V., and Cooper, T.A. (2001). Aberrant regulation of insulin receptor

alternative splicing is associated with insulin resistance in myotonic dystrophy. Nat. Genet. *29*, 40–47.

Schaffitzel, C., Berger, I., Postberg, J., Hanes, J., Lipps, H.J., and Plückthun, A. (2001). In vitro generated antibodies specific for telomeric guanine-quadruplex DNA react with Stylonychia lemnae macronuclei. PNAS *98*, 8572–8577.

Schiestl, R.H., and Prakash, S. (1988). RAD1, an excision repair gene of Saccharomyces cerevisiae, is also involved in recombination. Mol Cell Biol *8*, 3619–3626.

Schneider, L., Fumagalli, M., and d'Adda di Fagagna, F. (2012). Terminally differentiated astrocytes lack DNA damage response signaling and are radioresistant but retain DNA repair proficiency. Cell Death Differ *19*, 582–591.

Schornack, S., Meyer, A., Römer, P., Jordan, T., and Lahaye, T. (2006). Gene-for-gene-mediated recognition of nuclear-targeted AvrBs3-like bacterial effector proteins. Journal of Plant Physiology *163*, 256–272.

Schroth, G.P., Chou, P.J., and Ho, P.S. (1992). Mapping Z-DNA in the human genome. Computer-aided mapping reveals a nonrandom distribution of potential Z-DNA-forming sequences in human genes. J. Biol. Chem. *267*, 11846–11855.

Schweitzer, J.K., and Livingston, D.M. (1997). Destabilization of CAG trinucleotide repeat tracts by mismatch repair mutations in yeast. Hum. Mol. Genet. *6*, 349–355.

Semenova, E., Jore, M.M., Datsenko, K.A., Semenova, A., Westra, E.R., Wanner, B., van der Oost, J., Brouns, S.J.J., and Severinov, K. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U.S.A. *108*, 10098–10103.

Seznec, H., Lia-Baldini, A.-S., Duros, C., Fouquet, C., Lacroix, C., Hofmann-Radvanyi, H., Junien, C., and Gourdon, G. (2000). Transgenic mice carrying large human genomic sequences with expanded CTG repeat mimic closely the DM CTG repeat intergenerational and somatic instability. Hum Mol Genet *9*, 1185–1194.

Shelbourne, P., Winqvist, R., Kunert, E., Davies, J., Leisti, J., Thiele, H., Bachmann, H., Buxton, J., Williamson, B., and Johnson, K. (1992). Unstable DNA may be responsible for the incomplete penetrance of the myotonic dystrophy phenotype. Hum. Mol. Genet. *1*, 467–473.

Shi, T., Bunker, R.D., Mattarocci, S., Ribeyre, C., Faty, M., Gut, H., Scrima, A., Rass, U., Rubin, S.M., Shore, D., et al. (2013). Rif1 and Rif2 Shape Telomere Function and Architecture through Multivalent Rap1 Interactions. Cell *153*, 1340–1353.

Shin, H.Y., Wang, C., Lee, H.K., Yoo, K.H., Zeng, X., Kuhns, T., Yang, C.M., Mohr, T., Liu, C., and Hennighausen, L. (2017). CRISPR/Cas9 targeting events cause complex deletions and

insertions at 17 sites in the mouse genome. Nat Commun *8*, 15464.

Shishkin, A.A., Voineagu, I., Matera, R., Cherng, N., Chernet, B.T., Krasilnikova, M.M., Narayanan, V., Lobachev, K.S., and Mirkin, S.M. (2009). Large-Scale Expansions of Friedreich's Ataxia GAA Repeats in Yeast. Molecular Cell *35*, 82–92.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O.O., Gootenberg, J.S., Makarova, K.S., Wolf, Y.I., et al. (2017). Diversity and evolution of class 2 CRISPR–Cas systems. Nat Rev Microbiol *15*, 169–182.

Sikorski, R.S., and Hieter, P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics *122*, 19–27.

Silverman, J., Takai, H., Buonomo, S.B.C., Eisenhaber, F., and de Lange, T. (2004). Human Rif1, ortholog of a yeast telomeric protein, is regulated by ATM and 53BP1 and functions in the S-phase checkpoint. Genes Dev. *18*, 2108–2119.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

Smith, J., Bibikova, M., Whitby, F.G., Reddy, A.R., Chandrasegaran, S., and Carroll, D. (2000). Requirements for double-strand cleavage by chimeric restriction enzymes with zinc finger DNA-recognition domains. Nucleic Acids Res. *28*, 3361–3369.

Smith, J., Grizot, S., Arnould, S., Duclert, A., Epinat, J.-C., Chames, P., Prieto, J., Redondo, P., Blanco, F.J., Bravo, J., et al. (2006). A combinatorial approach to create artificial homing endonucleases cleaving chosen sequences. Nucleic Acids Res *34*, e149.

Sobczak, K., Wheeler, T.M., Wang, W., and Thornton, C.A. (2013). RNA Interference Targeting CUG Repeats in a Mouse Model of Myotonic Dystrophy. Mol Ther *21*, 380–387.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 110–113.

Strand, M., Prolla, T.A., Liskay, R.M., and Petes, T.D. (1993). Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature *365*, 274–276.

Strathern, J.N., Weinstock, K.G., Higgins, D.R., and McGill, C.B. (1991). A novel recombinator in yeast based on gene II protein from bacteriophage f1. Genetics *127*, 61–73.

Suenaga, K., Lee, K.-Y., Nakamori, M., Tatsumi, Y., Takahashi, M.P., Fujimura, H., Jinnai, K., Yoshikawa, H., Du, H., Ares, M., et al. (2012). Muscleblind-like 1 knockout mice reveal novel splicing defects in the myotonic dystrophy brain. PLoS ONE *7*, e33218.

Sugawara, N., and Haber, J.E. (2006). Repair of DNA double strand breaks: in vivo biochemistry. Meth. Enzymol. *408*, 416–429.

Sugawara, N., Pâques, F., Colaiácovo, M., and Haber, J.E. (1997). Role of Saccharomyces cerevisiae Msh2 and Msh3 repair proteins in double-strand break-induced recombination. PNAS *94*, 9214–9219.

Sugawara, N., Ira, G., and Haber, J.E. (2000). DNA length dependence of the single-strand annealing pathway and the role of Saccharomyces cerevisiae RAD59 in double-strand break repair. Mol. Cell. Biol. *20*, 5300–5309.

Sugisaki, H., and Kanazawa, S. (1981). New restriction endonucleases from Flavobacterium okeanokoites (FokI) and Micrococcus luteus (MluI). Gene *16*, 73–78.

Sukup-Jackson, M.R., Kiraly, O., Kay, J.E., Na, L., Rowland, E.A., Winther, K.E., Chow, D.N., Kimoto, T., Matsuguchi, T., Jonnalagadda, V.S., et al. (2014). Rosa26-GFP Direct Repeat (RaDR-GFP) Mice Reveal Tissue- and Age-Dependence of Homologous Recombination in Mammals In Vivo. PLoS Genet *10*.

Suwaki, N., Klare, K., and Tarsounas, M. (2011). RAD51 paralogs: Roles in DNA damage signalling, recombinational repair and tumorigenesis. Seminars in Cell & Developmental Biology *22*, 898–905.

Sy, S.M.H., Huen, M.S.Y., and Chen, J. (2009). PALB2 is an integral component of the BRCA complex required for homologous recombination repair. Proc. Natl. Acad. Sci. U.S.A. *106*, 7155–7160.

Syed, A., and Tainer, J.A. (2018). The MRE11–RAD50–NBS1 Complex Conducts the Orchestration of Damage Signaling and Outcomes to Stress in DNA Replication and Repair. Annu Rev Biochem *87*, 263–294.

Symington, L.S. (2016). Mechanism and regulation of DNA end resection in eukaryotes. Crit. Rev. Biochem. Mol. Biol. *51*, 195–212.

Symington, L.S., and Gautier, J. (2011). Double-strand break end resection and repair pathway choice. Annu. Rev. Genet. *45*, 247–271.

Tabebordbar, M., Zhu, K., Cheng, J.K.W., Chew, W.L., Widrick, J.J., Yan, W.X., Maesner, C., Wu, E.Y., Xiao, R., Ran, F.A., et al. (2016). In vivo gene editing in dystrophic mouse muscle and muscle stem cells. Science *351*, 407–411.

Tan, Y., Chu, A.H.Y., Bao, S., Hoang, D.A., Kebede, F.T., Xiong, W., Ji, M., Shi, J., and Zheng, Z. (2019). Rationally engineered Staphylococcus aureus Cas9 nucleases with high genome-wide specificity. PNAS *116*, 20969–20976.

Taneja, K.L., McCurrach, M., Schalling, M., Housman, D., and Singer, R.H. (1995). Foci of

trinucleotide repeat transcripts in nuclei of myotonic dystrophy cells and tissues. J. Cell Biol. *128*, 995–1002.

Thornton, C.A., Johnson, K., and Moxley, R.T. (1994). Myotonic dystrophy patients have larger CTG expansions in skeletal muscle than in leukocytes. Annals of Neurology *35*, 104–107.

Thornton, C.A., Wymer, J.P., Simmons, Z., McClain, C., and Moxley, R.T. (1997). Expansion of the myotonic dystrophy CTG repeat reduces expression of the flanking DMAHP gene. Nature Genetics *16*, 407.

Thyme, S.B., Akhmetova, L., Montague, T.G., Valen, E., and Schier, A.F. (2016). Internal guide RNA interactions interfere with Cas9-mediated cleavage. Nat Commun *7*, 11750.

Thys, R.G., and Wang, Y.-H. (2015). DNA Replication Dynamics of the GGGGCC Repeat of the C9orf72 Gene. J. Biol. Chem. *290*, 28953–28962.

Tichy, E.D., Pillai, R., Deng, L., Liang, L., Tischfield, J., Schwemberger, S.J., Babcock, G.F., and Stambrook, P.J. (2010). Mouse embryonic stem cells, but not somatic cells, predominantly use homologous recombination to repair double-strand DNA breaks. Stem Cells Dev. *19*, 1699–1711.

Timchenko, L.T., Miller, J.W., Timchenko, N.A., DeVore, D.R., Datar, K.V., Lin, L., Roberts, R., Caskey, C.T., and Swanson, M.S. (1996). Identification of a (CUG) n Triplet Repeat RNA-Binding Protein and Its Expression in Myotonic Dystrophy. Nucleic Acids Res *24*, 4407–4414.

Tomé, S., Holt, I., Edelmann, W., Morris, G.E., Munnich, A., Pearson, C.E., and Gourdon, G. (2009). MSH2 ATPase domain mutation affects CTG*CAG repeat instability in transgenic mice. PLoS Genet. *5*, e1000482.

Tran, N.-T., Bashir, S., Li, X., Rossius, J., Chu, V.T., Rajewsky, K., and Kühn, R. (2019). Enhancement of Precise Gene Editing by the Association of Cas9 With Homologous Recombination Factors. Front Genet *10*.

Trochet, D., de Pontual, L., Keren, B., Munnich, A., Vekemans, M., Lyonnet, S., and Amiel, J. (2007). Polyalanine expansions might not result from unequal crossing-over. Hum. Mutat. *28*, 1043–1044.

Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat. Biotechnol. *33*, 187–197.

Tsai, S.Q., Nguyen, N.T., Malagon-Lopez, J., Topkar, V.V., Aryee, M.J., and Joung, J.K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nature Methods *advance online publication*.

Umek, R.M., and Kowalski, D. (1990). The DNA unwinding element in a yeast replication origin functions independently of easily unwound sequences present elsewhere on a plasmid. Nucleic Acids Res *18*, 6601–6605.

Usdin, K., and Woodford, K.J. (1995). CGG repeats associated with DNA instability and chromosome fragility form structures that block DNA synthesis in vitro. Nucleic Acids Res *23*, 4202–4209.

Usdin, K., House, N.C.M., and Freudenreich, C.H. (2015). Repeat instability during DNA repair: Insights from model systems. Crit. Rev. Biochem. Mol. Biol. *50*, 142–167.

Valencia, M., Bentele, M., Vaze, M.B., Herrmann, G., Kraus, E., Lee, S.E., Schär, P., and Haber, J.E. (2001). NEJ1 controls non-homologous end joining in Saccharomyces cerevisiae. Nature *414*, 666–669.

Valton, J., Guyot, V., Marechal, A., Filhol, J.-M., Juillerat, A., Duclert, A., Duchateau, P., and Poirot, L. (2015). A Multidrug-resistant Engineered CAR T Cell for Allogeneic Combination Immunotherapy. Molecular Therapy *23*, 1507–1518.

Vannucci, L., Lai, M., Chiuppesi, F., Ceccherini-Nelli, L., and Pistello, M. (2013). Viral vectors: a look back and ahead on gene transfer technology. New Microbiol. *36*, 1–22.

Veeranagouda, Y., and Didier, M. (2017). Transposon Insertion Site Sequencing (TIS-Seq): An Efficient and High-Throughput Method for Determining Transposon Insertion Site(s) and Their Relative Abundances in a PiggyBac Transposon Mutant Pool by Next-Generation Sequencing. Curr Protoc Mol Biol *120*, 21.35.1-21.35.11.

Verkerk, A.J.M.H., Pieretti, M., Sutcliffe, J.S., Fu, Y.-H., Kuhl, D.P.A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M.F., Zhang, F., et al. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell *65*, 905–914.

Viterbo, D., Michoud, G., Mosbach, V., Dujon, B., and Richard, G.-F. (2016). Replication stalling and heteroduplex formation within CAG/CTG trinucleotide repeats by mismatch repair. DNA Repair (Amst.) *42*, 94–106.

Viterbo, D., Marchal, A., Mosbach, V., Poggi, L., Vaysse-Zinkhöfer, W., and Richard, G.-F. (2018). A fast, sensitive and cost-effective method for nucleic acid detection using non-radioactive probes. Biol Methods Protoc *3*.

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

Wagers, A.J. (2012). The Stem Cell Niche in Regenerative Medicine. Cell Stem Cell *10*, 362–369.

Wah, D.A., Bitinaite, J., Schildkraut, I., and Aggarwal, A.K. (1998). Structure of FokI has implications for DNA cleavage. Proc Natl Acad Sci U S A *95*, 10564–10569.

Wakamiya, M., Matsuura, T., Liu, Y., Schuster, G.C., Gao, R., Xu, W., Sarkar, P.S., Lin, X., and Ashizawa, T. (2006). The role of ataxin 10 in the pathogenesis of spinocerebellar ataxia type 10. Neurology *67*, 607–613.

Wang, A.H.-J., Quigley, G.J., Kolpak, F.J., Crawford, J.L., van Boom, J.H., van der Marel, G., and Rich, A. (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature *282*, 680–686.

Wang, E.T., Cody, N.A.L., Jog, S., Biancolella, M., Wang, T.T., Treacy, D.J., Luo, S., Schroth, G.P., Housman, D.E., Reddy, S., et al. (2012). Transcriptome-wide Regulation of Pre-mRNA Splicing and mRNA Localization by Muscleblind Proteins. Cell *150*, 710–724.

Ward, A.J., Rimer, M., Killian, J.M., Dowling, J.J., and Cooper, T.A. (2010). CUGBP1 overexpression in mouse skeletal muscle reproduces features of myotonic dystrophy type 1. Hum. Mol. Genet. *19*, 3614–3622.

Warren, S.T. (1997). Polyalanine expansion in synpolydactyly might result from unequal crossing-over of HOXD13. Science *275*, 408–409.

Watson, J.D., and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature *171*, 737–738.

Weber, N.D., Stone, D., Sedlak, R.H., Feelixge, H.S.D.S., Roychoudhury, P., Schiffer, J.T., Aubert, M., and Jerome, K.R. (2014). AAV-Mediated Delivery of Zinc Finger Nucleases Targeting Hepatitis B Virus Inhibits Active Replication. PLOS ONE *9*, e97579.

Wheeler, T.M., Sobczak, K., Lueck, J.D., Osborne, R.J., Lin, X., Dirksen, R.T., and Thornton, C.A. (2009). Reversal of RNA dominance by displacement of protein sequestered on triplet repeat RNA. Science *325*, 336–339.

Wheeler, T.M., Leger, A.J., Pandey, S.K., MacLeod, A.R., Nakamori, M., Cheng, S.H., Wentworth, B.M., Bennett, C.F., and Thornton, C.A. (2012). Targeting nuclear RNA for in vivo correction of myotonic dystrophy. Nature *488*, 111–115.

White, M., Xia, G., Gao, R., Wakamiya, M., Sarkar, P.S., McFarland, K., and Ashizawa, T. (2012). Transgenic mice with SCA10 pentanucleotide repeats show motor phenotype and susceptibility to seizure: a toxic RNA gain-of-function model. J. Neurosci. Res. *90*, 706–714.

White, M.C., Gao, R., Xu, W., Mandal, S.M., Lim, J.G., Hazra, T.K., Wakamiya, M., Edwards, S.F., Raskin, S., Teive, H.A.G., et al. (2010). Inactivation of hnRNP K by expanded intronic AUUCU repeat induces apoptosis via translocation of PKCdelta to mitochondria in spinocerebellar ataxia 10. PLoS Genet. *6*, e1000984.

Willemsen, R., Levenga, J., and Oostra, B.A. (2011). CGG repeat in the FMR1 gene: size matters. Clin Genet *80*, 214–225.

Williams, R.M., Senanayake, U., Artibani, M., Taylor, G., Wells, D., Ahmed, A.A., and Sauka-Spengler, T. (2018). Genome and epigenome engineering CRISPR toolkit for in vivo modulation of cis-regulatory interactions and gene expression in the chicken embryo. Development *145*.

Wright, A.V., Sternberg, S.H., Taylor, D.W., Staahl, B.T., Bardales, J.A., Kornfeld, J.E., and Doudna, J.A. (2015). Rational design of a split-Cas9 enzyme complex. PNAS *112*, 2984–2989.

Xiong, Y., and Sundaralingam, M. (2000). Crystal structure of a DNA·RNA hybrid duplex with a polypurine RNA r(gaagaagag) and a complementary polypyrimidine DNA d(CTCTTCTTC). Nucleic Acids Res *28*, 2171–2176.

Xu, Z., Poidevin, M., Li, X., Li, Y., Shu, L., Nelson, D.L., Li, H., Hales, C.M., Gearing, M., Wingo, T.S., et al. (2013). Expanded GGGGCC repeat RNA associated with amyotrophic lateral sclerosis and frontotemporal dementia causes neurodegeneration. Proc. Natl. Acad. Sci. U.S.A. *110*, 7778–7783.

Xue, Y., Rushton, M.D., and Maringele, L. (2011). A novel checkpoint and RPA inhibitory pathway regulated by Rif1. PLoS Genet. *7*, e1002417.

Yáñez, R.J., and Porter, A.C. (1999). Gene targeting is enhanced in human cells overexpressing hRAD51. Gene Ther. *6*, 1282–1290.

Yáñez, R.J., and Porter, A.C.G. (2002). Differential effects of Rad52p overexpression on gene targeting and extrachromosomal homologous recombination in a human cell line. Nucleic Acids Res *30*, 740–748.

Yang, Z., Lau, R., Marcadier, J.L., Chitayat, D., and Pearson, C.E. (2003). Replication inhibitors modulate instability of an expanded trinucleotide repeat at the myotonic dystrophy type 1 disease locus in human cells. Am. J. Hum. Genet. *73*, 1092–1105.

Yoshida, K., Matsushima, A., and Nakamura, K. (2017). Inter-generational instability of inserted repeats during transmission in spinocerebellar ataxia type 31. J. Hum. Genet. *62*, 923–925.

Zeigelboim, B.S., Mesti, J.C., Fonseca, V.R., Faryniuk, J.H., Marques, J.M., Cardoso, R.C., and Teive, H.A.G. (2017). Otoneurological Abnormalities in Patients with Friedreich's Ataxia. Int Arch Otorhinolaryngol *21*, 79–85.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., et al. (2015). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell *163*, 759–771.

Zhang, F., Ma, J., Wu, J., Ye, L., Cai, H., Xia, B., and Yu, X. (2009). PALB2 links BRCA1 and BRCA2 in the DNA-damage response. Curr. Biol. *19*, 524–529.

Zhang, X.-H., Tee, L.Y., Wang, X.-G., Huang, Q.-S., and Yang, S.-H. (2015). Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Molecular Therapy - Nucleic Acids *4*, e264.

Zhu, L., Chou, S.H., and Reid, B.R. (1996). A single G-to-C change causes human centromere TGGAA repeats to fold back into hairpins. PNAS *93*, 12159–12164.

Zierhut, C., and Diffley, J.F.X. (2008). Break dosage, cell cycle stage and DNA replication influence DNA double strand break response. EMBO J. *27*, 1875–1885.

Zimmermann, M., Lottersberger, F., Buonomo, S.B., Sfeir, A., and de Lange, T. (2013). 53BP1 regulates DSB repair using Rif1 to control 5' end resection. Science *339*, 700–704.

Zincarelli, C., Soltys, S., Rengo, G., and Rabinowitz, J.E. (2008). Analysis of AAV Serotypes 1–9 Mediated Gene Expression and Tropism in Mice After Systemic Injection. Molecular Therapy *16*, 1073–1080.

Zu, T., Gibbens, B., Doty, N.S., Gomes-Pereira, M., Huguet, A., Stone, M.D., Margolis, J., Peterson, M., Markowski, T.W., Ingram, M.A.C., et al. (2011). Non-ATG–initiated translation directed by microsatellite expansions. PNAS *108*, 260–265.

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

Zuris, J.A., Thompson, D.B., Shu, Y., Guilinger, J.P., Bessen, J.L., Hu, J.H., Maeder, M.L., Joung, J.K., Chen, Z.-Y., and Liu, D.R. (2015). Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. Nat. Biotechnol. *33*, 73–80.

Annexes

Annex 1: Trinucleotide repeat instability during double-strand break repair: from mechanisms to gene therapy

Article reviewing the mechanisms of DSB repairs in microsatellites and the advances in genome edition approaches to treat the associated disorders. I wrote the part II about gene edition.

Current Genetics (2019) 65:17–28
https://doi.org/10.1007/s00294-018-0865-1

REVIEW

Trinucleotide repeat instability during double-strand break repair: from mechanisms to gene therapy

Valentine Mosbach[1,3] · Lucie Poggi[1,2,3,4] · Guy-Franck Richard[1,3]

Received: 30 April 2018 / Revised: 25 June 2018 / Accepted: 1 July 2018 / Published online: 5 July 2018

Abstract
Trinucleotide repeats are a particular class of microsatellites whose large expansions are responsible for at least two dozen human neurological and developmental disorders. Slippage of the two complementary DNA strands during replication, homologous recombination or DNA repair is generally accepted as a mechanism leading to repeat length changes, creating expansions and contractions of the repeat tract. The present review focuses on recent developments on double-strand break repair involving trinucleotide repeat tracts. Experimental evidences in model organisms show that gene conversion and break-induced replication may lead to large repeat tract expansions, while frequent contractions occur either by single-strand annealing between repeat ends or by gene conversion, triggering near-complete contraction of the repeat tract. In the second part of this review, different therapeutic approaches using highly specific single- or double-strand endonucleases targeted to trinucleotide repeat loci are compared. Relative efficacies and specificities of these nucleases will be discussed, as well as their potential strengths and weaknesses for possible future gene therapy of these dramatic disorders.

Keywords Gene conversion · Break-induced replication · Single-strand annealing · ZFN · TALEN · CRISPR-Cas9

Abbreviations

BIR	Break-induced replication
CRISPR	Clustered regularly interspaced short palindromic repeats
SSA	Single-strand annealing
ZFN	Zinc-finger nucleases
TALEN	Transcription activator-like effector nuclease
DSB	Double-strand break
SDSA	Synthesis-dependent strand annealing
UAS	Upstream activating sequence
MRX complex	Mre11-Rad50-Xrs2 complex
PAM	Protospacer adjacent motif
iPSC	Induced pluripotent stem cells
sgRNA (or gRNA)	Single-guide RNA
*Sp*Cas9	*Streptococcus pyogenes* Cas9
*Sa*Cas9	*Staphylococcus aureus* Cas9
HNH	Homing endonuclease domain
HEK293	Human embryonic kidney cell line 293
K562	Human immortalized myelogenous leukemia cell line
AAV	Adenovirus-associated vector

Communicated by M. Kupiec.

✉ Guy-Franck Richard
gfrichar@pasteur.fr

[1] Department Genomes and Genetics, Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France

[2] Collège Doctoral, Sorbonne Université, 4 Place Jussieu, 75005 Paris, France

[3] CNRS, UMR3525, 75015 Paris, France

[4] Biologics Research, Sanofi R&D, 13 Quai Jules Guesde, 94403 Vitry sur Seine, France

Introduction

Trinucleotide repeats are a particular class of microsatellites whose large expansions are responsible for at least two dozen human neurological and developmental disorders, discovered over the past 27 years (Fu et al. 1991). Molecular mechanisms responsible for these dramatic large expansions are not totally understood. Yet, experiments in model organisms (mainly bacteria, yeast and mouse) have been fruitful in unraveling some of the key processes underlying trinucleotide repeat instability. These mechanisms involve two features: the ability for these repeats to form stable secondary

structures in a test tube (and most probably in vivo too; Liu et al. 2010) and the capacity to form DNA heteroduplex (or slipped-strand DNA) by slippage of the newly synthesized strand on the template strand, during DNA synthesis associated with replication, repair or recombination. These features have been extensively described and commented in a number of recent reviews on trinucleotide repeats (Richard et al. 2008; McMurray 2010; Kim and Mirkin 2013; Usdin et al. 2015; Neil Alexander et al. 2017; McGinty and Mirkin 2018). Here, we will specifically focus on recent developments involving double-strand breaks as a source of genetic variability for these unstable repeated sequences. The role of gene conversion, break-induced replication (BIR) and single-strand annealing (SSA) in trinucleotide repeat expansions and contractions will be discussed. In addition, several approaches using highly specific DNA endonucleases, such as zinc-finger nucleases (ZFN), TALE nucleases (TALEN) or CRISPR-Cas nucleases were undertaken as possible gene therapies for disorders associated to trinucleotide repeat expansions. Progresses as well as obstacles in each of these different approaches will be discussed.

Double-strand break repair triggers CAG/CTG repeat expansions and contractions by different mechanisms

Some trinucleotide repeats impair replication fork progression, leading to chromosomal fragility and double-strand breaks (DSB), like for example CGG repeats in the fragile X syndrome (Yudkin et al. 2014). Former experiments in yeast showed that some repeats exhibit a length-dependent propensity to break in vivo (Callahan et al. 2003; Freudenreich et al. 1998; Jankowski et al. 2000; Kim et al. 2008). In addition, the absence of either *MEC1, DDC2* or *RAD53*, which detect DNA damage during replication and transduce the checkpoint response, also led to an increase in chromosomal fragility. However, the strongest increase in fragility was observed when *RAD9*, a checkpoint gene signaling unprocessed DSBs, was deleted (Lahiri et al. 2004). These results suggest that both stalled forks and unrepaired DSBs occur in cells containing long CAG/CTG repeat tracts. Given all these observations, it was, therefore, legitimate to address the role of DSB-repair in trinucleotide repeat instability.

Gene conversion and BIR lead to CAG/CTG repeat expansions

Initial studies performed almost 20 years ago pointed out the role of gene conversion in CAG/CTG repeat expansions and contractions. The authors used the I-*Sce* I or HO endonucleases, to induce a single DSB into a yeast chromosome. Both nucleases were discovered in the yeast *Saccharomyces cerevisiae*. I-Sce I is a meganuclease encoded by a mitochondrial homing intron (Colleaux et al. 1986) and HO initiates mating type switching by making a double-strand break at the MAT locus (Kostriken et al. 1983). In experimental systems using these nucleases, the induced DSB was repaired using a CAG/CTG repeat-containing homologous template as the donor sequence (Richard et al. 1999, 2000, 2003). Frequent expansions and contractions were observed and suggested that they occurred through a Synthesis-Dependent Strand Annealing (SDSA) mechanism, a particular type of gene conversion that is never associated to crossover (Fig. 1; Richard and Pâques 2000).

Trinucleotide repeat instability may also occur by homologous recombination in the absence of an induced DSB. Such length changes arise from replication fork blocking and/or spontaneous breakage during S phase replication. It was shown that CAG/CTG repeat expansions occurred in a *srs2* yeast mutant, most probably by homologous recombination between sister chromatids (Kerrest et al. 2009). In the absence of the Srs2 helicase activity, recombination intermediates were increased, as visualized by 2D gel electrophoresis. They partly disappeared when *RAD51*, the main recombinase gene in yeast, was deleted, proving that they were bona fide recombining molecules (Nguyen et al. 2017).

Expansions were also studied in mice deficient for the RAD52 recombination gene, but no difference in the rate of instability of a $(CTG)_{300}$ repeat tract was found, as compared to control mice (Savouret et al. 2003). However, RAD52 does not play the same role in mammals as it is playing in *S. cerevisiae*. In yeast cells, it is the mediator of all homologous recombination events (SSA, BIR, gene conversion) whereas it is only an accessory recombination gene whose exact function is not totally understood in mammalian cells. Therefore, it would be interesting to address the effect of BRCA1 and/or BRCA2 mutants on CAG/CTG repeat expansions, since these two genes belong to the real recombination mediator complex in human cells (Moynahan et al. 1999, 2001).

Large CAG/CTG repeat expansions were also investigated in yeast using an experimental assay based on the insertion of a $(CTG)_{140}$ repeat tract between the *GAL1* UAS and its TATA box. Transcriptional activation of the downstream reporter no longer occurred if the repeat tract was too long. The average size of detected expansions ranged from 60 to more than 150 triplets. Expansions decreased in the absence of *RAD51* and *RAD52*, proving that homologous recombination was the key mechanism (Kim et al. 2017). *POL32* (a non-essential DNA polymerase δ subunit) and the *PIF1* helicase were also involved, suggesting that expansions were controlled by BIR (Llorente et al. 2008; Lydeard et al. 2007). A one-ended DSB occurring within the repeat tract could invade the sister chromatid out-of-register, creating a D-loop. BIR would progress until colliding a converging

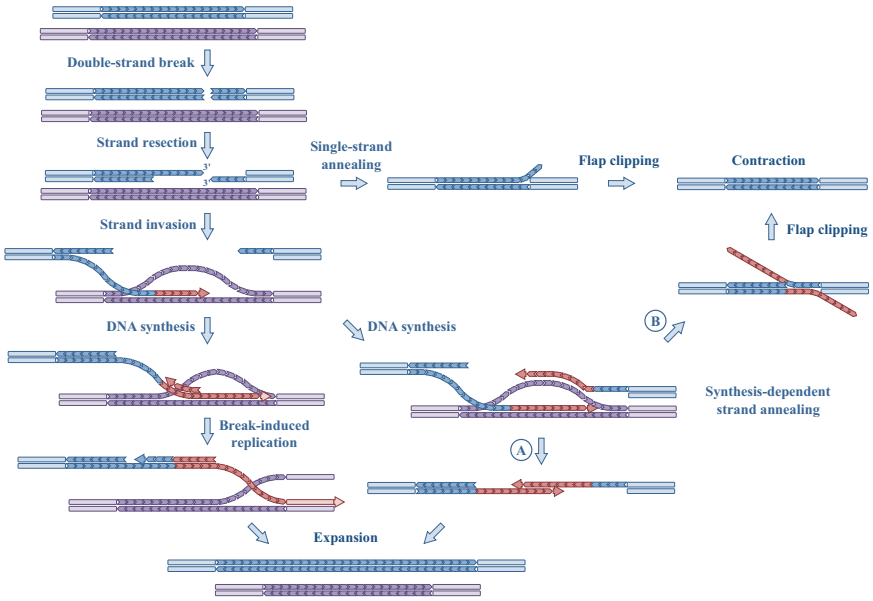


Fig. 1 Double-strand break repair mechanisms leading to repeat contraction or expansion. After a DSB was made into (or close to) a trinucleotide repeat tract, the broken molecule is resected by several nucleases and helicases leading to 3′-hydroxyl single-stranded ends. These ends may engage into different types of homologous recombination. Direct annealing of the two ends by SSA leads to repeat tract contraction after flap clipping (right). DNA synthesis during BIR generates repeat expansions (bottom). Synthesis-dependent strand annealing is resolved by unwinding and out-of-frame annealing of the recombination intermediate, possibly leading to repeat expansion (**a**) or repeat contraction (**b**). Note that none of these mechanisms requires crossover formation or resolution

fork or reaching the telomere, eventually resulting in an expansion (Fig. 1). Altogether these data tend to show that homologous recombination (gene conversion and BIR) may become a major source of CAG/CTG triplet repeat expansion if not properly controlled.

Gene conversion and SSA lead to CAG/CTG repeat contractions

Initial studies with the I-*Sce* I endonuclease suggested that DSB repair occurred in 67% of the cases by annealing between two short CAG/CTG repeats flanking the I-*Sce* I restriction site (Richard et al. 1999). More recently, a TALE nuclease (TALEN) was used to specifically induce a DSB within a $(CTG)_{80}$ repeat tract integrated in a yeast chromosome. Expression of this nuclease promoted repeat contraction at a high frequency (Mosbach et al. 2018; Richard et al. 2014). Repair was dependent on *RAD50, SAE2* and *RAD52*, but did not require *RAD51, POL32* or *LIG4*. It was, therefore, concluded that neither gene conversion nor BIR were the prefered contraction mechanism. It was instead proposed that progressive repeat contractions occurred through iterative cycles of DSB formation followed by SSA (Mosbach et al. 2018). In hamster CHO cells, CAG/CTG repeat contractions were also found to be associated to gene conversion and SSA events, at a frequency (5%) more than 10-fold increased as compared to replicating cells (Meservy et al. 2003).

In conclusion, trinucleotide repeat expansions and contractions appear to occur through different recombination mechanisms (Fig. 1). However, it is still unclear whether some of the spontaneous contractions observed during S phase replication in model systems may be triggered by a spontaneous DSB followed by SSA, or are mainly induced by gene conversion associated to DNA slippage.

Role of the SbcCD/MRX complex in CAG/CTG repeat instability

The Mre11-Rad50-Xrs2 (MRX) complex is one of the first players acting at a DSB. The complex triggers end trimming in such a way that resection enzymes-exonucleases and helicases- may be subsequently recruited to produce

recombinogenic 3′-hydroxyl single-strand extremities. The Sae2 protein works with the MRX complex in resection initiation, but it is still debated whether Sae2 exhibits a nuclease activity by itself or stimulates Mre11 nuclease activity to initiate resection (Zhu et al. 2008; Mimitou and Symington 2008; Lengsfeld et al. 2007). The MRX complex as well as Sae2 are also required to resolve hairpin-capped natural DSBs in yeast (Lobachev et al. 2002).

Repeat instability following an induced double-strand break

Repair by gene conversion of an HO-induced DSB using a homologous template containing a long CAG/CTG repeat tract led to longer repeat expansions when *MRE11* or *RAD50* were overexpressed (Richard et al. 2000). In addition, it was recently discovered that resection of a TALEN-induced DSB in a $(CTG)_{80}$ tract was completely abolished in the absence of Rad50, and that Sae2 was required to resect the DSB end containing the longest part of the triplet repeat tract (Mosbach et al. 2018). So the MRX complex, along with Sae2, are essential to process a DSB within a CTG trinucleotide repeat, suggesting the presence of secondary structures that need to be removed by the nuclease complex. These results are strengthened by previous evidences showing the accumulation of unrepaired natural chromosomal breaks within long CTG repeats in the absence of *RAD50* (Freudenreich et al. 1998).

