8120–8134 *Nucleic Acids Research, 2021, Vol. 49, No. 14* *Published online 7 July 2021*
https://doi.org/10.1093/nar/gkab569

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

Lucie Poggi[1,2,3], Lisa Emmenegger[1], Stéphane Descorps-Declère[1,4], Bruno Dumas[3] and 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 and [4]Institut Pasteur, Bioinformatics and Biostatistics Hub, Department of Computational Biology, USR3756 CNRS, F-75015 Paris, France

Received December 17, 2019; Revised June 11, 2021; Editorial Decision June 11, 2021; Accepted July 06, 2021

ABSTRACT

Microsatellite expansions are the cause of >20 neurological or developmental human disorders. Shortening expanded repeats using specific DNA endonucleases may be envisioned as a gene editing approach. Here, we measured the efficacy of several CRISPR–Cas nucleases to induce recombination within disease-related microsatellites, in *Saccharomyces cerevisiae*. Broad variations in nuclease performances were detected on all repeat tracts. Wild-type *Streptococcus pyogenes* Cas9 (*Sp*Cas9) was more efficient than *Staphylococcus aureus* Cas9 on all repeats tested, except $(CAG)_{33}$. Cas12a (Cpf1) was the most efficient on GAA trinucleotide repeats, whereas GC-rich repeats were more efficiently cut by *Sp*Cas9. The main genetic factor underlying Cas efficacy was the propensity of the recognition part of the sgRNA to form a stable secondary structure, independently of its structural part. This suggests that such structures form *in vivo* and interfere with sgRNA metabolism. The yeast genome contains 221 natural CAG/CTG and GAA/CTT trinucleotide repeats. Deep sequencing after nuclease induction identified three of them as carrying statistically significant low frequency mutations, corresponding to SpCas9 off-target double-strand breaks.

INTRODUCTION

A growing number of neurological disorders were identified to be linked to microsatellite expansions (1). Each disease is associated with 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 (2). Indeed, when a trinucleotide repeat contraction occurred during transmission from father to daughter of an expanded myotonic dystrophy type 1 (DM1) allele, clinical examination of the daughter showed no sign of the disease (3,4).

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 (5). Later on, ZFNs were used to induce DSBs into CAG or CTG repeats, which mostly led to contractions in CHO cells (6) and in a HEK293 cell GFP reporter assay (7). 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.

More recently, a TALEN was designed to recognize and cut an expanded CTG triplet repeat from a DM1 patient. It was very efficient at shortening it in yeast cells (>99% cells showed contraction) and highly specific as no other mutation was detected (8). The TALEN was shown to induce specific repeat contractions through single-strand annealing (SSA) by a *RAD52*, *RAD50* and *SAE2* dependent mechanism (9).

*To whom correspondence should be addressed. Tel: +33 1 45 68 84 36; Email: gfrichar@pasteur.fr
Present addresses:
Lucie Poggi, Institut Imagine, 24 boulevard du Montparnasse, F-75015 Paris, France.
Lisa Emmenegger, Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association, Robert-Rössle-Strasse 10, 13125 Berlin, Germany.

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	200–1700
$(CCTG)_n$	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	700–1600

The CRISPR–Cas system is the easiest to manipulate and to target any locus, as sequence recognition is based on the complementarity to a single guide RNA (sgRNA). This sgRNA is made of the fusion between a 20 or 21 nt sequence complementary to the target DNA (crRNA) and the trans-activating RNA (tracrRNA), serving as a scaffold for the Cas nuclease (10). 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) (11). Its 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 (12). Similarly, Cas9-HF1 was mutated on 4 residues interacting through hydrogen bonds with the target strand (13). *Staphylococcus aureus* harbors a smaller Cas9, its PAM is NNGRRT, having a similar structure to SpCas9 with two catalytic sites. Finally, type V CRISPR–Cas nucleases, like Cas12a (Cpf1), 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 (14). In the present experiments, the *Francisella novocida* gene (FnCpf1), whose PAM is TTN, was used (15).

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 with 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 crRNA stability, strongly suggesting that RNA 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 genome wide and found that three microsatellites with similar sequences were also edited by the nuclease. The mutation pattern was different depending on the microsatellite targeted.

MATERIALS AND 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 sequence (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 (16), 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 (17).These plasmids were cloned directly into yeast cells by homology-driven recombination (18) 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 (19). sgRNAs for SpCas9 (and variants) were ordered from ThermoFisher (GeneArt), flanked by *Eco* RI sites for subsequent cloning into pRS416 (16). FnCpf1 sgRNAs were ordered at Twist Biosciences, directly cloned into pRS416 (see Supplemental Table S1 for plasmid names). For SaCas9, GCCT, GGGGCC and CGG sgRNA were ordered at Eurogentec, the seven other guides come from Twist Biosciences. All sgRNAs were cloned in pRS416 at SpeI and XhoI sites. Each sgRNA was synthetized under the control of the *SNR52* promoter. All crRNA sequences can be found in Supplemental Table S3.

