sgRNA structural constraints and genetic limitations for efficient Cas9 genome editing to generate knock-outs

IRMGARD U. HAUSSMANN[1,2], THOMAS C. DIX[1], VERONICA DEZI[1], ABDULLAH I. HANS[1], ROLAND ARNOLD[3,4,5] AND MATTHIAS SOLLER[1, 4, 5]

[1]School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[2]Department of Life Science, School of Health Sciences, Birmingham City University, Birmingham, B5 3TN, United Kingdom

[3]Institute of Cancer and Genomics Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[4]Birmingham Centre for Genome Biology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

Running title: Constraints for efficient CRISPR/Cas9 mediated genome editing

Key Words: sgRNA/Cas9 optimization, secondary structure, retro-transposition

[5] Corresponding authors

m.soller@bham.ac.uk R.Arnold.2@bham.ac.uk

Tel: +44 121 414 5905

SUMMARY

Single guide RNA (sgRNA) directs Cas9 nuclease for gene-specific scission of double-stranded DNA. High Cas9 activity is essential for efficient gene editing to generate gene deletions and gene replacements by homologous recombination. Here we describe constraints in sgRNA design originating from maintaining the secondary structure of the sgRNA that is essential for high-efficiency DNA scission of Cas9, but aberrant with about 50% of sequences adjacent to PAM sites in *Drosophila*. We developed an sgRNA design tool (PlatinumCRISPR) to evaluate base-pairing for optimal design of highly efficient sgRNAs for Cas9 genome editing. Applied to generate gene deletions in *Drosophila Ythdc1* and *Ythdf*, that bind to N^6 methylated adenosines (m^6A) in mRNA we further discovered, that generating small deletions with sgRNAs and Cas9 leads to ectopic reinsertion of the deleted DNA fragment elsewhere in the genome. These insertions, however, can be removed by standard genetic recombination and chromosome exchange.

INTRODUCTION

Bacterially derived CRISPR/Cas9 from *Streptococcus pyogenes* provides a powerful tool for precise genome editing (Doudna, 2020; Garcia-Doval and Jinek, 2017; Hille et al., 2018; Jiang and Doudna, 2017). To induce double stand-breaks in DNA at desired locations, the DNA scission enzyme Cas9 is loaded with a single guide RNA (sgRNA) containing a complementary guide sequence (gRNA) to the genomic target site (protospacer), which also requires the protospacer adjacent motif (PAM) at the 3'end, comprised of NGG sequence (whereby N is any nucleotide and G is guanine). In addition to the target complementary sequence containing gRNA, two constant RNAs, the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) RNA (crRNA) and *trans*-activating crRNA (tracrRNA) are required for Cas9 to be able to cleave DNA, but they can be fused into a single sgRNA (Cong et al., 2013; Jinek et al., 2012).

High efficiency CRISPR/Cas9 mediated DNA scission is essential to generate mutants at high frequency in genetic screens and to provide the resource for efficient homologous recombination directed gene replacements. Large scale analysis of sgRNA efficiencies revealed the whole spectrum of on-target cleavage activities ranging from 0-100% arguing for a number of parameters that need to be correct for high efficiency cleavage leading to models incorporating weighing of features and /or thermodynamics of secondary structures (Abadi et al., 2017; Chari et al., 2015; Chuai et al., 2018; Doench et al., 2016; Doench et al., 2014; Farboud and Meyer, 2015; Gagnon et al., 2014; Graf et al., 2019; Hsu et al., 2013; Labuhn et al., 2018; Liu et al., 2016; Michlits et al., 2020; Moreno-Mateos et al., 2015; Ren et al., 2014; Sledzinski et al., 2020; Wang et al., 2014; Zhang et al., 2019). These studies identified that sequences with very low (≤35%) or very high (>80%) guanine-cytosine (GC) overall content were less effective. In addition, purines in the four nucleotides preceding the PAM substantially increased Cas9 cleavage efficiency, while pyrimidines