Repeat instability following spontaneous DNA damage

Spontaneous $(CTG)_{70}$ repeat expansions of moderate lengths were increased during S phase in a *mre11*Δ mutant, these expansions being dependent on the *RAD52* gene (Sundararajan et al. 2010). These moderate expansions were very frequent, reaching 8.6% of colonies analyzed. In comparison, large scale $(CTG)_{140}$ repeat expansions were decreased in a *mre11*Δ mutant, from 10^{-5} to 10^{-6} per cell per division. Differences in stability, as well as in the role of Mre11 may reflect differences in mechanisms underlying moderate and large scale CTG repeat expansions: replication-triggered recombination versus BIR. Interestingly, it was recently shown that the MRX complex drove expansions of short $(CTG)_{20}$ trinucleotide repeats (which are not prone to spontaneous breakage) by a process independent of the nuclease function of Mre11 and of the Rad51 recombinase (Ye et al. 2016). This suggests that MRX may promote CTG repeat expansions by recombination-dependent and -independent mechanisms, the relative importance of each during cell life remaining to be determined.

In *Escherichia coli*, it was found that a CAG/CTG repeat tract stimulates the instability of a 275-bp tandem repeat located up to 6.3 kb away (Blackwood et al. 2010). Interestingly, this stimulation required neither DSB-repair nor the hairpin endonuclease SbcCD (homologue of Mre11-Rad50), suggesting that the primary lesion generated at the CAG/CTG repeat was not a DSB. Instead, the authors showed that the mismatch repair machinery triggered the instability observed, probably by recognizing loops of a single triplet formed during replication, leading to the production of single-strand DNA nicks. In eukaryotes, although its precise role is not totally clear, the mismatch repair machinery appears to be an important player of repeat instability by its propensity to recognize mismatches in hairpins formed by trinucleotide repeats while being unable to repair them (Pearson et al. 1997; Owen et al. 2005; Tomé et al. 2009, 2013; Williams and Surtees 2015; Slean et al. 2016; Viterbo et al. 2016). It is reasonnable to assume that DNA nickases now available will help to study the possible involvement of single stranded DNA nicks on CAG/CTG trinucleotide repeat instability.

GAA/TTC repeat instability occurs by template switching

A genetic assay was designed in yeast to study large-scale expansions of a $(GAA)_{78-150}$ repeat tract inserted into an artificial intron of the *URA3* gene, larger repeat lengths inhibiting intron splicing, therefore, inactivating the gene (Shishkin et al. 2009). Expansions reaching more than 300 triplets were observed, as well as small insertions/deletions or substitutions outside the repeat tract. Large chromosomal deletions including the *URA3* gene and its flanking sequences were also detected. *RAD50* or *RAD52* deletion had no effect on the expansion rate, ruling out the implication of homologous recombination in this process. On the contrary, the absence of replication fork-stabilizing proteins increased the expansion rate while it was decreased in the absence of postreplication DNA repair proteins or the Sgs1 DNA helicase. This strongly suggests that template switching during replication fork progression through GAA repeats was responsible for the observed GAA expansions (Shishkin et al. 2009). More recently, advances in long-read DNA sequencing technologies allowed to identify complex genomic rearrangements originating from improper repair of naturally occurring DSBs at GAA repeats. Various chromosomal rearrangements involving gene conversion between Ty retrotransposons and the formation of neochromosomes by BIR were described. These rearrangements apparently originated from DSBs into the GAA repeat tract (McGinty et al. 2017).

It is worth noting that recombination-independent recognition of DNA homology associated to mutation in *Neurospora crassa* (and probably in *Ascobolus immersus* too)

is enhanced by GAC/GTC trinucleotides (Gladyshev and Kleckner 2017). It would be interesting to know if other triplets also interfer with homology recognition and whether such a mechanism could be involved in trinucleotide repeat instability.

In conclusion, although both CAG/CTG and GAA/TTC repeats are apparently able to trigger DSB formation in yeast, expansions involve different sets of genes, therefore, different molecular pathways. These differences may be due to: (i) distinct secondary structures formed by both types of triplet repeats, GAA tracts folding into triplex DNA whereas CTG repeats form imperfect hairpins; (ii) the nature of DNA damage triggered by these structures, double- vs single-strand breaks or gaps; (iii) the amount of single-stranded DNA exposed following such damage; (iv) differences in chromatin conformation depending on the repeat tract sequence and structure. All these assumptions being not mutually exclusive, understanding the genetic complexity of trinucleotide repeat instability will probably require alternative methods to those applied so far.

Gene editing of trinucleotide repeat expansions

No cure is available for any triplet repeat disorder, although several preclinical and clinical trials have been attempted. Given that microsatellite disorders are always associated to an expansion of the repeat array, deleting or shortening the expanded array to non-pathological lengths should suppress symptoms of the pathology. Indeed, when a trinucleotide repeat contraction occurred during transmission from father to daughter of an expanded myotonic dystrophy allele, clinical examination of the 17-year-old daughter showed no sign of the symptoms (O'Hoy et al. 1993). In another study, a reversible model of DM1 transgenic mice, was relying on a recombinant GFP gene under the control of the TetOn promoter, fused to the DMPK 3′ UTR. After doxycycline treatment arrest, the GFP-DMPK transgene expression was stopped and sick mice reverted to normal (Mahadevan et al. 2006). Reversible mouse models of Huntington's disease (Yamamoto et al. 2000) and Spinocerebellar Ataxia Type 1 (Zu et al. 2004) showed that suppressing the expression of the toxic mutant protein led to a reversion of severe phenotypes associated to both disorders, including complex motor tasks, even at late disease stages. Hence, gene editing trinucleotide repeat tracts stands as an appealing approach to partially or totally cure these disorders.

Four families of highly specific nucleases may be used to edit trinucleotide repeats: meganucleases, Zinc-Finger Nucleases (ZFN), Transcription Activator Like Effector Nucleases (TALEN) and CRISPR-Cas9. Meganucleases are highly specific DNA endonucleases whose recognition site covers more than 12 bp, originally discovered in group I self-splicing introns in *S. cerevisiae* mitochondria (Dujon 1989). ZFNs were engineered from the fusion of a zinc-finger DNA binding domain to the FokI nuclease domain (Kim et al. 1996). ZFNs are active as heterodimers in which two arms need to dimerize to induce a DSB. TALENs are fusion proteins between a TAL effector derived from *Xanthomonas* bacteria and FokI, and also function as heterodimers (Cermak et al. 2011). The Cas9 protein is an RNA-guided nuclease belonging to the CRISPR system of bacterial acquired immune system. It needs the presence of a Protospacer Adjacent Motif (PAM) next to its guide sequence to induce one single-strand break on each DNA strand, resulting in a DSB (Doudna and Charpentier 2014). *Streptococcus pyogenes* Cas9 (*Sp*Cas9) was engineered by an aspartate-to-alanine substitution (D10A) in the RuvC catalytic domain to convert the double-strand endonuclease into a single-strand nickase (Cong et al. 2013). The same approach was used at the HNH catalytic site to generate the symetrical nickase cutting the opposite DNA strand (N863A). Depending on their bacterial origin, Cas9 proteins recognize different PAM and exhibit different activities. ZFN, TALEN and Cas9 were used to delete or shorten trinucleotide repeats, using two different approaches: (i) induce two DSBs upstream and downstream the repeat tract to completely delete it, or (ii) induce a DSB inside the repeat tract to shorten it (Fig. 2).

Huntington's disease

Huntington's disease is a dominant disorder caused by the expansion of a CAG repeat tract in the first exon of the *HTT* gene. In a first study, iPSCs (induced pluripotent stem cells) derived from Huntington patients harboring 72 CAG triplets were electroporated with a modified bacterial artificial chromosome containing 11.5 kb of the genomic region surrounding *HTT* first exon harboring 21 CAG triplets as well as an eGFP reporter cassette and a neomycin resistance gene. Out of 203 analyzed clones, only two showed the incorporation of the wild-type locus by homologous recombination. In these two clones, there was no detectable toxic huntingtin and modified cells retained the modifications when differentiated into neurons (An et al. 2012) (Table 1).

In another study, patient derived fibroblasts of variable CAG length were transfected with the D10A nickase and two guide RNAs, each targeting upstream and downstream the CAG repeat tract. Excision of the CAG repeat in the transfected non-clonal population showed decreased levels of the *HTT* mRNA and protein, from 68 to 82% depending on the cell line, suggesting that at least one allele was efficiently deleted, on the average. Four out of 13 predicted exonic off-target sites were tested and no mutation was detected (Dabrowska et al. 2018).

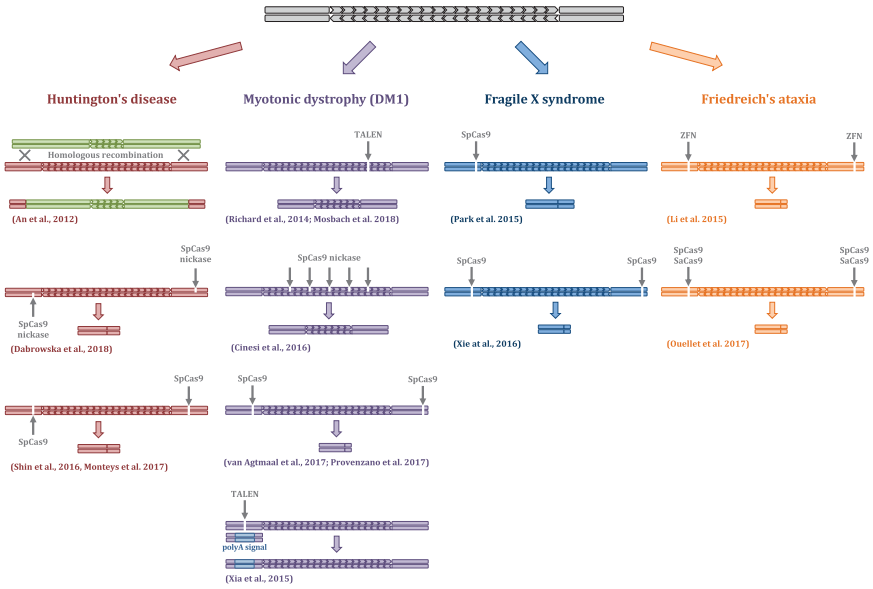


Fig. 2 Methods used for deleting or contracting trinucleotide repeats in human cells. Expanded trinucleotide repeat tracts were targeted by different nucleases in four human disorders. In each case, one or more approach was used to contract or delete the repeat tract. The nuclease expressed is shown in gray, along with arrows indicating whether the DSB (or SSB) was made within or outside the repeat tract. Repair outcomes following homologous recombination or non-homologous end joining are drawn. Corresponding references are shown under each approach

Alternative approaches exploited the presence of SNPs specific of the mutant CAG expanded allele. Two studies analyzing *HTT* haplotype were recently published, in which the authors took advantage of specific SNPs to remove the expanded allele in HD fibroblasts (Shin et al. 2016; Monteys et al. 2017). One of the studies also demonstrated that sgRNA/Cas9 complexes are also effective in vivo in an HD mouse model harboring the HD human allele. Viral delivery of sgRNA/*Sp*Cas9 complexes reduced human mutant HTT expression to 40% in the treated hemisphere as compared to the control untreated one (Monteys et al. 2017).

Myotonic dystrophy type I (DM1 or Steinert disease)

DM1 is an autosomal dominant disorder caused by an RNA-gain of function mutation: the expanded CTG repeat tract located at the 3′UTR of the DMPK gene is transcribed into a CUG-expanded RNA which accumulates into the nucleus and forms aggregates with splicing-effector proteins such as MBNL1 and CUG-BP1 (Miller et al. 2000). Deleting the CTG repeat tract should result in the suppression of the toxic RNA. The first work introducing the use of a highly specific nuclease to shorten a long CTG repeat from a DM1 patient, reported that a DSB made by a TALEN into the repeat tract induced a contraction of the repeat in 99% of cases, in yeast cells (Richard et al. 2014). In another study, a reporter assay was built in HEK293 cells to monitor contractions and expansions of a CTG repeat tract integrated into a synthetic intron interrupting a GFP gene. Efficacy of Cas9 D10A nickase, wild-type Cas9 and ZFNs cutting into the CTG repeat tract were compared. All induced contractions and expansions of the CTG repeat, but the nickase was the most efficient at inducing contractions (Cinesi et al. 2016).

Two proofs of concept of the removal of CTG repeats to cure DM1 were subsequently established. The introduction of Cas9 and a pair of guide RNAs each targeting a specific locus upstream and downstream the DM1 repeats in patient cells resulted in the deletion of the CTG repeats, the suppression of RNA foci and splicing defects (Van Agtmaal et al. 2017; Provenzano et al. 2017). Those two studies used different cell types, respectively, myogenic DM1 myoblast and DM1 fibroblasts and different target loci and achieved, respectively, 46 and 14% of successfully edited cells. Indels were found in both cases at cut sites and few loci were tested for off-target effects.

Table 1 Comparison between 12 gene editing studies aimed at correcting trinucleotide repeat disorders

Disease	Huntington's disease			
References	An et al. 2012	Dabrowska et al. 2018	Shin et al. 2016	Monteys et al.2017
Cell type	HD iPS cells	HD fibroblasts	HD fibroblasts	HD fibroblasts BacHD mice
Nuclease used	None (spontaneous homologous recombination)	Paired D10A nickases	*Sp*Cas9	*Sp*Cas9
Successful edition	1% (203 clones analyzed)	NA (bulk analysis)	NA (bulk analysis)	NA (bulk analysis)
Off-target analysis	NA	Indels at cut site. 4 off target sites analyzed: unchanged	None	11 top off target sites: unchanged
Phenotype of edited cells	No detectable toxic huntingtin	No detectable toxic huntingtin	HTT mRNA and protein levels decreased	HTT mRNA and protein levels decreased

Disease	Myotonic dystrophy type I				
Reference	Richard et al. 2014	Cinesi et al. 2016	Provenzano et al. 2017	Van Agtmaal et al. 2017	Xia et al. 2015
Cell type	*Saccharomyces cerevisiae*	HEK293 GFP$(CAG)_{89}$	Immortalized myogenic DM1 fibroblast	Immortalized DM1 myoblast (DM11)	DM1 neural stem cells
Nuclease	TALEN	ZFN *Sp*Cas9 D10A Cas9 nickase	e*Sp*Cas9	*Sp*Cas9	TALEN
Successful edition	99%	3%	14% (85 clones analyzed)	46% (103 clones analyzed)	After selection: 4 out of 10 colonies
Off-target analysis	Whole genome sequencing: no change	Number of CTG repeats at 7 different loci remained unchanged	Indels (1–151 bp) observed at cut sites Sequencing of the top 7 off-target sites of each sgRNA: unchanged	Indels at cut site Sequencing of the top 4 off-target loci: unchanged on model cell lines	Indels at cut site (40% cases)
Phenotype of edited cells	NA	NA	No foci Normal splicing of SERCA1 and INSR	No foci No MBNL1 aggregate Normal splicing of BIN1 and DMD	No foci Normal splicing of MBNL1&2 and MAPT

Disease	Fragile X syndrome		Friedreich's ataxia	
Reference	Park et al. 2015	Xie et al. 2016	Li et al. 2015	Ouellet et al. 2017
Cell type	FXS iPS cells	FXS iPS cells	FRDA fibroblasts and lymphoblasts	Transgenic mouse fibroblasts and whole animal muscles
Nuclease	*Sp*Cas9	*Sp*Cas9	ZFN	*Sp*Cas9 and *Sa*Cas9
Successful edition	2% (100 clones analyzed)	5 clones analyzed	6.7% (344 fibroblasts analyzed) 2.3% (305 lymphoblasts analyzed)	15% for the best gRNA combination in fibroblasts (33 clones analyzed) No quantification in vivo

Table 1 (continued)

Disease	Fragile X syndrome		Friedreich's ataxia	
Off-target analysis	49 and 112 bp deletion at cut site Sequencing of the 4 top off-target loci: unchanged on model cell lines	Indels at cut site	Indels at cut site. Ten off target analyzed: unchanged	Indels at cut site. No off-target study
Phenotype of edited cells	Decrease of FMR1 promoter methylation FMR1 mRNA and protein levels restored	Decrease of FMR1 promoter methylation for one clone. FMR1 mRNA and protein levels restored	FXN mRNA and protein levels restored. Neural cells showed restored levels of aconitase	Depending on the deletion event, FXN protein level was sometimes increased

NA not applicable

One last strategy consisted in inserting a polyA signal upstream the CTG tract to prevent its transcription. This was carried out by making a TALEN-induced DSB between exon 9 and 10 of the DMPK gene, while co-transfecting the polyA cassette (Xia et al. 2015). Successfully edited cells showed phenotype reversion including foci disappearance and normal splicing of MBNL1 and MBNL2.

Fragile X syndrome

The fragile X syndrome is caused by the expansion of a CGG repeat tract in the 5′ UTR of the *FMR1* gene which leads through an undetermined mechanism to the methylation of the *FMR1* promoter (Verkerk et al. 1991, Yu et al. 1991). FXS iPSCs (more than 450 CGG) were transfected with *Sp*Cas9 and a guide RNA targeting the region upstream the repeat tract (Park et al. 2015). Four potential off target sites were tested and no mutation was detected. Two successfully edited clones over 100 tested were obtained. In these two clones, promoter hypermethylation was abolished and *FMR1* expression was reactivated. A similar study was conducted by cutting upstream and downstream the CGG repeats using *Sp*Cas9. The authors observed a decrease in the methylation profile of the *FMR1* promoter in one of their analyzed clones along with partial restoration of the FMR1 protein (Xie et al. 2016).

Friedreich's ataxia (FRDA)

FRDA is a recessive disorder caused by an expanded GAA (up to 2000 triplets) located in intron 1 of the frataxin gene, inducing a heterochromatization of the *FXN* locus leading to low frataxin levels (Campuzano et al. 1996). Heterozygous carriers are asymptomatic. Two ZFNs were designed to specifically cut upstream and downstream the GAA repeat tract. FRDA lymphoblasts and fibroblasts were transfected with both ZFN arms. Successful edition was achieved for 7 out of 305 lymphoblasts (2.3% efficiency) and 23 out of 344 fibroblasts (6.7% efficiency). Heterozygous modifications were observed as well as large deletions at ZFN cut sites. Edited cells exhibited increased expression of frataxin. When differentiated into neurons the cells retained the corrections. Ten top off-target sites were studied in established cell line K562 cells and no mutation was detected (Li et al. 2015). *Sp*Cas9 was targeted in transgenic mice fibroblasts and whole animal muscles, upstream and downstream GAA repeats to remove them (Ouellet et al. 2017). Successful in vitro edition ranged from 4 to 15% depending on the couple of gRNA used. Indels were found at sequenced junctions in successfully edited clones. Gene editing events were observed by PCR in fibroblasts, as well as in vivo. SaCas9 was also transfected in mice fibroblasts but its expression level was much lower than *Sp*Cas9 and editing was not very efficient.

Limitations of nuclease approaches: off-target effects

One major concern about specific nucleases is the potential effect of off-target mutations due to a lack of specificity. In silico programs are poor predictors of real off-target sites and there is no simple rule so far to accurately predict off-targets. The first genome-wide assessment of Cas9 off-target sites was carried out using the GUIDE-seq method. Briefly, double-stranded modified oligonucleotides are transfected alongside the nuclease and integrate in the genome at all DSB sites generated by the nuclease. They can subsequently be amplified and serve as primers for genome-wide sequencing of their insertion sites. This analysis revealed that off-targets are difficult to predict, ranging from little cleavage outside the target to as many off-target as on-target DSBs, depending on the gRNA chosen. Cleavage can occur on sites bearing up to seven mismatches and no canonical PAM (Tsai et al. 2015). CIRCLE-seq is a simpler and more sensitive method to detect off-target sites in vitro, but requires the purified nuclease (Tsai et al. 2017). Using this approach, genomic DNA that was cleaved by the nuclease in a test tube was amplified and sequenced. This method is very sensitive but may not be relevant for in vivo assays and may depend on each cell type and chromatin state. Recently, the VIVO method was set up for in vivo validation of off-target sites found by CIRCLE-seq, demonstrating that careful choice of the gRNA may strongly reduce off-target effects, while keeping a good on-target efficacy (Akcakaya et al. 2018). The same team engineered a more specific version of *Sp*Cas9, called HF1, by mutating residues involved in the binding to the target DNA strand. Cas9-HF1 retains on-target activity comparable to wild-type on 85% of gRNAs tested and rendered all or nearly all off-target events not detectable by GUIDE-seq (Kleinstiver et al. 2016). No such extensive off-target study was carried out in any of the aforementioned articles. Such approaches must be encouraged in future assessments of gene therapy strategies for trinucleotide repeat disorders.

Limitations of nuclease approaches: vectorization

Nuclease vectorization is clearly a problem that also needs to be addressed. Adenovirus-associated vectors (AAV) are popular in gene therapy because they exhibit low integration frequency, but they have a limited cargo capacity making it impossible to deliver a full length *Sp*Cas9 with its cognate guide, or a TALEN. In this case, each of the two TALEN arms must be delivered by two different vectors, lowering the efficacy of the transduction. Alternative non-viral delivery systems such as cationic lipid transfection particles was efficient to deliver a Cas9-gRNA complex as well as a TALEN both in vitro and in vivo, achieving 20% efficacy in genome modification in mice (Zuris et al. 2015). AAV-based delivery could also potentially increase the rate of off-target site cleavage due to prolonged expression of the nuclease. To circumvent this problem, a self-limiting CRISPR-Cas9 system was implemented in vivo by inserting the sequence recognized by the nuclease on the plasmid encoding it such that the expression plasmid would be cut and eliminated following *Sp*Cas9 expression (Ruan et al. 2017).

An alternative approach would solve the vectorization as well as the immune response issues: in vitro modification of patient induced pluripotent stem cells, followed by reprogrammation of nuclease-treated iPSC into the desired cell type (neuron, myoblast, etc.). However, such an advance in regenerative medicine is still hampered by the need for expressing four transcription factors from retroviral vectors to induce pluripotency, with all the risks associated to retrovirus integration into human cells (Takahashi et al. 2007; Yu et al. 2007).

Conclusion

Little is known yet about the immune response toward these nucleases. A very recent work identified pre-existing immunity against Cas9 from *Streptococcus pyogenes* and *Staphylococcus aureus* (Charlesworth et al. 2018). The authors showed that 70% of healthy adults have antibodies directed to the nuclease and that *Sa*Cas9 induced a T-cell response in adult blood. A strong immune response may be a potential drawback to the use of Cas9 in future gene therapy.

An additional difficulty is raised by checkpoint effectors, such as p53, controlling the cellular response to double-strand breaks. Two studies have recently shown that during gene editing, cells with a functional p53 pathway were counterselected, due to cell arrest triggered by p53 upon DSB formation. Therefore, checkpoint activity should be tightly controlled when developing cell-based therapies utilizing CRISPR–Cas9 (Haapaniemi et al. 2018; Ihry et al. 2018).

These first reports of gene therapy attempts of trinucleotide repeat disorders are certainly promising and already give us insights into crucial factors to be considered when evaluating the success of a gene therapy approach: off-target sites number and frequency, nuclease efficacy, cell type to be targeted and vectorization method. Successful gene editing was achieved in a mouse model for Duchenne muscular dystrophy, by three independent teams. Using AAV delivery of Cas9, they obtained partial restoration of dystrophin

levels that were sufficient to allow partial muscle strength recovery (Long et al. 2016; Nelson et al. 2016; Tabebordbar et al. 2016). Forthcoming experiments in a mouse model for trinucleotide repeat disorders will establish if a similar success may be achieved.

Acknowledgements The authors wish to thank the continuous support of the Institut Pasteur and of the Centre National de la Recherche Scientifique (CNRS). L. P. is the recipient of a Cifre PhD fellowship from Sanofi. V. M. was the recipient of two post-doctoral fellowships from Fondation Guy Nicolas and from Fondation Hardy.

References

Agtmaal EL van, André LM, Willemse M et al (2017) CRISPR/Cas9-Induced (CTG·CAG)n repeat instability in the myotonic dystrophy type 1 locus: implications for therapeutic genome editing. Mol Ther 25:24–43. https://doi.org/10.1016/j.ymthe.2016.10.014

Akcakaya P, Bobbin ML, Guo JA et al (2018) In vivo CRISPR-Cas gene editing with no detectable genome-wide off-target mutations. bioRxiv. https://doi.org/10.1101/272724

An MC, Zhang N, Scott G et al (2012) Genetic correction of Huntington's disease phenotypes in induced pluripotent stem cells. Cell Stem Cell 11:253–263. https://doi.org/10.1016/j.stem.2012.04.026

Blackwood JK, Okely EA, Zahra R et al (2010) DNA tandem repeat instability in the *Escherichia coli* chromosome is stimulated by mismatch repair at an adjacent CAG·CTG trinucleotide repeat. Proc Natl Acad Sci USA 107:22582–22586. https://doi.org/10.1073/pnas.1012906108

Callahan JL, Andrews KJ, Zakian VA, Freudenreich CH (2003) Mutations in yeast replication proteins that increase CAG/CTG expansions also increase repeat fragility. Mol Cell Biol 23:7849–7860

Campuzano V, Montermini L, Molto MD et al (1996) Friedreich's Ataxia: autosomal recessive disease caused by an intronic GAA triplet repeat expansion. Science 271:1423–1427

Cermak T, Doyle EL, Christian M et al (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res 39:e82. https://doi.org/10.1093/nar/gkr218

Cinesi C, Aeschbach L, Yang B, Dion V (2016) Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. Nat Commun 7:13272. https://doi.org/10.1038/ncomms13272

Charlesworth CT, Deshpande PS, Dever DP et al (2018) Identification of Pre-Existing Adaptive Immunity to Cas9 Proteins in Humans. bioRxiv. https://doi.org/10.1101/243345

Colleaux L, d'Auriol L, Betermier M et al (1986) Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into *E. Coli* as a specific double strand break endonuclease. Cell 44:521–533

Cong L, Ran FA, Cox D et al (2013) Multiplex genome engineering using CRISPR/Cas systems. Science 339:819–823

Dabrowska M, Juzwa W, Krzyzosiak WJ, Olejniczak M (2018) Precise excision of the CAG tract from the Huntingtin gene by Cas9 nickases. Front Neurosci. https://doi.org/10.3389/fnins.2018.00075

Doudna JA, Charpentier E (2014) Genome editing. The new frontier of genome engineering with CRISPR-Cas9. Science 346:1258096. https://doi.org/10.1126/science.1258096

Dujon B (1989) Group I introns as mobile genetic elements: facts and mechanistic speculations—a review. Gene 82:91–114

Freudenreich CH, Kantrow SM, Zakian VA (1998) Expansion and length-dependent fragility of CTG repeats in yeast. Science 279:853–856

Fu Y-H, Kuhl DPA, Pizzuti A et al (1991) Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. Cell 67:1047–1058

Gladyshev E, Kleckner N (2017) Recombination-independent recognition of DNA homology for repeat-induced point mutation. Curr Genet 63:389–400. https://doi.org/10.1007/s00294-016-0649-4

Haapaniemi E, Botla S, Persson J et al (2018) CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. Nat Med. https://doi.org/10.1038/s41591-018-0049-z

Ihry RJ, Worringer KA, Salick MR et al (2018) p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. Nat Med. https://doi.org/10.1038/s41591-018-0050-6

Jankowski C, Nasar F, Nag DK (2000) Meiotic instability of CAG repeat tracts occurs by double-strand break repair in yeast. Proc Natl Acad Sci USA 97:2134–2139. https://doi.org/10.1073/pnas.040460297

Kerrest A, Anand R, Sundararajan R et al (2009) SRS2 and SGS1 prevent chromosomal breaks and stabilize triplet repeats by restraining recombination. Nat Struct Mol Biol 16:159–167

Kim JC, Mirkin SM (2013) The balancing act of DNA repeat expansions. Curr Opin Genet Dev 23:280–288. https://doi.org/10.1016/j.gde.2013.04.009

Kim YG, Cha J, Chandrasegaran S (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc Natl Acad Sci USA 93:1156–1160

Kim H-M, Narayanan V, Mieczkowski PA et al (2008) Chromosome fragility at GAA tracts in yeast depends on repeat orientation and requires mismatch repair. EMBO J 27:2896–2906. https://doi.org/10.1038/emboj.2008.205

Kim JC, Harris ST, Dinter T et al (2017) The role of break-induced replication in large-scale expansions of (CAG)n/(CTG)nrepeats. Nat Struct Mol Biol 24:55–60. https://doi.org/10.1038/nsmb.3334

Kleinstiver BP, Pattanayak V, Prew MS et al (2016) High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. Nature 529:490–495. https://doi.org/10.1038/nature16526

Kostriken R, Strathern JN, Klar AJ et al (1983) A site-specific endonuclease essential for mating-type switching in *Saccharomyces cerevisiae*. Cell 35:167–174. https://doi.org/10.1016/0092-8674(83)90219-2

Lahiri M, Gustafson TL, Majors ER, Freudenreich CH (2004) Expanded CAG repeats activate the DNA damage checkpoint pathway. Mol Cell 15:287–293. https://doi.org/10.1016/j.molcel.2004.06.034

Lengsfeld BM, Rattray AJ, Bhaskara V et al (2007) Sae2 Is an endonuclease that processes hairpin DNA Cooperatively with the Mre11/Rad50/Xrs2 complex. Mol Cell 28:638–651. https://doi.org/10.1016/j.molcel.2007.11.001

Li Y, Polak U, Bhalla AD et al (2015) Excision of expanded GAA repeats alleviates the molecular phenotype of Friedreich's Ataxia. Mol Ther 23:1055–1065. https://doi.org/10.1038/mt.2015.41

Liu G, Chen X, Bissler JJ et al (2010) Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. Nat Chem Biol 6:652–659. https://doi.org/10.1038/nchembio.416

Llorente B, Smith CE, Symington LS (2008) Break-induced replication: what is it and what is it for? Cell Cycle 7:859–864. https://doi.org/10.4161/cc.7.7.5613

Lobachev KS, Gordenin DA, Resnick MA (2002) The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. Cell 108:183–193

Long C, Amoasii L, Mireault AA et al (2016) Postnatal genome editing partially restores dystrophin expression in a mouse model of muscular dystrophy. Science 351:400–403. https://doi.org/10.1126/science.aad5725

Lydeard JR, Jain S, Yamaguchi M, Haber JE (2007) Break-induced replication and telomerase-independent telomere maintenance require Pol32. Nature 448:820–823. https://doi.org/10.1038/nature06047

Mahadevan MS, Yadava RS, Yu Q et al (2006) Reversible model of RNA toxicity and cardiac conduction defects in myotonic dystrophy. Nat Genet 38:1066–1070. https://doi.org/10.1038/ng1857

McGinty RJ, Mirkin SM (2018) Cis- and trans-modifiers of repeat expansions: blending model systems with human genetics. Trends Genet 34:448–465. https://doi.org/10.1016/j.tig.2018.02.005

McGinty RJ, Rubinstein RG, Neil AJ et al (2017) Nanopore sequencing of complex genomic rearrangements in yeast reveals mechanisms of repeat-mediated double-strand break repair. Genome Res 27:2072–2082. https://doi.org/10.1101/gr.228148.117

McMurray CT (2010) Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 11:786–799. https://doi.org/10.1038/nrg2828

Meservy JL, Sargent RG, Iyer RR et al (2003) Long CTG tracts from the myotonic dystrophy gene induce deletions and rearrangements during recombination at the APRT locus in CHO cells. Mol Cell Biol 23:3152–3162. https://doi.org/10.1128/MCB.23.9.3152-3162.2003

Miller JW, Urbinati CR, Teng-umnuay P et al (2000) Recruitment of human musclebind proteins to (CUG)n expansions associated with myotonic dystrophy. EMBO J 19:4439–4448

Mimitou EP, Symington LS (2008) Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature 455:770–774. https://doi.org/10.1038/nature07312

Monteys AM, Ebanks SA, Keiser MS, Davidson BL (2017) CRISPR/Cas9 editing of the mutant huntingtin allele in vitro and in vivo. Mol Ther 25:12–23. https://doi.org/10.1016/j.ymthe.2016.11.010

Mosbach V, Poggi L, Viterbo D et al (2018) TALEN-induced double-strand break repair of CTG trinucleotide repeats. Cell Rep 22:2146–2159. https://doi.org/10.1016/j.celrep.2018.01.083

Moynahan ME, Chiu JW, Koller BH, Jasin M (1999) Brca1 controls homology-directed DNA repair. Mol Cell 4:511–518

Moynahan ME, Pierce AJ, Jasin M (2001) BRCA2 is required for homology-directed repair of chromosomal breaks. Mol Cell 7:263–272

Neil Alexander J, Kim Jane C, Mirkin Sergei M (2017) Precarious maintenance of simple DNA repeats in eukaryotes. BioEssays 39:1700077. https://doi.org/10.1002/bies.201700077

Nelson CE, Hakim CH, Ousterout DG et al (2016) In vivo genome editing improves muscle function in a mouse model of Duchenne muscular dystrophy. Science 351:403–407. https://doi.org/10.1126/science.aad5143

Nguyen JHG, Viterbo D, Anand RP et al (2017) Differential requirement of Srs2 helicase and Rad51 displacement activities in replication of hairpin-forming CAG/CTG repeats. Nucleic Acids Res 45:4519–4531. https://doi.org/10.1093/nar/gkx088

O'Hoy KL, Tsilfidis C, Mahadevan MS et al (1993) Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission. Science 259:809–812

Ouellet DL, Cherif K, Rousseau J, Tremblay JP (2017) Deletion of the GAA repeats from the human frataxin gene using the CRISPR-Cas9 system in YG8R-derived cells and mouse models of Friedreich ataxia. Gene Ther 24:265–274. https://doi.org/10.1038/gt.2016.89

Owen BAL, Yang Z, Lai M et al (2005) (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. Nat Struct Mol Biol 12:663–670. https://doi.org/10.1038/nsmb965

Park C-Y, Halevy T, Lee DR et al (2015) Reversion of FMR1 methylation and silencing by editing the triplet repeats in fragile X iPSC-derived neurons. Cell Rep 13:234–241. https://doi.org/10.1016/j.celrep.2015.08.084

Pearson CE, Ewel A, Acharya S et al (1997) Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. Hum Mol Genet 6:1117–1123

Provenzano C, Cappella M, Valaperta R et al (2017) CRISPR/Cas9-mediated deletion of CTG expansions recovers normal phenotype in myogenic cells derived from myotonic dystrophy 1 patients. Mol Ther Nucleic Acids 9:337–348. https://doi.org/10.1016/j.omtn.2017.10.006

Richard G-F, Pâques F (2000) Mini- and microsatellite expansions: the recombination connection. EMBO Rep 1:122–126

Richard GF, Dujon B, Haber JE (1999) Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. Mol Gen Genet 261:871–882

Richard GF, Goellner GM, McMurray CT, Haber JE (2000) Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11-RAD50-XRS2 complex. EMBO J 19:2381–2390. https://doi.org/10.1093/emboj/19.10.2381

Richard GF, Cyncynatus C, Dujon B (2003) Contractions and expansions of CAG/CTG trinucleotide repeats occur during ectopic gene conversion in yeast, by a MUS81-independent mechanism. J Mol Biol 326:769–782

Richard GF, Kerrest A, Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev 72:686–727