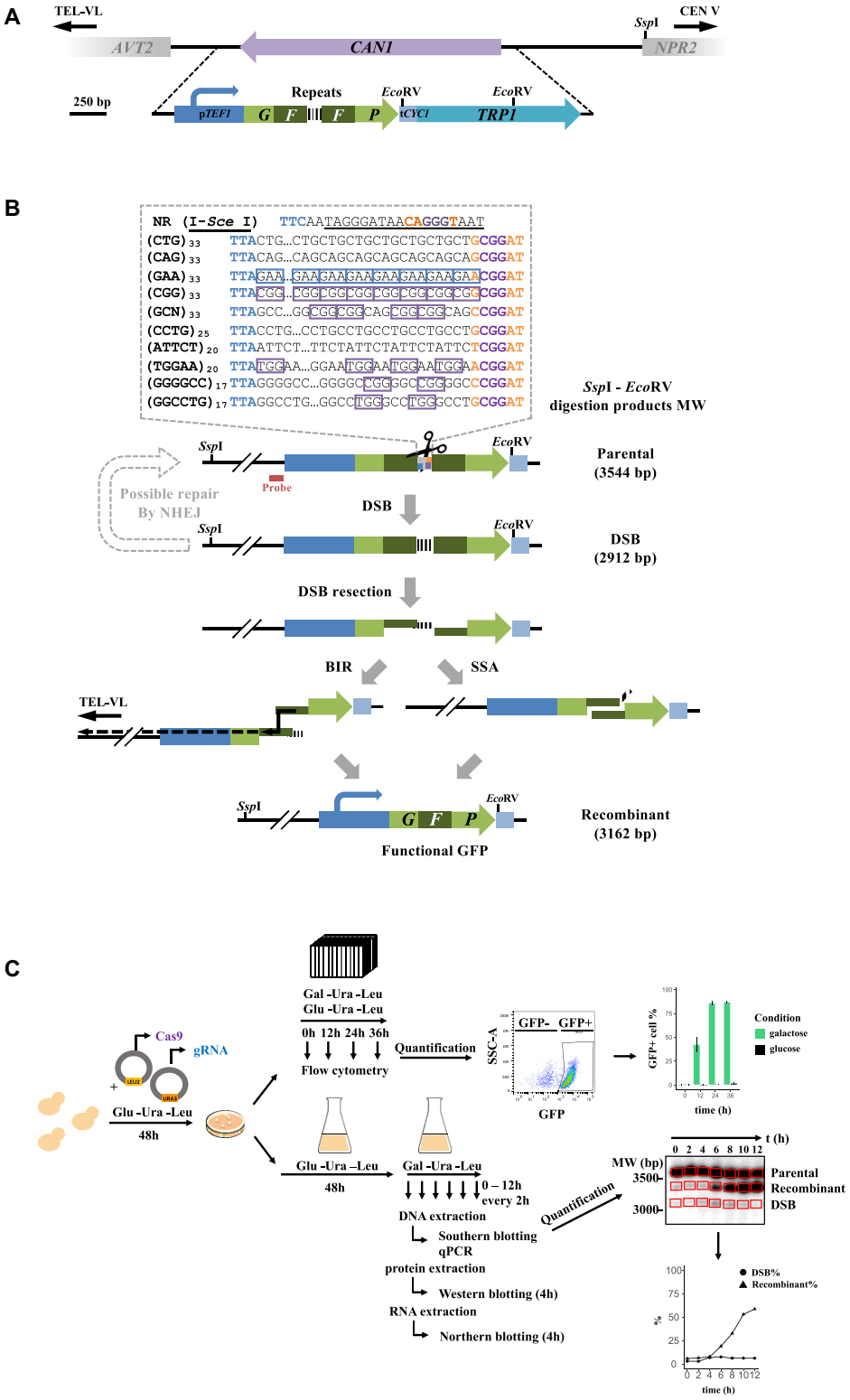


Figure 1. GFP reporter assay. (**A**) The *CAN1* locus was replaced by recombinant GFP cassettes. Each synthetic GFP contains the constitutive *TEF1* promoter, followed by the bipartite eGFP interrupted by one of the ten microsatellites or by the I-*Sce* I recognition sequence and the *CYC1* terminator. The *TRP1* gene was used as an auxotrophic marker. (**B**) Upon double-strand break induction, cells may repair by single-strand annealing or 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). FACS dot-plot axes are FSC-A/SSC-A.

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 (20). 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 S4).

Yeast mutants

The three deletions (*rad51*Δ, *pol32*Δ and *dnl4*Δ) were built as follows. DNA was extracted from strains VMY104, VMY432 and VMY551, previously described (9) (Supplemental Table S4). The KANMX deletion cassette was amplified using dedicated primers located upstream and downstream (Supplemental Table S2). PCR products were transformed into LPY111 and selected on G418 plates. Three to five transformants were checked by PCR for correct integrations.

Flow cytometry assay

Cells were transformed using standard lithium-acetate protocol (20) with both sgRNA and nuclease and selected on 2% glucose SC -URA -LEU plates and grown for 36 h. 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 (0, 12, 24, 36 h) 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 (21).

Time courses of DSB inductions

Cells were transformed using standard lithium-acetate protocol (20) with both sgRNA and nuclease and selected on 2% glucose SC -URA -LEU plates and grown for 36 h. Each colony was seeded into 2 ml of 2% glucose SC -URA -LEU for 24 h and then diluted into 10 ml of 2% glucose SC -URA -LEU for 24 h as a pre-culture step. Cells were washed twice in water and diluted at ca. 7×10^6 cells/ml in 2% galactose SC-URA -LEU, before being harvested at each time point (0, 2, 4, 6, 8, 10, 12 h) for subsequent DNA extractions. Due to the *pol32*Δ mutant growth defect, each colony was seeded into 2 ml of 2% glucose SC -URA -LEU for 24 h and then diluted into 10 mL of 2% glucose SC -URA -LEU for 48 h as a pre-culture step. Cells were washed twice in water and diluted at ca. 7×10^6 cells/ml in 2% galactose SC-URA -LEU, before being harvested at each time point (0, 24, 48, 72, 96, 120 h) for DNA extractions. The same cultures were used for cytometry analyses.

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 302 bp ^{32}P-randomly labeled *CAN1* probe amplified from primers CAN133 and CAN135 (Supplemental Table S2) (22). Each probe was purified on a G50 column (ProbeQuant G50 microcolumn, GE Healthcare) and specific activities were verified to be above 2.4×10^8 cpm/μg. 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, after background subtraction.

Agarose plug DNA preparation

During time courses of DSB induction (see above), 2×10^9 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 min. 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, plug-containing 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). Incubation was performed 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 microtubes for 1 h at 4°C, repeated four times. TE was replaced by 1 ml restriction enzyme buffer (Invitrogen REACT 2) for 1 h, 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 min without removing the buffer, 100 units of each enzyme *Eco* RV and *Ssp* I was added and left at 37°C for 1 h. Then, 2 μl of β-agarase (NEB M0392S) and 2 μl of RNAse A (Roche 1 119 915) were added and left for 1 h at 37°C. Microtubes were centrifuged at maximum speed in a tabletop centrifuge for 1 min 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 sgRNA and nucleases was grown for 4 h in 2% galactose SC-URA-LEU. Total RNAs were extracted using standard phenol–chloroform procedure (23) 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 1 h. 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, purified on a Sephadex G25 column (MicroSpin G25 column, GE Healthcare) and its specific activity was verified to be at least 1.2×10^8 cpm/μg, and denatured (5′ at 95°C) before hybridization.