and in particular uridine resulted in a lower efficiency (Graf et al., 2019; Ren et al., 2014; Wang et al., 2014). The lower efficiency of two uridine preceding the PAM site was further associated with premature termination of RNA Pol III (Graf et al., 2019). Moreover, changes in internal structure of sgRNA has been found associated with low activity (Jensen et al., 2017; Moreno-Mateos et al., 2015; Thyme et al., 2016). These observations have helped to improve the design of sgRNAs to yield higher efficiencies and incorporated into sgRNA design tools (Labun et al., 2016) (Sledzinski et al., 2020). We noticed, however, that these rules were not sufficient to guarantee high cleavage efficiency and many sgRNA candidates scoring high fail to cleave efficiently (Labun et al., 2016) (Labuhn et al., 2018; Sledzinski et al., 2020). In particular, the impact of sgRNA folding has not yet been analysed in detail and incorporated in web tools for sgRNA design.

The x-ray crystal structure of Cas9 bound to sgRNA has been determined and indicates four regions of base-pairing termed tetraloop (tetraloop forms due to fusion of crRNA and tracrRNA) and stem loops 1-3 that might be important for its function (Anders et al., 2014; Nishimasu et al., 2014). This structure revealed many points of close interactions of the folded sgRNA with Cas9, but whether disruptions in the sgRNA structure would impact on Cas9 cleavage efficiency has not systematically been analyzed. In particular highly GC-rich guide RNAs could disrupt the rather weak secondary structure of the sgRNA bound by Cas9.

The sgRNA/Cas9 genome editing tool is widely used to generate knock-out mutants by introducing frameshifts and for site directed mutagenesis based on homologous recombination from a mutated template. It has been recognized that introducing premature termination codons (PTCs) can induce use of alternative translation initiation sites (Tuladhar et al., 2019). Likewise, compensatory responses have been observed involving upregulating genes and as a consequence causing stronger phenotypes than compared to removing a gene entirely (El-Brolosy et al., 2019; Ma et al., 2019). In

addition, some genes have dual functions as protein and RNA (Hachet and Ephrussi, 2004). Hence, introducing a frameshift will only remove the protein function. Likewise, many non-coding RNAs are present in introns suggestion that their expression is connected to the expression of the host gene (Boivin et al., 2018), but such relationships have not yet been explored comprehensively as they require more sophisticated genome editing (Deveson et al., 2017). Thus, for generating gene knock-outs, removing the transcription start site or the entire gene should be considered.

Here, we discovered that maintaining the secondary structure of the sgRNA is essential for high-efficiency DNA scission of Cas9. Likewise, excessive base-pairing in the seed region also impacts on Cas9 cleavage efficiency. Accordingly optimal design of sgRNA for high efficiency DNA cleavage of Cas9 requires analysis of sgRNA secondary structure which is aberrant with about 50% of PAM adjacent sequences in *Drosophila*. To facilitate design of optimal sgRNAs, we developed an online tool incorporating all currently known parameters for sgRNA design including correct sgRNA folding (https://platinum-crispr.bham.ac.uk/predict.pl). Further, we devised a strategy to generate gene deletions in *Drosophila* using transposon markers. Using these novel techniques we generated deletions of *Ythdc1* and *Ythdf*, that bind to N^6 methylated adenosines (m^6A) in mRNA important in development (Anreiter et al., 2021; Balacco and Soller, 2019; Dezi et al., 2016; Haussmann et al., 2016). We further discovered, that generating small deletions with sgRNAs and Cas9 leads to ectopic reinsertion of the deleted DNA fragment elsewhere in the genome, but such inserts can be removed by standard genetic recombination and chromosome exchange.

RESULTS

RNA secondary structure constraints limit sgRNA/Cas9 activity

Although sgRNA/Cas9 can cleave DNA efficiently, the first sgRNAs (L11GC and R13GC) we designed according to previously published guidelines to cleave the plasmid *pUC 3GLA Dscam 3-5* for gap repair recombineering did not cut (**Fig 1 A and B, Supplementary Fig S1**) (Haussmann et al., 2019). Similarly, the first sgRNAs flanking the *Drosophila Ythdf* gene did not result in a deletion of the locus based on screening for loss of an RFP-marked transposon even though we had validated the target sequence in the strain used (n=103, **Supplementary Fig S2**). Hence, we devised an in vitro assay to test the DNA scission efficiency of these two sgRNA based on in vitro transcribed sgRNAs and commercially available Cas9 using oligonucleotide and plasmid substrates.