Richard G-F, Viterbo D, Khanna V et al (2014) Highly specific contractions of a single CAG/CTG trinucleotide repeat by TALEN in yeast. PLoS One 9:e95611. https://doi.org/10.1371/journal.pone.0095611

Ruan G-X, Barry E, Yu D et al (2017) CRISPR/Cas9-mediated genome editing as a therapeutic approach for leber congenital amaurosis 10. Mol Ther 25:331–341. https://doi.org/10.1016/j.ymthe.2016.12.006

Savouret C, Brisson E, Essers J et al (2003) CTG repeat instability and size variation timing in DNA repair-deficient mice. Embo J 22:2264–2273

Shin JW, Kim K-H, Chao MJ et al (2016) Permanent inactivation of Huntington's disease mutation by personalized allele-specific CRISPR/Cas9. Hum Mol Genet 25:4566–4576. https://doi.org/10.1093/hmg/ddw286

Shishkin AA, Voineagu I, Matera R et al (2009) Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. Mol Cell 35:82–92. https://doi.org/10.1016/j.molcel.2009.06.017

Slean MM, Panigrahi GB, Castel AL et al (2016) Absence of MutSβ leads to the formation of slipped-DNA for CTG/CAG contractions at primate replication forks. DNA Repair 42:107–118. https://doi.org/10.1016/j.dnarep.2016.04.002

Sundararajan R, Gellon L, Zunder RM, Freudenreich CH (2010) Double-strand break repair pathways protect against CAG/CTG repeat expansions, contractions and repeat-mediated chromosomal fragility in *Saccharomyces cerevisiae*. Genetics 184:65–77. https://doi.org/10.1534/genetics.109.111039

Tabebordbar M, Zhu K, Cheng JKW et al (2016) In vivo gene editing in dystrophic mouse muscle and muscle stem cells. Science 351:407–411. https://doi.org/10.1126/science.aad5177

Takahashi K, Tanabe K, Ohnuki M et al (2007) Induction of pluripotent stem cells from adult human fibroblasts by defined factors. Cell 131:861–872. https://doi.org/10.1016/j.cell.2007.11.019

Tomé S, Holt I, Edelmann W et al (2009) MSH2 ATPase domain mutation affects CTG*CAG repeat instability in transgenic mice. PLoS Genet 5:e1000482. https://doi.org/10.1371/journal.pgen.1000482

Tomé S, Manley K, Simard JP et al (2013) MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. PLoS Genet 9:e1003280. https://doi.org/10.1371/journal.pgen.1003280

Tsai SQ, Zheng Z, Nguyen NT et al (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas

nucleases. Nat Biotechnol 33:187–197. https://doi.org/10.1038/nbt.3117

Tsai SQ, Nguyen NT, Malagon-Lopez J et al (2017) CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat Meth 14:607–614. https://doi.org/10.1038/nmeth.4278

Usdin K, House NC, Freudenreich CH (2015) Repeat instability during DNA repair: insights from model systems. Crit Rev Biochem Mol Biol. https://doi.org/10.3109/10409238.2014.999192

Verkerk AJMH, Pieretti M, Sutcliffe JS et al (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. Cell 65:905–914

Viterbo D, Michoud G, Mosbach V et al (2016) Replication stalling and heteroduplex formation within CAG/CTG trinucleotide repeats by mismatch repair. DNA Repair 42:94–106. https://doi.org/10.1016/j.dnarep.2016.03.002

Williams GM, Surtees JA (2015) MSH3 promotes dynamic behavior of trinucleotide repeat tracts in vivo. Genetics 200:737–754. https://doi.org/10.1534/genetics.115.177303

Xia G, Gao Y, Jin S et al (2015) Genome modification leads to phenotype reversal in human myotonic dystrophy type 1 induced pluripotent stem cell-derived neural stem cells. Stem Cells 33:1829–1838. https://doi.org/10.1002/stem.1970

Xie N, Gong H, Suhl JA et al (2016) Reactivation of FMR1 by CRISPR/Cas9-mediated deletion of the expanded CGG-repeat of the fragile X chromosome. PLoS One 11:e0165499. https://doi.org/10.1371/journal.pone.0165499

Yamamoto A, Lucas JJ, Hen R (2000) Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. Cell 101:57–66. https://doi.org/10.1016/S0092-8674(00)80623-6

Ye Y, Kirkham-McCarthy L, Lahue RS (2016) The *Saccharomyces cerevisiae* Mre11-Rad50-Xrs2 complex promotes trinucleotide repeat expansions independently of homologous recombination. DNA Repair 43:1–8. https://doi.org/10.1016/j.dnarep.2016.04.012

Yu S, Pritchard M, Kremer E et al (1991) Fragile X genotype characterized by an unstable region of DNA. Science 252:1179–1181

Yu J, Vodyanik MA, Smuga-Otto K et al (2007) Induced pluripotent stem cell lines derived from human somatic cells. Science 318:1917–1920. https://doi.org/10.1126/science.1151526

Yudkin D, Hayward BE, Aladjem MI et al (2014) Chromosome fragility and the abnormal replication of the FMR1 locus in fragile X syndrome. Hum Mol Genet 23:2940–2952. https://doi.org/10.1093/hmg/ddu006

Zhu Z, Chung WH, Shim EY et al (2008) Sgs1 helicase and two nucleases Dna2 and Exo1 resect DNA double-strand break ends. Cell 134:981–994. https://doi.org/10.1016/j.cell.2008.08.037

Zu T, Duvick LA, Kaytor MD et al (2004) Recovery from polyglutamine-induced neurodegeneration in conditional SCA1 transgenic mice. J Neurosci 24:8853–8861. https://doi.org/10.1523/JNEUROSCI.2978-04.2004

Zuris JA, Thompson DB, Shu Y et al (2015) Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. Nat Biotechnol 33:73–80. https://doi.org/10.1038/nbt.3081

Annex 2: Monitoring double-strand break repair of trinucleotide repeats using a yeast fluorescence reporter assay.

Method article explaining in detail the protocol used to test nucleases in the yeast reporter assay. I wrote the manuscript.

Chapter 7

Monitoring Double-Strand Break Repair of Trinucleotide Repeats Using a Yeast Fluorescent Reporter Assay

Lucie Poggi, Bruno Dumas, and Guy-Franck Richard

Abstract

Cells can repair a double-strand break (DSB) by homologous recombination if a homologous sequence is provided as a template. This can be achieved by classical gene conversion (with or without crossover) or by single-strand annealing (SSA) between two direct repeat sequences flanking the DSB. To initiate SSA, single-stranded regions are needed adjacent to the break, extending up to the direct repeats in such a way that complementary strands can anneal to each other to repair the DSB. In the present protocol, we describe a GFP reporter assay in *Saccharomyces cerevisiae* allowing for the quantification of nuclease efficacy at inducing a DSB, by monitoring the reconstitution of a functional GFP gene whose expression can be rapidly quantified by flow cytometry.

Key words Endonuclease, CRISPR-Cas9, Flow cytometry, GFP-based reporter assay, Yeast, Homologous recombination

1 Introduction

Double-strand DNA endonucleases have been used for decades as biotechnological tools to engineer DNA sequences, first on plasmids, later on in whole genomes. Their activities were determined mainly using biochemical approaches. Among the four large families of highly specific DNA endonucleases [1], meganucleases were the first to have been used to engineer mammalian genomes [2] The first meganuclease to be characterized was I-*Sce* I which was isolated from a mitochondrial intron in *S. cerevisiae* [3] and its activity was assessed in vitro on a plasmid carrying its recognition site [4]. The development of new target specificities for meganucleases also relied on in vitro assays to assess cleavage efficiencies of synthetic meganucleases [5]. Zinc-finger nucleases (ZFN) were engineered from Zinc-finger transcription activators fused to the FokI endonuclease domain, and in vitro assays were carried out on λ DNA to test cleavage efficacy, sequence recognition specificity and time course of cleavage [6]. Fifteen years later, Transcription

Guy-Franck Richard (ed.), *Trinucleotide Repeats: Methods and Protocols*, Methods in Molecular Biology, vol. 2056, https://doi.org/10.1007/978-1-4939-9784-8_7,

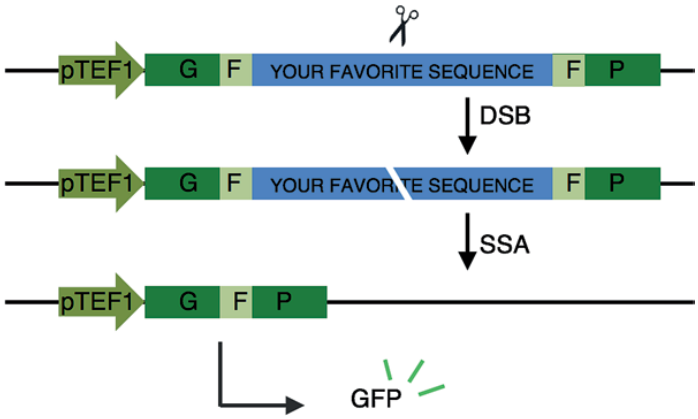


Fig. 1 Cassette integrated into the FYBL1-4D strain**.** pTEF1 is a strong constitutive promoter driving the expression of the GFP cassette. Upon nuclease induction, a double-strand break is made within the CTG repeat tract, generating recombination intermediates that lead to annealing between the two GFP moieties by Single-Strand Annealing, regenerating a functional protein. Readouts of nuclease efficacy is made by counting GFP-positive cells using flow cytometry

Activator-Like Effectors (TALE) were also fused to FokI to make TALE Nucleases (TALEN). This new family of nucleases was validated as a bona fide DNA double-strand strand endonuclease in vivo, on a yeast plasmid [7]. More recently, *Streptococcus pyogenes* Cas9 (SpCas9) and variants that were engineered to be more specific were characterized using a cutting-edge technology based on FRET experiments to identify off-target sites in vitro [8]. Here, we propose a simple in vivo assay, in the yeast *Saccharomyces cerevisiae* to test cleavage efficacy of any DNA endonuclease on any target sequence. The assay uses homologous recombination as a proxy for nuclease efficacy. Homologous recombination assays have already been used to quantify repair efficacy after the induction of a DSB [9]. Here we introduced in a *S. cerevisiae* S288C derivative strain a reporter gene made of a CTG trinucleotide repeat integrated in a GFP bipartite gene containing 100 bp overlapping homology regions flanking the repeat (Fig. 1). Upon nuclease induction, cells repair the DSB by single-strand annealing (SSA) between the two overlapping GFP moieties, reconstituting a functional gene that can be easily quantified by flow cytometry. This method is adapted to high throughput data acquisition to study DSB repair efficacy at any given target sequence integrated in the GFP cassette, using any endonuclease targeting this region. Here we present the protocol developed to study the efficacy of the *Streptococcus pyogenes* Cas9 nuclease (SpCas9) at repairing a DSB made into CTG trinucleotide repeats, but the nuclease as well as the target sequence may be easily changed, to make it a versatile assay.

2 Materials

Use good laboratory practices to avoid contaminating yeast cultures. Always manipulate cells in a sterile environment.

2.1 Yeast Media

1. YPD medium: 20 g Bacto peptone, 10 g yeast extract, 20 g glucose, H_2O to 1 L, autoclave for 20 min at 120 °C. For solid medium, add 20 g agar before autoclaving and pour 25 mL in each plate.
2. SC −Ura −Leu medium: 6.7 g yeast nitrogen base without amino acids, 20 g glucose, 2 g −Ura −Leu dropout mix, H_2O to 1 L, autoclave for 20 min at 120 °C. For solid medium, add 20 g agar before autoclaving and pour 25 mL in each plate. −Ura −Leu dropout mix contains 1 g adenine, 2 g alanine, 2 g arginine, 2 g aspartic acid, 2 g asparagine, 2 g cysteine, 2 g glutamic acid, 2 g glutamine, 2 g glycine, 2 g histidine, 2 g isoleucine, 2 g lysine, 2 g methionine, 2 g phenylalanine, 5 g proline, 2 g serine, 2 g threonine, 2 g tryptophan, 2 g tyrosine, and 2 g valine.
3. GAL −Ura −Leu medium: 6.7 g yeast nitrogen base without amino acids, 20 g galactose, 2 g −Ura −Leu dropout mix, H_2O to 1 L, autoclave for 20 min at 120 °C.

2.2 Yeast Transformation

1. FYBL1-4D strain (or another S288C derivative) into which the bipartite GFP-CTG cassette was integrated.
2. Sterile toothpicks.
3. Sterile 1.5 mL microtubes.
4. Sterile micropipette blue and yellow tips.
5. Sterile 2 and 25 mL disposable pipettes.
6. Sterile glass beads (or a sterile spreader).
7. TE/LiAc solution: 10 mM Tris–HCl pH 7.5, 1 mM EDTA pH 8.0, 100 mM lithium acetate pH 7.5. Sterilize the solution on a 0.22 μm filter unit.
8. PEG/TE/LiAc solution: 10 mM Tris–HCl pH 7.5, 1 mM EDTA, 100 mM Lithium acetate pH 7.5, 40% PEG3350. Sterilize on a 0.22 μm filter unit.
9. Carrier DNA: 10 mg/mL salmon sperm DNA denatured at 100 °C for 5 min then put on ice (*see* **Note 1**).
10. Water bath set at 42 °C.

2.3 Flow Cytometry

1. 1.3 mL sterile 96 deep-well plates.
2. Breathable sealing membrane.
3. 96 well plates.
4. Incubator with a 25 mm shaking diameter (*see* **Note 2**).

5. PBS buffer: 137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4.
6. Flow cytometer buffers (storage, running, washing).
7. Flow cytometer (*see* **Note 3**). It must be turned on 30 min before starting experiments, in order to allow lasers to warm up. Calibration should be carried out every day.

2.4 Vectors

1. pGAL-Cas9 [10]: This centromeric plasmid contains a *GAL1* promoter driving the expression of *Streptococcus pyogenes* Cas9 and *LEU2* as a selectable marker.
2. pRS416-CTG: This plasmid was constructed by integrating a synthetic cassette containing the *SNR52* promoter, the guide RNA target sequence TGCTGCTGCTGCTGCTGCTG, the guide RNA scaffold and the *SUP4* terminator into the pRS416 plasmid [11]. The sequence of the full cassette can be found at the end this chapter. This centromeric plasmid carries *URA3* as a selectable marker.

3 Methods

3.1 Transformation of FYBL1-4D Containing the Bipartite GFP-CTG Cassette

1. Thaw cells from −80 °C freezer, plate on YPD plate and let them grow for 48 h.
2. Pick a single colony using a sterile toothpick and start an overnight culture in 3 mL YPD. Incubate at 30 °C, 160 rpm.
3. Dilute saturated culture 1:50 in YPD and incubate at 30 °C, 160 rpm until cell density reaches 1×10^7–3×10^7 cells/mL.
4. Preheat water bath at 42 °C.
5. Wash cells with 20 mL TE/LiAc solution.
6. Resuspend pellet in TE/LiAc so that 50 µL contain 10^8 cells. Example: 50 mL at 1×10^7 cells/mL is resuspended after wash in 250 µL TE/LiAc.
7. Incubate cells for 15 min at 30 °C without shaking.
8. Prepare a 1.5 mL sterile microtube containing 300 µL PEG/TE/LiAc solution, 100 ng pGAL-Cas9 plasmid, 100 ng pRS416-CTG, 50 µg boiled carrier DNA. Vortex.
9. Add 50 µL of cells and gently mix using micropipette and a sterile yellow tip.
10. Incubate at 30 °C for 30 min without shaking.
11. Heat-shock at 42 °C for 20 min without shaking.
12. Centrifuge for 3 min at 3220 × *g*. Resuspend cells in sterile H_2O.
13. Plate cells on GLU −Ura −Leu plate.
14. Incubate at 30 °C for 40 h.

3.2 Flow Cytometry Assay

Each test should be carried out in triplicate, both in GLU −Ura −Leu and GAL −Ura −Leu media.

1. Fill one 96 deep-well plate with 300 µL of liquid GLU −Ura −Leu or GAL −Ura −Leu medium.
2. Using a toothpick seed one colony into each well. Be careful not to pick untransformed cells close to the colony (*see* **Note 4**). Gently agitate the toothpick so that the cells are transferred into the liquid medium.
3. Prepare another 96 well plate and add 100 µL of PBS in each well.
4. Transfer 100 µL of each well from the culture plate using a multichannel pipette to the 96-well plate containing the PBS.
5. Seal deep-well plate with breathable sealing membrane and incubate in a rotating incubator at 300 rpm, 30 °C.
6. On the flow cytometer, use the appropriate parameters for 96 well plates.
7. Select channels FSC, SSC both Area and Height. Voltages set up are as follows: 316 V (linear) for FSC, 350 V (linear) for SSC. To read FITC use channel B1 from the 488 nm laser with 525/50 nm filter. Use 390 V (log3) in voltage settings for this channel.
8. Select population on the FSC-A/SSC-A window to eliminate cellular debris (*see* **Notes 5** and **6**) then eliminate doublets on the SSC-A/SSC-H window. Quantify number of GFP-positive cells on the FITC-A histogram (Fig. 2).
9. After 12 h, prepare a new 96 well plate in which 100 µL of PBS is added. Harvest 30 µL of each well in the 96 deep-well culture plate and add them to the PBS plate.
10. Proceed from **step 6** to **8**, as above.
11. Repeat from **step 9** at 24 and 36 h (*see* **Note 7**) (Fig. 3).
12. At the end of the experiment, export all data to FCS files for subsequent analysis using FlowJo (or any appropriate software).

3.3 Data Analysis

Quantify the number of GFP-positive cells over time, using your favorite software (we use FlowJo 10.0). Use the following guidelines:

1. Select a homogeneous population on FSC-A/SSC-A graph to remove cell debris. Debris are usually only visible at T0 and most probably correspond to cell mortality due to the transformation process (Fig. 2a).
2. Select single cells on SSC-H/SSC-A graph (Fig. 2b).
3. On a glucose control population, select GFP- cells. Use this gate to define the same GFP- population in other samples. Cells exhibiting fluorescence above this limit are considered GFP+ (Fig. 2c).

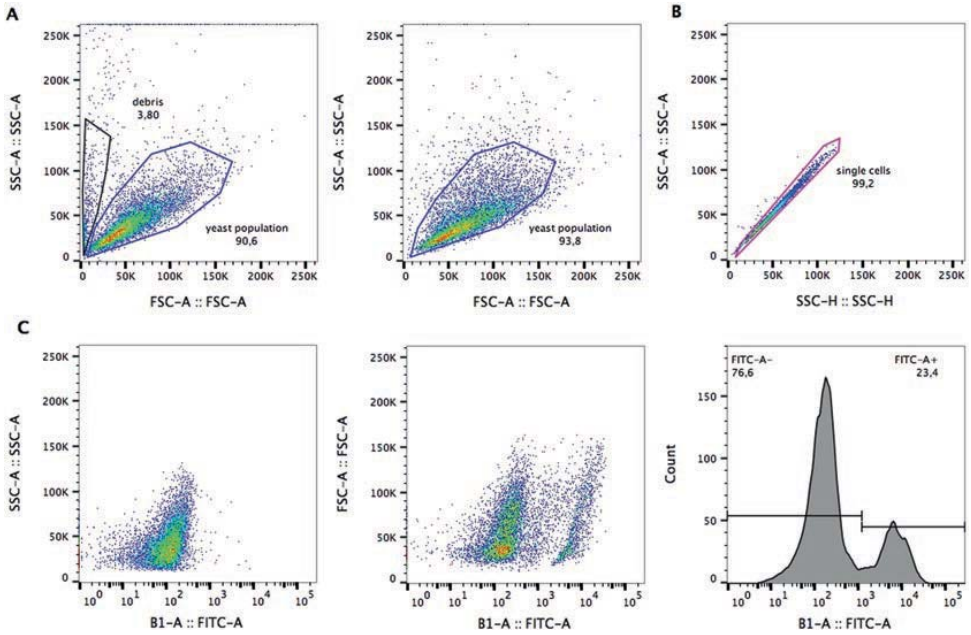


Fig. 2 Flow cytometry (**a**) Yeast total population analysis. Left**:** The FYBL1-4D-CTG strain was transformed with pGAL-Cas9 and pRS416-CTG plasmids, freshly picked from GLU −Ura −Leu plate 40 h after transformation. Yeast population is selected in a gate on the FSC-A/SSC-A graph and circled in blue. Debris are circled in black and not considered for subsequent analyses. Right: The same strain analyzed 12 h later. Note that cell debris are not visible anymore (**b**) Single cells are selected on the SSC-H/SSC-A graph (circled in pink on the figure) (**c**) Analysis of GFP expression. Single cells are now analyzed for GFP fluorescence. Left: Yeast cells at T0. Middle: Cells after 36 h in galactose medium. Note the shift of fluorescence corresponding to GFP-expressing cells. Right: GFP-positive cells are counted in the selected population: 23.4% of cells are GFP+ and have therefore repaired the DSB by Single-Strand Annealing

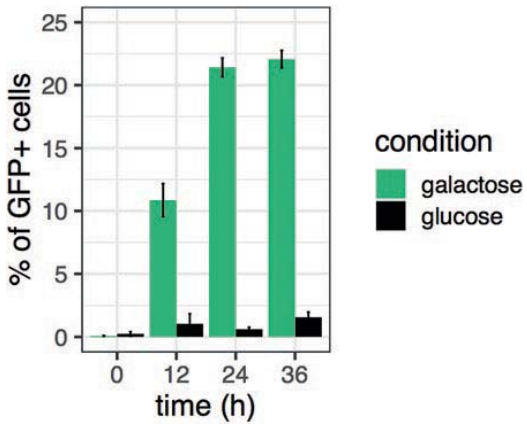


Fig. 3 GFP positive cells in glucose and galactose media during a 36 h time course**.** A low level of GFP+ cells is detected in glucose medium due to the *GAL1* promoter leakiness

4. Apply the same gating rules to all samples.
5. Export data collection to a spreadsheet.

3.4 Conclusion

In the present protocol, each time course was performed at only four time points (0, 12, 24 and 36 h), but additional points may be added. Note that the limiting step will be the time taken by the flow cytometer to process each sample. In addition, this assay is easily scalable. Here, the repair following a DSB made into a CTG repeat tract by SpCas9 was assessed, but it can be used to test a wide variety of nucleases [1] on many different target sequences. However, the experimenter should keep in mind the technical issue due to frequent leakiness of the *GAL1* promoter, resulting in an elevated background level of GFP+ cells in glucose medium. This is why we recommend carrying the experiment right after plasmid transformation, and as soon as small colonies can be picked.

Sequence of the *SNR52*-CTG-*SUP4* cassette for SpCas9:

TCTTTGAAAAGATAATGTATGATTATGCTTT
CACTCATATTTATACAGAAACTTGATGTTTTCTTTC
GAGTATATACAAGGTGATTACATGTACGTTT
GAAGTACAACTCTAGATTTTGTAGTGCC
CTCTTGGGCTAGCGGTAAAGGTGCGCATTTTTT
CACACCCTACAATGTTCTGTTCAAAAGATTTTGGTC
AAACGCTGTAGAAGTGAAAGTTGGTGCGCA
TGTTTCGGCGTTCGAAACTTCTCCGCAGTGA
AAGATAAATGATCTGCTGCTGCTGCTGCTGCTG
GTTTTAGAGCTAGAAATAGCAAGTTAAAATAAGGCTAG
TCCGTTATCAACTTGAAAAAGTGGCACCGAGTC
GGTGC*TTTTTTTGTTTTTTATGTCT*.

Bold: SNR52 promoter.
Plain: CTG guideRNA.
Underlined: Scaffold RNA.
Italics: SUP4 terminator.

4 Notes

1. Denaturation step is necessary to ensure that carrier DNA is single stranded and will bind to the cell wall and be degraded by DNases in the cytosol instead of the transformed plasmid. It greatly enhances transformation efficiency.
2. We use an incubator with a 25 mm shaking diameter to provide some oxygenation to yeast cells in 96 deep-well plates.
3. We work with a MACSQuant Analyzer (Miltenyi Biotec), but other cytometers may be used.
4. Very important: we recommend picking colonies right after they start being visible on a plate, 40 h after transformation and without preliminary subcloning. The *GAL1* promoter is

slightly leaky and waiting for a longer time will increase background noise as some cells will start expressing Cas9 and undergoing DSB-repair. This would result in an unwanted high proportion of GFP+ cells at the start of the experiment. Therefore, picking yeast colonies as early as 40 h after transformation is crucial.

5. At $t = 0$ h there can be cell debris due to mortality rate during transformation. These debris will be outnumbered by growing cells in future time points.
6. It is also possible to do a viability staining, using propidium iodide or a commercial viability dye like viability dye from Miltenyi 405/452. Be careful to choose a fluorophore that does not overlap with FITC signal.
7. We do not proceed after 36 h, since the cell culture is are saturated and cells stop dividing and expressing the GFP.

Acknowledgments

L.P. was supported by a CIFRE PhD fellowship from SANOFI. Work in G.-F. Richard laboratory was generously supported by the Institut Pasteur and the Centre National de la Recherche Scientifique (CNRS).

References

1. Richard G-F (2015) Shortening trinucleotide repeats using highly specific endonucleases: a possible approach to gene therapy? Trends Genet 31(4):177–186
2. Choulika A, Perrin A, Dujon B, Nicolas JF (1995) Induction of homologous recombination in mammalian chromosomes by using the I-SceI system of Saccharomyces cerevisiae. Mol Cell Biol 15(4):1968–1973
3. Colleaux L et al (1986) Universal code equivalent of a yeast mitochondrial intron reading frame is expressed into E. Coli as a specific double strand break endonuclease. Cell 44:521–533
4. Colleaux L, D'Auriol L, Galibert F, Dujon B (1988) Recognition and cleavage site of the intron-encoded omega transposase. Proc Natl Acad Sci U S A 85(16):6022–6026
5. Arnould S et al (2006) Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. J Mol Biol 355(3):443–458
6. Kim YG, Cha J, Chandrasegaran S (1996) Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc Natl Acad Sci U S A 93:1156–1160
7. Christian M et al (2010) Targeting DNA double-strand breaks with TAL effector nucleases. Genetics 186(2):757–761
8. Chen JS, Dagdas YS, Kleinstiver BP, Welch MM, Sousa AA, Harrington LB, Sternberg SH, Joung JK, Yildiz A, Doudna JA (2017) Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. Nature 550:407–410
9. Moynahan ME, Pierce AJ, Jasin M (2001) BRCA2 is required for homology-directed repair of chromosomal breaks. Molecular Cell 7:263–272
10. DiCarlo JE et al (2013) Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res 41(7):4336–4343
11. Sikorski RS, Hieter P (1989) A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics 122:19–27

Annex 3: TALEN-induced double-strand break repair of CTG trinucleotide repeats

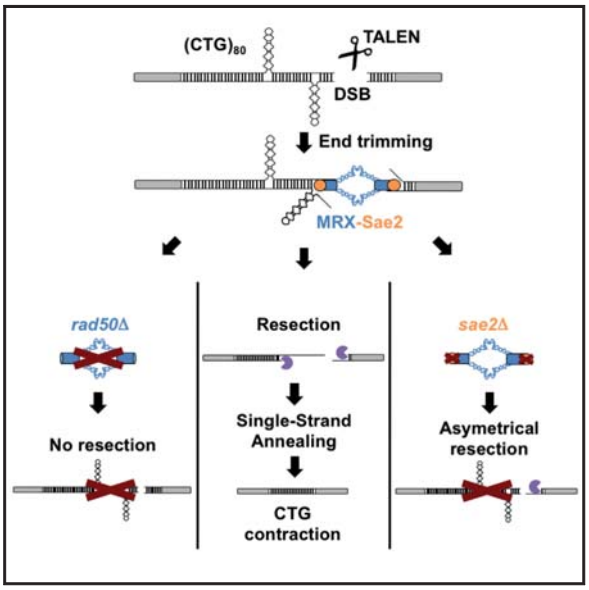
Article about the mechanism of CTG shortening after TALEN induction. I carried out terminal transferase-mediated PCR to amplify DSB at CTG repeats (Figure 1D) and did Southern blots including the one Figure 6.

Cell Reports

Article

TALEN-Induced Double-Strand Break Repair of CTG Trinucleotide Repeats

Graphical Abstract



Authors

Valentine Mosbach, Lucie Poggi, David Viterbo, Marine Charpentier, Guy-Franck Richard

Correspondence

gfrichar@pasteur.fr

In Brief

Mosbach et al. show that a TALEN could successfully contract an expanded CTG repeat tract below pathological length. The TALEN double-strand break needs Rad50, Rad52, and Sae2 to be repaired. A double-strand break end containing a long trinucleotide repeat needs both Rad50 and Sae2 to be efficiently resected.

Highlights

- *RAD50*, *SAE2*, and *RAD52* are involved in repairing a DSB within a CTG repeat
- *POL32*, *DNL4*, and *RAD51* do not play a role in repairing a DSB within a CTG repeat
- Resection of a DSB within a CTG repeat needs the Mre11-Rad50 complex, as well as Sae2
- The double-strand break is repaired by an iterative single-strand annealing process

Mosbach et al., 2018, Cell Reports *22*, 2146–2159
February 20, 2018
https://doi.org/10.1016/j.celrep.2018.01.083

CellPress

OPEN ACCESS
CellPress

Cell Reports
Article

TALEN-Induced Double-Strand Break Repair of CTG Trinucleotide Repeats

Valentine Mosbach,[1,2,3] Lucie Poggi,[1,2,3] David Viterbo,[1,3] Marine Charpentier,[4] and Guy-Franck Richard[1,3,5,*]
[1]Institut Pasteur, Département Génomes & Génétique, 25 rue du Dr Roux, 75015 Paris, France
[2]Sorbonne Universités, UPMC Univ Paris 06, IFD, 4 Place Jussieu, 75252 Paris Cedex 05, France
[3]CNRS, UMR3525, 75015 Paris, France
[4]INSERM U1154, CNRS UMR7196, Museum National d'Histoire Naturelle, 75005 Paris, France
[5]Lead Contact
*Correspondence: gfrichar@pasteur.fr
https://doi.org/10.1016/j.celrep.2018.01.083

SUMMARY

Trinucleotide repeat expansions involving CTG/CAG triplets are responsible for several neurodegenerative disorders, including myotonic dystrophy and Huntington's disease. Because expansions trigger the disease, contracting repeat length could be a possible approach to gene therapy for these disorders. Here, we show that a TALEN-induced double-strand break was very efficient at contracting expanded CTG repeats in yeast. We show that *RAD51*, *POL32*, and *DNL4* are dispensable for double-strand break repair within CTG repeats, the only required genes being *RAD50*, *SAE2*, and *RAD52*. Resection was totally abolished in the absence of *RAD50* on both sides of the break, whereas it was reduced in a *sae2*Δ mutant on the side of the break containing the longest repeat tract, suggesting that secondary structures at double-strand break ends must be removed by the Mre11-Rad50 complex and Sae2. Following the TALEN double-strand break, single-strand annealing occurred between both sides of the repeat tract, leading to repeat contraction.

INTRODUCTION

Microsatellite expansions are responsible for more than two dozen neurological or developmental disorders in humans. Among the most common sequences involved are CAG/CTG trinucleotide repeat tracts, whose expansions are the cause of Huntington's disease, myotonic dystrophy type 1 (DM1 or Steinert disease), and several spinocerebellar ataxias (Orr and Zoghbi, 2007). Despite having been under investigation for more than two decades, the molecular mechanism(s) leading to large expansions is not completely understood, although it is generally accepted that secondary structures formed by these microsatellites may be triggering or amplifying the expansion process (McMurray, 1999). It was shown that CAG and CTG trinucleotide repeats form imperfect hairpins *in vitro* (Gacy et al., 1995; Yu et al., 1995a, 1995b). In addition, there is biochemical and genetical evidence that CAG and CTG hairpins interfere with the mismatch repair machinery, an important player of the expansion process, although its precise role is not totally clear (Foiry et al., 2006; Manley et al., 1999; Owen et al., 2005; Pearson et al., 1997; Pinto et al., 2013; Savouret et al., 2004; Slean et al., 2016; Tian et al., 2009; Tomé et al., 2009, 2013; Viterbo et al., 2016; Williams and Surtees, 2015). Most trinucleotide repeat transmissions from parents to children lead to repeat tract expansion. However, it seldom happens that a large allele contracts to a shorter one. Indeed, in a family affected by DM 1, it was reported that a daughter inherited a contracted allele from her father by a mechanism likely to be gene conversion (O'Hoy et al., 1993). The daughter was followed until the age of 17 and did not develop any of the symptoms of the pathology, showing that a large repeat contraction prevented this kind of disease.

Recent attempts were made to cure trinucleotide repeat disorders by gene therapy. In Huntington's disease pluripotent stem cells, an expanded CAG repeat in the Huntington's disease (HD) gene was replaced by a smaller allele by homologous recombination. In corrected cells, HD disease phenotypes were reversed (An et al., 2012). Two independent groups used SpCas9 either to induce one single double-strand break upstream of the *FMR1* CGG repeat (Park et al., 2015) or two double-strand breaks (DSBs) upstream and downstream of the repeat tract (Xie et al., 2016). In both cases, *FMR1* reactivation was observed in edited cells. More recently, SpCas9 was used to delete the expanded CTG triplet repeat at the DM1 locus by making a DSB upstream and/or downstream of the repeat tract. Again, disease phenotypes were partially suppressed in DM1 myoblasts (van Agtmaal et al., 2017). In all of these cases, DSBs were always induced outside and never inside the trinucleotide repeat tract.

DSB repair is one of the molecular processes leading to trinucleotide repeat contractions and expansions. It was formerly shown that a DSB made by the I-*Sce* I meganuclease within a short CTG repeat tract often led to the loss of the nuclease recognition site and contraction of the repeat tract (Richard et al., 1999). In less frequent cases, it led to both expansions and contractions of the repeat tract by gene conversion during mitosis (Richard et al., 1999, 2000) or meiosis (Richard et al., 2003). Following these early experiments, zinc-finger nucleases (ZFNs) were used to direct a DSB within a CAG or CTG trinucleotide repeat tract. In two separate studies from the same lab, induction of a ZFN in Chinese hamster ovary (CHO) cells led to a 15-fold increase in repeat contractions. However, deletions in

one or both flanking regions were observed in 20% of the cases, whereas insertions of exogenous DNA at the DSB site were found in another 24% of the cases (Mittelman et al., 2009; Santillan et al., 2014). Different authors used another ZFN expressed in HeLa cells containing CAG/CTG trinucleotide repeats integrated in the two possible orientations compared with replication fork progression. They observed contractions as well as expansions of the repeat tract when both ZFN arms were expressed, but only contractions were recovered when one single arm was expressed (Liu et al., 2010). This suggested that one arm of the ZFN was able to homodimerize and induce a DSB by itself. Using a *Saccharomyces cerevisiae* strain in which a large CTG triplet repeat from a DM1 patient was integrated in a yeast chromosome, we were recently able to show that induction of a transcription activator-like effector (TALE) nuclease (TALEN) induced contractions of a CTG triplet repeat tract at a high frequency. Pulse-field gel electrophoresis and genome-wide deep sequencing showed that no other mutation, duplication, or chromosomal rearrangement was induced by the TALEN outside of the repeat tract (Richard et al., 2014). These experiments demonstrated that this new family of nucleases was efficient and specific enough to envision their possible use as a future gene therapy tool in human cells (Richard, 2015). Using a different approach, Cinesi et al. (2016) recently showed that inducing single-strand breaks within a CTG repeat tract using the Cas9D10A mutant nickase also promoted contractions of the repeat tract in model human cells.