Western blot analyses

Total proteins were extracted in 2× 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 1 h 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 min 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 $t = 12$ h 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 (24). 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

Cells were grown overnight in YPGal medium and diluted for two more hours. Cells were incubated in 20 ml 0.1 M of lithium acetate/TE buffer for 45 min at 30°C. 500 μl of 1M DTT was added and cells were incubated for a further 15 min at the same temperature. Cells were washed in water, then in 1 M sorbitol and resuspended in 120 μl ice-cold 1 M sorbitol. Then, 40 μl of competent cells were mixed with 150 ng of sgRNA-expressing plasmid, 300 ng of nuclease-expressing plasmid and 100 μM of dsODN (5′-P G*T*TTAATTGAGTTGTCATAT GTTAATAACGGT*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 h. 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 (6 × 16 mm) using the following setup: peak incident power: 105 W, duty factor: 5%, 200 cycles, 80 s. 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 30 s, 50°C for 30 s, 72°C for 30 s repeated 30 times; followed by 72°C for 7 min). A second round of PCR was performed using primers PE1 and GSP2-PE2 (First denaturation step: 98°C for 30 s; 98°C for 30 s, 65°C for 30 s, 72°C for 30 s repeated 30 times; followed by 72°C for 7 min) A final round of PCR was performed using primers PE1 and PE2 (First denaturation step: 98°C for 30 s; 98°C for 30 s, 65°C for 30 s, 72°C for 30 s repeated 15 times; followed by 72°C for 7 min). 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 (25) using '-ax sr –secondary = no' parameters. Sam files resulting from mappings were then all sorted and in-

dexed by the samtools software suite (26). 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 (27). 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 nucleic acid sequences

The Gibbs free energy formation for crRNA, sgRNA and DNA secondary structures were determined using the mfold RNA 2.3 (or mfold DNA) (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 sgRNA 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 crRNA, sgRNA or DNA 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.

RESULTS

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

The goal of the present experiments was to determine efficacy and specificity of different Cas nucleases on various microsatellites, using a yeast recombination reporter assay. 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 chromatin 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, six out of ten also harbor internal PAM sequences (Figure 1B).

All experiments were performed as follows: independent yeast colonies expressing each Cas nuclease and its cognate sgRNA were picked from glucose plates and seeded in 96-deep well plates for flow cytometry measurements over a 36-h time period, either in glucose or galactose medium. 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 (subsequently called NR, for Non Repeated).

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-h 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).

Upon DSB induction, haploid yeast cells may fix the break by three different pathways (Figure 1B). 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. Alternatively, the DSB may be repaired by non-homologous end-joining (NHEJ) between the two DNA ends. However, this is unlikely, since NHEJ is much less efficient than homologous recombination in budding yeast. Perfectly religated DSB ends could be recut by the nuclease, until a functional GFP could be reconstituted by homologous recombination. In order to discriminate between the different pathways used to repair this DSB, we built *rad51*Δ, *pol32*Δ and *dnl4*Δ deleted strains. *DNL4* encodes the DNA Ligase IV protein, responsible for all end-joining religations, *POL32* encodes a polymerase δ subunit involved in all long-range homologous recombination events such as BIR, and *RAD51* encodes the yeast recombinase protein (RecA homologue). These mutations were individually introduced into the GFP-NR containing strain (LPY111). The *rad51*Δ strain showed reduced levels of DSB and GFP-positive cells as compared to wild type, 12 h after induction (Supplemental Figure S1A). But at later time points, amounts of GFP-positive cells are comparable to wild type. This shows that homologous recombination between sister chromatids is not a preferred pathway to generate GFP-positive cells in our experimental assay. This is not unexpected, since both chromatids should be cut by the nuclease and therefore no intact template should be available to repair the DSB. The same pattern was observed for the *dnl4*Δ mutant (Supplemental Figure S1B). This shows that NHEJ is not a preferred pathway

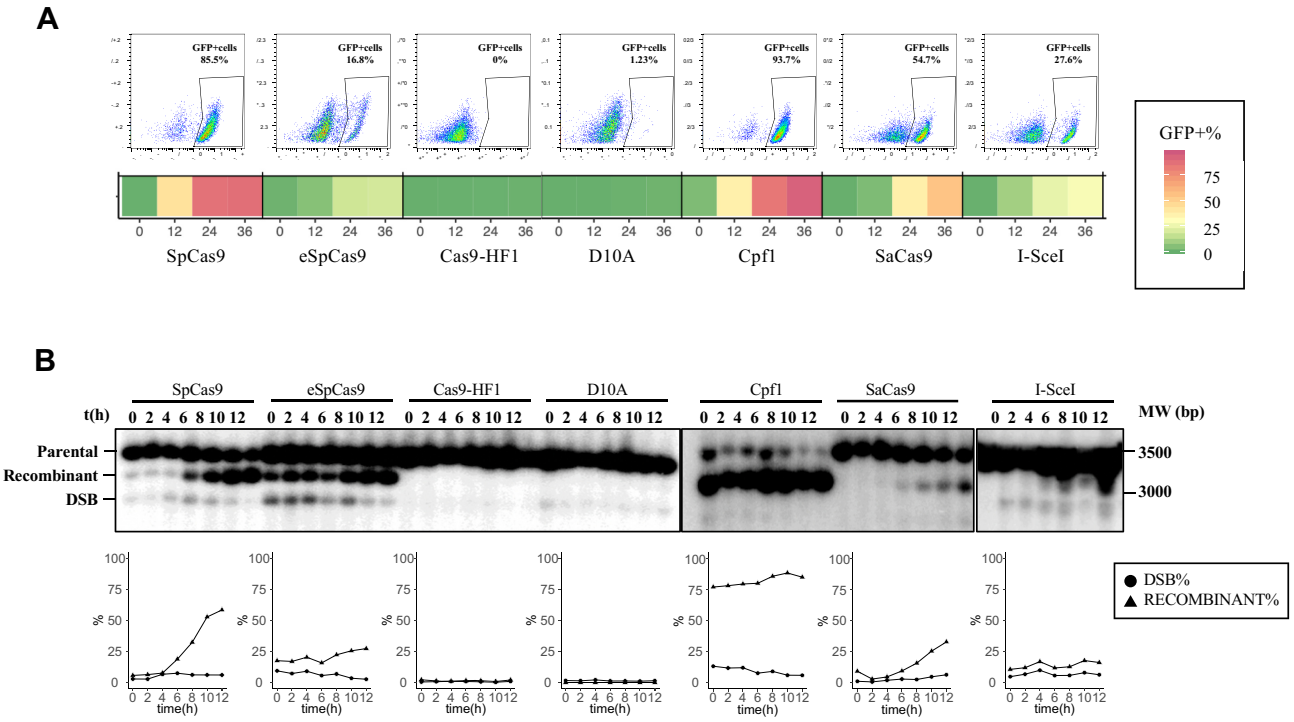


Figure 2. CRISPR–Cas nuclease induction on non-repeated sequence containing an I-*Sce* I recognition site. (**A**) Top: Percentage of GFP+ cells was measured throughout a time course of 36 h. Dot plots indicate final populations at 36 h. X-axis: FITC, Y-axis: SSC. Bottom: GFP+ cells are represented by 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 h. Parental (3500 bp), recombinant (3100 bp) and DSB (2900 bp) products were quantified, as described in Materials and Methods. Bottom: DSB and recombinant products are represented as a percentage of the total signal in each lane.