It has previously been shown that extending the tetraloop in the constitutive component of sgRNAs constituted by tracrRNA and crRNA enhances cleavage efficiency in vitro using oligonucleotide substrates (**Fig 1C**) (Jinek et al., 2012). Introducing the extended sequence present in tracrRNA and crRNA into sgRNA (L7GCext) did not increase efficiency of cleaving a plasmid at 37ºC, but increasing the temperature to 42ºC enhanced cleavage by L7GCext (**Fig 1B**). Increasing the salt concentration to 200 mM also did not result in enhanced cleavage by L7GC/R3G (**Fig 1D**). In contrast, both L7GC and L7GCext could cleave an oligonucleotide, while R13GC did not (**Fig 1E**). Using this assay also confirms the previous observation that sgRNAs of shorter length will lead to cleavage. In addition, three guanosines introduced at the 5'end of sgRNAs required for efficient in vitro transcription are tolerated (**Fig 1F and 1G**) (Jinek et al., 2012).

The sgRNA scaffold adopts a typical fold when bound to Cas9 consisting of the bulged tetraloop, followed by small loop 1 and the more extended loops 2 and 3, which form a protective 3'end structure (**Fig 1C**) (Nishimasu et al., 2014). The loop2/3 structure does not involve the uridines incorporated for termination of RNA Pol III driven expression from plasmids (**Fig 1C**). When comparing the secondary structures of the four sgRNAs, we noticed that the two well-cutting

sgRNAs L7GC or R3G maintained the secondary structure of the constitutive RNA part, while the non-cutting sgRNA L11GC disrupted the structure of the tetraloop (**Fig 1H-K**). The effect of R13GC seems more subtle as it could cut in the oligonucleotide assay suggesting that the repeated bulge structure is the cause for its inefficiency, which is supported by x-ray crystal structure of the Cas9-sgRNA-DNA complex. Here, the bulge structure is recognized by Cas9 where Tyr_{359} base-stacks with G_{43}, that also forms hydrogen bonds with Asp_{364} and Phe_{351} , and Phe_{351} forms a hydrogen bond with A_{42} (Nishimasu et al., 2014). In addition, the sgRNAs initially used for deleting the *Ythdf* gene have severely disrupting the secondary structure (**Supplementary Fig S2**).

When comparing the sequences of the cutting sgRNAs L7GC or R3G with the non-cutting L11GC and R13GC sgRNAs, we further noticed that L11GC and R13GC sgRNAs contained more guanosines, which in RNA can base-pair with C and U. To test if guanosines in the sgRNA limit Cas9 activity we increased their number in R3G to 13 to make sgRNA 13G (**Fig 2A**). For the design of the R13 sgRNA, care was taken not to disrupt the tetraloop, but we noticed the potential to interfere with loop 2 (**Fig 2D**, see below). Intriguingly, sgRNA 13G is not capable of directing Cas9 cleavage if the target sequence is present in a 3 kb plasmid, but is active with a short oligonucleotide substrate (**Fig 2A and 2B**). Likewise, adding a restriction enzyme together with sgRNA/Cas9 inhibited Cas9 in cleaving plasmid DNA suggesting that Cas9's ability to scan DNA can be impaired separately from its ability to cleave DNA.

Since the increased number of Gs in sgRNA R13G lead to enhanced base-pairing, we exchanged the Gs with Cs leading to an open structure in the seed region. This sgRNA R13C cleaved the test-plasmid efficiently (**Fig 2B and 2E**). Introducing Cs in the left or right half of sgRNA R13G lead to short stem loops and inefficient cleavage of the test-plasmid (**Fig 2B, 2F and 2G**).

To further test to what extent base-pairing impacts on Cas9 activity, we generated sgRNAs L10ds6G and R10ds6GC, where the proximal or the distal half leads to complementary base-pairing of the gRNA with the constant part, respectively (**Fig 3A and 3B**). Although both sgRNAs supported Cas9 cleavage of oligonucleotide substrate, the R10ds6GC sgRNA base-pairing with the proximal part was mostly inactive in cleaving the plasmid indicating an impaired ability of Cas9 to scan DNA (**Figs 3E and 3F**).