Mechanisms of DSB repair have been studied in yeast for several decades, and the main proteins involved in this process have been identified (Krogh and Symington, 2004). A large part of these advances was made possible by the use of highly specific meganucleases such as HO or I-*Sce* I (Fairhead and Dujon, 1993; Haber, 1995; Plessis et al., 1992). However, the fate of a single DSB made within a long repeated and structured DNA sequence was never addressed before.

One of the goals of the present work was to study the role of several recombination genes (namely, *RAD50*, *RAD51*, *RAD52*, *DNL4*, *SAE2*, and *POL32*) in the repair of a single DSB made within a long CTG trinucleotide repeat. *RAD52* encodes a mediator multimeric protein controlling homologous recombination pathways (gene conversion, single-strand annealing [SSA], and break-induced replication [BIR]) (Davis and Symington, 2004; Krogh and Symington, 2004; Sugawara and Haber, 1992). *RAD51* is a RecA homolog responsible for nucleofilament formation and subsequent strand exchange and gene conversion (Shinohara et al., 1992; Sung, 1994). *RAD50* belongs to the multifunctional Mre11-Rad50-Xrs2 complex involved, along with Sae2, in DSB end clipping and resection during meiosis as well as mitosis (Borde et al., 2004; Lee et al., 1998; Mimitou and Symington, 2008; Zhu et al., 2008). *DNL4* encodes ligase IV, the protein responsible for the ligation step during non-homologous end joining (Wilson et al., 1997), and *POL32* is part of the polymerase δ complex and was shown to be essential for BIR (Lydeard et al., 2007) as well as to be an important player in microhomology-mediated repair (Villarreal et al., 2012).

RAD50 was found to be essential to resect both DSB ends, whereas *SAE2* was needed to resect only the DSB end that contains most of the triplet repeat tract. This observation supports the presence of secondary structures that need a functional Sae2 activity to be removed. *RAD52* was also required to repair the DSB but not *RAD51*, *POL32*, or *DNL4*, suggesting an iterative SSA process that progressively leads to repeat shortening.

RESULTS

A DSB Induced within CTG Repeats Requires the Mre11-Rad50 Complex to be Processed

In the present work, two TALENs were used. The $TALEN_{CTG}$ was the same as the one used in our former publication, designed to induce a DSB within a modified *SUP4* allele containing expanded CTG triplets from the dystrophia myotonica protein kinase (DMPK) human locus (Richard et al., 2014). The $TALEN_{noCTG}$ was designed to induce a DSB within a modified *SUP4* allele containing an I-*Sce* I recognition site (Richard et al., 1999). The trinucleotide repeat tract lengths used here ranged from 20–50 triplets for short alleles to 70–90 triplets for long alleles. In a first series of experiments, the $TALEN_{CTG}$ was expressed in wild-type yeast cells and in isogenic strains mutated for DSB repair genes. Both $TALEN_{CTG}$ arms were carried on centromeric vectors, and their expression was under control of an inducible TetOFF promoter. DSB formation was followed during a time course by Southern blot analysis. When the $TALEN_{CTG}$ was repressed, uncut chromosomes containing CTG repeats of two different lengths were visible. When the $TALEN_{CTG}$ was expressed, signals corresponding to DSB formation were detected (Figure 1A). By using probes specific to each side of the repeat tract, it was possible to distinguish between signals corresponding to 5′ or 3′ ends of the DSB (Figure S1). The 5′ end of the break, containing the long CTG tract appears like a smear. This smear corresponds to different repeat lengths because of progressive CTG repeat contraction over time. The 3′ end of the DSB appears as a sharper band because it contains only a few triplets. The DSB was not visible before 14 hr after $TALEN_{CTG}$ induction (time point labeled 0, Figure 1A). Quantification showed that the maximum of broken molecules was reached 4–6 hr after T0 for all strains except *rad50*Δ and *sae2*Δ (Figure 1B).

To determine how long it would take for cells to completely repair the DSB, a longer time course was run in wild-type cells over 72 hr after TALEN induction (58 hr after DSB formation; Figure 2). This experiment was set up in haploid cells to discriminate between the parental (uncontracted) allele and the contracted allele recovered after DSB repair. Cells were collected at several time points, with particular attention to the 34–46 hr time range. Total genomic DNA was extracted, and Southern blot was run as described previously (Figure 2A). Signal quantification showed that, during the first 40 hr in which ~12% of chromosomes were broken, the DSB signal stayed stable. After that time, it increased to ~20% of broken molecules and stayed at the same level until the end of the time course (Figure 2B). This result may be interpreted in two ways: (1) the nuclease was not active in all cells at the same time. Therefore, only a subfraction of repeat tracts was cleaved in the first 40 hr and another, larger, subfraction was cleaved later on. (2) A first burst of DSBs partially contracted repeat tracts in all cells. A second round of DSBs cleaved shortened repeats more efficiently because they

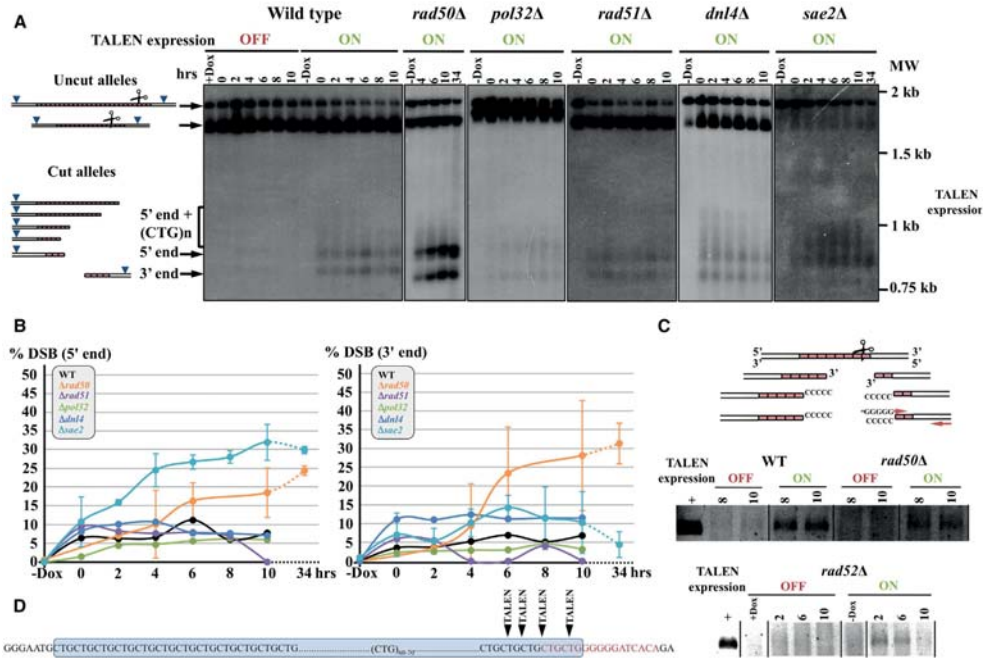


Figure 1. DSB Induction by a TALEN$_{CTG}$ within CTG Repeats

(A) Southern blots of yeast strains during DSB induction. For each wild-type and mutant strain, cells were collected at different time points after induction (+Dox or −Dox). The time point labeled "0" represents the first time point which the DSB was detectable. For all experiments, it corresponds to 14 hr after induction (+Dox or −Dox). When the TALEN$_{CTG}$ was repressed (OFF), no band corresponding to the DSB was visible. When the TALEN$_{CTG}$ was induced (ON), several signals were detected: a smear corresponding to successive contractions of the large trinucleotide repeat tract located 5′ of the DSB, a band corresponding to the 5′ end of the DSB with only a few triplets left, and another band corresponding to the 3′ end of the DSB containing 1–4 triplets. The cartoons at the left describe the different molecules detected. Blue triangles indicate the location of EcoRV sites used for restriction digestion before Southern blotting. Scissors indicate the DSB location.

(B) Quantification of 5′ and 3′ DSB signals. Note that the time course was run during 34 hr only in the *rad50*Δ and *sae2*Δ strains.

(C) Terminal transferase-mediated PCR. After DSB induction, dCTP was added to both 3′ strands by terminal transferase. The 3′ end of the break was subsequently amplified with a poly-dG oligonucleotide and another primer specific of the 3′ end of the DSB. Note that additional time points and controls were present on the same gels, but only significant time points are shown here.

(D) Results of terminal transferase-mediated PCR sequences. The repeat tract is shown in the blue box, and the right TALEN$_{CTG}$ binding site is indicated by red letters. The locations of the four TALEN$_{CTG}$ DSBs sequenced are indicated by black arrowheads.

One time course was performed for wild-type, *lig4*Δ, *rad51*Δ, and *pol32*Δ strains. For the *rad50*Δ and *sae2*Δ strains, values are the average of two or three independent experiments, depending on the time points considered. Error bars correspond to one SD.

were more accessible to the nuclease. This experiment also showed that smear length progressively decreased over time, although the 5′ smear intensity was too low to be reliably quantified (Figure 2A). Given the rate of decrease of the parental band, it is expected to see its complete disappearance after 6 days of induction (Figure 2C).

Time courses for *dnl4*Δ, *pol32*Δ, and *rad51*Δ mutants were similar to the wild-type strain, with a maximum of 12.4% of broken molecules in the *dnl4*Δ mutant (Figures 1A and 1B). We concluded that none of these mutants showed a detectable effect on DSB repair kinetics. On the contrary, in the *rad50*Δ mutant, an accumulation of DSBs was observed (Figure 1A). On both sides of the break, a signal increase was clearly detected (Figure 1B). This suggested that a DSB induced in CTG repeats was not correctly processed in this mutant, leading to an accumulation of unrepaired broken molecules. In the *sae2*Δ mutant, the 5′ and 3′ ends of the DSB showed an asymmetric increase compared with *rad50*Δ cells. The DSB end containing most of the trinucleotide repeat tract (5′ end) shows the same accumulation rate as in *rad50*Δ, whereas the other DSB end, containing only a few triplets, accumulates more slowly (Figures 1A and 1B).

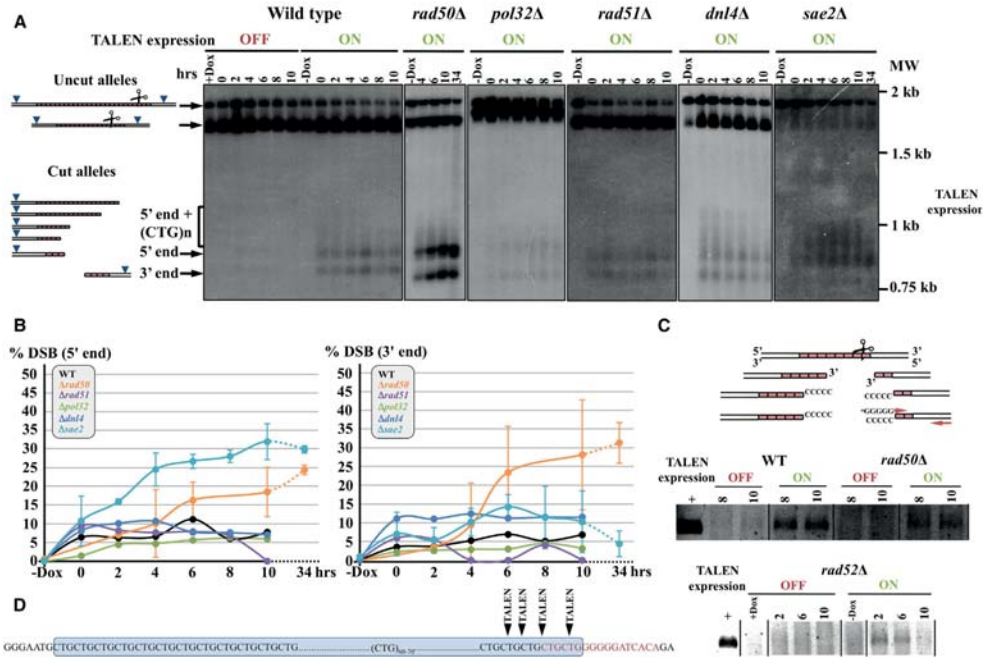


Figure 1. DSB Induction by a TALEN$_{CTG}$ within CTG Repeats

(A) Southern blots of yeast strains during DSB induction. For each wild-type and mutant strain, cells were collected at different time points after induction (+Dox or −Dox). The time point labeled "0" represents the first time point which the DSB was detectable. For all experiments, it corresponds to 14 hr after induction (+Dox or −Dox). When the TALEN$_{CTG}$ was repressed (OFF), no band corresponding to the DSB was visible. When the TALEN$_{CTG}$ was induced (ON), several signals were detected: a smear corresponding to successive contractions of the large trinucleotide repeat tract located 5′ of the DSB, a band corresponding to the 5′ end of the DSB with only a few triplets left, and another band corresponding to the 3′ end of the DSB containing 1–4 triplets. The cartoons at the left describe the different molecules detected. Blue triangles indicate the location of EcoRV sites used for restriction digestion before Southern blotting. Scissors indicate the DSB location.

(B) Quantification of 5′ and 3′ DSB signals. Note that the time course was run during 34 hr only in the *rad50*Δ and *sae2*Δ strains.

(C) Terminal transferase-mediated PCR. After DSB induction, dCTP was added to both 3′ strands by terminal transferase. The 3′ end of the break was subsequently amplified with a poly-dG oligonucleotide and another primer specific of the 3′ end of the DSB. Note that additional time points and controls were present on the same gels, but only significant time points are shown here.

(D) Results of terminal transferase-mediated PCR sequences. The repeat tract is shown in the blue box, and the right TALEN$_{CTG}$ binding site is indicated by red letters. The locations of the four TALEN$_{CTG}$ DSBs sequenced are indicated by black arrowheads.

One time course was performed for wild-type, *lig4*Δ, *rad51*Δ, and *pol32*Δ strains. For the *rad50*Δ and *sae2*Δ strains, values are the average of two or three independent experiments, depending on the time points considered. Error bars correspond to one SD.

were more accessible to the nuclease. This experiment also showed that smear length progressively decreased over time, although the 5′ smear intensity was too low to be reliably quantified (Figure 2A). Given the rate of decrease of the parental band, it is expected to see its complete disappearance after 6 days of induction (Figure 2C).

Time courses for *dnl4*Δ, *pol32*Δ, and *rad51*Δ mutants were similar to the wild-type strain, with a maximum of 12.4% of broken molecules in the *dnl4*Δ mutant (Figures 1A and 1B). We concluded that none of these mutants showed a detectable effect on DSB repair kinetics. On the contrary, in the *rad50*Δ mutant, an accumulation of DSBs was observed (Figure 1A). On both sides of the break, a signal increase was clearly detected (Figure 1B). This suggested that a DSB induced in CTG repeats was not correctly processed in this mutant, leading to an accumulation of unrepaired broken molecules. In the *sae2*Δ mutant, the 5′ and 3′ ends of the DSB showed an asymmetric increase compared with *rad50*Δ cells. The DSB end containing most of the trinucleotide repeat tract (5′ end) shows the same accumulation rate as in *rad50*Δ, whereas the other DSB end, containing only a few triplets, accumulates more slowly (Figures 1A and 1B).

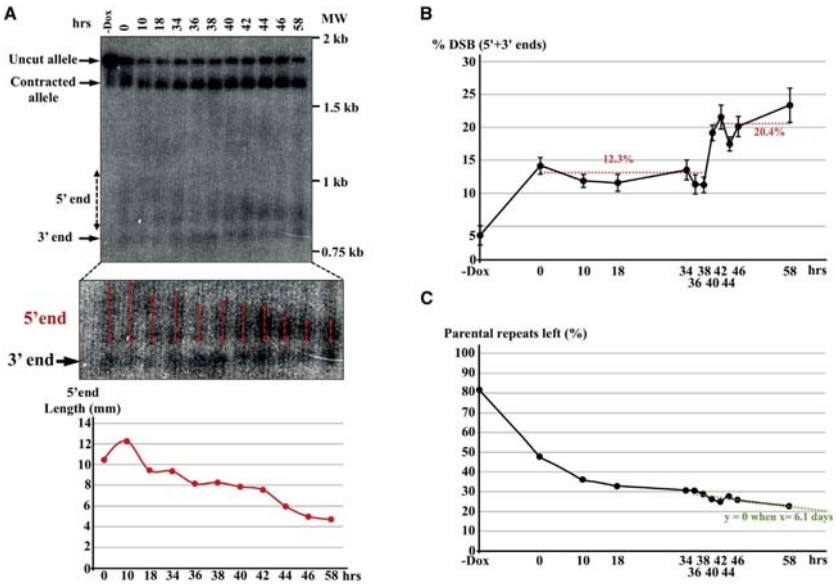


Figure 2. Expression of a TALEN$_{CTG}$ during a 72-hr Time Course

(A) Southern blot of a time course in a haploid strain containing only one CTG repeat allele, for 3 days.

(B) Quantification of the DSB level during the time course. For this experiment, 5′ and 3′ end signals were added and compared with the total signal in each lane. The two horizontal red dotted lines correspond to average DSB levels in each time frame.

(C) Quantification of the amount of parental repeat length left during TALEN$_{CTG}$ induction. The ratio of the parental allele over total signal is represented. The amount of parental allele decreased rapidly during the first day, corresponding to cells that repaired the DSB by SSA and, therefore, contracted the repeat tract. Then parental allele reduction occurs more slowly, when new repeat tracts are cut and contracted over time. Linear regression of the slower part of the graph (green dotted line) shows that the parental allele would have completely disappeared 6 days after the beginning of the induction. Note that the graphs in both (B) and (C) show the average of two independent experiments. Error bars correspond to one SD.

In the *rad52*Δ mutant, no signal corresponding to DSB formation could be detected by Southern blot (Figure S2A). TALEN expression was verified by western blot analysis (Figure S3). In the presence of doxycycline, no signal was detected in any of the strains. In the absence of doxycycline, the hemagglutinin (HA)-tagged TALEN was clearly detected in all strains. Its relative level was similar in *rad50*Δ, *rad51*Δ, and *rad52*Δ strains but ~10-fold higher in wild-type cells. We concluded that the absence of visible DSB in the *rad52*Δ strain was not due to a lack of expression of the nuclease because it was present in similar amounts in the two other mutant strains, in which the DSB was clearly detected. To check whether the absence of detectable DSB was due to some mutation unrelated to the *RAD52* deletion itself, we performed the same experiment in a *rad52*Δ/*RAD52* heterozygote. In this strain, the DSB was clearly visible, showing that complementing the *rad52*Δ deletion with a *RAD52* gene restored the wild-type phenotype (Figure S2B).

Subsequently, a terminal deoxynucleotidyl transferase-mediated PCR approach was used to amplify the DSB (Förstemann et al., 2000). PCR products were visible in the wild-type and *rad50*Δ strains used as positive controls at 8 hr and 10 hr, but no product was detected when the TALEN$_{CTG}$ was repressed (Figure 1C). PCR products corresponding to DSB amplification were also visible in the *rad52*Δ mutant when the TALEN$_{CTG}$ was expressed, but very faintly. We concluded that DSBs occurred as expected within the repeat tract in the *rad52*Δ strain but that their level was too low to be detected by Southern blot. Sanger sequencing of the terminal end of the PCR product generated by terminal deoxynucleotidyl transferase-mediated PCR products showed that the DSB occurred in the very last 1–4 CTG triplets of the repeat tract (Figure 1D).

DSB Accumulation in *rad50*Δ Depends on CTG Repeats

To determine whether DSB accumulation was only dependent on the presence of CTG repeats at the end of the DSB, a second TALEN was designed to recognize a *SUP4* allele that did not contain a repeat tract. This TALEN was called TALEN$_{noCTG}$ to distinguish it from the TALEN$_{CTG}$. In the wild-type strain, the DSB signal was weaker because only one of the two chromosomes could be cut by the TALEN$_{noCTG}$ (Figures 3A and 3B).

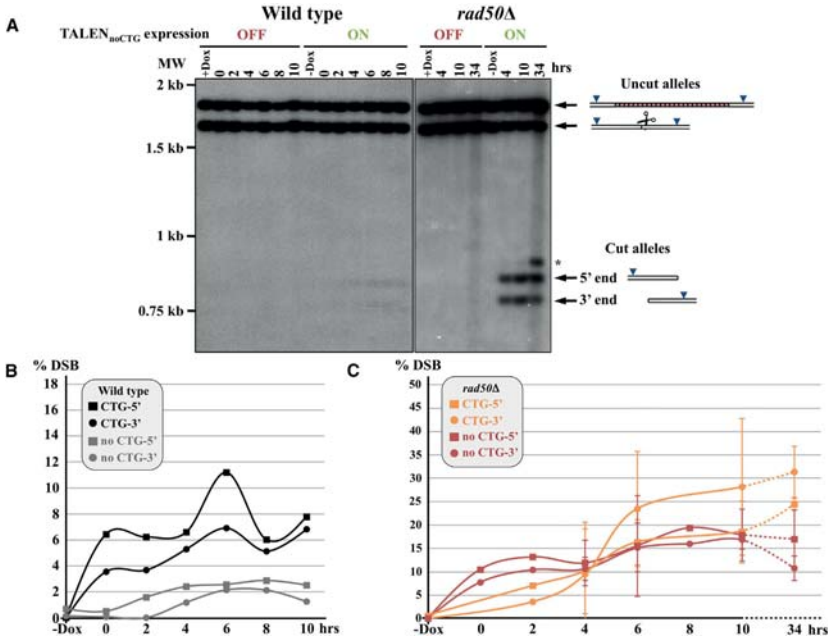


Figure 3. DSB Induction by a TALEN$_{noCTG}$ within a Non-repeated Region

(A) Southern blots of time courses during DSB induction of the TALEN$_{noCTG}$ in wild-type and *rad50*Δ strains. The asterisk indicates an extra band only visible in the *rad50*Δ strain, probably corresponding to some chromosomal rearrangement specific to this mutant.

(B) Quantifications of the TALEN$_{noCTG}$ 5′ and 3′ DSB signals and comparisons with the TALEN$_{CTG}$. For each time point, the amounts of 5′ or 3′ signals were quantified and plotted as a ratio of the total signal in the lane. One time course was performed for the wild-type strain. For the *rad50*Δ strain, values are the average of two or three independent experiments, depending on the time points. Error bars correspond to one SD.

The smear detected when the TALEN$_{CTG}$ was induced was not visible with the TALEN$_{noCTG}$, proving that is corresponds to different repeat lengths because of progressive repeat contraction over time. The number of broken molecules increased at a slower rate over time compared with the TALEN$_{CTG}$. Six hours after DSB, non CTG-containing ends are four times less abundant than CTG-containing ends. Repeat-containing broken molecules are more persistent, suggesting that non CTG-containing ends are repaired faster than CTG-containing ends. In the *rad50*Δ mutant, the DSB also accumulates in the TALEN$_{noCTG}$ strain compared with the TALEN$_{CTG}$ strain (Figure 3C). However, the amount of non CTG-containing ends decreases slowly after 10 hr, whereas CTG-containing ends keep on accumulating. We concluded that cut fragments were greatly stabilized in the absence of *RAD50* but that repair of non-repetitive ends eventually occurs at the last time point, whereas it is definitely compromised when CTG-containing ends need to be repaired. This strongly suggests that these repeats form secondary structures at DSB ends that need the Mre11-Rad50-Xrs2 (MRX) complex to be removed for repair to occur.

DSB Resection within CTG Repeats Is Almost Completely Abolished in *rad50*Δ and *sae2*Δ

DSB resection was determined by qPCR of total genomic DNA preliminary digested by a restriction endonuclease (EcoRV in the present case) (Chen et al., 2013). Restriction sites that were resected during the course of the experiment could not be digested by EcoRV because of their single-stranded nature and could therefore be amplified by PCR primers located around the restriction site. Comparisons of cycle threshold (Ct) obtained in the fraction digested by EcoRV with Ct obtained in the undigested fraction was indicative of the amount of resection at this particular restriction site. Four EcoRV sites were studied: two of them located 800–900 base pairs (bp) upstream and downstream of the repeat tract and two located further away, 1.8–2.9 kb upstream and downstream of the CTG repeat (Figure 4). In all experiments, raw resection and relative resection values were calculated. Raw resections were computed as a percentage of PCR product amplified in EcoRV digested compared with non-digested fractions (Experimental Procedures). Relative resection values were calculated as the ratio of raw values to DSB amounts quantified on Southern blots.

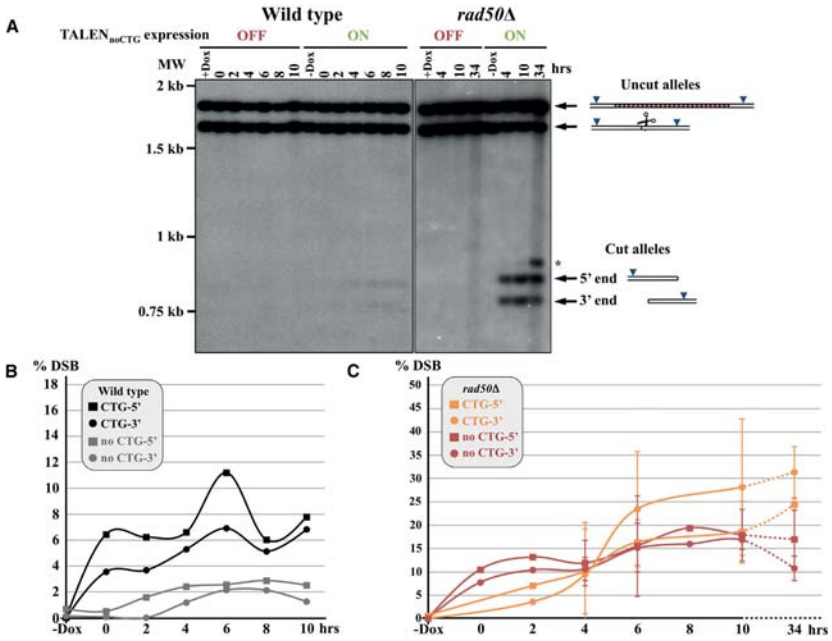


Figure 3. DSB Induction by a TALEN$_{noCTG}$ within a Non-repeated Region

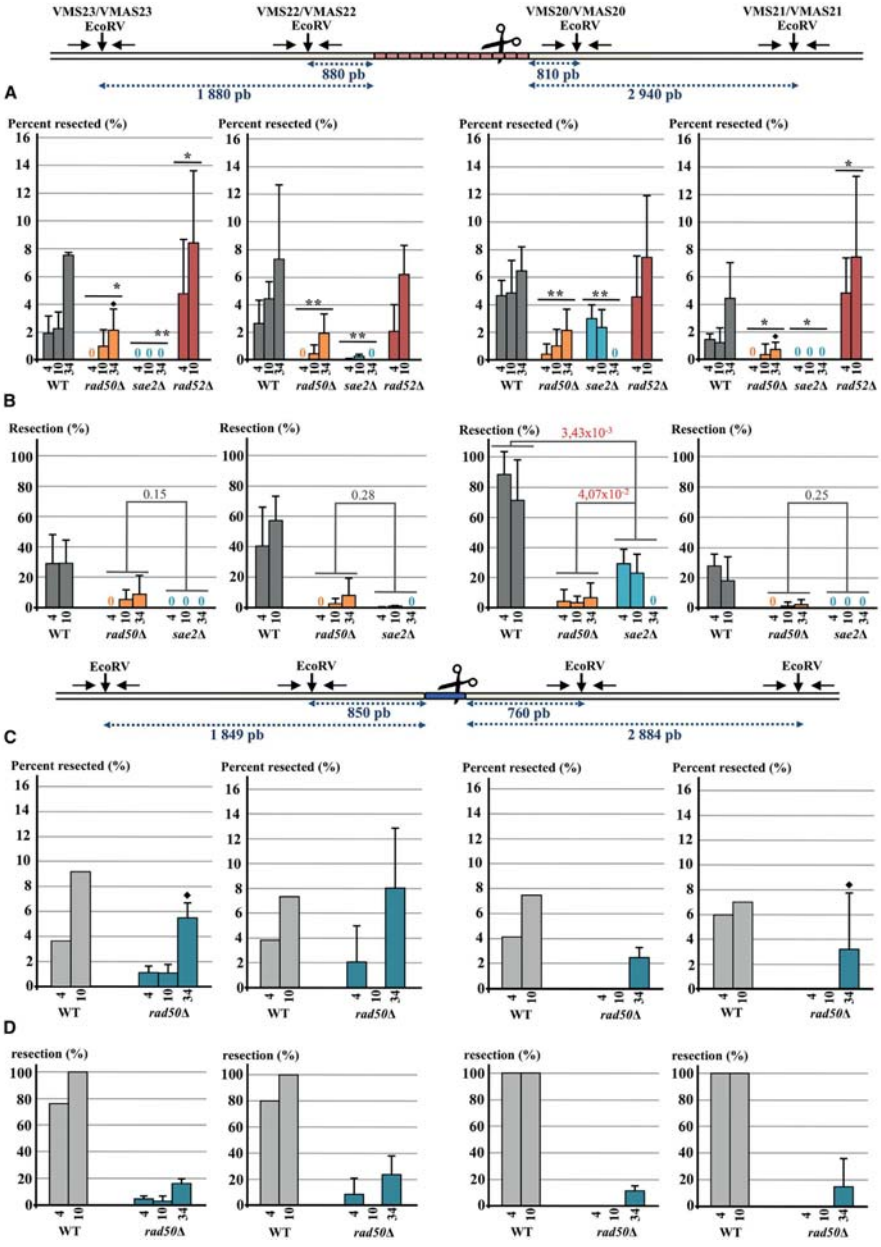
(A) Southern blots of time courses during DSB induction of the TALEN$_{noCTG}$ in wild-type and *rad50*Δ strains. The asterisk indicates an extra band only visible in the *rad50*Δ strain, probably corresponding to some chromosomal rearrangement specific to this mutant.

(B) Quantifications of the TALEN$_{noCTG}$ 5′ and 3′ DSB signals and comparisons with the TALEN$_{CTG}$. For each time point, the amounts of 5′ or 3′ signals were quantified and plotted as a ratio of the total signal in the lane. One time course was performed for the wild-type strain. For the *rad50*Δ strain, values are the average of two or three independent experiments, depending on the time points. Error bars correspond to one SD.

The smear detected when the TALEN$_{CTG}$ was induced was not visible with the TALEN$_{noCTG}$, proving that is corresponds to different repeat lengths because of progressive repeat contraction over time. The number of broken molecules increased at a slower rate over time compared with the TALEN$_{CTG}$. Six hours after DSB, non CTG-containing ends are four times less abundant than CTG-containing ends. Repeat-containing broken molecules are more persistent, suggesting that non CTG-containing ends are repaired faster than CTG-containing ends. In the *rad50*Δ mutant, the DSB also accumulates in the TALEN$_{noCTG}$ strain compared with the TALEN$_{CTG}$ strain (Figure 3C). However, the amount of non CTG-containing ends decreases slowly after 10 hr, whereas CTG-containing ends keep on accumulating. We concluded that cut fragments were greatly stabilized in the absence of *RAD50* but that repair of non-repetitive ends eventually occurs at the last time point, whereas it is definitely compromised when CTG-containing ends need to be repaired. This strongly suggests that these repeats form secondary structures at DSB ends that need the Mre11-Rad50-Xrs2 (MRX) complex to be removed for repair to occur.

DSB Resection within CTG Repeats Is Almost Completely Abolished in *rad50*Δ and *sae2*Δ

DSB resection was determined by qPCR of total genomic DNA preliminary digested by a restriction endonuclease (EcoRV in the present case) (Chen et al., 2013). Restriction sites that were resected during the course of the experiment could not be digested by EcoRV because of their single-stranded nature and could therefore be amplified by PCR primers located around the restriction site. Comparisons of cycle threshold (Ct) obtained in the fraction digested by EcoRV with Ct obtained in the undigested fraction was indicative of the amount of resection at this particular restriction site. Four EcoRV sites were studied: two of them located 800–900 base pairs (bp) upstream and downstream of the repeat tract and two located further away, 1.8–2.9 kb upstream and downstream of the CTG repeat (Figure 4). In all experiments, raw resection and relative resection values were calculated. Raw resections were computed as a percentage of PCR product amplified in EcoRV digested compared with non-digested fractions (Experimental Procedures). Relative resection values were calculated as the ratio of raw values to DSB amounts quantified on Southern blots.



(legend on next page)

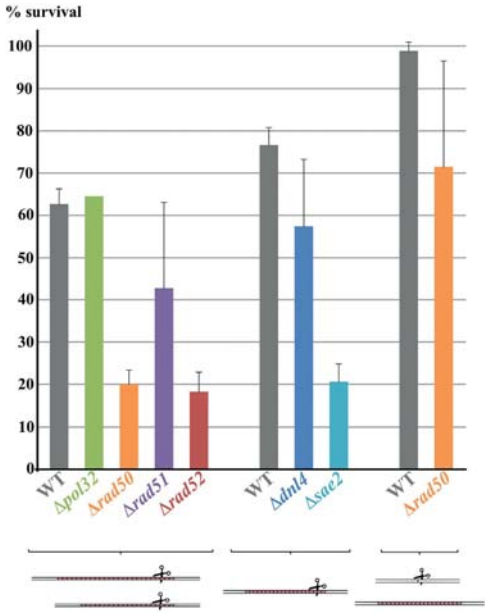


Figure 5. Survival of TALEN-Induced DSBs

Left: survival in diploid cells carrying two alleles of different trinucleotide repeat tract lengths. In these strains, both chromosomes carry trinucleotide repeats and were cut by the nuclease, as shown in the bottom cartoon. Center: survival in haploid cells carrying only one repeat tract. Right: survival in diploid cells carrying only one repeat tract. In these strains, only the chromosome that did not carry trinucleotide repeats was cut by the nuclease. Survival values are the average of two to five experiments, expect for *pol32*Δ. Error bars are equal to one SD.

Because DSB signals in the *rad52*Δ strain were too low to be quantified, relative resection values could not be calculated for this mutant.

In the wild-type strain, raw resection as well as relative resection values were significantly higher at sites proximal to the DSB (VMS20/VMAS20 and VMS22/VMAS22) than at distal sites (VMS21/VMAS21 and VMS23/VMAS23) (Figures 4A and 4B; two-tailed t test p = 0.0085). Resection values dramatically dropped in the *rad50*Δ strain as well as in the *sae2*Δ mutant to become barely detectable at early time points. Note that *SAE2* was found to be more important for relative resection on the DSB end containing the longer repeat tract, whereas *RAD50* was required on both DSB ends (Figure 4B; VMS20/VMAS20, p = 0.04).

In the *rad52*Δ strain, resection was not different from the wild-type at proximal sites but increased at distal sites, suggesting that, in the absence of *RAD52*, resecting enzymes have a better access to DNA ends.

In the wild-type strain expressing the $TALEN_{noCTG}$, raw as well as relative resection values were much higher compared with the $TALEN_{CTG}$ values, particularly at distal EcoRV sites (Figures 4C and 4D). In addition, there was no detectable difference between resection values at proximal versus distal sites. In the *rad50*Δ mutant expressing the $TALEN_{noCTG}$, resection was increased 34 hr after DSB formation compared with the $TALEN_{CTG}$ *rad50*Δ strain, but this increase was found to be significant only at distal restriction sites (Figures 4A and 4C; two-tailed t test, p = 0.0197). This shows that, in the absence of *RAD50*, resection is less efficient when CTG repeats are present at the DSB.