to repair the DSB. The *pol32*Δ strain exhibited a known growth defect (28). After 36 h in galactose medium, the wild type strain reached a concentration of 2.7×10^8 cells/ml, whereas it took 120 h for the *pol32*Δ mutant to reach the same concentration. Therefore, experiments performed in this mutant background were made over a five days period (120 h) instead of 36 h, in order to compare results. At identical cell concentrations (36 h for WT, 120 hrs for *pol32*Δ), 72.7% (±8.4%) of cells were GFP-positive, as compared to 87% of cells in the wild type. This shows that ∼14% of GFP-positive cells are the product of a *POL32*-dependent pathway, most probably BIR (Supplemental Figure S1C). Given that *DNL4* and *RAD51* have no effect on the generation of GFP-positive cells, and *POL32* only a limited effect, we therefore concluded that most cells must repair the DSB using the SSA pathway.

Streptococcus pyogenes Cas9 variants exhibit a wide range of efficacies

Once the experimental setup was optimized with the NR sequence, the same exact assay was performed using seven different nucleases on ten microsatellites, including tri-, tetra-, penta- and hexanucleotide repeats (Supplemental Figures S2 and S3 and Supplemental Table S5). SpCas9 was able to cut every repeated sequence although $(GGCCTG)_{17}$, $(GAA)_{33}$, $(CAG)_{33}$ and $(CTG)_{33}$ were less efficiently cut (Figure 3A). Surprisingly, the G-quadruplex forming sequence GGGGCC was one of the most efficiently cut, although it is supposed to form stable secondary structures *in vitro* (29). 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 than eSpCas9, itself consistently 2-10 times more efficient than Cas9-HF1 (Figure 3A). The NR sequence was also less efficiently cut, confirming the 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. The Tm of a $(CTG)_{25}$ repeat is 58–61°C (depending on the method used for the measurement), and the T_m of a $(CAG)_{25}$ repeat is only 54°C (30). However, to the best of our knowledge, there is no evidence at the present time for formation of such secondary structures in living cells (31). However, given that we found opposite results with CAG and CTG repeat tracts, it is possible that such secondary structures occur *in vivo*. Since eSpCas9 exhibits 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.

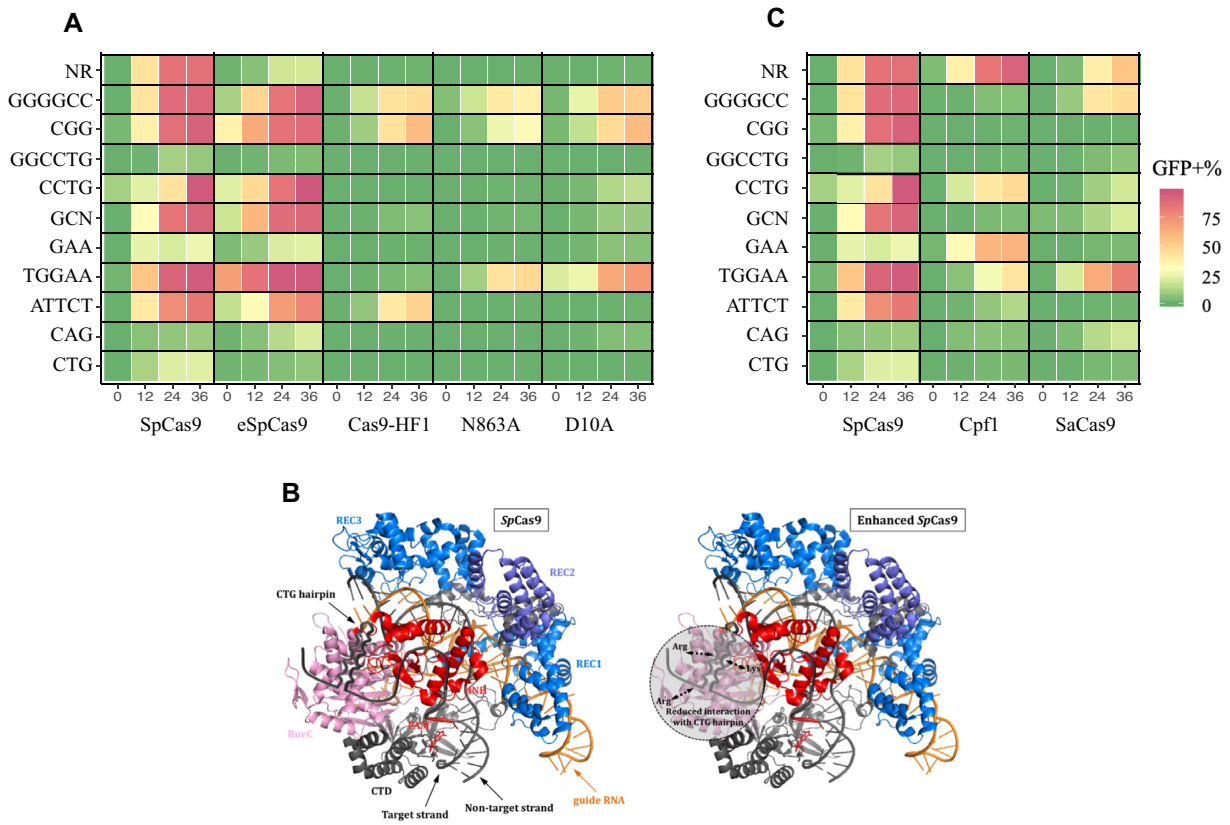


Figure 3. GFP-positive cells after DSB repair. (**A**) SpCas9 and variants. NR: I-*Sce*I recognition site. Each microsatellite is shown on a horizontal line and called by its sequence motif. Recombination efficacies are indicated by the same color code as in Figure 2A. (**B**) Reconstructed models of 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.