Taken together, these results demonstrate that the structure of the sgRNA is important for efficient Cas9 mediated DNA scission. Furthermore, high G content and base-pairing in the distal part of the gRNA also impairs DNA scission, while base-pairing in the proximal part is tolerated.

To further substantiate these findings, we analyzed the structures of sgRNAs from previous studies in mammalian cells and *Drosophila* with regard to their cleavage efficiency of previous attempts to define rules for sgRNA cleavage efficiency in vivo (Graf et al., 2019; Ren et al., 2014). Indeed, in 39 sgRNAs designed for use in *Drosophila* reduced cleavage efficiency in nine sgRNAs is associated with disturbances of the sgRNA secondary structure resulting in a cleavage efficiency below 35% (**Supplementary Figs S3 and S4**)(Ren et al., 2014). Similarly, from 22 sgRNAs designed for use in mammalian cells, 13 had a cleavage efficiency below 35% associated with disturbances of the sgRNA secondary structure (**Supplementary Fig S5**) (Graf et al., 2019).

Given the requirement for correct folding of the sgRNA for efficient Cas9 mediated DNA scission, we further examined the x-ray crystal structure of the Cas9-sgRNA-DNA complex to see whether this would provide additional instructions to design sgRNAs (Nishimasu et al., 2014). Indeed, the first two nucleotides, adenosine 51 and 52 (A_{51} and A_{52}, **Supplementary Fig S6**) after the tetraloop form an aromatic base-stacking interaction with phenylalanine 1105 (Phe_{1105}) of Cas9. Furthermore, these interactions are stabilized by guanosine 62 (G_{62}) forming non-Watson Crick hydrogen bonds

with A_{51} A_{52} and Phe_{1105}, and uracil 63 (U_{63}) forms a base-stacking interaction with A_{52}. These interactions indicate that base-pairing of the gRNA with these nucleotides of the constant part of the sgRNA reduce Cas9 activity in Cas9 cleavage assays in vitro and mutagenesis in vivo.

To facilitate design of sgRNAs we have incorporated these findings in a freely available web based tool https://platinum-crispr.bham.ac.uk/predict.pl and a stand-alone code is available for non-commercial use upon request from the authors. When we systematically analysed genome sequences from Drosophila for correct folding and activity of sgRNAs using the above parameters, 50% of sgRNAs (241 from 481) did not fold properly or were otherwise compromised in exerting maximal activity.

A *Drosophila* transformation vector for expression of two sgRNAs

Next, we applied these novel sgRNA design rules to generate *Drosophila* gene deletions. For this purpose we generated a fly transformation vector that harbors two sgRNAs expressed from *U6.1* and *U6.3* promoters (**Fig 4A-C**), as the only requirement for expression is a G as first nucleotide (Paule and White, 2000; Ren et al., 2013). This plasmid can be generated by incorporating the two sgRNA sequences in PCR primers for single step cloning into the plasmid, that can be injected into *Drosophila* expressing Cas9 in the germline for CRISPR. Alternatively, this vector can be used to generate a transgenic line via the attB site using phiC31 integrase mediated transformation. This fly strain is then crossed to a line expressing Cas9 in the germline for generation of the desired genetic lesion. Transgenically provided sgRNA/Cas9 generally results in a higher efficiency, because the sgRNAs are been provided maternally.