RAD50, *RAD52*, and *SAE2* Are Needed to Repair a TALEN-Induced DSB

Survival of the DSB was determined in each strain by the ratio of colony-forming units (CFUs) when the $TALEN_{CTG}$ was expressed to CFUs when the $TALEN_{CTG}$ was repressed. In wild-type diploid cells, survival was 62.5% ± 3.7%, not significantly different from *pol32*Δ and *rad51*Δ mutant strains (64.4% and 42.6% ± 20.4%, respectively), although a somewhat higher SD was observed for *rad51*Δ (Figure 5). Therefore, the absence of these genes did not significantly affect DSB repair efficacy, consistent with similarities in DSB formation and processing during time courses (Figure 1). On the contrary, the absence of *RAD50* or *RAD52* led to a higher mortality because only 19.9% ± 3.4% of cells survived in the *rad50*Δ strain and 18.2% ± 4.7% survived in *rad52*Δ cells, proving that the product of both genes was required to repair the break. Survival in diploid strains heterozygous for the *rad50*Δ or the *rad52*Δ deletion was not significantly different from the wild-type (Figure S2C). When a chromosome that did not contain CTG repeats was cut by the $TALEN_{noCTG}$ in the *rad50*Δ strain, survival was not significantly different from the wild-type, confirming that this gene product was not essential to repair a DSB in non-repeated DNA (Figure 5).

Figure 4. Quantification of DSB Resection

Resection graphs for each primer pair are plotted under each EcoRV site.

(A) Raw values of resection with the $TALEN_{CTG}$. Statistical comparisons between the wild-type and each mutant strain were determined by two-tailed t tests and are shown above each graph (*p < 0.05, **p < 0.01). Diamonds show significant differences between $TALEN_{CTG}$ and $TALEN_{noCTG}$ in *rad50*Δ strains at both distal sites at 34 hr (t test, p = 0.0197, comparison with C).

(B) Relative values of resection with the $TALEN_{CTG}$. Same as (A), except that resection values were divided by the amount of DSBs detected on Southern blots. Statistical comparisons between the wild-type, *rad50*Δ, and *sae2*Δ were determined by two-tailed t tests and are shown by vertical gray lines along with corresponding p values. Because DSB signals were not detectable by Southern blots in *rad52*Δ, relative resection values could not be calculated.

(C) Raw values of resection with the $TALEN_{noCTG}$. Diamonds show significant differences between $TALEN_{CTG}$ and $TALEN_{noCTG}$ in *rad50*Δ strains at both distal sites at 34 hr (t test, p = 0.0197, comparison with A). Proximal sites do not show any significant difference.

(D) Relative values of resection with the $TALEN_{noCTG}$.

Raw and resection values are the average of two to four independent experiments for each strain and each time point, except in the wild-type strain with the $TALEN_{noCTG}$ (only one experiment), but in this strain, 100% of broken molecules were resected at each EcoRV site. Error bars are equal to one SD.

In haploid wild-type cells that are unable to repair the DSB by homologous recombination, 76.5% ± 4.3% of cells survived. This frequency slightly decreased in the *dnl4*Δ mutant (57.3% ± 16%) but was not significantly different from the wild-type (t test, p = 0.06). A significant decrease in survival was observed in haploid *sae2*Δ cells (20.5% ± %4.3%), proving that this gene was also needed to repair such a break.

DSB Repair within CTG Repeat Tracts Is Mainly an Intramolecular Mechanism

Repeat lengths were analyzed in several surviving colonies after $TALEN_{CTG}$ induction in wild-type and mutant strains by two different techniques. First, Southern blots were run on yeast colonies for each strain to determine the overall range of allele contractions (Figure 6A). Then a subset of these colonies was PCR-amplified at the repeat locus and Sanger-sequenced. When both alleles carried repeat tracts of the same exact length, the sequence was very clear before and after the repeat tract, as previously demonstrated (Richard et al., 2014). On the contrary, a sequence becoming fuzzy after the repeat tract was the signature of two alleles of different lengths. Sequences were therefore classified in two categories: homozygous (when both repeat tract alleles shared the same length) or heterozygous (different lengths). Distribution of repeat lengths in wild-type survivors showed that a large majority of clones (86.1%) carried contracted repeats with less than 20 CTG triplets, most of them (51.6%) exhibiting very large contractions (only 4–10 CTG triplets left; Figure 6B). Among survivors, only 11.1% were homozygous, all of them exhibiting large contractions. These homozygous survivors may correspond to a minority of cells that have repaired the DSB by gene conversion using an already contracted repeat tract as a template or, alternatively, to cells that independently repaired both alleles to the same length by chance.

Distribution of repeat lengths observed in *pol32*Δ and *rad51*Δ mutants was not statistically different from the wild-type (homogeneity chi-square test = 5.10 and 5.58, respectively) (Figure 6B). In *rad51*Δ, no homozygous clone was found, consistent with the hypothesis that the few homozygous clones observed in wild-type cells indeed corresponded to gene conversion events. Distributions of repeat lengths in *rad50*Δ and *rad52*Δ were significantly different from the wild-type (chi-square test = 31.33 and 20.08, respectively). In both strains, all survivors were heterozygous, and a large majority of them harbored short repeat lengths (20–25 triplets in *rad50*Δ, 4–10 triplets in *rad52*Δ). The high mortality in these two mutant backgrounds suggests that surviving clones had spontaneously contracted CTG repeat tracts below the minimal length required for TALEN nuclease activity (less than 17 triplets; Richard et al., 2014) and, therefore, did not receive any DSB. Alternatively, survivors may correspond to a subset of cells that were able to repair the break in the absence of Rad50 or Rad52, both hypotheses being not mutually exclusive. The few homozygous events detected in *rad52*Δ (5.7%) probably correspond to cells that independently contracted both alleles to the same length by chance. There is no statistical difference between repeat tract length distribution in *dnl4*Δ or *sae2*Δ strains compared with the wild-type, most survivors being contracted to 4–10 CTG triplets (Figure 6C). However, we must note that only 18 of 48 *sae2*Δ survivors showed a clear sequencing product, suggesting that the length distribution observed might represent only a subset of all repair events obtained in this mutant background.

All of these results showed that repair of a DSB induced in CTG repeats involved *RAD50*, *SAE2*, and *RAD52*. We ruled out that this repair could occur by BIR or gene conversion because neither *POL32* nor *RAD51* was involved. Instead, SSA, which is a non-conservative intrachromosomal recombination mechanism depending on *RAD52*, appeared to be the favored pathway to repair the break. Recombination may occur between CTG repeats present on both sides of the break, eventually resulting in large repeat contractions. Two arguments support iterative cycles of repeat contractions. First, progressive contractions of repeat tracts occurred between time points during TALEN induction (Figure 7A). Second, the reduction of smear length was clearly visible on Southern blots over the duration of a time course (Figures 2 and 7B).

DISCUSSION

Integrity of Sae2 and of the Mre11-Rad50 Complex Is Essential for DSB Processing within CTG Repeats

Former studies showed that Mre11 was not required to process "clean ends" such as those resulting from multiple HO DSBs in yeast (Llorente and Symington, 2004), although resection and single-strand DNA formation were delayed when the Mre11-Rad50 complex was not functional (Lee et al., 1998; Sugawara and Haber, 1992). Here we show that, when a DSB was induced in a *rad50*Δ strain in non-repeated DNA ($TALEN_{noCTG}$), resection and repair were delayed, but survival was not significantly decreased, confirming that the Mre11-Rad50 complex was not essential to resect clean ends (Figure 4C).

On the contrary, resection and repair of a TALEN-induced DSB within a CTG trinucleotide repeat was completely abolished in a *rad50*Δ strain (Figures 4A and 4B). This result is consistent with a former work in which a natural chromosomal break within a long CTG repeat tract in yeast was left unrepaired and accumulated in a *rad50*Δ strain in such proportions that it was possible to detect the broken chromosome by pulse-field gel electrophoresis (Freudenreich et al., 1998). It is also compatible with a recent study using *Xenopus* egg extracts in which Liao et al. (2016) showed that Mre11 was essential for resection of DNA with 3′ damaged nucleotides and 3′ or 5′ bulky adducts.

Sae2 as well as the Mre11-Rad50 complex were previously shown to be required to resolve hairpin-capped natural DSBs in yeast cells (Lobachev et al., 2002). Consistent with this study, the purified Sae2 protein was shown to exhibit endonuclease activity on DNA gaps close to a hairpin structure, and this activity was stimulated by the Mre11-Rad50 complex (Lengsfeld et al., 2007). Later on, Sae2 was shown to be involved in resection at the *MAT* locus following HO DSB only in the absence of the single-strand binding protein Rfa1. In that particular case, Sae2 was required to remove hairpin-like folded back structures at DSB ends (Chen et al., 2013). In our present experiments, a clear absence of resection was observed in the *sae2*Δ mutant on the 5′ DSB end that contained most of the repeat tract

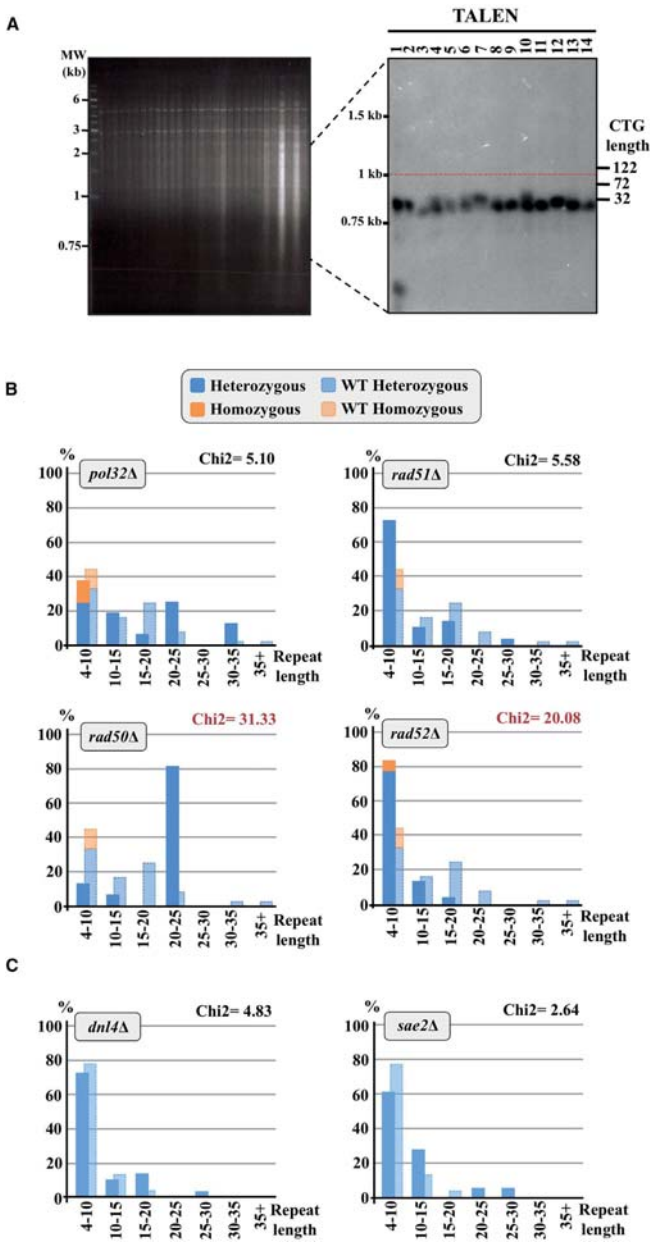


Figure 6. Molecular Analysis of CTG Repeat Length after DSB Repair

(A) Representative Southern blot showing 14 haploid colonies in which the $TALEN_{CTG}$ was induced. In each lane, total genomic DNA was extracted from one single yeast colony and analyzed. Two molecular weight markers were used: the GeneRuler 1-kb ladder (Sigma-Aldrich) on the left and a homemade CTG repeat tract length on the right (Viterbo et al., 2016). The red dotted line shows the parental CTG repeat tract length. In colonies after $TALEN_{CTG}$ induction, one or more bands containing contracted CTG repeat tracts were detected.

(B) Molecular analysis of cells after $TALEN_{CTG}$ induction. DNA was extracted from colonies after DSB induction, and the repeat containing-locus was sequenced. After Sanger sequencing of the PCR product, two outcomes could be obtained. when the two alleles contained the same exact number of triplets, one unique sequence was clearly read (homozygous, in orange); when the two alleles contained different numbers of triplets, the sequence was blurry and unreadable after the shortest of the two repeat tracts (heterozygous, in blue). Repeat lengths are given in number of triplets. The number of colonies sequenced in each strain was as follows: WT, 36; *pol32*Δ, 16; *rad50*Δ, 31; *rad51*Δ, 29; *rad52*Δ, 53. Chi-square test values (degrees of freedom [ddl] = 3) of comparisons between wild-type and mutant distributions are indicated above each graph. Only two distributions (in red) are significantly different from the wild-type: *rad50*Δ and *rad52*Δ.

(C) The same as (B), except that repeat tract lengths were compared between the wild-type strain and *dnl4*Δ or *sae2*Δ haploid strains, so only one trinucleotide repeat allele was present. The number of colonies sequenced in each strain was as follows: WT, 32; *dnl4*Δ, 40; *sae2*Δ, 18. chi-square test values of comparisons between wild-type and mutant distributions are indicated above each graph (ddl = 2).

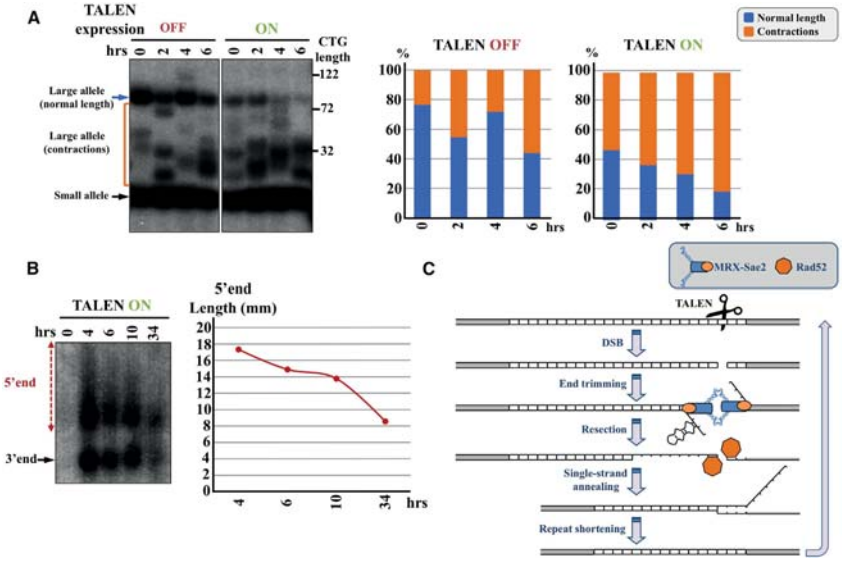


Figure 7. A Model of Progressive CTG Repeat Contractions following Iterative DSB Repair

(A) Diploid yeast cells were collected at several time points during TALEN induction and plated on non-inducible medium so that the nuclease was turned off during colony growth. From 20 to 30 colonies were picked and pooled before DNA extraction and analysis. Repeat tract lengths were analyzed by Southern blot. In each lane, three types of signals were detected: the large uncontracted allele (normal length), large contracted alleles, and the small allele. Note that the small allele is too small to be efficiently cut and contracted by the TALEN in this experiment. Graphs show quantifications of large allele signals, when the TALEN was repressed or expressed during the time course.

(B) Total smear length was measured in a wild-type time course (left), showing length reduction over time (right).

(C) Mechanistic model for CTG repeat contractions following a TALEN-induced DSB. After DSB formation, the ends of the break require the integrity of the MRX complex and Sae2 to be trimmed. Following resection, Rad52 binds to DSB ends and catalyzes SSA between the two ends of the DSB. With the DSB occurring very close to the repeat tract end, only a few triplet repeats may be involved in SSA, leading to a moderate repeat length shortening. The shortened trinucleotide repeats may be the substrate of one or more other round(s) of breakage and SSA until the repeat tract is too short for the TALEN to bind and induce a DSB.

(Figure 4B). On the 3′ end, though, resection was reduced at the proximal site compared with the wild-type but was higher than in the *rad50*Δ mutant. It is possible that the few triplet repeats left at the 3′ end after DSB induction are sufficient to form a small hairpin that decreases or delays resection. This delay might also explain why resection at the distal EcoRV site is decreased on the 3′ end of the DSB. We concluded that DSB ends of CTG trinucleotide repeats most probably harbor some kind of secondary structures, are therefore not clean, and absolutely require a functional Mre11-Rad50 complex as well as Sae2 to be processed for repair to start.

On the opposite, resection was increased at longer distances in the *rad52*Δ mutant (Figure 4A). This suggests that binding of the Rad52 protein on early recombination intermediates interferes with DNA end resection, as already shown in former publications (Van Dyck et al., 1999; Frank-Vaillant and Marcand, 2002; Parsons et al., 2000; Ristic et al., 2003; Sugawara and Haber, 1992; White and Haber, 1990). It is at the present time unclear how competition occurs and is resolved between Rad52p and resection proteins.

Differences between Yeast and Mammalian Systems

Former works showed the role of the Mre11-Rad50 complex on natural CTG trinucleotide repeat expansions in yeast (Sundararajan et al., 2010; Ye et al., 2016). When an I-*Sce* I DSB was repaired using a long CTG trinucleotide repeat as a template, expansions occurring during DSB repair were larger in strains overexpressing *MRE11* or *RAD50* (Richard et al., 2000). In the present experiments, DSB induction in long CTG repeats only led to contractions of the repeat tract, in wild-type as well as in *rad50*Δ strains (Figure 6B), and no expansion was ever observed. Cinesi et al. (2016) reported some expansions when inducing a DSB within a CTG repeat tract in human cells using either wild-type Cas9 or the mutant Cas9D10A nickase. This suggests that DSB repair mechanisms within CTG repeats exhibit subtle differences between yeast and human cells. The chromatin environment in human cells is different and may affect the way a DSB within a CTG repeat tract will be repaired (reviewed in House et al., 2014). In addition, although human cells contain *RAD51* and *RAD52* homologs, two additional genes, *BRCA1* and *BRCA2*, involved in breast cancer, play a central

role in homologous recombination, whereas yeast cells lack these genes (Moynahan et al., 1999, 2001). Comparing results obtained in yeast and in human cells will also hopefully help our understanding of the respective roles of these factors during DSB repair of CTG repeats.

Single-Strand Annealing Is the Main Mechanism of DSB Repair within a CTG Repeat Tract

Former studies of SSA requirements on direct repeats showed that its efficacy relied on three factors: homology length between the two repeated sequences, resection rate, and proximity on the DNA molecule, with closer sequences recombining more easily than distant ones (Lazzaro et al., 2008; Sugawara and Haber, 1992). In addition, *RAD52* was shown to be important for SSA reaction between 15- to 18-bp microhomologous sequences but strongly inhibited SSA between 6- to 13-bp microhomologies (Villarreal et al., 2012). In the present case, the DSB was made close to the 3′ end of the repeat tract, leaving only 1–4 repeat units (3–12 bp) on the 3′ end of the break but a much longer stretch of repeats on the 5′ end (around 70 triplets). SSA between triplet repeats was partially *RAD52*-dependent (survival was 3-fold decreased), suggesting that 15-bp microhomologies were sometimes present and used for SSA between triplet repeats. These results are in good accordance with our former work in which repair occurred in 67% of the cases by annealing between two short repeats flanking an I-*Sce* I restriction site (Richard et al., 1999).

Although no effect of *POL32* and *RAD51* on DSB repair of a CTG repeat tract was detected in the present experiments, it is interesting to mention that both genes were involved in spontaneous expansions of CAG repeats, probably by a BIR-related mechanism. However, expansions rates were low (10^{-5}–10^{-6} per cell per division), and the authors could not test the possible role of these two genes in repeat contractions in their experimental system (Kim et al., 2017).

A Model Supporting Progressive Repeat Contractions Associated with TALEN-Induced DSB Repair

We propose a model involving iterative SSA between short repeat-containing DNA ends after DSB induction (Figure 7C). In this model, the Mre11-Rad50 complex and Sae2 are essential to process DSB ends, after which Rad52 annealing activity catalyzes the SSA reaction. Given that the DSB occurs only a few triplets before the end of the repeat tract, homology available to anneal both ends is very small. This ''short SSA'' can hardly lead to large repeat contractions in one single step. We thus propose that, following repair of this first DSB, the repeat tract may still be a substrate for the nuclease and could receive a second DSB, leading to further contraction of the repeat, and so on, until it is too small to be efficiently cut by the $TALEN_{CTG}$. Progressive contractions of repeat tracts (Figure 7A) as well as reduction of smear length between time points (Figure 7B) support this model. However, we cannot completely exclude near-complete contraction of repeat tracts in a subpopulation of cells receiving a DSB, given that TALEN efficacy is very low (~10% broken molecules at each time point). Hence, this apparent progressive repeat contraction could indeed represent complete contraction in this subpopulation at each time point. It is unclear at the present time why the TALEN is so inefficient at inducing a DSB compared with known meganucleases such as HO or I-*Sce* I expressed in yeast cells. Further experiments are needed to address this specific question.

EXPERIMENTAL PROCEDURES

Yeast Strains and Plasmids

All mutant strains were built from strains GFY6162-14A and GY6162-3D by classical gene replacement method (Orr-Weaver et al., 1981) using KANMX4 as a marker (Table S1), amplified from the European *Saccharomyces cerevisiae* archive for functional analysis (EUROSCARF) deletion library, using primers located 1 kb upstream and downstream of the cassette (Table S2). The VMY350 strain was used to construct the VMY650 strain by mating-type switching, using the pJH132 vector carrying an inducible HO endonuclease (Holmes and Haber, 1999). Plasmid pCLS9996 carrying the $TALEN_{CTG}$ right arm was digested by NcoI (New England Biolabs) and EagI (Takara). The fragment containing the right arm was cloned in the centromeric pCMha182, digested by BamHI (NEB) and PstI (NEB) using two oligomeric adaptors of 16 bp (BamHI-NcoI) and 19 bp (EagI-PstI). The resulting vector, pCMha182KN9996, was transformed in the haploid strain GFY6162-14A and its mutant derivatives. Plasmid pCLS16715 carrying the $TALEN_{CTG}$ left arm was digested by NcoI (NEB) and EagI (Takara). The fragment containing the left arm was cloned in the centromeric pCMha188, digested by BamHI (NEB) and PstI (NEB) using two oligomeric adaptors of 16 bp (BamHI-NcoI) and 19 bp (EagI-PstI). The resulting vector, pCMha188KN16715, was transformed in the GFY6161-3D haploid strain and its mutant derivatives. Haploid transformants were crossed on rich medium (yeast extract peptone dextrose medium [YPD]) supplemented with doxycycline (10 μg/mL), and diploids containing both $TALEN_{CTG}$ arms were selected on synthetic complete medium lacking uracil and tryptophan (SC-Ura-Trp) with doxycycline (10 μg/mL). For the *dnl4*Δ mutant, both $TALEN_{CTG}$ arms were transformed in haploid strains GFY6162-3D and VMY104 because NHEJ is downregulated in diploid cells (Frank-Vaillant and Marcand, 2001; Valencia et al., 2001).

The $TALEN_{noCTG}$ was designed by the Muséum National d'Histoire Naturelle platform. The target sequence was chosen to be an I-*Sce* I recognition site integrated in the *SUP4* gene (Richard et al., 1999). Plasmid pR1 was used as a template to PCR-amplify the $TALEN_{noCTG}$ right arm using primer pairs VMS25/VMAS25, containing a 50-bp tail homologous to sequences flanking a KpnI site on pCMha182KN. The PCR product and pCMha182KN linearized by KpnI (NEB) were directly cloned in yeast cells. Plasmid pL1 was used as a template to PCR-amplify the $TALEN_{noCTG}$ left arm using primer pairs VMS25/VMAS25, containing a 50-bp tail homologous to sequences flanking a KpnI site on pCMha188KN. The PCR product and pCMha188KN linearized by KpnI (NEB) were also directly cloned in yeast. Centromeric vectors pCMha182KNR1 and pCMha188KNL1, carrying respectively, the $TALEN_{noCTG}$ right arm and the $TALEN_{noCTG}$ left arm, were transformed in diploid strains VMY001 and VMY002.

TALEN Inductions

Before nuclease induction, Southern blot analyses were conducted on several independent subclones to select cells with different CTG repeat tract lengths on both chromosomes. Yeast cells were grown at 30°C in liquid SC-Ura-Trp medium complemented with 10 μg/ml of doxycycline. Cells were washed with sterile water to eliminate doxycycline and then split in two cultures at 9×10^6 cells/mL, one in SC-Ura-Trp medium complemented with 10 μg/mL of doxycycline (TALEN-repressed) and the other in SC-Ura-Trp (TALEN-induced). For each time point, 2×10^8 cells were collected at time point (T) = 0, 14, 16, 18, 20, 22, 24, or 48 hr afterward, rapidly centrifuged, washed with water, and frozen in dry ice before DNA extraction. To determine viability after DSB induction, cells were plated at 24 hr on SC-Ura-Trp plates supplemented with doxycycline (10 μg/mL) for the TALEN-repressed culture and on SC-Ura-Trp plates for the TALEN-induced culture. CFUs were counted after 3–5 days of growth at 30°C.

DSB Analysis and Quantification

Total genomic DNA (4 μg) of cells collected at each time point was digested for 6 hr by EcoRV (20 U) (NEB) and analyzed by Southern blot as described previously (Viterbo et al., 2016). Alternatively, a terminal transferase-mediated PCR assay (Förstemann et al., 2000) was used to amplify the TALEN-induced DSB. Genomic DNA (100 ng) of cells collected at different time points was heat-denatured and treated with 7 U of terminal deoxynucleotidyl transferase (Takara) in a volume of 10 μL (100 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid [HEPES], pH 7.2; 40 mM $MgCl_2$; 0.5 mM DTT; 0.1% BSA; and 1 mM deoxycytidine triphosphate [dCTP]) for 30 min at 37°C to add polyC tails to 3′OH free ends. The enzyme was inactivated for 10 min at 65°C and 5 min at 94°C. Then 30 μL of PCR mix was added to each reaction to obtain a final volume of 40 μL containing 67 mM Tris HCl (pH 8.8), 16 mM $(NH_4)_2SO_4$, 5% glycerol, 0.01% Tween, 200 μM each deoxynucleotide triphosphate (dNTP), and 40 nM of each primer (G18 and VMS14). The following PCR program was used: 94°C for 2 min (94°C for 20 s, 62°C for 12 s, and 72°C for 20 s) for 45 cycles and then 72°C for 5 min. For each reaction, 20 μL was loaded on a 1% analytical agarose gel, and 20 μL was sent for Sanger sequencing.

Trinucleotide Repeat Length Analysis

Several colonies from each induced or repressed plates were picked, total genomic DNA was extracted, and 4 μg was digested for 6 hr by SspI (20 U) (NEB) and analyzed by Southern blot as described previously (Viterbo et al., 2016). Repeat tracts in some of the survivors were subsequently amplified using primers su3 and su7 and sequenced using a third internal primer, su2 (Table S2). Sanger sequencing was performed by GATC Biotech.

Analysis of DSB End Resection

A real-time PCR assay using primer pairs flanking EcoRV sites 0.81 kb and 2.94 kb away from the 3′ end of the CTG repeat tract (VMS20/VMAS20 and VMS21/VMAS21, respectively) and 0.88 kb and 1.88 kb away from the 5′ end of the CTG repeat tract (VMS22/VMAS22 and VMS23/VMAS23, respectively) was used to quantify end resection. Another pair of primers was used to amplify a region of chromosome X near the *ARG2* gene to serve as an internal control of the DNA amount (JEM1f-JEM1r). Genomic DNA of cells collected at T = 0 hr, T = 18 hr, T = 24 hr, and T = 48 hr was split in two fractions; one was used for EcoRV digestion and the other one for a mock digestion in a final volume of 15 μL. Samples were incubated for 5 hr at 37°C and then the enzyme was inactivated for 20 min at 80°C. DNA was subsequently diluted by adding 55 μL of ice-cold water, and 4 μL was used for each real-time PCR reaction in a final volume of 25 μL. PCRs were performed with Absolute SYBR Green Fluorescein Mix (Thermo Scientific) in a Mastercycler S Realplex (Eppendorf) using the following program: 95°C for 15 min, 95°C for 15 s, 55°C for 30 s, and 72°C for 30 s repeated 40 times, followed by a 20-min melting curve. Reactions were performed in triplicate, and the mean value was used to determine the amount of resected DNA using the following formula: raw resection = $2/(1+2^{\Delta Ct})$ with $\Delta Ct = C_{t,EcoRV} - C_{t,mock}$. Relative resection values were calculated by dividing raw resection values by the percentage of DSB quantified at the corresponding time point. All t tests were performed using the R package (Millot, 2011).

Western Blots

Liquid cultures were grown to exponential phase in the presence or absence of 10 μg/mL doxycycline. Proteins were extracted on 2 × 10^8 cells in 200 μL Laemmli solution with 100 μL glass beads. Proteins were separated on a 10% acrylamide gel under standard conditions and blotted to a nitrocellulose membrane (Optitran BA-S 83 reinforced NC, Schleicher & Schuell). For TALEN detection, a polyclonal anti-HA antibody was used (ab9110, Abcam, 0.25 μg/mL final concentration). For Msh2 detection, the primary antibody was a polyclonal rabbit antibody directed against an internal part of the yeast Msh2 protein (N3C2, GeneTex, 1 μg/mL final concentration) (Viterbo et al., 2016). A secondary goat anti-rabbit antibody conjugated to horseradish peroxidase was used for detection in both cases (Thermo Scientific, 0.16 μg/mL final concentration). Quantification was performed using a ChemiDoc MP Imager (Bio-Rad) with the dedicated Image Lab software. The molecular weight marker used was the Precision Plus Protein Standards All Blue (Bio-Rad).

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures and two tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.01.083.

ACKNOWLEDGMENTS

V.M. was the recipient of a graduate student award from the Fondation pour la Recherche Médicale (PLP20131028794) and was supported by Fondation Guy Nicolas and Fondation Hardy. L.P. is the recipient of a graduate student CIFRE fellowship from SANOFI. This project was supported by the ValoExpress program of the Institut Pasteur (project TRINUCONTRACT).

AUTHOR CONTRIBUTIONS

V.M. designed and performed most of the TALEN experiments and all resection experiments, analyzed the data, and wrote the manuscript. L.P. performed some TALEN experiments, analyzed the data, and contributed to the manuscript. D.V. contributed to building mutant strains. M.C. designed and built the $TALEN_{noCTG}$. G.-F.R. designed the project, analyzed the data, and wrote the manuscript.

DECLARATION OF INTERESTS

There is a patent related to this work, held by the Institut Pasteur (WO 2015/078935 A1).

Received: February 27, 2017
Revised: December 19, 2017
Accepted: January 25, 2018
Published: February 20, 2018

REFERENCES

An, M.C., Zhang, N., Scott, G., Montoro, D., Wittkop, T., Mooney, S., Melov, S., and Ellerby, L.M. (2012). Genetic correction of Huntington's disease phenotypes in induced pluripotent stem cells. Cell Stem Cell *11*, 253–263.

Borde, V., Lin, W., Novikov, E., Petrini, J.H., Lichten, M., and Nicolas, A. (2004). Association of Mre11p with double-strand break sites during yeast meiosis. Mol. Cell *13*, 389–401.

Chen, H., Lisby, M., and Symington, L.S. (2013). RPA coordinates DNA end resection and prevents formation of DNA hairpins. Mol. Cell *50*, 589–600.

Cinesi, C., Aeschbach, L., Yang, B., and Dion, V. (2016). Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. Nat. Commun. *7*, 13272.

Davis, A.P., and Symington, L.S. (2004). RAD51-dependent break-induced replication in yeast. Mol. Cell. Biol. *24*, 2344–2351.

Fairhead, C., and Dujon, B. (1993). Consequences of unique double-stranded breaks in yeast chromosomes: death or homozygosis. Mol. Gen. Genet. *240*, 170–178.

Foiry, L., Dong, L., Savouret, C., Hubert, L., te Riele, H., Junien, C., and Gourdon, G. (2006). Msh3 is a limiting factor in the formation of intergenerational CTG expansions in DM1 transgenic mice. Hum. Genet. *119*, 520–526.

Förstemann, K., Höss, M., and Lingner, J. (2000). Telomerase-dependent repeat divergence at the 3′ ends of yeast telomeres. Nucleic Acids Res. *28*, 2690–2694.

Frank-Vaillant, M., and Marcand, S. (2001). NHEJ regulation by mating type is exercised through a novel protein, Lif2p, essential to the ligase IV pathway. Genes Dev. *15*, 3005–3012.

Frank-Vaillant, M., and Marcand, S. (2002). Transient stability of DNA ends allows nonhomologous end joining to precede homologous recombination. Mol. Cell *10*, 1189–1199.

Freudenreich, C.H., Kantrow, S.M., and Zakian, V.A. (1998). Expansion and length-dependent fragility of CTG repeats in yeast. Science *279*, 853–856.

Gacy, A.M., Goellner, G., Juranić, N., Macura, S., and McMurray, C.T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell *81*, 533–540.

Haber, J.E. (1995). In vivo biochemistry: physical monitoring of recombination induced by site-specific endonucleases. BioEssays *17*, 609–620.

Holmes, A.M., and Haber, J.E. (1999). Double-strand break repair in yeast requires both leading and lagging strand DNA polymerases. Cell *96*, 415–424.

House, N.C.M., Koch, M.R., and Freudenreich, C.H. (2014). Chromatin modifications and DNA repair: beyond double-strand breaks. Front. Genet. *5*, 296.

Kim, J.C., Harris, S.T., Dinter, T., Shah, K.A., and Mirkin, S.M. (2017). The role of break-induced replication in large-scale expansions of (CAG)n/(CTG)n repeats. Nat. Struct. Mol. Biol. *24*, 55–60.

Krogh, B.O., and Symington, L.S. (2004). Recombination proteins in yeast. Annu. Rev. Genet. *38*, 233–271.

Lazzaro, F., Sapountzi, V., Granata, M., Pellicioli, A., Vaze, M., Haber, J.E., Plevani, P., Lydall, D., and Muzi-Falconi, M. (2008). Histone methyltransferase Dot1 and Rad9 inhibit single-stranded DNA accumulation at DSBs and uncapped telomeres. EMBO J. *27*, 1502–1512.

Lee, S.-E., Moore, J.K., Holmes, A., Umezu, K., Kolodner, R.D., and Haber, J.E. (1998). Saccharomyces Ku70, mre11/rad50 and RPA proteins regulate adaptation to G2/M arrest after DNA damage. Cell *94*, 399–409.

Lengsfeld, B.M., Rattray, A.J., Bhaskara, V., Ghirlando, R., and Paull, T.T. (2007). Sae2 is an endonuclease that processes hairpin DNA cooperatively with the Mre11/Rad50/Xrs2 complex. Mol. Cell *28*, 638–651.

Liao, S., Tammaro, M., and Yan, H. (2016). The structure of ends determines the pathway choice and Mre11 nuclease dependency of DNA double-strand break repair. Nucleic Acids Res. *44*, 5689–5701.