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).

SaCas9 belongs to the same structural family as SpCas9, but shows reduced efficacy on all repeat tracts tested, except $(CAG)_{33}$. Given that SpCas9 and SaCas9 share very similar biochemical structures, this increase may come from the SaCas9 sgRNA structure, that might increase interactions with the CTG target strand.

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-h time periods for each nuclease-repeat couple. GFP-positive cell percentage at 12, 24 or 36 h was plotted as a function of cumulative DSB over 12 h (Figure 4). Given the number of different strains and nucleases tested, only one time course was performed in each condition (Supplemental Figure S4). 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 h and the total signal of DSB accumulated during the same time period was found (linear regression test P-value $= 4.7 \times 10^{-13}$, $R^2 = 0.67$) (Figure 4A). A good linear correlation was also found at later time points, 24 h (P-value $= 3.3 \times 10^{-11}$, $R^2 = 0.60$) and 36 h (P-value $= 1.6 \times 10^{-9}$, $R^2 = 0.53$) (Figure 4B, C). 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.

sgRNA 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 sgRNA 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 S5A, top). For SpCas9, sgRNA levels were different among the 10 strains. In one strain, $(GCN)_{33}$, smaller species were detected, around 75 nt, that may correspond to degradation or abortive transcription. No correlation was found between

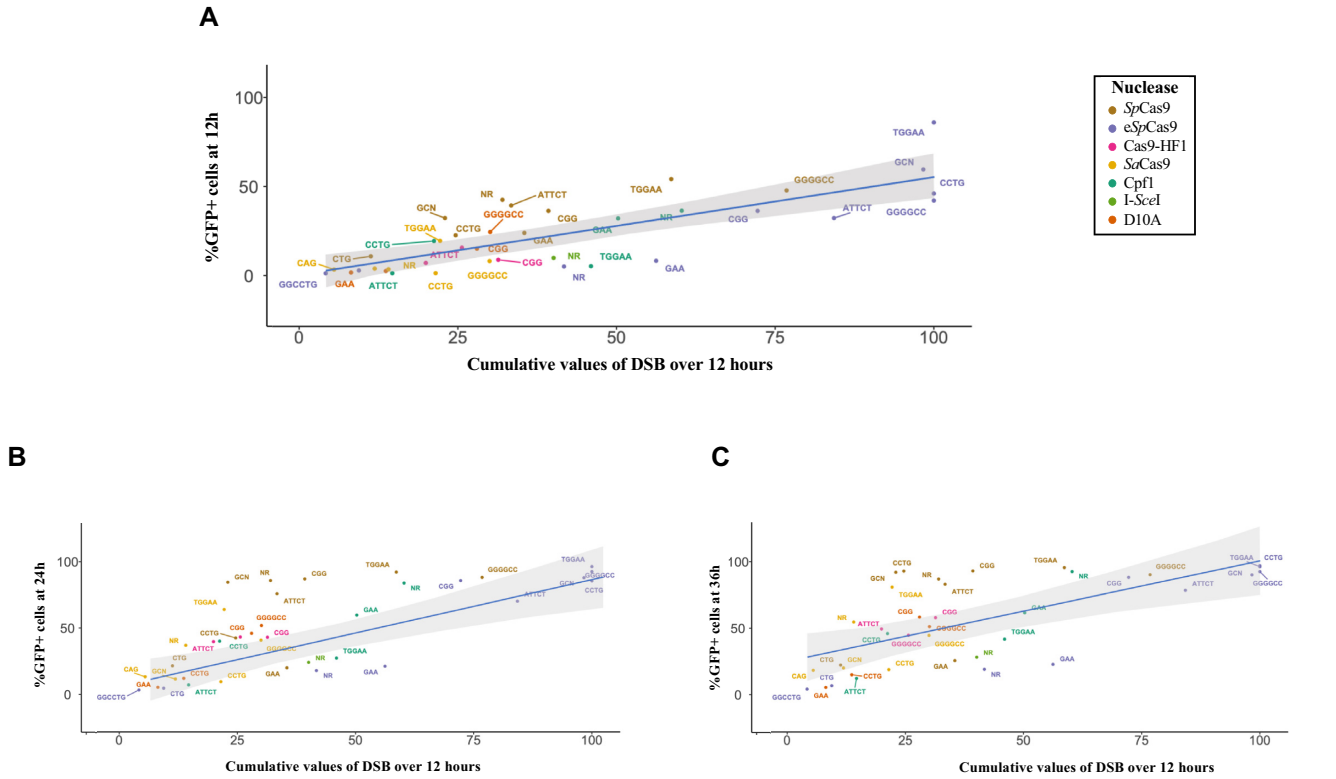


Figure 4. GFP-positive cell percentages as a function of DSB. Correlation between cumulative DSB level at 12 h and GFP-positive cell percentage at 12, 24 and 36 h. Each color represents a nuclease. The blue line corresponds to the linear regression model. In grey: 95% confidence interval of the model. (**A**) After 12 h. (**B**) After 24 h. (**C**) After 36 h.

sgRNA levels and GFP-positive cells, showing that sgRNA steady state level was not the limiting factor in this reaction (Supplemental Figure S5A, bottom). For SaCas9, levels of sgRNAs were lower than SpCas9 and the *SNR44* control RNA. Given that signal over background ratios were low (<2 in all strains except CAG), it was not possible to reliably quantify them (32). It is therefore possible that the lower activity of SaCas9 in all strains is due to a lower expression of its cognate sgRNA, although this does not explain its higher activity on the $(CAG)_{33}$ repeat tract. Using the classical phenol-glass beads protocol and despite numerous attempts, the FnCpf1 sgRNA could not be detected. We hypothesized that it may be so tightly associated with its nuclease that phenol could not extract it. 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 sgRNAs. It is possible that their amounts were too low to be detected by Northern blot. However, given that FnCpf1 efficiently cuts some repeats we consider this an unlikely hypothesis, and favor the possibility that the oligonucleotide used as a probe forms a secondary structure that impedes its proper binding to the target RNA on the membrane. This was not further investigated.