Efficient generation of gene deletions by sgRNA/Cas9 using transposon markers

To generate deletions of the YTH protein genes *Ythdc1* and *Ythdf*, which are located on the third chromosome, transposon inserts *Mi{MIC}YT521-B*^{MI02006} and *PBac{SAstopDsRed}*^{LL04081} marked with *GFP* or *RFP*, respectively, were combined with an X-linked *vasCas9* or for *nosCas9* germline expression of Cas9. To allow for detecting loss of the transposon in the YTH protein genes the GFP and RFP markers of the *vasCas9* insert had been removed. These flies were then crossed to the GFP marked sgRNA construct inserted on the 3rd chromosome (**Fig 5 A and F**). The sgRNA insert has a weak GFP marker and can generally be distinguished from *Mi{MIC}* inserts. Females from this cross were mated with males containing *TM3 Sb/TM6 Tb* double-balancers in single crosses to recover the sgRNA induced individual deletions and avoid analysis of clonal events. The male progeny was then screened for loss of the GFP or RFP marker. Males which had lost the marker were detected in 100 % (n=9) of the crosses for *Ythdc1* and in 88 % (n=9) for *Ythdf*, respectively. In those crosses with loss of markers, all males for *Ythdc1* had lost the GFP marker, while for *Ythdf* the average frequency of marker loss was 42 % (n=8). In the reverse cross for *Ythdf*, the frequency was 0% (n=5) indicating that the low frequency in males is linked to the absence of recombination in males. The identified single marker-less males were then crossed to *TM3 Sb/TM6 Tb* double-balancers to establish a line and analysed with PCR using primers next to the deletion breakpoints yielding a short PCR product (**Fig 5B and G**). A PCR product had been obtained in all lines were the marker had been lost indicating that the expected deletion had indeed been generated. To generate *Ythdc1* excision lines *nosCas9* was used for germline expression of Cas9 as *vasCas9* together with sgRNAs targeting *Ythdc1* resulted in female sterility.

sgRNA/Cas9 scissioned fragments insert elsewhere in the genome

After establishing the lines, we noticed that all lines (n=10) established for the *Ythdc1* deletion did not show the flightless phenotype previously reported (Haussmann et al., 2016). Therefore, we selected four lines for further analysis by RT-PCR from RNA (*Ythdc1* excision lines, **Fig 5C**) or of genomic DNA (*Ythdf* excision lines, **Fig 5H**) of homozygous flies with primers that were within the deletion and also flanked an intron. Unexpectedly, a copy of the gene was still present in all *Ythdc1* and *Ythdf* excision lines analyzed suggesting that the deleted fragment had been inserted elsewhere in the genome.

To remove this ectopic insert(s), positive lines were crossed to *w*+ marked deficiencies and out-crossed for two generations. The X and 2nd chromosomes were then exchanged to establish null-mutant lines from single chromosomes for *Ythdc1* and *Ythdf* that were confirmed by RT-PCR to be free of any ectopic inserts (**Fig 5D, E and Fig 5I, J**). To avoid such complications in the future we generated a PBac w+ containing vector which can be efficiently inserted into a locus by cloning left and right homology arms either to induce a partial deletion upon insertion. Alternatively, mutations in sgRNA cleavage sites can be introduced into the homology arms to insert point mutations followed by scareless removal of the PBac w+ by transposase.

DISCUSSION

DNA scission by the sgRNA/Cas9 complex is highly specific and requires complete base-pairing between the sgRNA and the target DNA generally not tolerating single miss-matches (Farboud and Meyer, 2015; Ren et al., 2014). This feature makes the sgRNA/Cas9 complex an ideal tool for genome editing, but its use is currently limited by the low predictability to cut its target in the genome (Labuhn et al., 2018; Sledzinski et al., 2020).

Here, we discovered that the structure of the sgRNA is a key determinant for the scission efficiency of the sgRNA/Cas9 complex, but that about 50% of sequences adjacent to PAM sites possibly constituting gRNAs do not fold properly in the sgRNA context or are compromised by base-pairing. Our data suggest support these two levels of interference. First, sgRNA R13GC did not result in cleavage of a short oligonucleotide substrate. This indicates that the structure of this sgRNA blocks the catalytic activity of the sgRNA/Cas9 complex. Second, some sgRNAs allowed cleavage of short oligonucleotide substrates (e.g. L7GC and R10ds6GC), but did not support efficient DNA scission of the target sequence in the context of a 3 kb test-plasmid. Likely, these sgRNAs interfere with the ability of the sgRNA/Cas9 complex to scan the DNA for target sites.