Liu, G., Chen, X., Bissler, J.J., Sinden, R.R., and Leffak, M. (2010). Replication-dependent instability at (CTG) x (CAG) repeat hairpins in human cells. Nat. Chem. Biol. *6*, 652–659.

Llorente, B., and Symington, L.S. (2004). The Mre11 nuclease is not required for 5′ to 3′ resection at multiple HO-induced double-strand breaks. Mol. Cell. Biol. *24*, 9682–9694.

Lobachev, K.S., Gordenin, D.A., and Resnick, M.A. (2002). The Mre11 complex is required for repair of hairpin-capped double-strand breaks and prevention of chromosome rearrangements. Cell *108*, 183–193.

Lydeard, J.R., Jain, S., Yamaguchi, M., and Haber, J.E. (2007). Break-induced replication and telomerase-independent telomere maintenance require Pol32. Nature *448*, 820–823.

Manley, K., Shirley, T.L., Flaherty, L., and Messer, A. (1999). Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat. Genet. *23*, 471–473.

McMurray, C.T. (1999). DNA secondary structure: a common and causative factor for expansion in human disease. Proc. Natl. Acad. Sci. USA *96*, 1823–1825.

Millot, G. (2011). Comprendre et réaliser les tests statistiques à l'aide de R (de boeck).

Mimitou, E.P., and Symington, L.S. (2008). Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature *455*, 770–774.

Mittelman, D., Moye, C., Morton, J., Sykoudis, K., Lin, Y., Carroll, D., and Wilson, J.H. (2009). Zinc-finger directed double-strand breaks within CAG repeat tracts promote repeat instability in human cells. Proc. Natl. Acad. Sci. USA *106*, 9607–9612.

Moynahan, M.E., Chiu, J.W., Koller, B.H., and Jasin, M. (1999). Brca1 controls homology-directed DNA repair. Mol. Cell *4*, 511–518.

Moynahan, M.E., Pierce, A.J., and Jasin, M. (2001). BRCA2 is required for homology-directed repair of chromosomal breaks. Mol. Cell *7*, 263–272.

O'Hoy, K.L., Tsilfidis, C., Mahadevan, M.S., Neville, C.E., Barceló, J., Hunter, A.G., and Korneluk, R.G. (1993). Reduction in size of the myotonic dystrophy trinucleotide repeat mutation during transmission. Science *259*, 809–812.

Orr, H.T., and Zoghbi, H.Y. (2007). Trinucleotide repeat disorders. Annu. Rev. Neurosci. *30*, 575–621.

Orr-Weaver, T.L., Szostak, J.W., and Rothstein, R.J. (1981). Yeast transformation: a model system for the study of recombination. Proc. Natl. Acad. Sci. USA *78*, 6354–6358.

Owen, B.A., Yang, Z., Lai, M., Gajec, M., Badger, J.D., 2nd, Hayes, J.J., Edelmann, W., Kucherlapati, R., Wilson, T.M., and McMurray, C.T. (2005). (CAG)(n)-hairpin DNA binds to Msh2-Msh3 and changes properties of mismatch recognition. Nat. Struct. Mol. Biol. *12*, 663–670.

Park, C.-Y., Halevy, T., Lee, D.R., Sung, J.J., Lee, J.S., Yanuka, O., Benvenisty, N., and Kim, D.-W. (2015). Reversion of FMR1 Methylation and Silencing by Editing the Triplet Repeats in Fragile X iPSC-Derived Neurons. Cell Rep. *13*, 234–241.

Parsons, C.A., Baumann, P., Van Dyck, E., and West, S.C. (2000). Precise binding of single-stranded DNA termini by human RAD52 protein. EMBO J. *19*, 4175–4181.

Pearson, C.E., Ewel, A., Acharya, S., Fishel, R.A., and Sinden, R.R. (1997). Human MSH2 binds to trinucleotide repeat DNA structures associated with neurodegenerative diseases. Hum. Mol. Genet. *6*, 1117–1123.

Pinto, R.M., Dragileva, E., Kirby, A., Lloret, A., Lopez, E., St Claire, J., Panigrahi, G.B., Hou, C., Holloway, K., Gillis, T., et al. (2013). Mismatch repair genes Mlh1 and Mlh3 modify CAG instability in Huntington's disease mice: genome-wide and candidate approaches. PLoS Genet. *9*, e1003930.

Plessis, A., Perrin, A., Haber, J.E., and Dujon, B. (1992). Site-specific recombination determined by I-SceI, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. Genetics *130*, 451–460.

Richard, G.F. (2015). Shortening trinucleotide repeats using highly specific endonucleases: a possible approach to gene therapy? Trends Genet. *31*, 177–186.

Richard, G.-F., Dujon, B., and Haber, J.E. (1999). Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. Mol. Gen. Genet. *261*, 871–882.

Richard, G.-F., Goellner, G.M., McMurray, C.T., and Haber, J.E. (2000). Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11-RAD50-XRS2 complex. EMBO J. *19*, 2381–2390.

Richard, G.-F., Cyncynatus, C., and Dujon, B. (2003). Contractions and expansions of CAG/CTG trinucleotide repeats occur during ectopic gene conversion in yeast, by a MUS81-independent mechanism. J. Mol. Biol. *326*, 769–782.

Richard, G.F., Viterbo, D., Khanna, V., Mosbach, V., Castelain, L., and Dujon, B. (2014). Highly specific contractions of a single CAG/CTG trinucleotide repeat by TALEN in yeast. PLoS ONE *9*, e95611.

Ristic, D., Modesti, M., Kanaar, R., and Wyman, C. (2003). Rad52 and Ku bind to different DNA structures produced early in double-strand break repair. Nucleic Acids Res. *31*, 5229–5237.

Santillan, B.A., Moye, C., Mittelman, D., and Wilson, J.H. (2014). GFP-based fluorescence assay for CAG repeat instability in cultured human cells. PLoS ONE *9*, e113952.

Savouret, C., Garcia-Cordier, C., Megret, J., te Riele, H., Junien, C., and Gourdon, G. (2004). MSH2-dependent germinal CTG repeat expansions are produced continuously in spermatogonia from DM1 transgenic mice. Mol. Cell. Biol. *24*, 629–637.

Shinohara, A., Ogawa, H., and Ogawa, T. (1992). Rad51 protein involved in repair and recombination in S. cerevisiae is a RecA-like protein. Cell *69*, 457–470.

Slean, M.M., Panigrahi, G.B., Castel, A.L., Tomkinson, A.E., and Pearson, C.E. (2016). Absence of MutSβ leads to the formation of slipped-DNA for CTG/CAG contractions at primate replication forks. DNA Repair *42*, 107–118.

Sugawara, N., and Haber, J.E. (1992). Characterization of double-strand break-induced recombination: homology requirements and single-stranded DNA formation. Mol. Cell. Biol. *12*, 563–575.

Sundararajan, R., Gellon, L., Zunder, R.M., and Freudenreich, C.H. (2010). Double-strand break repair pathways protect against CAG/CTG repeat expansions, contractions and repeat-mediated chromosomal fragility in Saccharomyces cerevisiae. Genetics *184*, 65–77.

Sung, P. (1994). Catalysis of ATP-dependent homologous DNA pairing and strand exchange by yeast RAD51 protein. Science *265*, 1241–1243.

Tian, L., Hou, C., Tian, K., Holcomb, N.C., Gu, L., and Li, G.-M. (2009). Mismatch recognition protein MutSbeta does not hijack (CAG)n hairpin repair in vitro. J. Biol. Chem. *284*, 20452–20456.

Tomé, S., Holt, I., Edelmann, W., Morris, G.E., Munnich, A., Pearson, C.E., and Gourdon, G. (2009). MSH2 ATPase domain mutation affects CTG*CAG repeat instability in transgenic mice. PLoS Genet. *5*, e1000482.

Tomé, S., Manley, K., Simard, J.P., Clark, G.W., Slean, M.M., Swami, M., Shelbourne, P.F., Tillier, E.R., Monckton, D.G., Messer, A., and Pearson, C.E. (2013). MSH3 polymorphisms and protein levels affect CAG repeat instability in Huntington's disease mice. PLoS Genet. *9*, e1003280.

Valencia, M., Bentele, M., Vaze, M.B., Herrmann, G., Kraus, E., Lee, S.-E., Schär, P., and Haber, J.E. (2001). NEJ1 controls non-homologous end joining in Saccharomyces cerevisiae. Nature *414*, 666–669.

van Agtmaal, E.L., André, L.M., Willemse, M., Cumming, S.A., van Kessel, I.D.G., van den Broek, W.J.A.A., Gourdon, G., Furling, D., Mouly, V., Monckton, D.G., et al. (2017). CRISPR/Cas9-Induced (CTG•CAG)n Repeat Instability in the Myotonic Dystrophy Type 1 Locus: Implications for Therapeutic Genome Editing. Mol. Ther. *25*, 24–43.

Van Dyck, E., Stasiak, A.Z., Stasiak, A., and West, S.C. (1999). Binding of double-strand breaks in DNA by human Rad52 protein. Nature *398*, 728–731.

Villarreal, D.D., Lee, K., Deem, A., Shim, E.Y., Malkova, A., and Lee, S.E. (2012). Microhomology directs diverse DNA break repair pathways and chromosomal translocations. PLoS Genet. *8*, e1003026.

Viterbo, D., Michoud, G., Mosbach, V., Dujon, B., and Richard, G.-F. (2016). Replication stalling and heteroduplex formation within CAG/CTG trinucleotide repeats by mismatch repair. DNA Repair (Amst.) *42*, 94–106.

White, C.I., and Haber, J.E. (1990). Intermediates of recombination during mating type switching in Saccharomyces cerevisiae. EMBO J. *9*, 663–673.

Williams, G.M., and Surtees, J.A. (2015). MSH3 Promotes Dynamic Behavior of Trinucleotide Repeat Tracts In Vivo. Genetics *200*, 737–754.

Wilson, T.E., Grawunder, U., and Lieber, M.R. (1997). Yeast DNA ligase IV mediates non-homologous DNA end joining. Nature *388*, 495–498.

Xie, N., Gong, H., Suhl, J.A., Chopra, P., Wang, T., and Warren, S.T. (2016). Reactivation of FMR1 by CRISPR/Cas9-Mediated Deletion of the Expanded CGG-Repeat of the Fragile X Chromosome. PLoS ONE *11*, e0165499.

Ye, Y., Kirkham-McCarthy, L., and Lahue, R.S. (2016). The Saccharomyces cerevisiae Mre11-Rad50-Xrs2 complex promotes trinucleotide repeat expansions independently of homologous recombination. DNA Repair (Amst.) *43*, 1–8.

Yu, A., Dill, J., and Mitas, M. (1995a). The purine-rich trinucleotide repeat sequences d(CAG)15 and d(GAC)15 form hairpins. Nucleic Acids Res. *23*, 4055–4057.

Yu, A., Dill, J., Wirth, S.S., Huang, G., Lee, V.H., Haworth, I.S., and Mitas, M. (1995b). The trinucleotide repeat sequence d(GTC)15 adopts a hairpin conformation. Nucleic Acids Res. *23*, 2706–2714.

Zhu, Z., Chung, W.H., Shim, E.Y., Lee, S.E., and Ira, G. (2008). Sgs1 helicase and two nucleases Dna2 and Exo1 resect DNA double-strand break ends. Cell *134*, 981–994.

Annex 4: Resection and repair of a Cas9 double-strand break at CTG trinucleotide repeats induces local and extensive chromosomal rearrangements

Article about the mechanism of the repair of CTG repeats after SpCas9 induction. I did the first experiments of SpCas9 inductions, and performed southern blots from figure 1B and figure 3A and C. The article was deposited on bioarchive (Mosbach et al., 2019b).

Resection and repair of a Cas9 double-strand break at CTG trinucleotide repeats induces local and extensive chromosomal rearrangements

Valentine Mosbach*[1, 5], David Viterbo*[1], Stéphane Descorps-Declère[1, 4], Lucie Poggi[1, 2, 3], Wilhelm Vaysse-Zinkhöfer[1, 2], and Guy-Franck Richard[1]

[1] Institut Pasteur, CNRS, UMR3525, 25 rue du Dr Roux, F-75015 Paris, France

[2] Sorbonne Universités, Collège doctoral, F-75005 Paris, France

[3] Biologics Research, Sanofi R&D, 13 Quai Jules Guesde, 94403 Vitry sur Seine, France

[4] Institut Pasteur, Center of Bioinformatics, Biostatistics and Integrative Biology (C3BI), F-75015 Paris, France

[5] Present address : Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), UMR7104 CNRS/Unistra, INSERM U1258, 1, rue Laurent Fries BP 10142, 67404 Illkirch, France

* These authors equally contributed to this work

Corresponding author: Guy-Franck Richard

Tel: +33-1-40-61-34-54

e-Mail: gfrichar@pasteur.fr

Summary

Microsatellites are short tandem repeats, ubiquitous in all eukaryotes and represent ~2% of the human genome. Among them, trinucleotide repeats are responsible for more than two dozen neurological and developmental disorders. Targeting microsatellites with dedicated DNA endonucleases could become a viable option for patients affected with dramatic neurodegenerative disorders. Here, we used the *Streptococcus pyogenes* Cas9 to induce a double-strand break within the expanded CTG repeat involved in myotonic dystrophy type 1, integrated in a yeast chromosome. Repair of this double-strand break generated unexpected large chromosomal rearrangements around the repeat tract. These rearrangements depended on *RAD52*, *DNL4* and *SAE2*, and both non-homologous end-joining and single-strand annealing pathways were involved. Resection and repair of the double-strand break (DSB) were totally abolished in a *rad50*Δ strain, whereas they were impaired in a *sae2*Δ mutant, only on the DSB end containing most of the repeat tract. This proved that Sae2 plays significant different roles in resecting a DSB end containing a repeated and structured sequence as compared to a non-repeated DSB end.

In addition, we also discovered that gene conversion was less efficient when the DSB could be repaired using a homologous template, suggesting that the trinucleotide repeat may interfer with gene conversion too. Altogether, these data show that *Sp*Cas9 is probably not a good choice when inducing a double-strand break at or near a microsatellite, especially in mammalian genomes that contain many more dispersed repeated elements than the yeast genome.

Keywords

Trinucleotide repeats, CRISPR-Cas9, double-strand break resection, chromosomal rearrangements, gene conversion

Introduction

Microsatellites are short tandem repeats ubiquitously found in all eukaryotic genomes sequenced so far (Richard et al., 2008). Altogether, they cover ~2% of the human genome, a figure similar to the whole protein-coding sequence (International Human Genome Sequencing Consortium, 2004). Naturally prone to frequent repeat length polymorphism, some microsatellites are also prone to large expansions that lead to human neurological or developmental disorders, such as trinucleotide repeats involved in Huntington disease, myotonic dystrophy type 1 (Steinert disease), fragile X syndrome or Friedreich ataxia (Orr and Zoghbi, 2007). These expansion-prone microsatellites share the common property to form secondary DNA structures *in vitro* (Gacy et al., 1995) and genetic evidences suggest that similar structures may also form *in vivo*, transiently stalling replication fork progression (Anand et al., 2012; Nguyen et al.; Pelletier et al., 2003; Samadashwily et al., 1997; Viterbo et al., 2016). Among those, CCG/CGG trinucleotide repeats are fragile sites in human cells, forming frequent double-strand breaks when the replication machinery is slowed down or impaired (Sutherland et al., 1998). Similarly, CAG/CTG and CCG/CGG microsatellites are also fragile sites in *Saccharomyces cerevisiae* cells (Balakumaran et al., 2000; Freudenreich et al., 1998). Therefore, microsatellite abundance and the natural fragility of some of them make these repeated sequences perfect targets to generate chromosomal rearrangements potentially leading to cancer.

Double-strand break (DSB) repair mechanisms have been studied for decades in model organisms as well as in human cells and led to the identification of the main genes involved in this process (reviewed in Krogh and Symington, 2004). Many of these advances were made possible by the use of highly specific DNA endonucleases, such as the meganucleases I-*Sce* I or HO (Fairhead and Dujon, 1993; Haber, 1995; Plessis et al., 1992). Other frequently used methods involved ionizing radiations making genome-wide DSBs (Nelms et al., 1998).

However, the fate of a single double-strand break within a repeated and structured DNA sequence has never been addressed, until recently. In a former work, we used a TALE Nuclease (TALEN) to induce a unique DSB into a long CTG trinucleotide repeat integrated into a *S. cerevisiae* chromosome. We showed that 100% of yeast cells in which the TALEN was expressed exhibited a large contraction of the repeat tract, going from an initial length of ~80 CTG triplets to less than 35. *POL32*, *DNL4* and *RAD51* were shown to play no detectable role in repairing this DSB. On the contrary, *RAD50*, *RAD52* and *SAE2* were required for proper repair of the DSB, and a functional Sae2 protein was found to be essential for efficient DSB resection, suggesting that repeat contraction occurred by a single-strand annealing (SSA) process, involving preliminary resection of the break, followed by annealing of the two DSB ends carrying the repeat tract (Mosbach et al., 2018; Richard et al., 2014).

In the present work, we used the *Streptococcus pyogenes* Cas9 endonuclease (*Sp*Cas9) to induce a DSB within the same long CTG trinucleotide repeat integrated in the yeast genome. The break was made at the 3' end of the repeat tract (Figure 1A), using a guide RNA that targets the repeat tract. Frequent rearrangements were found in surviving cells, with local deletions as well as more extensive ones involving recombination between retrotransposon LTRs. Survival and repair depended on *RAD50*, *RAD52*, *SAE2* and *DNL4* and double-strand break resection was abolished in *rad50*Δ and *sae2*Δ mutants. A more specific version of the nuclease, Enhanced *Sp*Cas9, generated the same rerrangements. In addition, we also discovered that gene conversion was less efficient when *Sp*Cas9 was used to induce a DSB within a CTG repeat tract that could be repaired with a homologous template, suggesting that the trinucleotide repeat may interfer with gene conversion too.

Results

A Cas9-induced double-strand break within CTG repeats induces cell death and chromosomal rearrangements around the repeat tract

In previous works, we showed that a TALEN targeted at the CTG trinucleotide repeat from the human *DMPK* gene 3' UTR integrated in a yeast chromosome, was extremely efficient at contracting the repeat tract below the pathological length (Mosbach et al., 2018; Richard et al., 2014). In order to determine whether the CRISPR-Cas9 system could be used in the same manner, a plasmid-borne *Streptococcus pyogenes* Cas9 nuclease (*Sp*Cas9) was expressed in *Saccharomyces cerevisiae* from a *GAL1*-inducible promoter (DiCarlo et al., 2013). The same plasmid also carried a CTG guide RNA (hereafter named gRNA#1) under the control of the constitutive *SNR52* promoter. The PAM used in this experiment was the TGG sequence located right at the border between the CTG tract and non-repeated DNA (Figure 1A). As controls, we used the same *Sp*Cas9-containing plasmid without the gRNA or a frameshift mutant of the *Sp*Cas9 gene resulting in a premature stop codon (*Sp*Cas9ΔNdeI) and the gRNA#1. The same genetic assay as previously was used (Mosbach et al., 2018; Richard et al., 2014). It is based on a modified suppressor tRNA gene (*SUP4*) in which a CTG trinucleotide repeat was integrated. The length of the CTG repeat at the start of the experiment was determined to be approximately 80 triplets. Four hours after transition from glucose to galactose medium, two faint bands were visible on a Southern blot, corresponding to the 5' and 3' ends of the *Sp*Cas9 DSB. No signal was detected in both control strains (Figure 1B). The DSB was quantified to be present in ca. 10% of the cells at any given time point and remained the same for the duration of the time course (4 hours). No evidence for repeat tract contraction was visible by Southern blot. Survival to the *Sp*Cas9 break was low (17.9%±4%), as calculated from CFU on galactose plates over CFU on glucose plates (Figure 2, see Materials & Methods). Surviving colonies were picked, total genomic DNA was extracted and the *SUP4*::CTG-locus was analyzed by Southern blot. Patterns observed were

remarkably different among clones, most of them showing bands of aberrant molecular weight, either much larger or much shorter than the repeat tract. In some clones, a total absence of signal suggested that the probe target was deleted and in other cases weakness of the signal was compatible with partial deletion (Figure 3A). To understand these abnormal patterns, the genome of nine surviving clones were totally sequenced by paired-end Illumina. As a control, one clone in which *Sp*Cas9 had not been induced was also sequenced. In all nine cases, a deletion around the repeat tract was found, extending from a few nucleotides to several kilobases (Figure 3B). Some of the rearrangements involved flanking Ty1 retrotransposon LTRs, and in one case (clone #2), a complex event between a distant LTR (δ16) and the δ20 LTR close to the repeat tract was detected. Following this discovery, total genomic DNA was extracted from more surviving colonies, analyzed by Southern blot and rearrangement junctions were amplified by PCR and Sanger sequencing. Two different sets of primers were used to amplify the junction, su47/su48 allowing to amplify local rearrangements around the repeat tract, whereas su23/su42 were used to amplify larger rearrangements between Ty1 LTRs (Figure 4, Table 2). Surviving colonies were classified into 12 different types, according to the *SUP4*::CTG locus after *Sp*Cas9 induction: type I corresponded to a colony in which the repeat tract was unchanged, types II-V corresponded to local rearrangements around the repeat tract and types VI-XII corresponded to more extensive rearrangements, the last category involving a complex event between δ16 and δ20. A few examples of junction sequences are shown in Supplemental Figure 1.

Chromosomal rearrangements are under the control of RAD52, DNL4 and SAE2

We next decided to investigate the role of several genes known to be involved in DSB repair on chromosomal rearrangements generated by the *Sp*Cas9 nuclease. In a *rad52*Δ strain, in which all homologous recombination was abolished, survival significantly decreased below

wild type (7.6%±0.7%, Figure 2). Molecular analysis of the survivors by Southern blot showed that rearrangements seemed to be less extensive than in the wild-type strain, fewer lanes showing a partial or total absence of signal (Figure 3C). Sequencing of the junctions confirmed that rearrangements between Ty LTRs were lost (Type VI events), except for two cases in which the deletion occurred through annealing of eight or nine nucleotides and was therefore *RAD52* independent (Figure 4 and Supplemental Figure 1, clones #C4 and #C7). This result shows that about 50% of colonies growing on galactose plates survived to the DSB by *RAD52*-dependent homologous recombination between two LTR elements flanking the trinucleotide repeat tract.

The possible role of non-homologous end-joining (NHEJ) in the observed rearrangements was also addressed by deleting the gene encoding yeast Ligase IV (*DNL4*). In the *dnl4*Δ strain, the level of detected DSBs was significantly lower than in wild type (Figures 1B and 1C), suggesting that DSB ends may be partially protected by the presence of Dnl4p, and were processed more rapidly when Ligase IV was absent. Survival was slightly decreased, but not significantly different from wild type (10.5%±6.3%, Figure 2). Molecular analysis of the survivors showed that local rearrangements were totally lost, whereas extensive rearrangements involving Ty LTR represented 84% of all events (Figure 4 and Supplemental Figure 1). Hence, we concluded that all local rearrangements were NHEJ dependent.

In a recent work, we showed that *SAE2* was essential to repair a DSB induced by a TALEN within a long trinucleotide repeat. In its absence, unrepaired breaks accumulated and DSB resection was lost on the trinucleotide repeat-containing end (Mosbach et al., 2018). We therefore tested the effect of a *sae2*Δ mutation on a *Sp*Cas9 DSB in the same experimental system. Southern blot analysis of repair intermediates showed that DSB ends accumulated twice as much in the *sae2*Δ mutant as compared to wild type (Figures 1B and 1C). In addition, a smear was detected below the 5' DSB end (Figure 1B, orange bracket), hallmark of

a resection defect (Chen et al., 2013). Survival was similar to wild type (21.5%±2.9%, Figure 2). Southern blot analysis of surviving colonies displayed very little size changes as compared to uninduced controls (Figure 3D). However, sequencing showed that the most frequent event was an insertion (or sometimes a small deletion) of one to eight nucleotides between the PAM and the repeat tract (Type III events, Figure 4). These local insertions represented 78% of all survivors, whereas only one Ty LTR recombination (Type VI) was detected (Supplemental Figure 1). This showed us that in the absence of *SAE2*, long range rearrangements were lost, probably due to the inability to resect the DSB into single-stranded DNA prone for homologous recombination.

The double mutant *sae2*Δ *dnl4*Δ was also built and showed an additive effect on survival, with a 30-fold reduction in CFU on galactose plates (0.6%±0.9%, Figure 2). This proved that in the absence of one of the two genes repair could occur by the other pathway, but absence of both genes was almost lethal to yeast cells receiving a *Sp*Cas9 DSB. Southern analysis showed that DSB levels were similar to *sae2*Δ levels (ca. 24% after 8 hrs versus 28% for *sae2*Δ), showing that *SAE2* was epistatic to *DNL4* for this phenotype. The smear corresponding to resection defects was also visible (Figure 1B, orange bracket). Interestingly, 21% of survivors exhibited zero to two triplets lost, which could be attributed to natural microsatellite instability. These were classified as Type I events and were specific of the *sae2*Δ *dnl4*Δ double mutant (Figure 4 and Supplemental Figure 1). It is possible that given the low survival rate, cells in which *Sp*Cas9 and/or the gRNA was mutated were positively selected during the time course in liquid culture and were therefore subsequently recovered on galactose plates. It is however surprising that such events were not recovered in *rad50*Δ cells. Remarkably, to the exception of the Type I events hereabove mentioned, all but one event corresponded to extensive rearrangements around the repeat tract, similarly to the single *dnl4*Δ mutant. This shows that although effects of both mutations were additive on cell

survival, *DNL4* was epistatic to *SAE2*, when chromosomal rearrangements induced by *Sp*Cas9 were considered.

Finally, in a *rad50*Δ strain, the DSB accumulated over the duration of the time course at levels similar to *sae2*Δ mutants (Figures 1B and 1C). No smear was detected in this strain background, suggesting that the *sae2*Δ resection defect was specific of this gene and did not involve the integrity of the MRX-Sae2 complex. However, no survivor could be found on galactose plates, showing that the DSB was lethal in this mutant background (Figure 2).

In conclusion, when a *Sp*Cas9 DSB was induced into a long CTG trinucleotide repeat, cell survival was low and depended on *RAD50*, *RAD52*, *SAE2* and *DNL4*. Two classes of repair events were found: local rearrangements under the control of *DNL4* and therefore the NHEJ pathway, and extensive rearrangements under the control of *SAE2* and *RAD52*. In addition, the deletion of *RAD50* almost completely recapitulated the *sae2*Δ *dnl4*Δ double mutation, except that the smear was not visible in *rad50*Δ and no surviving cell could be found, suggesting that in the absence of this gene DSB repair cannot occur at all on CTG trinucleotide repeats, following a *Sp*Cas9-induced DSB.

Enhanced SpCas9 generates the same chromosomal rearrangements as SpCas9

Over the last three years, several mutants of the widely used *Sp*Cas9 have been engineered or selected by genetic screens. *Sp*Cas9-HF1 and e*Sp*Cas9 were built to exhibit less off-target DSBs (Kleinstiver et al., 2016; Slaymaker et al., 2016), HypaCas9 was made to be even more accurate (Chen et al., 2017), Sniper-Cas9 also showed reduced off-target effects (Lee et al., 2018), while evoCas9 was selected in yeast for improved specificity (Casini et al., 2018). We decided to explore the possibility that chromosomal rearrangements observed in our experimental system were partly due to the fact that *Sp*Cas9 exhibited a high off-target activity on long CTG trinucleotide repeat tract, perhaps by generating more than one DSB

within the repeat tract, or within the surrounding loci. In order to test this hypothesis, Enhanced *Sp*Cas9 (e*Sp*Cas9) was expressed in yeast, along with the same guide RNA as previously (gRNA #1, Figure 1A). Survival was slightly higher than with *Sp*Cas9 (26.3%±3.0%), but not significantly different (t test p-value= 0.06). DSB end accumulation was lower than *Sp*Cas9 (Figure 1B, 1C). Molecular analysis of surviving yeast cells did not show any statistical difference between types of rearrangements observed with e*Sp*Cas9 as compared to *Sp*Cas9 (Chi2 p-value= 0.14) (Figure 5 and Supplemental Figure 1). A second guide RNA (gRNA#2) was designed, so that the DSB would be made two nucleotides closer to the repeat tract end (Figure 1A). Interestingly, the number of rearrangements involving a LTR was lower than with gRNA#1 (19% with gRNA#2 vs 92% with gRNA#1) but the proportion of very large deletions (Type XI) significantly increased from 4% to 31% (Chi2 p-value= $1.6\ 10^{-3}$). We concluded that moving the DSB cut site two nucleotides toward non-repeated DNA increased the outcome of very large deletions. Altogether, these results show that using a more specific version of *Sp*Cas9 did not decrease chromosomal rearrangements, but actually increased extensive deletions, suggesting that the effects seen were probably not due to extra off-target DSBs within the CTG repeat tract or the surrounding loci.

Gene conversion efficacy is decreased when a Cas9 DSB is made within a long CTG trinucleotide repeat

Gene conversion is a very efficient DSB-repair mechanism in *S. cerevisiae*. We previously showed that a single DSB induced by the I-*Sce* I meganuclease in a yeast chromosome was efficiently repaired using a CTG repeat-containing homologous template as a donor (Richard et al., 1999a, 2000, 2003). In order to determine whether a Cas9-induced DSB within a CTG repeat was properly repaired by the recombination machinery, we reused a similar experimental system in which two copies of the *SUP4* allele were present on yeast

chromosome X, one containing a $(CTG)_{60}$ repeat tract and the other copy containing an I-*Sce* I recognition site (Figure 6A). In this ectopic gene conversion assay, 80.2%±2.3% of yeast cells survived to an I-*Sce* I DSB and 100% of survivors were repaired by gene conversion using the ectopic *SUP4*::(CTG)n copy as a donor (Richard et al., 2003). When *Sp*Cas9 was induced in the same yeast strain along with gRNA#1, only 32.6%±3.8% of CFU formed on galactose plates (Figure 2). Molecular analysis of surviving cells showed that 89% (34 out of 38) repaired by ectopic gene conversion, as expected, and now contain two I-*Sce* I recognition sites, one in each *SUP4* copy (Figure 6B, GC events). However, one expansion event was also detected, as well as one local rearrangement (Type IV) and two rearrangements involving a deletion and a DNA insertion (Type V). Intriguingly, the DNA insertion is a 211 bp piece of DNA from the *YAK1* gene, located 158 kilobases upstream the *ARG2* locus, on chromosome X left arm. This gene contains a long and imperfect CAG/CTG repeat within its reading frame, like many yeast genes (Field and Wills, 1998; Malpertuy et al., 2003; Richard and Dujon, 1996; Richard et al., 1999b). An unusual non-homologous recombination event occurred between the *YAK1* CAG/CTG repeat and the *ARG2* repeat, leading to a chimeric insertion (Figure 6C). This may be the result of an off-target DSB generated by *Sp*Cas9 within the *YAK1* repeat, or be due to an abnormal recombination event between the two CTG repeats following *Sp*Cas9 induction. Using the CRISPOR *in silico* tools, three off-target sites were found for *Sp*Cas9 gRNA#1 if no mismatch were allowed and 80 sites when up to three mismatches were permitted (Haeussler et al., 2016). The second best off-target score was *YAK1*. We concluded that gene conversion efficacy was reduced when the DSB was made by *Sp*Cas9 within a CTG repeat, as compared to an I-*Sce* I DSB made in a non-repeated sequence, partly due to occasional off-target DSBs in other CTG repeats of the yeast genome.

Resection of a Cas9-induced double-strand break

Quantitative PCR experiments were performed in order to determine the resection level in strains in which Cas9 was induced. The nuclease generates a DSB in the very last CTG triplets of the repeat tract (Figure 1A). Therefore, the 5' end of the break contains most of the 80 triplets whereas the 3' end contains only two triplets. This allows to compare resection of a repeated and structured DNA end versus non-repeated DNA, concomitantly and in the same experimental setting. We took advantage of the convenient position of four *Eco*RV restriction sites, two on each side of the DSB, at different distances from the break (Figure 7A). Primers were designed in such a way that *Eco*RV digested DNA could not be PCR amplified. However, if DNA resection reached an *Eco*RV site, the resulting single-stranded DNA became resistant to digestion and therefore susceptible to amplification. In wild-type cells after eight hours, resection of the Cas9 DSB was always 100% at all *Eco*RV sites, except at the 3' distal site in which it was a little lower, around 70% (Figure 7A). In *dnl4*Δ cells, resection was not statistically different from wild type. In the *rad50*Δ mutant, DSB resection was totally abolished on the 5' end of the break that contains most of the repeat tract and severely impaired on the other end, showing that the MRX-Sae2 complex was essential in this process, on both DSB ends. Interestingly, the *sae2*Δ mutant exhibited a resection defect on the 5' end of the break but not on the other side. This was also true for the double mutant *sae2*Δ *dnl4*Δ. All these data prove that: i) Ligase IV plays no role in DSB resection; ii) Sae2 is essential to resect a long CTG trinucleotide repeat but is dispensable to resect a non-repeated DSB end.

Genome-wide mutation spectrum in cells expressing SpCas9

When carefully looking at deletion borders in haploid strains in which *Sp*Cas9 was induced, they were found to be more extensive on the 5' side of the break than on the 3' side (Figure 4). This suggests that larger 3' deletions encompassing the essential gene *CDC8* or its promoter

may not have been recovered because they would be lethal. This could also account for the lethality observed with *Sp*Cas9 in haploids. In order to check this hypothesis, we analyzed diploids containing *SUP4*::(CTG)n repeat tracts on both homologues, in which *Sp*Cas9 was induced. In these cells, both chromosomes could be cut by the nuclease. We quantified by qPCR *CDC8* copy number in six independent diploid survivors. In all cases, it was reduced by half as compared to a control qPCR on another chromosome (Supplemental Figure 2A). This showed that in these cells, only one of the two *CDC8* alleles was present, suggesting that the other was often deleted during DSB repair. In order to check if the whole chromosome could have been lost, we also amplified a region near the *JEM1* gene, on the other chromosomal arm, near the *ARG2* gene. Surviving clones showed a significantly higher signal, compatible with the presence of two chromosomes (Mann-Whitney-Wilcoxon rank test, p-value= 10^{-3}). Therefore, it was concluded that the mortality observed in haploid cells expressing SpCas9 was at least partly due to the frequent deletion of the essential *CDC8* gene, but did not induce significant chromosome loss.

In order to detect possible off-target mutations, independent haploid and diploid colonies in which *Sp*Cas9 had been induced were deep-sequenced, using Illumina paired-end technology. In diploid cells, two nucleotide substitutions were detected out of five independent clones when *Sp*Cas9 was repressed (Supplemental Figure 2B). When the nuclease was induced, six mutations were detected out of 18 sequenced clones, a similar proportion. One 36-bp deletion was found in the *FLO11* minisatellite and two deletions of one repeat unit were found in AT dinucleotide repeats, but no mutation was found in any other CAG/CTG repeat tract. Altogether, we concluded that *Sp*Cas9 expression in diploid cells did not significantly increase genome-wide mutation frequency. The genome of 10 independent haploid cells in which *Sp*Cas9 was induced was also completely sequenced. Eight mutations were detected among six of these survivors, all of them being nucleotide substitutions in non-repeated DNA

(Supplemental Figure 2B). This is statistically not different from what was observed in diploids (Fisher exact test p-value= 0.003). We concluded that besides chrosomosomal rearrangements around the *SUP4* locus observed in these haploids, *Sp*Cas9 did not induce other mutations in yeast cells in which it was expressed.