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 S5B, top). 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 S5B, bottom). 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 clearly correlated with sgRNA levels, 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 more abundant than SaCas9 and/or FnCpf1 in our experiments.

crRNA secondary structure stability determines 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 (33–39). Secondary structures are known to form both at

DNA and at RNA levels (40, 41). 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 (31). In our present experiments, secondary structures may possibly form on target DNA. 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 5A, ANOVA test P-value = 0.389) (see Materials & Methods).

Alternatively, secondary structures due to microsatellites may form on the crRNA. We subsequently performed the same calculation for the 20 nt crRNA with or without their cognate tracrRNA. Predicted structures of crRNA are shown in Figure 5B. When tracrRNA scaffolds were taken into account, theoretical Gibbs energies were very low and comparable to each other, except for FnCpf1 sgRNA, which is much smaller than the others and for ATTCT that does not form secondary structure. Therefore, there was no correlation between sgRNA stability and GFP-positive cells (Figure 5C, ANOVA test P-value = 0.822). However, a statistically significant inverse correlation was found between the 20 nt crRNA stability and GFP-positive cells (Figure 5D, ANOVA test P-value = 0.0116). We concluded that despite the great thermal stability of the SpCas9 tracrRNA, secondary structure formation on the crRNA is the main determinant of nuclease efficacy. Gibbs free energies were also calculated at 37°C and the same correlations were observed.

DSB-resection of microsatellites

Resection rate at *Sty* I restriction sites was measured by qPCR as previously described (42,24). 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 h. 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 (Supplemental Figure S6).

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 $(GGGGCC)_{17}$, $(GCC)_{33}$, $(CCTG)_{25}$, $(GCN)_{33}$, $(GAA)_{33}$ and $(TGGAA)_{20}$ repeats. SSBs are repaired by a specific machinery involving Base Excision Repair (BER) (43). 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 S4). 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 some nicks were indeed converted into DSBs *in vivo* (Supplemental Figure S7).

Genome-wide determination of off-target mutations

Microsatellites are very common elements of all eukaryotic genomes and the yeast genome contains 1818 di-, tri- and tetranucleotide repeats (44). 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 relying on NHEJ to detect genome-wide DSBs. NHEJ is less efficient than homologous recombination in yeast, but haploid yeast cells have no other way of repairing a single DSB in a unique region than by religating the two DSB ends. This inefficient mechanism is mutagenic, making frequent small insertions and deletions, and less frequent capture of mitochondrial DNA (45, 46). Therefore, it was decided to adapt the GUIDE-Seq method to budding yeast (see Materials & Methods). Shortly, cells were transformed with SpCas9 or FnCpf1, a sgRNA 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 million reads, only 1.5 million (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 S8).

A threshold was set at twice the median coverage of each library. From 7 to 2103 gene loci whose coverage was above this threshold were identified in each library. These gene loci were compared to predicted off-target loci using the CRISPOR web tool (27). The overlap between both sets of gene loci was very small, suggesting 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

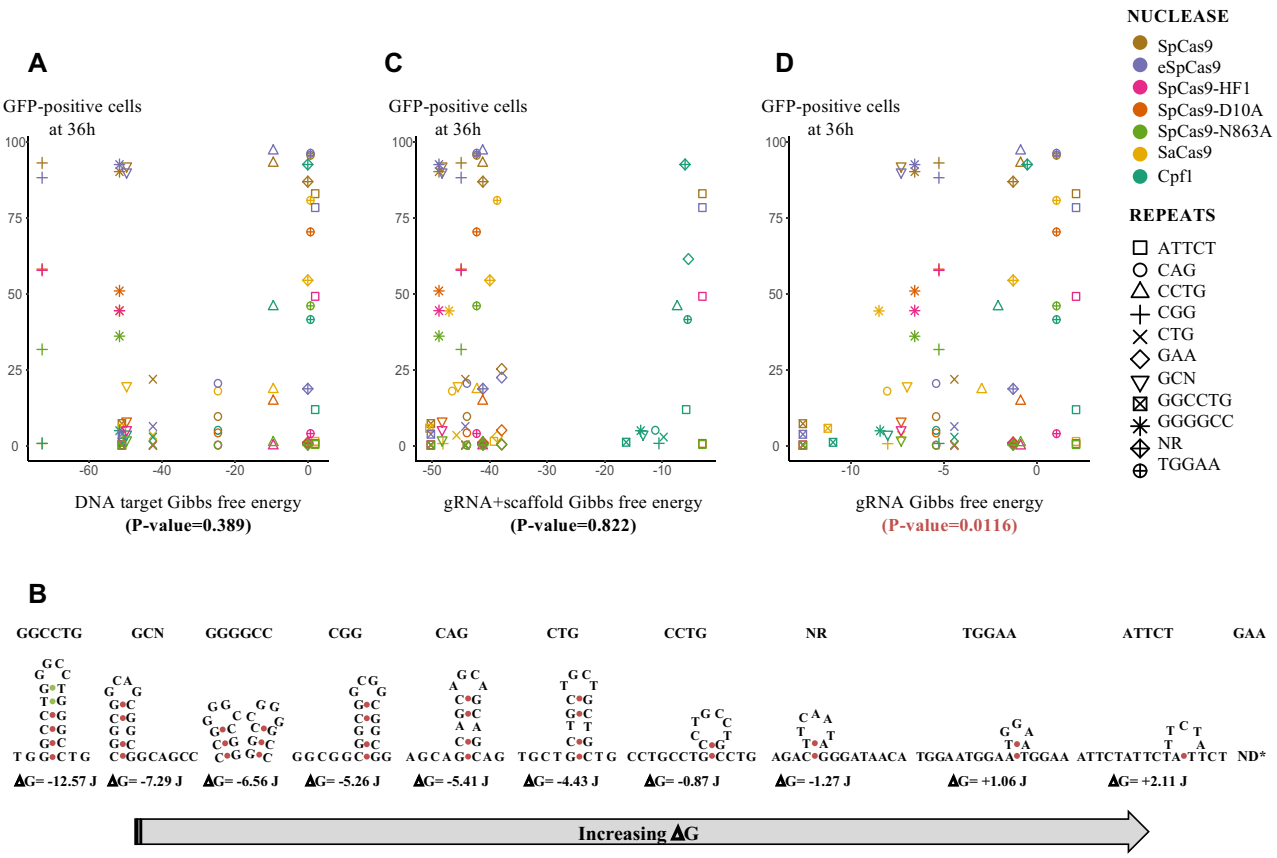


Figure 5. Secondary structures and Gibbs free energy. (**A**) GFP-positive cells as a function of Gibbs energy calculated for each DNA sequence. Each nuclease is represented by a different color code and each microsatellite by a different shape. (**B**) Predicted secondary structure of each SpCas9 crRNA. 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 sgRNA. (**D**) GFP-positive cells as a function of Gibbs energy calculated for each crRNA. *P*-values are given below each graph.