When we analyzed sgRNAs from previous studies where the efficiency had been determined we observed a good overlap with the ability of those sgRNAs to adopt the correct structure and having a high cleavage efficiency. Further refinement to the design of sgRNAs comes from the recognition that RNA Pol III can terminate at the boundary of the tracrRNA if two UU precede the GUUUU of the start of the tracrRNA (Arimbasseri and Maraia, 2015; Graf et al., 2019). However, analysis of the x-ray crystal structure of the Cas9-sgRNA-DNA complex also revealed that the two Us at the end of the gRNA (N_{19} and N_{20}) before the start of the tracrRNA have key roles in folding of the sgRNA/CAS9 complex (Nishimasu et al., 2014).

In bacteria, the CRISPR/Cas9 system consists of two RNAs (crRNA and trRNA) that assemble with Cas9 to form the active complex. crRNA and trRNA base-pair through sequence complementarity to form the tetra loop in the active Cas9 complex (Garcia-Doval and Jinek, 2017; Hille et al., 2018; Jiang and Doudna, 2017). In type II systems the tracrRNA is required for crRNA maturation suggesting that base-pairing takes place while being assembled with Cas9. After the crRNA is trimmed, the entire CRISPR/Cas9 complex can scan genomic DNA for DNA scission sites.

Alternatively, the crRNA could hybridize first to genomic DNA and recruit tracrRNA and Cas9 to form a complex for DNA scission on site. In this scenario, the crRNA would not be able to interfere with tracrRNA/Cas9 complex activity by forming an aberrant RNA secondary structure, but whether this second scenario could be applied to more efficient genome editing with reduced off-target cleavage needs to be tested. In this context, it would be worth to explore how much longer the gRNA can be to still support Cas9 DNA scission as longer RNAs would more stably hybridize to DNA. In either case, however, understanding sgRNA/CAS9 complex assembly will inform how to prevent off-target DNA scission.

Taken together, optimizing the structural constraints in sgRNA design is key to their efficiency in cleaving DNA. Accordingly, optimized sgRNAs show very little base-pairing with sequences adjacent to the loop 1 region, or are not complementary to the tetraloop structure and/or the loop2/3 structure. In addition, avoiding G-rich sequences and base-pairing in the 10 nt seed region prior to the PAM site predicts high efficiency of sgRNAs for DNA scission. Furthermore, avoiding two Us before the PAM site prevents interference with the sgRNA/Cas9 structure. In any case, however, by introducing the target sequence into a plasmid the efficiency of a particular sgRNA can be reliably determined in an in vitro cleavage assay. Although this assay will determine the cleavage efficiency of an sgRNA, cellular features such as chromatin state can also impact on Cas9 mediated DNA scission (Singh et al., 2015).

Generating deletions of entire genes, or essential parts of them is the preferred way to generate a null allele. This approach will avoid complications arising from introducing frameshifts at the beginning of the ORF, as translation could reinitiate from later AUG or CUG start codons (Koushika et al., 1999). In addition, in some genes the RNA has functions on its own, as shown for *oscar* RNA that forms a large RNP particle with Oscar and other RNA binding proteins at the

posterior pole of a *Drosophila* oocyte (Hachet and Ephrussi, 2004; Haussmann et al., 2011). Thus, deletion of the entire gene region will discover such additional functions harbored in gene transcripts.

When generating deletions of entire genes, we were very surprised to discover that the deleted DNA fragment was inserted into the genome and transcribed, resulting in the expression of protein that rescued the flightless phenotype in *Ythdc1* deletion allele. Although the mechanism for the generation of these new inserts is not known, it seemed not to have led to complex chromosomal aberrations, because the inserts could be removed by standard recombination and/or exchange of chromosomes. Retro-transposition has been observed in the ELAV gene, which led to the loss of all introns in *Drosophila* (Samson, 2008). Since intron loss and gain occurs during evolution, the mechanism underlying insertion of DNA fragments from sgRNA induced deletions might be responsible for these changes. Likewise, the *amyloid-β precursor protein* (*APP*) gene displays copy number variation in the human brain, which are increased by retro-transposition in sporadic forms of Alzheimer's disease suggesting evolutionary conserved mechanisms for reinsertion of genomic information (Lee et al., 2018). Thus, when generating sgRNA induced gene deletions, it is essential to test for the absence of any transcripts by RT-PCR, but also to validate a deletion at the DNA level by PCR using flanking primers or Southern blots. Alternatively, a GFP cassette with a polyA site can be inserted to terminate the ORF in the beginning, but it needs to be evaluated whether the polyA site in the beginning of the gene is used or whether the exon containing the GFP cassette is skipped (Soller, 2006; Wierson et al., 2020). For a more reliable way to generate gene knock-outs we developed PBac w+ marker to be inserted alongside a deletion of the 5'part of a protein that on its own is not functional. However, a marked locus also allows for rigorous cleaning of the genetic background.