Discussion

SpCas9-induced DSB repair within CTG repeats generates chromosomal rearrangements

In previous works, in which we induced a DSB within a CTG repeat using a dedicated TALEN, 100% of surviving yeast colonies repaired the break by contracting the repeat tract (Richard et al., 2014). These contractions occurred by an iterative single-strand annealing mechanism, that depended on *RAD52*, *RAD50* and *SAE2*, but was independent of *LIG4*, *POL32* and *RAD51* (Mosbach et al., 2018). It was therefore striking and completely unexpected that a DSB made by the *Sp*Cas9 nuclease (or by its more specific mutant version, e*Sp*Cas9) at exactly the same location within the very same CTG trinucleotide repeat induced frequent chromosomal rearrangements around the repeat tract and almost no repeat contraction. With the TALEN, repeat contraction was shown to be an iterative phenomenon, involving several rounds of cutting and contraction until the repeat tract was too short for the two TALEN arms to dimerize and induce a DSB (Mosbach et al., 2018; Richard et al., 2014). A similar outcome was expected with *Sp*Cas9, iterative rounds of cutting and contraction could occur until the remaining CTG repeat tract would be too short for the gRNA to bind and induce a DSB. However, this was not observed here, surprisingly proving that a *Sp*Cas9-induced DSB was differently repaired from a TALEN-induced DSB targeting the same exact repeated sequence.

Spontaneous homologous recombination events between delta elements surrounding *SUP4* were already described by the past (Rothstein et al., 1987). However, in the present case,

Cas9-induced deletions also involved microhomology sequences or no homology at all, in addition to delta LTR elements, suggesting that the initiating damage was different in both experimental systems (replication-induced single-strand nicks vs. nuclease-induced double-strand breaks, for example). Our data are more reminiscent of a previous work in which spontaneous deletions around the *URA2* gene were classified in seven different classes, six of them harboring microhomologies at their junctions and one showing no obvious homology (Welcker et al., 2000). In a recent work, using a GFP reporter system in human cells, it was shown that SpCas9 induced contractions as well as expansions of long CTG trinucleotide repeats, whereas the nickase mutant SpCas9-D10A only induced contractions. However, it was not possible to determine whether local rearrangements could be present in some of the surviving cells (Cinesi et al., 2016). In a recent work looking at the effect of a Cas9-induced DSB at the *LYS2* locus in *S. cerevisiae*, the authors found frequent *POL4*-dependent small insertions (1-3 bp) in 42-68% of the survivors (depending on the PAM used) and local deletions (1-17 bp) in the remaining cases. In the present experiments, local rearrangements account for 19.6% with 20% of those being small insertions (Figure 4). The remaining events (80.4%) corresponded to extensive rearrangements involving retrotransposon LTRs. However, given that there is no transposon or transposon remnant in the close proximity of the *LYS2* locus, the authors could not retrieve LTR rearrangements (Lemos et al., 2018). This strongly suggests that rearrangements observed heavily depend on the surrounding chromosomal location where the DSB is made.

When an I-*Sce* I DSB was induced within a *SUP4* allele, the break could be repaired by gene conversion with a CTG repeat-containing homologous donor at the *ARG2* locus. All yeast cells repaired by gene conversion with the donor, generating repeat contractions and expansions in the process (Richard et al., 2003). Here, the exact reverse reaction was induced, the break was made within CTG repeats and repaired with a non-repeated sequence. DSB

v

repair was much less efficient, since only 32.6% of the cells survived (Figure 2) and less specific since 10% of the repair events were unfaithful recombination (Figure 6B). This shows that when a Cas9 DSB was made into a CTG repeat, gene conversion was impaired, either by the repeat tract or by the Cas9 protein, or by both.

Ligase IV and Sae2 are respectively driving local and extensive chromosomal rearrangements

Yeast Ligase IV is encoded by the *DNL4* gene and is the enzyme used to ligate DSB ends during non homologous end-joining (Wilson et al., 1997). It was previously shown that *RAD50* and *SAE2* were essential to resect and process a TALEN-induced DSB but a *DNL4* deletion had no effect on break processing, cell survival or repair efficacy (Mosbach et al., 2018). On the contrary, repair of a Cas9, an HO or an I-*Sce* I DSB at the *MAT* locus, in the absence of any homologous donor casette, was shown to be dependent on the product of the *DNL4* gene (Frank-Vaillant and Marcand, 2001; Lemos et al., 2018). *Sp*Cas9 DSB repair has also been studied in human cells in the presence of a drug (NU7441), acting as a chemical inhibitor of non-homologous end-joining. In these conditions, the frequency of single-base insertions and small deletions decreased whereas larger deletions increased, suggesting that these repair events occurred by an alternative end-joining mechanism (alt-EJ/MMEJ) involving microhomologies flanking the DSB (Charpentier et al., 2018; van Overbeek et al., 2016). Here, we showed that when *DNL4* was inactivated, local deletions were totally lost. However, survival was not significantly decreased because yeast cells could repair the DSB using LTR recombination, generating extensive deletions around the repeat tract (Figure 4). Supporting this model, the absence of any resection defect in the *dnl4*Δ mutant proved that in the absence of end-joining, resection may take place very efficiently to repair the DSB by homologous recombination, using flanking homologies.

SAE2 is associated to the *MRE11-RAD50-XRS2* complex, whose roles are multiple during DSB repair (Haber, 1998) and it was proposed to encode an endonuclease activity essential to process DNA hairpins (Lengsfeld et al., 2007), as well as to resect I-*Sce* I double-strand breaks (Mimitou and Symington, 2008). We previously showed that it was essential to resect a TALEN-induced DSB end containing a long CTG trinucleotide repeat, but less important to resect non-repeated DNA (Mosbach et al., 2018). In the present experiments, extensive rearrangements involving LTR elements were lost in a *sae2*Δ mutant, and 97% of yeast cells repaired the DSB by local rearrangements, most of them resulting in insertions or deletions between the PAM and the gRNA sequence (Figure 4), inactivating *Sp*Cas9 capacity to induce another DSB. Small insertions of a few nucleotides were also frequently detected following *Sp*Cas9 DSB induction at the VDJ locus in human B cells (So and Martin, 2019) or at the *MAT* locus in *S. cerevisiae* (Lemos et al., 2018). However, in our experiments, all nucleotides inserted were C, T or G, all three encoded by the gRNA. No insertion of an adenosine residue was found out of 28 insertions sequenced (Supplemental Figure 1). This intriguing observation suggests the possibility that the gRNA could be used as a template to repair the DSB, as it was demonstrated that a single-stranded RNA could be used to repair an HO-induced DSB into the *LEU2* gene (Storici et al., 2007).

Although *DNL4* and *SAE2* trigger different types of chromosomal rearrangements and none of the single mutants significantly decreased survival, the *dnl4*Δ *sae2*Δ double mutant abolished repair, almost to the point of the *rad50*Δ mutant, since only 0.6% of the cells survived (Figure 2), showing the synthetic effect of both mutations. However, repair events in the double mutant were similar to those observed in *dnl4*Δ, proving that *DNL4* was epistatic on *SAE2* (Figure 4). All these results are compatible with a model in which a Cas9 DSB was tentatively repaired by NHEJ first (Figure 7B). Then, if repair was unsuccessful or if *DNL4* was

inactivated, resection proceeded with MRX-Sae2 in charge of removing secondary structures present at DSB ends. When resection reached flanking LTRs, repair occured by *RAD52*-mediated SSA. In the absence of this gene, the break was repaired by *RAD52*-independent local rearrangements. Finally, when *SAE2* was inactivated, resection was impeded on the 5' DSB end containing most of the CTG repeat tract (Figure 7A), and mutagenic NHEJ was favored, leading to local insertions and deletions. It is unknown whether Sae2 would play the same essential role on other secondary structure-forming trinucleotide repeats, like GAA or CGG triplets, or if its activity is specific of CTG triplets, hence of a structure rather than a repeat. It is also unclear whether other nucleases, like *EXO1* or *DNA2*, would be important to perform long range resection on a long CTG repeat tract (Mimitou and Symington, 2008; Zhu et al., 2008), but the present experimental system allows us to address this question in a unique and unbiased way.

Methods

Yeast strains and plasmids

All mutant strains were built from strain GY6162-3D by classical gene replacement method (Orr-Weaver et al., 1981), using KANMX4 or *HIS3* as marker (Supplemental Table 1). KANMX4 cassettes were amplified from the EUROSCARF deletion library, using primers located 1kb upstream and downstream the cassette. VMS1/VMAS1 were used to amplify *rad52*Δ::KANMX, VMS2/VMAS2 were used to amplify *rad51*Δ::KANMX, VMS3/VMAS3 were used to amplify *pol32*Δ::KANMX, VMS4/VMAS4 were used to amplify *dnl4*Δ::KANMX, VMS6/VMAS6 were used to amplify *rad50*Δ::KANMX and SAE2up/SAE2down were used to amplify *sae2*Δ::KANMX (Supplemental Table 2). VMY350 and VMY352 strains were respectively used to construct VMY650 and VMY352 by mating-type switching, as follows: the pJH132 vector (Holmes and Haber, 1999) carrying

the HO endonuclease under the control of an inducible *GAL1-10* promoter was transformed in the haploid *MAT*α strains. After 5h of growth in lactate medium, HO expression was induced by addition of 2% galactose (final concentration) and grown for 1.5 hour. Cells were then plated on YPD and mating type was checked three days later by crosses with both *MAT*a and *MAT*α tester strains.

For *Sp*Cas9 inductions, addgene plasmid #43804 containing the nuclease under the control of the GalL promoter and the *LEU2* selection marker was digested with HpaI and cloned into yeast by homology-driven recombination (Muller et al., 2012) with a single PCR amplified fragment containing the *SNR52* promoter, the gRNA#1 and the *SUP4* terminator, using primers SNR52Left and SNR52Right (Supplemental Table 2) to give plasmid pTRi203. A frameshift was then introduced in this plasmid by NdeI digestion followed by T4 DNA polymerase treatment and religation of the plasmid on itself, to give plasmid pTRi206. In this plasmid, the *Sp*Cas9 gene is interrupted by a stop codon after amino acid Ile_{161}. The haploid GFY6162-3D strain (or its mutant derivatives), was subsequently transformed with pTRi203 or pTRi206 and transformants were selected on SC-Leu. The plasmid containing Enhanced *Sp*Cas9 (version 1.1, Addgene #71814, Slaymaker et al., 2016) was a generous gift of Carine Giovannangeli from the *Museum National d'Histoire Naturelle*. The e*Sp*Cas9 gene was amplified using primers LP400 and LP401 (Supplemental Table 2) and cloned into yeast cells in the Addgene#43804 plasmid digested with BamHI, by homology-driven recombination, with 34-bp homology on one side and 40-bp homology on the other side (Muller et al., 2012), to give plasmid pLPX11. For the gRNA#1, plasmid pLPX11 was digested with HpaI and cloned into yeast by homology-driven recombination (Muller et al., 2012) with a single PCR amplified fragment containing the *SNR52* promoter, the gRNA#1 and the *SUP4* terminator, as above to give plasmid pTRi207. For the gRNA#2, a guide RNA cassette was ordered from

ThermoFisher (GeneArt), flanked by *Eco*RI sites and was cloned in pRS416 (Sikorski and Hieter, 1989) using standard procedures to give plasmid pLPX210.

In silico simulations of off-target sites

To assess the number of off-target sites for *Sp*Cas9 in *Saccharomyces cerevisiae*, online tools were used. CRISPOR is a software that evaluates the specificity of a guide RNA through an alignment algorithm that maps sequences to a reference genome to identify putative on- and off- target sites (Li and Durbin, 2009). To predict off-target sites, the online tool sequentially introduces changes in the sequence of the gRNA and checks for homologies in the specified genome (Haeussler et al., 2016).

Cas9 inductions

Before nuclease induction, Southern blot analyses were conducted on several independent subclones to select one containing ca. 80 CTG triplets. For Cas9 inductions, yeast cells were grown overnight at 30°C in liquid SC-Leu medium, then washed with sterile water to remove any trace of glucose. Cells were split in two cultures, half of the cells were grown in synthetic -Leu medium supplemented with 2% galactose (final concentration) and the other half were grown in synthetic -Leu medium supplemented with 2% glucose (final concentration). Around $4x10^8$ cells were collected at different time points (T=0, 4, 5, 6, 7 and 8 hours) and killed by addition of sodium azide (0.01% final). Cells were washed with water, and frozen in dry ice before DNA extraction. To determine survival to Cas9 induction, 24 hours after the T0 time point, cells were diluted to an appropriate concentration, then plated on SC-Leu plates containing either 20 g/l glucose or galactose. After 3-5 days of growth at 30°C, ratio of CFU on galactose plates over CFU on glucose plates was considered to be the survival rate.

Double-strand break analysis and quantification

Total genomic DNA (4 µg) of cells collected at each time point was digested for 6h with EcoRV (40 U) (NEB) loaded on a 1% agarose gel (15x20 cm) and run overnight at 1 V/cm. The gel was vaccum transfered in alkaline conditions to a Hybond-XL nylon membrane (GE Healthcare) and hybridized with two randomly-labeled probes specific of each side of the repeat tract, upstream and downstream the *SUP4* gene (Viterbo et al., 2018). After washing, the membrane was overnight exposed to a phosphor screen and signals were read and quantified on a FujiFilm FLA-9000.

SUP4 locus analysis after Cas9 induction

Several colonies from each induced or repressed plates were picked, total genomic DNA (4 µg) was extracted with Zymolyase, digested for 6h by SspI (20 U) (NEB), loaded on a 1% agarose gel (15x20 cm) and run overnight at 1V/cm. The gel was vaccum transfered in alkaline conditions to a Hybond-XL nylon membrane (GE Healthcare) and hybridized with a randomly-labeled PCR fragment specific of a region downstream the *SUP4* gene, amplified from the su8-su9 primer couple (Supplemental Table 2). After washing, the membrane was overnight exposed on a phosphor screen and signals were revealed on a FujiFilm FLA-9000. Genomic DNA of each clone for which a signal was detected by Southern blot was subsequently amplified with su47-su48 primers and sequenced using su47 (Supplemental Table 2). Genomic DNA of clones for which no signal was detected by Southern blot of no PCR product was obtained with su47-su48 were subsequently amplified with su23-su42 primers and sequenced using su42 (Supplemental Table 2). Sanger sequencing was performed by GATC biotech.

Analysis of Cas9-induced DSB end resection by qPCR

A real-time PCR assay, using primer pairs flanking EcoRV sites 0.81 kb and 2.94 kb away from the 3' end of the CTG repeat tract (VMS20/VMAS20 and VMS21/VMAS21 respectively) and 0.88 kb and 1.88 kb away from the 5'end of the CTG repeat tract (VMS22/VMAS22 and VMS23/VMAS23 respectively), was used to quantify end resection. Another pair of primers was used to amplify a region of chromosome X near the *ARG2* gene (Viterbo et al., 2016), to serve as an internal control of DNA amount (JEM1f-JEM1r). Genomic DNA of cells collected at T=0h, T=6h and T=8h was split in two fractions, incubated at 80°C for 10 minutes in order to inactivate any remaining active DNA nuclease, then one fraction was used for EcoRV digestion and the other one for a mock digestion in a final volume of 15 µl. Samples were incubated for 5h at 37°C, then the enzyme was inactivated for 20 min at 80°C. DNA was subsequentltly diluted by adding 55 µl of ice-cold water, and 4 µl was used for each real-time PCR reaction in a final volume of 25 µl. PCRs were performed with the Absolute SYBR Green Fluorescein mix (Thermo Scientific) in the Mastercycler S realplex (Eppendorf), using the following program: 95°C 15min, 95°C 15sec, 55°C 30 sec, 72°C 30 sec repeated 40 times, followed by a 20 min melting curve. Reactions were performed in triplicates and the mean value was used to determine the amount of resected DNA, using the following formula: raw resection=$2/(1+2^{\Delta Ct})$ with $\Delta Ct=C_{t,EcoRV}-C_{t,mock}$. Relative resection values were calculated by dividing raw resection values by the percentage of DSB quantified at the corresponding time point.

The same protocol was used to determine the relative amount of *CDC8* and chromosome X in surviving clones after Cas9 induction, except that total genomic DNA was not digested prior to real-time PCR. Primer couples VMS23-VMAS23 were used to amplify *CDC8* and JEM1f-JEM1r for chromosome X left arm. Primers Chromo4_f and Chromo4_r were were used to

amplify a region of chromosome IV as an internal control for total DNA amount. See Supplemental Table 2 for all primer sequences.

Library preparation for deep-sequencing

Approximately 10 µg of total genomic DNA was extracted and sonicated to an average size of 500 bp, on a Covaris S220 (LGC Genomics) in microtubes AFA (6x16 mm) using the following setup: Peak Incident Power: 105 Watts, Duty Factor: 5%, 200 cycles, 80 seconds. DNA ends were subsequently repaired with T4 DNA polymerase (15 units, NEBiolabs) and Klenow DNA polymerase (5 units, NEBiolabs) and phosphorylated with T4 DNA kinase (50 units, NEBiolabs). Repaired DNA was purified on two MinElute columns (Qiagen) and eluted in 16 µl (32 µl final for each library). Addition of a 3' dATP was performed with Klenow DNA polymerase (exo-) (15 units, NEBiolabs). Home-made adapters containing a 4-bp unique tag used for multiplexing, were ligated with 2 µl T4 DNA ligase (NEBiolabs, 400,000 units/ml). DNA was size fractionated on 1% agarose gels and 500-750 bp DNA fragments were gel extracted with the Qiaquick gel extraction kit (Qiagen). DNA was PCR amplified for 12 cycles with Illumina primers PE1.0 and PE2.0 and Phusion DNA polymerase (1 unit, Thermo Scientific). Six PCR reactions were pooled for each library, and purified on a Qiagen purification column. Elution was performed in 30 µl and DNA was quantified on a spectrophotometer and on agarose gel.

Analysis of paired-end Illumina reads

Multiplexed libraries were loaded on a HiSeq2500 (Illumina), 110 bp paired-end reads for haploids and 260 bp paired-end reads for diploids were generated. Reads quality was evaluated by FastQC v.0.10.1 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped along S288C chromosome reference sequence (*Saccharomyces Genome*

Database, release R64-2-1, November 2014), using the paired-end mapping mode of BWA v0.7.4-r385 with default parameters (Li and Durbin, 2009). The output SAM files were converted and sorted to BAM files using SAMtools v0.1.19-44428cd (Li et al., 2009). The command *IndelRealigner* from GATK v2.4-9 (DePristo et al., 2011) was used to realign the reads. Duplicated reads were removed using the option "*MarkDuplicates"* implemented in Picard v1.94 (http://picard.sourceforge.net/). Reads uniquely mapped to the reference sequence with a minimum mapping quality of 20 (Phred-scaled) were kept. Mpileup files were generated by SAMtools without BAQ adjustments. SNPs and INDELs were called by the options "*mpileup2snp*" and "*mpileup2indel*" of *Varscan2 v*2.3.6 (Koboldt et al., 2012) with a minimum depth of 10 reads for haploids and 20 reads for diploids. Average read coverage was 255X for diploid cells (σ= 187X) and 190X for haploids (σ= 43X). Diploid strains are homozygous except for selection markers and some specific loci like *MAT* and *SUP4*. Therefore, *de novo* heterozygous mutations should represent 50% of reads, on the average. Taking that into account, lower and upper thresholds for variant allele frequency were respectively set between 30% and 70% in diploids. For haploids, the threshold for minimum variant allele frequency was set at 70%. Mutations less than 10 bp away from each other were discarded to avoid mapping problems due to paralogous genes or repeated sequences. To assess microsatellite mutations, we only retained reads uniquely anchored at least 20 bp on each side of the microsatellite (Fungtammasan et al., 2015). All detected mutations were manually examined using the IGV software (version 2.3.77), and compared between all sequenced libraries for interpretation. All the scripts used in order to process data are available on github (https://github.com/sdeclere/nuclease). All Illumina sequences were uploaded in the European Nucleotide Archive (ENA), accession number PRJEB16068.

Ackowledgments

We thank Carine Giovannangeli for the generous gift of the Enhanced *Sp*Cas9 plasmid. V. M. was supported by Fondation Guy Nicolas and Fondation Hardy. L. P. was the recipient of a graduate student CIFRE fellowship from SANOFI. W. V.-Z. is the recipient of a PhD fellowship from la Ligue Nationale Contre le Cancer. This work was generously supported by the Institut Pasteur and by the Centre National de la Recherche Scientifique (CNRS).

Author contributions

V. M. built yeast mutant strains, performed time courses, Southern blots, survival experiments, Illumina library constructions and the initial resection assays. D. V. built yeast mutant strains, performed time courses, Southern blots and survival experiments. L. P. built pLPX11 and pLPX210 plasmids and did the first *Sp*Cas9 induction as well as survival experiments. W. V.-Z. performed resection assays. S. D. D. analyzed Illumina sequences. G.-F. R. designed the experiments, analyzed rearrangement junction sequences and wrote the manuscript.

Declaration of interest

The authors declare no competing interests.

References

Anand, R.P., Shah, K.A., Niu, H., Sung, P., Mirkin, S.M., and Freudenreich, C.H. (2012). Overcoming natural replication barriers: differential helicase requirements. Nucleic Acids Res *40*, 1091–1105.

Balakumaran, B.S., Freudenreich, C.H., and Zakian, V.A. (2000). CGG/CCG repeats exhibit orientation-dependent instability and orientation-independent fragility in Saccharomyces cerevisiae. Hum Mol Genet *9*, 93–100.

Casini, A., Olivieri, M., Petris, G., Montagna, C., Reginato, G., Maule, G., Lorenzin, F., Prandi, D., Romanel, A., Demichelis, F., et al. (2018). A highly specific SpCas9 variant is identified by *in vivo* screening in yeast. Nat. Biotechnol. *36*, 265–271.

Charpentier, M., Khedher, A.H.Y., Menoret, S., Brion, A., Lamribet, K., Dardillac, E., Boix, C., Perrouault, L., Tesson, L., Geny, S., et al. (2018). CtIP fusion to Cas9 enhances transgene integration by homology-dependent repair. Nat. Commun. *9*, 1133.

Chen, H., Lisby, M., and Symington, L.S. (2013). RPA coordinates DNA end resection and prevents formation of DNA hairpins. Mol Cell *50*, 589–600.

Chen, J.S., Dagdas, Y.S., Kleinstiver, B.P., Welch, M.M., Sousa, A.A., Harrington, L.B., Sternberg, S.H., Joung, J.K., Yildiz, A., and Doudna, J.A. (2017). Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. Nature *550*, 407–410.

Cinesi, C., Aeschbach, L., Yang, B., and Dion, V. (2016). Contracting CAG/CTG repeats using the CRISPR-Cas9 nickase. Nat. Commun. *7*, 13272.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet *43*, 491–498.

DiCarlo, J.E., Norville, J.E., Mali, P., Rios, X., Aach, J., and Church, G.M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. Nucleic Acids Res *41*, 4336–4343.

Fairhead, C., and Dujon, B. (1993). Consequences of unique double-stranded breaks in yeast chromosomes: death or homozygosis. Mol Gen Genet *240*, 170–180.

Field, D., and Wills, C. (1998). Abundant microsatellite polymorphism in Saccharomyces cerevisiae, and the different distributions of microsatellites in eight prokaryotes and S. cerevisiae, result from strong mutation pressures and a variety of selective forces. Proc Natl Acad Sci USA *95*, 1647–1652.

Frank-Vaillant, M., and Marcand, S. (2001). NHEJ regulation by mating type is exercised through a novel protein, Lif2p, essential to the Ligase IV pathway. Genes&Development *15*, 3005–3012.

Freudenreich, C.H., Kantrow, S.M., and Zakian, V.A. (1998). Expansion and length-dependent fragility of CTG repeats in yeast. Science *279*, 853–856.

Fungtammasan, A., Ananda, G., Hile, S.E., Su, M.S.-W., Sun, C., Harris, R., Medvedev, P., Eckert, K., and Makova, K.D. (2015). Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. Genome Res.

Gacy, A.M., Goellner, G., Juranic, N., Macura, S., and McMurray, C.T. (1995). Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell *81*, 533–540.

Haber, J.E. (1995). In vivo biochemistry: physical monitoring of recombination induced by site-specific endonucleases. BioEssays *17*, 609–620.

Haber, J.E. (1998). The many interfaces of Mre11. Cell *95*, 583–586.

Haeussler, M., Schönig, K., Eckert, H., Eschstruth, A., Mianné, J., Renaud, J.-B., Schneider-Maunoury, S., Shkumatava, A., Teboul, L., Kent, J., et al. (2016). Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. *17*.

Holmes, A.M., and Haber, J.E. (1999). Double-strand break repair in yeast requires both leading and lagging strand DNA polymerases. Cell *96*, 415–424.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016). High-fidelity CRISPR–Cas9 nucleases with no detectable genome-wide off-target effects. Nature *529*, 490–495.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res *22*, 568–576.

Krogh, B.O., and Symington, L.S. (2004). Recombination proteins in yeast. Annu. Rev. Genet. *38*, 233–271.

Lee, J.K., Jeong, E., Lee, J., Jung, M., Shin, E., Kim, Y., Lee, K., Jung, I., Kim, D., Kim, S.,

et al. (2018). Directed evolution of CRISPR-Cas9 to increase its specificity. Nat. Commun. *9*, 3048.

Lemos, B.R., Kaplan, A.C., Bae, J.E., Ferrazzoli, A.E., Kuo, J., Anand, R.P., Waterman, D.P., and Haber, J.E. (2018). CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. Proc. Natl. Acad. Sci. *115*, E2040–E2047.

Lengsfeld, B.M., Rattray, A.J., Bhaskara, V., Ghirlando, R., and Paull, T.T. (2007). Sae2 Is an Endonuclease that Processes Hairpin DNA Cooperatively with the Mre11/Rad50/Xrs2 Complex. Mol. Cell *28*, 638–651.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinforma. Oxf. Engl. *25*, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Malpertuy, A., Dujon, B., and Richard, G.-F. (2003). Analysis of microsatellites in 13 hemiascomycetous yeast species: mechanisms involved in genome dynamics. J. Mol. Evol. *56*, 730–741.

Mimitou, E.P., and Symington, L.S. (2008). Sae2, Exo1 and Sgs1 collaborate in DNA double-strand break processing. Nature *455*, 770–774.

Mosbach, V., Poggi, L., Viterbo, D., Charpentier, M., and Richard, G.-F. (2018). TALEN-induced double-strand break repair of CTG trinucleotide repeats. Cell Rep. *22*, 2146–2159.

Muller, H., Annaluru, N., Schwerzmann, J.W., Richardson, S.M., Dymond, J.S., Cooper, E.M., Bader, J.S., Boeke, J.D., and Chandrasegaran, S. (2012). Assembling large DNA segments in yeast. Methods Mol. Biol. Clifton NJ *852*, 133–150.

Nelms, B.E., Maser, R.S., MacKay, J.F., Lagally, M.G., and Petrini, J.H.J. (1998). In Situ Visualization of DNA Double-Strand Break Repair in Human Fibroblasts. Science *280*, 590–592.

Nguyen, J.H.G., Viterbo, D., Anand, R.P., Verra, L., Sloan, L., Richard, G.-F., and Freudenreich, C.H. Differential requirement of Srs2 helicase and Rad51 displacement activities in replication of hairpin-forming CAG/CTG repeats. Nucleic Acids Res. *45*, 4519–4531.

Orr, H.T., and Zoghbi, H.Y. (2007). Trinucleotide repeat disorders. Annu Rev Neurosci *30*, 575–621.

Orr-Weaver, T.L., Szostak, J.W., and Rothstein, R.J. (1981). Yeast transformation: a model system for the study of recombination. Proc. Natl. Acad. Sci. U. S. A. *78*, 6354–6358.

van Overbeek, M., Capurso, D., Carter, M.M., Thompson, M.S., Frias, E., Russ, C., Reece-Hoyes, J.S., Nye, C., Gradia, S., Vidal, B., et al. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. Mol. Cell *63*, 633–646.

Pelletier, R., Krasilnikova, M.M., Samadashwily, G.M., Lahue, R., and Mirkin, S.M. (2003). Replication and expansion of trinucleotide repeats in yeast. Mol Cell Biol *23*, 1349–1357.

Plessis, A., Perrin, A., Haber, J.E., and Dujon, B. (1992). Site-specific recombination determined by I-Sce I, a mitochondrial group I intron-encoded endonuclease expressed in the yeast nucleus. Genetics *130*, 451–460.

Richard, G.-F., and Dujon, B. (1996). Distribution and variability of trinucleotide repeats in the genome of the yeast Saccharomyces cerevisiae. Gene *174*, 165–174.

Richard, G.-F., Dujon, B., and Haber, J.E. (1999a). Double-strand break repair can lead to high frequencies of deletions within short CAG/CTG trinucleotide repeats. Mol Gen Genet *261*, 871–882.

Richard, G.-F., Hennequin, C., Thierry, A., and Dujon, B. (1999b). Trinucleotide repeats and other microsatellites in yeasts. Res Microbiol *150*, 589–602.

Richard, G.-F., Goellner, G.M., McMurray, C.T., and Haber, J.E. (2000). Recombination-induced CAG trinucleotide repeat expansions in yeast involve the MRE11/RAD50/XRS2 complex. EMBO J *19*, 2381–2390.

Richard, G.-F., Cyncynatus, C., and Dujon, B. (2003). Contractions and expansions of CAG/CTG trinucleotide repeats occur during ectopic gene conversion in yeast, by a MUS81-independent mechanism. J. Mol. Biol. *326*, 769–782.

Richard, G.-F., Kerrest, A., and Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol Mol Biol Rev *72*, 686–727.

Richard, G.-F., Viterbo, D., Khanna, V., Mosbach, V., Castelain, L., and Dujon, B. (2014). Highly specific contractions of a single CAG/CTG trinucleotide repeat by TALEN in yeast. PLoS ONE *9*, e95611.

Rothstein, R., Helms, C., and Rosenberg, N. (1987). Concerted deletions and inversions are caused by mitotic recombination between delta sequences in Saccharomyces cerevisiae. Mol Cell Biol *7*, 1198–1207.

Samadashwily, G., Raca, G., and Mirkin, S.M. (1997). Trinucleotide repeats affect DNA replication in vivo. Nat. Genet *17*, 298–304.

Sikorski, R.S., and Hieter, P. (1989). A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae. Genetics *122*, 19–27.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. Science *351*, 84–88.

So, C.C., and Martin, A. (2019). DSB structure impacts DNA recombination leading to class switching and chromosomal translocations in human B cells. PLOS Genet. *15*, e1008101.

Storici, F., Bebenek, K., Kunkel, T.A., Gordenin, D.A., and Resnick, M.A. (2007). RNA-templated DNA repair. Nature *447*, 338–341.

Sutherland, G.R., Baker, E., and Richards, R.I. (1998). Fragile sites still breaking. Trends Genet *14*, 501–506.

Viterbo, D., Michoud, G., Mosbach, V., Dujon, B., and Richard, G.-F. (2016). Replication stalling and heteroduplex formation within CAG/CTG trinucleotide repeats by mismatch repair. DNA Repair *42*, 94–106.

Viterbo, D., Marchal, A., Mosbach, V., Poggi, L., Vaysse-Zinkhöfer, W., and Richard, G.-F. (2018). A fast, sensitive and cost-effective method for nucleic acid detection using non-radioactive probes. Biol. Methods Protoc. *3*.

Welcker, A.J., de Montigny, J., Potier, S., and Souciet, J.L. (2000). Involvement of very short DNA tandem repeats and the influence of the RAD52 gene on the occurrence of deletions in Saccharomyces cerevisiae. Genetics *156*, 549–557.

Wilson, T.E., Grawunder, U., and Lieber, M.R. (1997). Yeast DNA ligase IV mediates non-homologous DNA end joining. Nature *388*, 495–498.

Zhu, Z., Chung, W.H., Shim, E.Y., Lee, S.E., and Ira, G. (2008). Sgs1 helicase and two nucleases Dna2 and Exo1 resect DNA double-strand break ends. Cell *134*, 981–994.

Figure Legends

Figure 1: *Sp*Cas9 and e*Sp*Cas9 DSB induction in wild type and mutant strains

A: Sequence of the *SUP4*::(CTG)n locus. The CTG trinucleotide repeat tract comes from a human DM1 patient and is shown in blue. The flanking non-repeated DNA is in black. For each guide RNA, the PAM, the gRNA sequence as well as the expected DSB site are indicated. **B**: Southern blots of yeast strains during the time course. Lanes labeled *Sp*Cas9-gRNA and *Sp*Cas9ΔCter+gRNA are control strains in which no DSB was visible. In the strain expressing both *Sp*Cas9 and the gRNA, two bands are visible in addition to the parental allele (1966 bp). One band corresponds to the 3' end of the DSB containing a small number of triplets (821 bp), the other one corresponds to the 5' end of the DSB containing most of the repeat tract (1145 bp). **C:** Quantification of 5' and 3' DSB signals. For each time points, the total 5' + 3' signals were quantified and plotted as a ratio of the total signal in the lane. Three independent time courses were run in each strain background (except *rad50*Δ for which two time courses were run) and plots show the average of three (or two) time courses.

Figure 2: Yeast survival to Cas9 induction

For each strain, the same number of cells were plated on galactose and glucose plates and the survival was expressed in CFU number on galactose plates over CFU number on glucose plates. The mean and the 95% confidence interval are plotted for each strain. Significant t-test p-values when compared to wild-type *Sp*Cas9 survival are indicated by asterisks, as shown on the figure.

Figure 3: Chromosomal rearrangements following *Sp*Cas9 induction

A: Southern blot of genomic DNA at the *SUP4* locus in the wild-type strain. The probe hybridizes ~300 bp downstream the repeat tract (see Figure 3B). The dotted red line shows

the initial length of the CTG repeat tract. The lane labeled "Glucose" contains a clone in which Cas9 was not induced. Lanes numbered #1 through #19 contain independent clones in which Cas9 was induced. Asterisks point to lanes in which no signal was detected, meaning that the probe containing sequence was deleted. Note that signal intensities varies among lanes, showing that the probe did not fully bind to its target sequence, due to its partial deletion. **B**: Some examples of chromosome rearrangements following Cas9 induction in the wild-type strain. The genomic locus surrounding *SUP4* is shown on top, ARS1018 is drawn in red, delta elements are in grey, protein-coding genes are colored in blue and tRNA genes in purple. The DSB (vertical purple arrow) is induced within *SUP4*::(CTG)n. Chromosome coordinates are indicated above and the probe used for hybridization is represented by an horizontal red bar. The locus sequence was retrieved from the Saccharomyces Genome Database (http://yeastgenome.org/, genome version R64-2-1, released 18th November 2014). Under the reference locus are cartooned the different chromosomal structures observed in some of the survivors. A yeast colony that was grown in glucose was also sequenced as a control. For each clone, vertical dotted lines represent junctions of rearrangements observed, with deletion sizes indicated in base pairs. Asterisk: clone #2 showed a complex rearrangement with a local inverted duplication involving the δ16 LTR and the 3' end of the *KCH1* gene 5 kb upstream *SUP4*. Two clones (#8 and #9) exhibit exactly the same chromosomal rearrangement at precisely the same nucleotides. Note that *CDC8* is an essential gene. **C**: Southern blot of genomic DNA at the *SUP4* locus in the *rad52*Δ strain. Legend as for Figure 3A. Note that for this Southern blot genomic DNA was digested with EcoRV (instead of Ssp I, see Methods), therefore the expected CTG repeat length was around 1.8 kb, instead of 1 kb. **D**: Southern blot of genomic DNA at the *SUP4* locus in the *sae2*Δ strain. Legend as for Figure 3A.