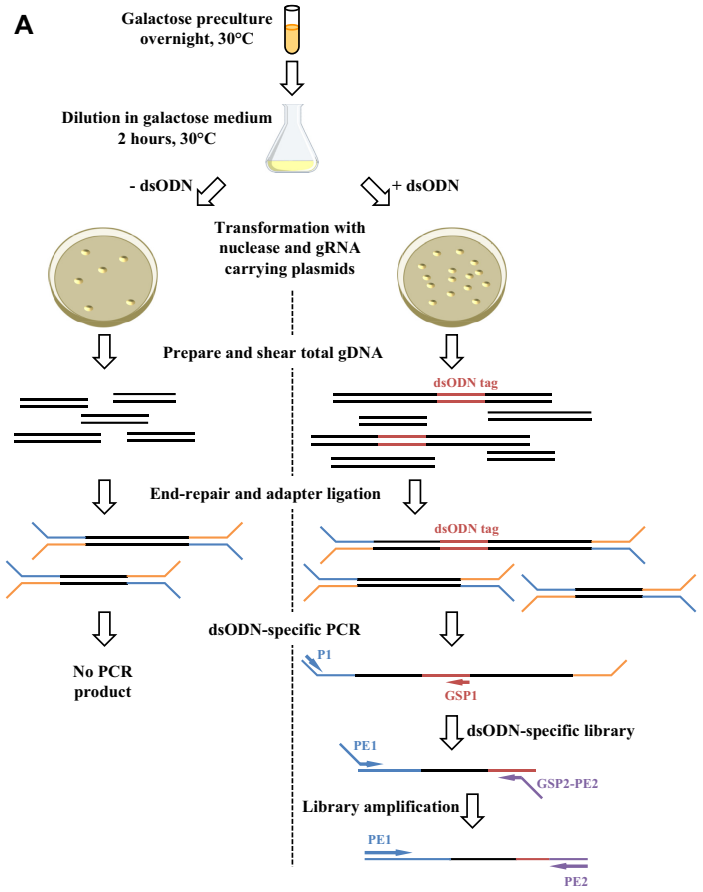
Table S6). Next, among these 56 positions, all positions containing only one mutant read were discarded. This left us with 14 genes containing at least two mutant reads at a predicted off-target position. In order to determine whether these mutant reads were statistically significant, they were compared 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 targeted by SpCas9. Interestingly, all but one mutations in *YMR124w* were deletions of one or more triplets, but the two validated off-targets in the GAA library were all point mutations or 1 nt insertions (Supplemental Figure S9). In conclusion, in the present experiments only nucleases targeted to CTG and GAA repeats exhibited detectable off-target effects. The type of mutations induced in these triplet repeats was strikingly different, suggesting alternative modes of repair following breakage.

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>SaCas9>Cas9-HF1. FnCpf1, the only type V nuclease tested, was shown to exhibit substrate preferences different from type II nucleases. We also demonstrated that sgRNA and protein levels did not generally correlate with 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 with 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 (47). Cas9-HF1 was more efficiently trapped in this inactive state than eSpCas9, showing more drastic impairment of cleavage. In our experiments, SpCas9 was more



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 12.5×10^6 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 mutant reads in each library to mutant reads in the NR library.

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 (48). 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.* (49) 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 4). 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 (9). Comparison of resection values between repeated and non-repeated ends demonstrated that all repeats inhibit resection to some extent (Supplemental Figure S6). Note that in the present work, much shorter microsatellites (17-33 repeated units) were used as compared to our previous experiments with 80 CTG triplets.

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 (50). Surprisingly, both nickases induced homologous recombination into GGGGCC, GCC, TGGAA repeat tracts and to a lower extent into CCTG, GCN and GAA repeat tracts (Figure 3). Nicks are usually formed in the course of the BER pathway and trigger specific protein recruitment (43). Nicks are therefore normally not processed by double-strand break repair machineries. However, there is some evidence supporting the hypothesis that nicks are recombinogenic (51,52) 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 (53,54). 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 S4). 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 (55). In our experiment, gaps due to multiple nicks may arise in GGGGCC, CGG and TGGAA repeat tracts, 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. Alternatively, nicks could be directly used to trigger homologous recombination, as was shown in some experimental systems using single-stranded oligonucleotide as template (reviewed by Maizels and Davis, 2018).

Correlation between secondary structure formation and nuclease efficacy

The sgRNA plays a crucial role in orchestrating conformational rearrangements of Cas9 (56). Stable secondary structure of the sgRNA as well as close state of the chromatin negatively affect Cas9 efficiency (57, 58). Possible secondary structures formed by the crRNA 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 (59). We found that more stable crRNAs were correlated with less efficient DSBs (Figure 5D). This is consistent with former studies on non-repeated sgRNAs and crRNAs showing that stable structured crRNAs (less than -4 kcal/mol) were not efficient at inducing cleavage (58). In addition, improperly folded inactive sgRNAs could be competing with active and properly folded sgRNAs within the same cell, to form inactive or poorly active complexes with Cas9, that will inefficiently induce a DSB (59). Differential folding of CAG and CTG crRNA 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 (60). 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 (32, 61). This difference in hairpin stability may impede Cas9-sgRNA 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 and SaCas9. 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 (57). Other nucleotide preferences were found (62) 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 (2, 9). Other approaches may be envisioned to specifically target toxic repeats in human using the CRISPR toolkit: (i) Cas9-D10A induced CAG contractions (55), (ii) dCas9 targeting microsatellites was able to partially block transcription, reversing partly phenotype in DM1, DM2 and ALS cell models (63), (iii) efficient elimination of microsatellite-containing toxic RNA using RNA-targeting Cas9 was also reported (64). 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.

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 (44). The GUIDE-seq method was very successful at identifying off-target sites in the human genome, following Cas9 expression. In *S. cerevisiae*, we showed that this approach was not efficient, most probably because NHEJ is not as active as in human cells. 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 (65, 66). 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. 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 (67,68). 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 thoroughly investigated.

Very recently, a similar assay based on a bipartite GFP reporter gene was built in U2OS cells. It was efficient at detecting and quantifying DSB repair following Cas9 induction (69). Altogether, our results give a new insight into which nuclease could be efficiently used to induce a DSB into a microsatellite in other eukaryotes.

DATA AVAILABILITY

All sequencing data have been deposited in the European Nucleotide Archive, under accession number PRJEB35597.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

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.

FUNDING

Sanofi, the Institut Pasteur; Centre National de la Recherche Scientifique (CNRS); L.P. was supported by a CIFRE PhD fellowship from Sanofi; Off-target studies were supported by the AFM-Telethon [AFM 21431]. Funding for open access charge: Institut Pasteur.
Conflict of interest statement. None declared.

REFERENCES

1. Orr,H.T. and Zoghbi,H.Y. (2007) Trinucleotide repeat disorders. *Annu. Rev. Neuro Sci.*, **30**, 575–621.
2. Richard,G.-F. (2015) Shortening trinucleotide repeats using highly specific endonucleases: a possible approach to gene therapy? *Trends Genet. TIG*, **31**, 177–186.
3. 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.
4. 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.
5. 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. MGG*, **261**, 871–882.
6. 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.
7. 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.
8. 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.
9. 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.
10. Doudna,J.A. and Charpentier,E. (2014) Genome editing. The new frontier of genome engineering with CRISPR–Cas9. *Science*, **346**, 1258096.
11. 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.
12. 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.
13. 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.
14. 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.
15. Świat,M.A., Dashko,S., den Ridder,M., Wijsman,M., van der Oost,J., Daran,J.-M. and Daran-Lapujade,P. (2017) FnCpf1: a novel and efficient genome editing tool for Saccharomyces cerevisiae. *Nucleic Acids Res.*, **45**, 12585–12598.
16. 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.
17. 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.
18. 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.
19. Richard,G.-F., Cyncynatus,C. and Dujon,B. (2003) Contractions and gexpansions of CAG/CTG trinucleotide repeats occur during ectopic gene conversion in yeast, by a MUS81-independent mechanism. *J. Mol. Biol.*, **326**, 769–782.
20. 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 Chichester Engl.*, **11**, 355–360.
21. Poggi,L., Dumas,B. and Richard,G.-F. (2020) Monitoring double-strand break repair of trinucleotide repeats using a yeast fluorescent reporter assay. In: Richard,G.-F. (ed). *Trinucleotide Repeats: Methods and Protocols, Methods in Molecular Biology*. Springer, NY, pp. 113–120.
22. 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**, bpy006.
23. 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.
24. 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.
25. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma. Oxf. Engl.*, **34**, 3094–3100.
26. 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. *Bioinforma. Oxf. Engl.*, **25**, 2078–2079.
27. 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**, 148.
28. Huang,M.E., Le Douarin,B., Henry,C. and Galibert,F. (1999) The Saccharomyces cerevisiae protein YJR043C (Pol32) interacts with the catalytic subunit of DNA polymerase α and is required for cell cycle progression in G2 / M. *Mol. Gen. Genet. MGG*, **260**, 541–550.
29. Parkinson,G.N., Lee,M.P.H. and Neidle,S. (2002) Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature*, **417**, 876–880.
30. Amrane,S., Sacca,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**, 4065–77.
31. Poggi,L. and Richard,G.-F. (2021) Alternative DNA structures in vivo: molecular evidence and remaining questions. *MicroBiol. Mol. Biol. Rev.*, **85**, e00110-20.
32. 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.
33. 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.
34. 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.
35. McMurray,C.T. (2010) Mechanisms of trinucleotide repeat instability during human development. *Nat. Rev. Genet.*, **11**, 786–799.
36. Mirkin,S.M. (2006) DNA structures, repeat expansions and human hereditary disorders. *Curr. Opin. Struct. Biol.*, **16**, 351–358.
37. Pearson,C.E., Nichol Edamura,K. and Cleary,J.D. (2005) Repeat instability: mechanisms of dynamic mutations. *Nat. Rev. Genet.*, **6**, 729–742.
38. Richard,G.-F., Kerrest,A. and Dujon,B. (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *MicroBiol. Mol. Biol. Rev. MMBR*, **72**, 686–727.

39. 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.
40. Kiliszek,A. and Rypniewski,W. (2014) Structural studies of CNG repeats. *Nucleic Acids Res.*, **42**, 8189–8199.
41. 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.
42. 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.
43. Krokan,H.E. and Bjørås,M. (2013) Base excision repair. *Cold Spring Harb. Perspect. Biol.*, **5**, https://doi.org/10.1101/cshperspect.a012583.
44. 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.
45. Moore,J.K. and Haber,J.E. (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, **16**, 2164–2173.
46. Ricchetti,M., Fairhead,C. and Dujon,B. (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature*, **402**, 96–100.
47. 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.
48. 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.
49. 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 l. *Nat. Commun.*, **6**, 10124.
50. 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.
51. Maizels,N. and Davis,L. (2018) Initiation of homologous recombination at DNA nicks. *Nucleic Acids Res.*, **46**, 6962–6973.
52. 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.
53. Arcangioli,B. (1998) A site- and strand-specific DNA break confers asymmetric switching potential in fission yeast. *EMBO J.*, **17**, 4503–4510.
54. 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.
55. Cinesi,C., Aeschbach,L., Yang,B. and Dion,V. (2016) Contracting CAG/CTG repeats using the CRISPR–Cas9 nickase. *Nat. Commun.*, **7**, 13272.
56. 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. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, 2984–2989.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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. *Mol. Cell*, **68**, 479–490.
64. 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.
65. Mirkin,E.V. and Mirkin,S.M. (2007) Replication fork stalling at natural impediments. *MicroBiol. Mol. Biol. Rev. MMBR*, **71**, 13–35.
66. Rothstein,R., Michel,B. and Gangloff,S. (2000) Replication fork pausing and recombination or 'gimme a break'. *Genes Dev.*, **14**, 1–10.
67. Kozlowski,P., de Mezer,M. and Krzyzosiak,W.J. (2010) Trinucleotide repeats in human genome and exome. *Nucleic Acids Res.*, **38**, 4027–4039.
68. 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.
69. Eki,R., She,J., Parlak,M., Benamar,M., Du,K.-P., Kumar,P. and Abbas,T. (2020) A robust CRISPR–Cas9-based fluorescent reporter assay for the detection and quantification of DNA double-strand break repair. *Nucleic Acids Res.*, **48**, e126.