Figure 4: Summary of chromosomal rearrangements observed in wild-type and mutant strains, following *Sp*Cas9 induction

Left: The twelve different possible outcomes following *Sp*Cas9 induction are shown, subdivided in local and extensive rearrangements (see text for details). The *SUP4* locus is pictured and shows the position of each genetic element on yeast chromosome X. The probe used on Southern blots is shown, as well as both primer couples used to amplify the locus. In order to assess a given clone to a rearrangement type, the following rules were followed: i) when a band was detected by Southern blot, primers su47 and su48 were used to amplify the locus an sequence it. These events corresponded to types I-V. The absence of a PCR product indicated that primer su48 genomic sequence was probably deleted and therefore primers su23 and su42 were used to amplify and sequence the locus. These were classified as types IV-V events; ii) when no band was detected by Southern blot, primers su 23 and su42 were directly used to amplify and sequence the locus. These events were classified as types VI-X and XII. When no PCR product was obtained, it meant that at least one of the two primers genomic sequence was probably deleted and these events were classified as type XI. Note that this last category may also contain rare -but possible- chromosomal translocations that ended up in puting each primer in a separate chromosome, making unobtainable the PCR product. The extent of type XI deletions cannot go downstream the su42 primer, since the *CDC8* gene is essential. **Right**: The proportion of each type or event recovered is represented for wild type and mutants. Altogether, 220 surviving clones were sequenced, distributed as follows: WT: 51, *rad52*Δ: 29, *dnl4*Δ: 61, *sae2*Δ: 32, *dnl4*Δ *sae2*Δ: 47.

Figure 5: Summary of chromosomal rearrangements observed in following *Sp*Cas9 and enhanced *Sp*Cas9 (e*Sp*Cas9) inductions

See Figure 4 for legend. e*Sp*Cas9 #1 and #2 refer respectively to guide RNAs #1 and #2 described in Figure 1A. The following number of surviving clones were analyzed: *Sp*Cas9: 51, e*Sp*Cas9 #1: 49, e*Sp*Cas9 #2: 48. See text for details.

Figure 6: Chromosomal rearrangements observed at the *ARG2* locus, following *Sp*Cas9 induction

A: *ARG2* and *SUP4* loci drawn to scale. A 2.6 kb piece of DNA containing 1.8 kb of the *SUP4* locus in which a CTG repeat was integrated, as well as the *TRP1* selection marker were integrated at ARG2 (Richard et al., 2003). The *TRP1* gene is not represented here but is centromere-proximal located. **B**: Types of rearrangements observed. Types IV and V are explained in Figure 5. GC: gene conversion with *SUP4*::I-Sce I. Exp.: CTG repeat expansion. **C**: Type V rearrangements involving the *YAK1* gene. The imperfect repeat in *YAK1* and the CTG repeat in *SUP4* are underlined. The junction of the rearrangement contains the green sequence from *YAK1* ligated to the red sequence from *SUP4*::(CTG)n.

Figure 7: Quantification of double-strand break resection

A: Couples of primers used to amplify each EcoRV site are indicated above and can be found in Supplemental Table 2. EcoRV sites are shown by vertical arrows. Resection graphs are plotted for each primer pair. Average relative values of resection as compared to the total DSB amount detected on Southern blots are shown at 6 hours (in blue) and 8 hours (in red), along with standard deviations. **B**: Mechanistic model for chromosomal rearrangements following a Cas9-induced DSB. See text for details. Resulting rearrangement types are indicated in red near each pathway.

Supplemental Figure 1: Sequences at the left and right of junctions in rearranged haploid clones. Junctions were deduced from Illumina read mapping (when available) and confirmed by subsequent PCR and Sanger sequencing. Nucleotides in red are those used to anneal each DSB end, and are therefore present in only one copy in the genomic sequence. The extent of calculated deletions (Δ) is indicated in parentheses. Nucleotides in red in parentheses correspond to small deletions. Nucleotides in green correspond to insertions. The length of Ty insertions is indicated along with the LTR it comes from. Nucleotides in purple (*Sp*Cas9 at the *ARG2* locus) correspond to the I-*Sce* I site (see text). Nucleotides in light blue correspond to homeologies between the left and right junction sequences that were lost after rearrangement (the junction sequence shows the nucleotide in blue, not the one in red). Nucleotides in light blue in parentheses correspond to homeologies that were deleted during the rearrangement. Note that extended homologies between LTRs does not always allow to determine the exact breakpoint with a high precision.

Supplemental Figure 2: Genome- wide mutation spectrum observed in haploid and diploid cells following Cas9 induction

A: Real-time PCR quantification of *CDC8* and *JEM1* amounts relative to an internal control on chromosome IV, in diploid cells in which Cas9 was induced. Half the amount of *CDC8* product was detected in each clone analyzed. This was significantly different from the amount of product amplified from the *JEM1* gene located on the other chromosome X arm.

B: Illumina results for diploid and haploid cells. For each clone, the number of mutations detected is shown. Substit.: nucleotide substitution; Indel: insertion or deletion; Indel micro.: insertion or deletion of one repeat unit in a microsatellite. The asterisk corresponds to a 36 bp deletion in the *FLO11* minisatellite (36 bp repeat).

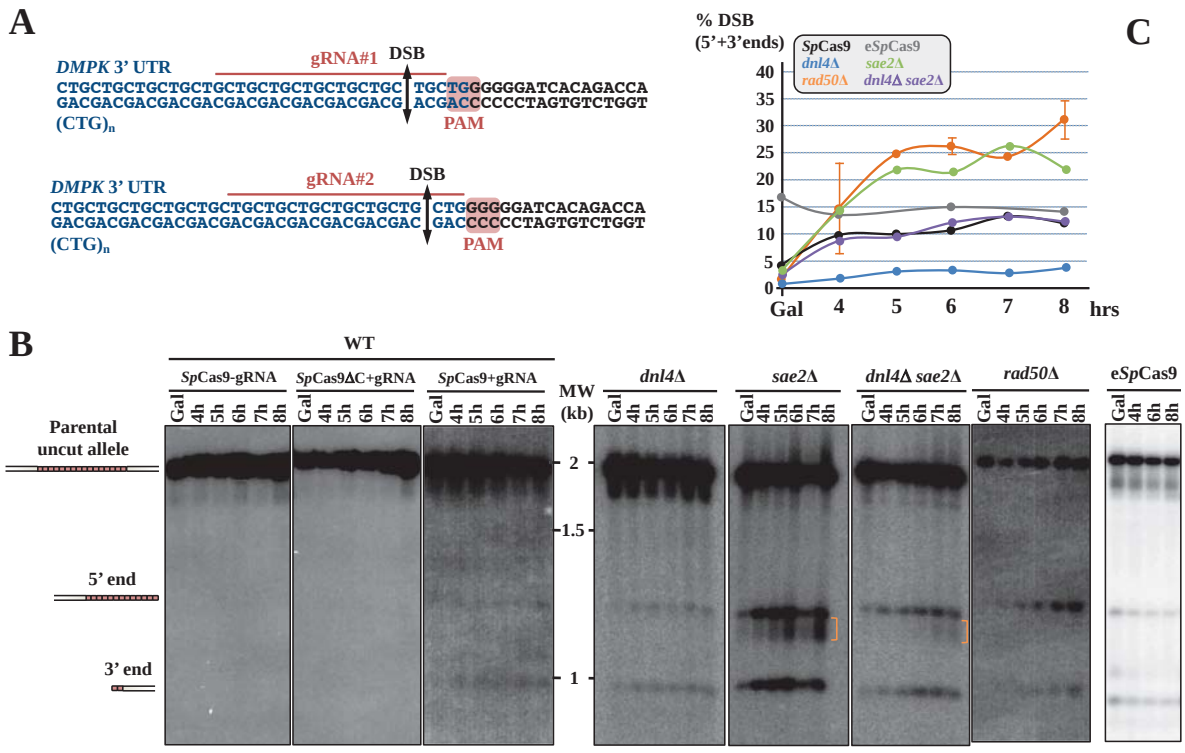


Figure 1
Mosbach *et al.*

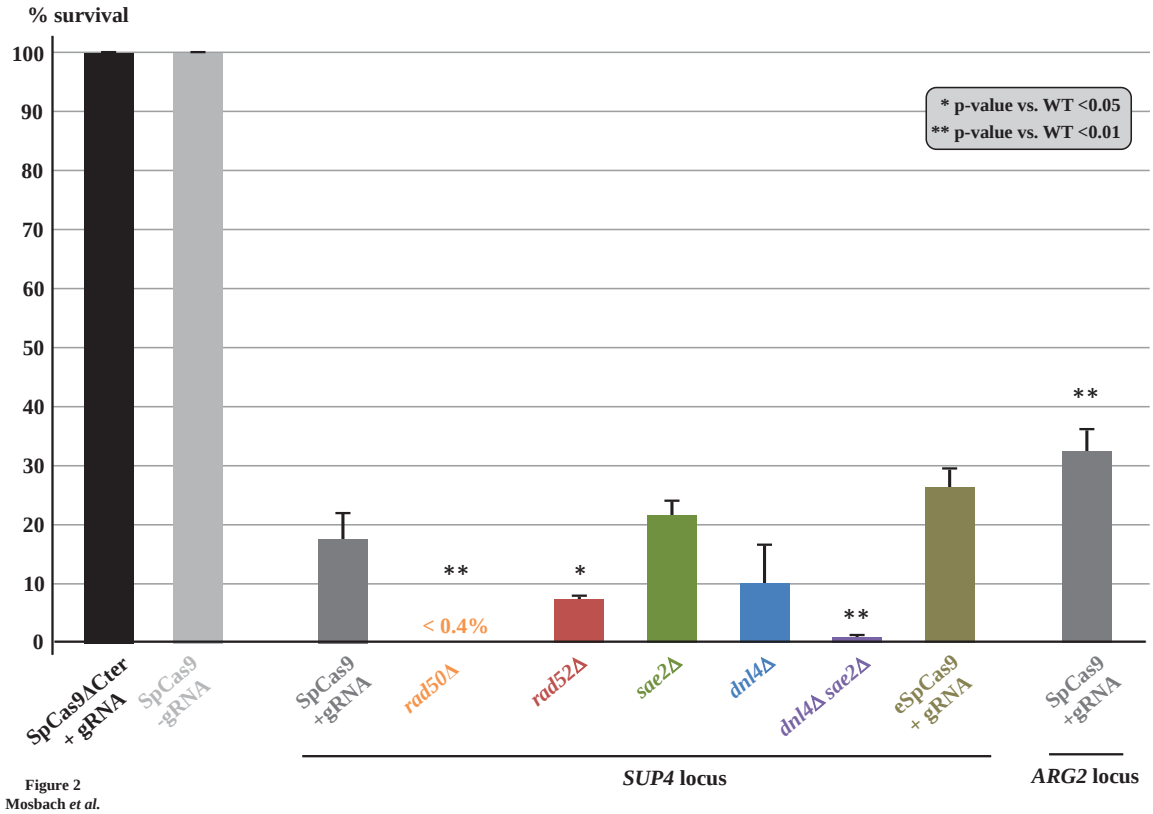


Figure 2
Mosbach *et al.*

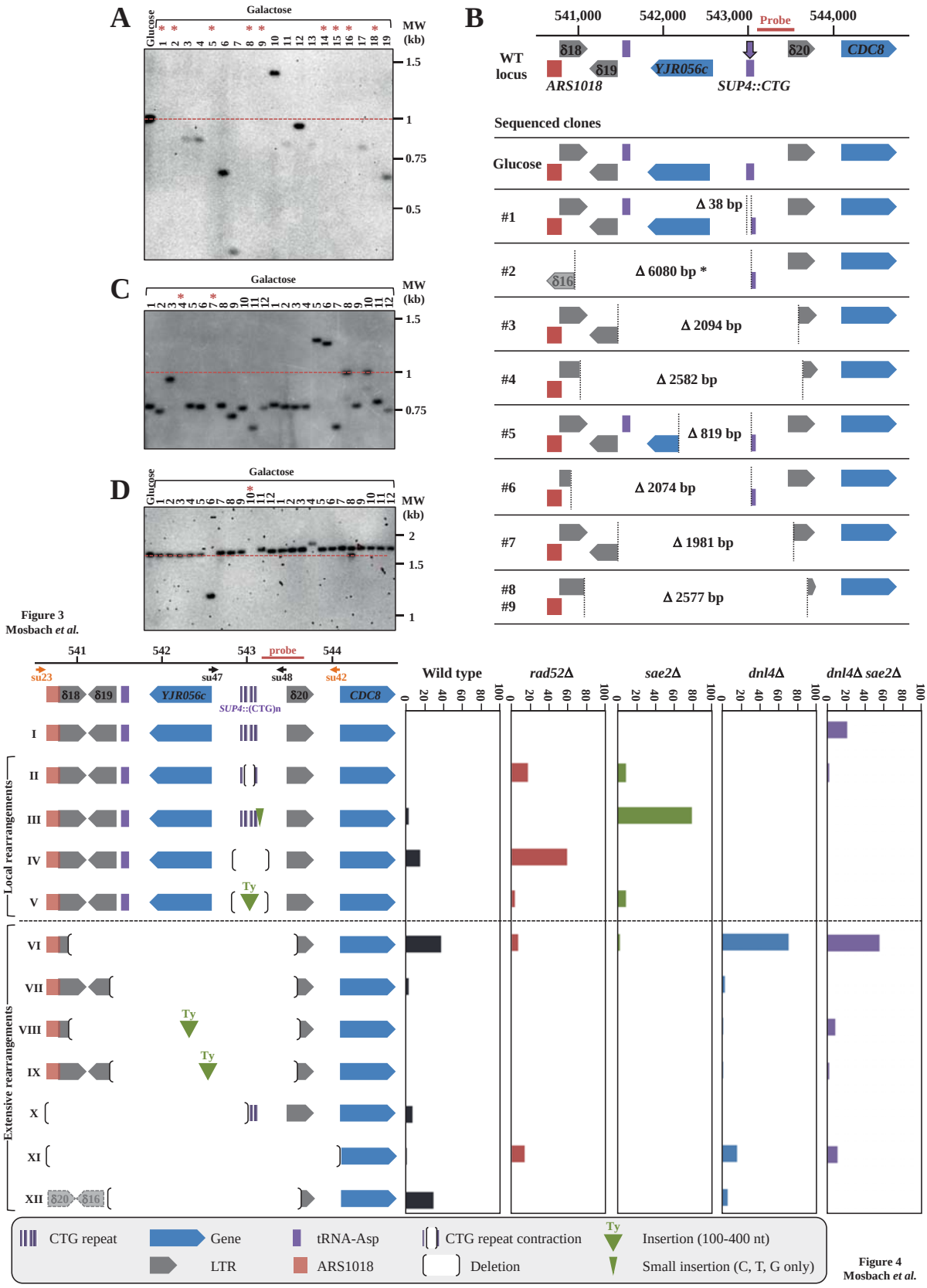


Figure 3
Mosbach *et al.*

Figure 4
Mosbach *et al.*

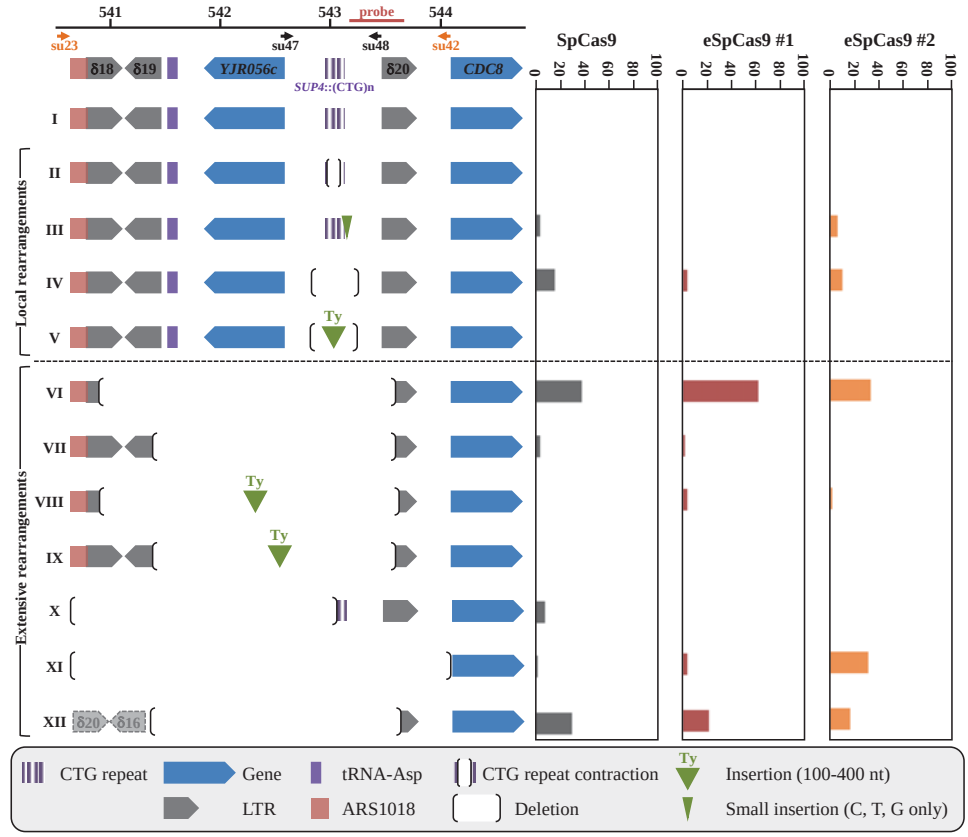


Figure 5
Mosbach *et al.*

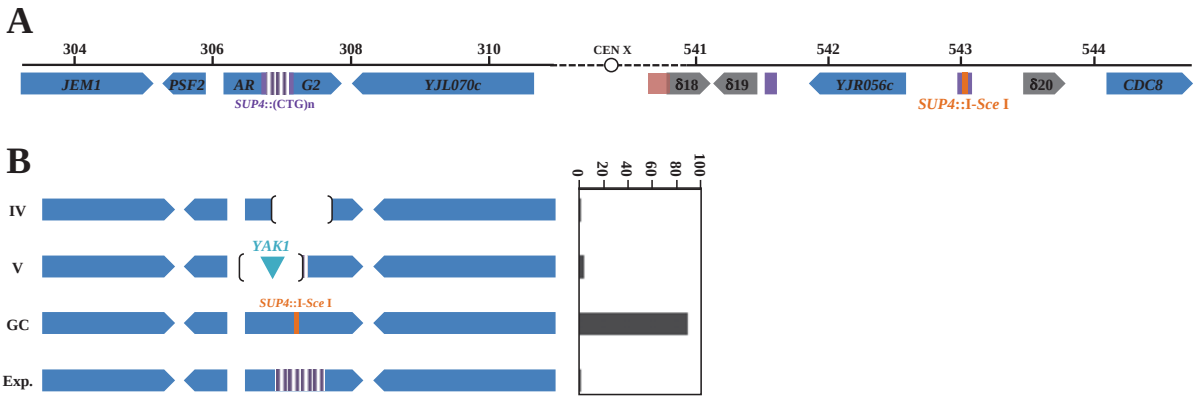


Figure 6
Mosbach *et al.*

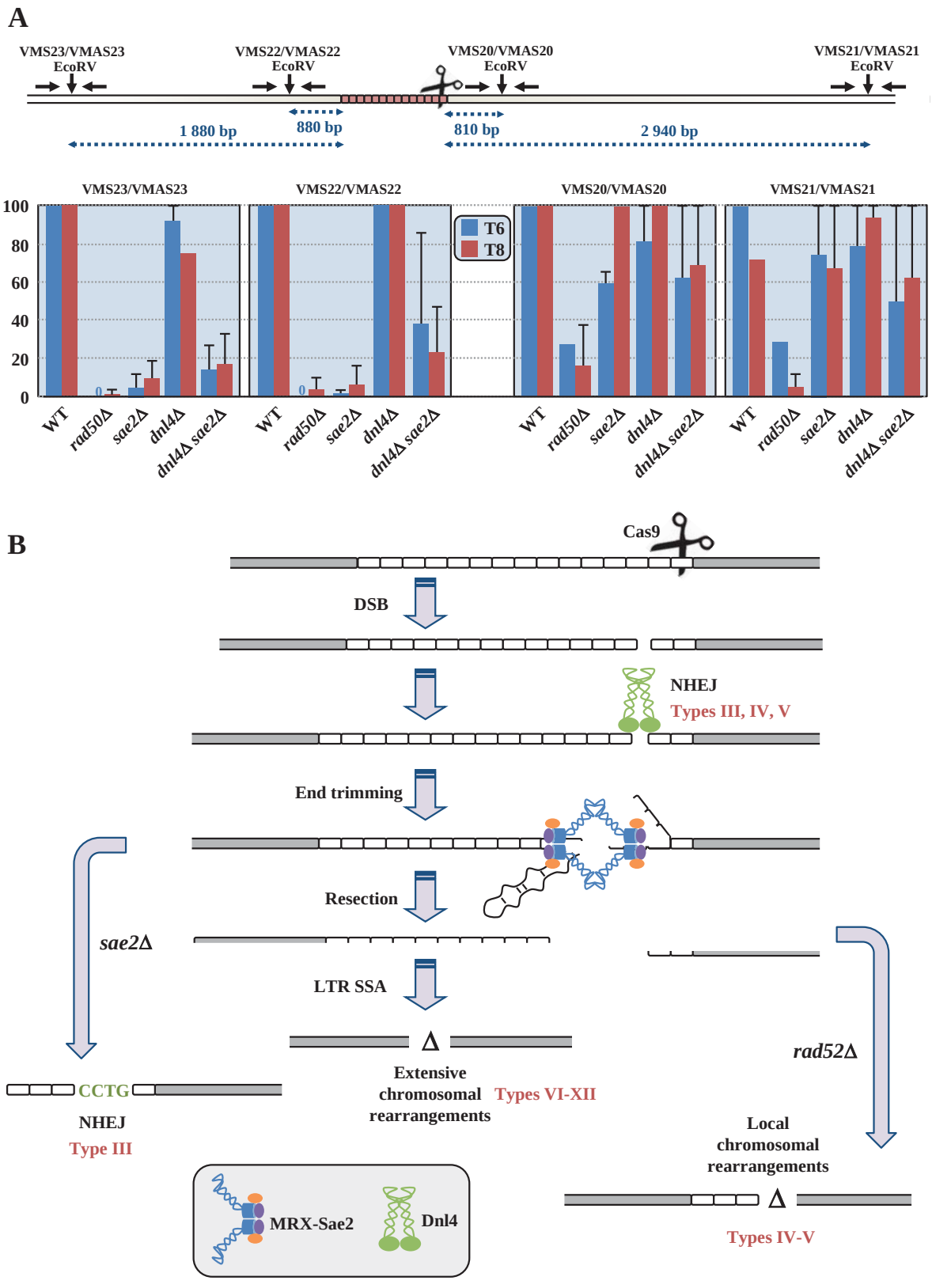
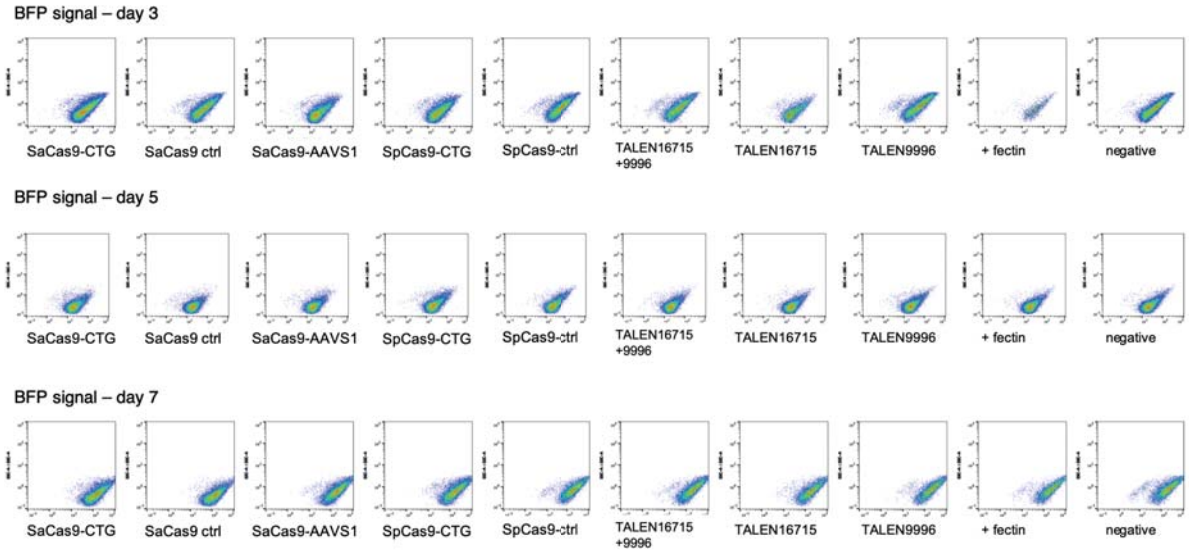


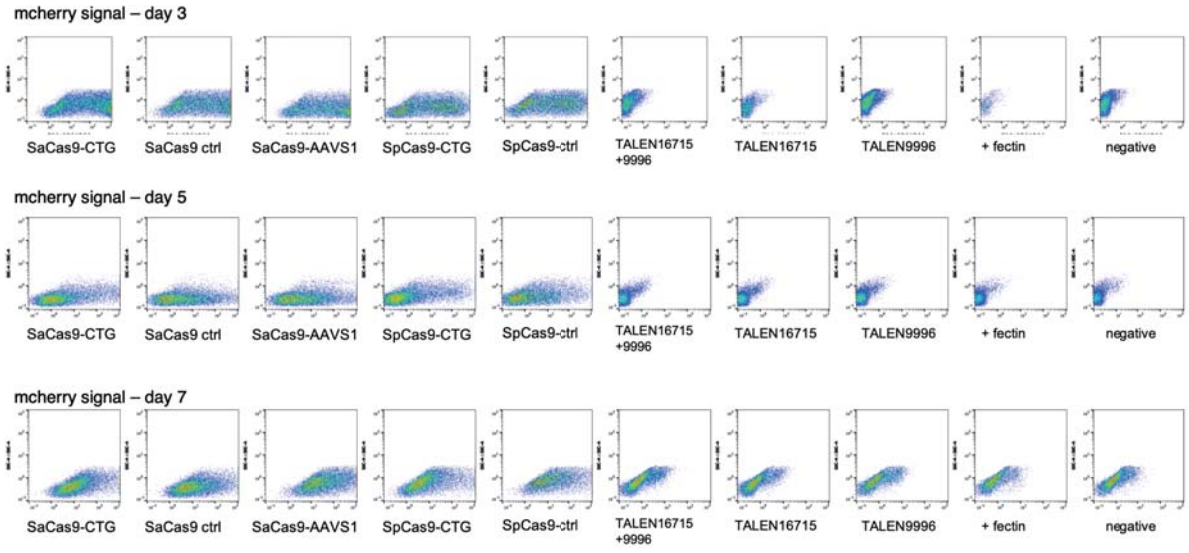
Figure 7
Mosbach *et al.*

Annex 5: Dot plots of transfected L320 clone with different nucleases over time, at days 3, 5 and 7 after transfection of the nuclease

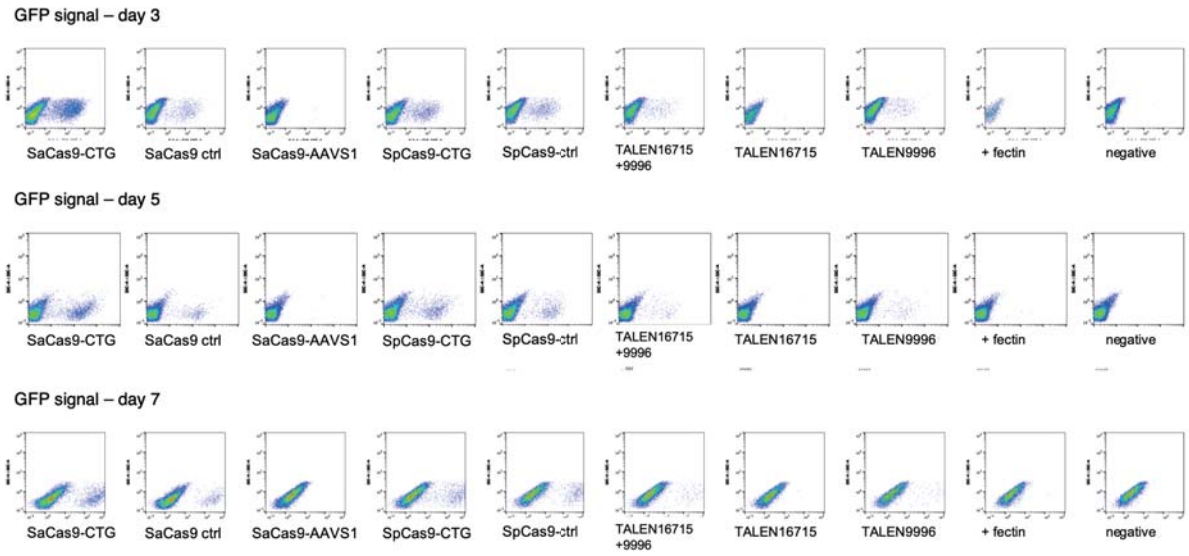
SaCas9 and SpCas9 were expressed on plasmids encoding an mcherry gene which expression was linked by T2A to Cas9 expression. TALEN expression vector consisted of vectors used for AAV production, respectively pAAV16715 and pAAV9996.



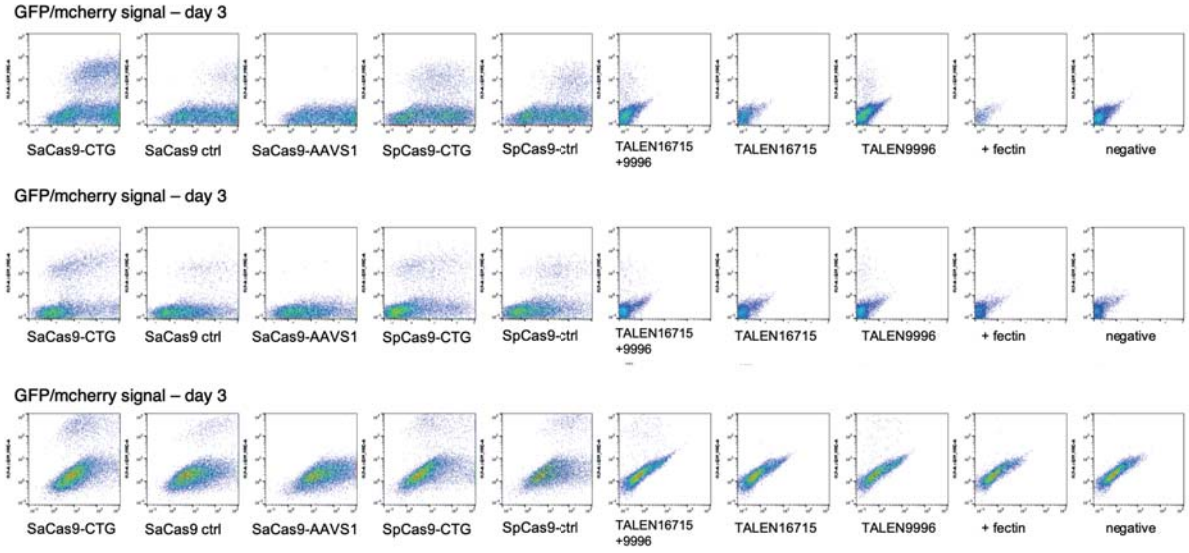
(x-axis=BFP, y-axis=SSC-A)



(x-axis=mcherry, y-axis=SSC-A)



(x-axis=GFP, y-axis=SSC-A)



(x-axis=mcherry, y-axis=GFP)

Abstract:
Microsatellite disorders are a specific class of human diseases that are due to the expansion of repeated sequences above pathological thresholds. These disorders have varying symptoms and pathogenic mechanisms, caused by the expanded repeat. No cure exists for any of these dramatic conditions. This thesis is investigating new gene editing approaches to remove pathological expansions in the human genome. In a first part, a yeast-based screen was constructed to identify potent CRISPR-associated nucleases that can cut these microsatellites. The second part focuses on myotonic dystrophy type 1 (DM1), which is due to and expanded CTG repeat tract located at the 3'UTR of the DMKP gene. A nuclease, $TALEN_{CTG}$ was designed to induce a double strand break into the CTG repeats. It was previously shown to be active in yeast cells, inducing contractions of CTG repeats from a DM1 patient integrated into the yeast genome. The TALEN was tested in DM1 patient cells. The nuclease was found to trigger some contraction events in patient cells. *In vivo* experiments were carried out in a mouse model of myotonic dystrophy type 1 containing a human genomic fragment from a patient and 1000 CTG. Intramuscular injections of recombinant AAV encoding the $TALEN_{CTG}$ revealed that the nuclease is toxic and/or immunogenic in muscle cells in the tested experimental conditions. Finally, the reporter assay integrated in yeast to screen nucleases was transposed in HEK293FS cell line. The integrated cassette contains a CTG expansion from a myotonic dystrophy type 1 patient flanked by two halves of GFP genes. This system would enable to find nucleases active in human cells.

Keywords: [microsatellite disorders ; gene editing ; gene therapy ; myotonic dystrophy type 1]

[Edition de génome pour de nouvelles approches de thérapie génique des maladies à triplets]

Les maladies à triplet sont dues à des expansions de trinucléotides dans l'ADN. Aucun traitement n'existe pour les soigner. Le but de cette thèse est de mettre au point de nouvelles approches de thérapie génique pour supprimer les expansions pathologiques dans le génome humain. Dans une première partie, un système expérimental dans la levure a été construit afin d'évaluer l'efficacité de différentes nucléases associées au système CRISPR sur des microsatellites. La seconde partie est concentrée sur une maladie à triplet en particulier ; la dystrophie myotonique de type 1 (DM1), qui est due à une expansion d'une répétition de triplets CTG dans la région 3'UTR du gène DMPK. Une nucléase, $TALEN_{CTG}$, construite pour induire une cassure double-brin dans les répétitions CTG en 3'UTR du gène DMPK, induit de manière très efficace des contractions de triplets CTG dans la levure. Je me suis intéressée à l'effet de cette TALEN dans des cellules de patient atteint de dystrophie myotonique de type 1. Des événements de contraction ont été observés lorsque cette nucléase est exprimée. Des expériences *in vivo* dans un modèle de souris contenant un fragment d'ADN génomique humain de patient contenant 1000 CTG ont été menées. Des particules virales AAV recombinantes portant le gène de la TALEN ont été produites. Après injection intramusculaire, les cellules musculaires expriment la nucléase, mais dû à une toxicité ou immunogénicité de la protéine, l'expression est perdue. Enfin, le système mis au point dans la levure a été transposé dans une lignée cellulaire humaine établie, les HEK293FS. La cassette introduite contient 200 triplets CTG d'un patient flanqué de deux moitiés de GFP. Ce système pourra servir à sélectionner des nucléases actives dans les cellules humaines.

Mots clés : [maladies à triplet ; édition de génomes ; thérapie génique ; dystrophie myotonique de type 1]