In essence, we have established the rules for designing highly efficient sgRNAs and established methodology to efficiently generate gene deletions. These findings have implications for other RNA based methodologies including prime editing (Anzalone et al., 2019; Bosch et al., 2021).

ACKNOWLEDGMENTS

We thank Bloomington and Kyoto stock centers for fly lines, A. Handler for the BiggyBac clone, BacPAc for Bac clones and plasmids, the University of Cambridge Department of Genetics Fly Facility for embryo injections for generating transgenic lines and Pawel Grzechnik for comments on the manuscript. For this work we acknowledge funding from the Biotechnology and Biological Science Research Council, the Leverhulme Trust and a Genetics Society summer studentship to AIH.

AUTHOR CONTRIBUTIONS

IUH, TCD and MS performed biochemistry and genetic experiments, VD constructed vectors and AH performed genetic experiments. RA performed bioinformatics analysis, programming and server implementation. IUH, RA and MS conceived the project and wrote the manuscript with help from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

FIGURE LEGENDS

Figure 1: Sequence dependent in vitro cleavage of oligonucleotides and plasmid DNA by the sgRNA/Cas9 complex.

(A) Sequences of sgRNAs with observed cleavage sites indicated by arrow heads. Small letter guanosines used for in vitro transcription are not present in the target DNA sequence. The seed sequence is indicated by a line at the bottom.

(B) Agarose gel showing Cas9 mediated cleavage of the 11.3 kb Dscam 3-5 plasmid for 24 h with indicated sgRNAs. Plasmids were cut with either Acc65I (lanes 2, 3, 6 and 7) or BspEI (lanes 4, 5, 8 and 9) after Cas9 cleavage. The line at the bottom shows a map of the plasmid with restriction sites indicated. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

(C) Structure of the sgRNA scaffold from co-crystallization with Cas9 (Nishimasu et al., 2014). Vertical or horizontal lines indicate Watson-Crick base-pairing, and dots or a dashed lines indicate non-Watson-Crick base-pairing. Nucleotides base-pairing in loop 1 are bold. Additional base-pairing found in the tracrRNA-crRNA heterodimer is indicated in the extended scaffold (Jinek et al., 2012).

(D) Agarose gel showing Cas9 mediated cleavage of the 11.3 kb Dscam 3-5 plasmid for 24 h with indicated sgRNAs L7GC and R3G. Plasmids were cut with PstI and NotI. The star denotes incomplete cleavage by NotI and the line at the bottom shows a map of the plasmid with restriction sites indicated. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

(E) Denaturing acrylamide gel showing Cas9 mediated cleavage of synthetic oligonucleotides with indicated sgRNAs.

(F) Sequences of sgRNAs with variable length. Small letter guanosines used for in vitro transcription are not present in the target DNA sequence.

(G) Denaturing acrylamide gel showing Cas9 mediated cleavage for 1 h of synthetic oligonucleotides with indicated sgRNAs of variable length.

(H-K) Structure of sgRNAs. Nucleotides base-pairing in loop 1 are bold. Red lines in J and K indicate potential base-pairing with nucleotides in loop 2. The red arrow in J indicates the complementariness leading to the nearby loop-extension in the tetraloop. The red arrows in K indicate a duplication of the bulge structure present in the tetraloop.

Figure 2: Secondary structure of sgRNAs affects Cas9 mediated cleavage efficiency.

(A) Sequences of sgRNAs. Small letter guanosines used for in vitro transcription are not present in the target DNA sequence.

(B) Agarose gel showing Cas9 mediated cleavage after 6 h of 3 kb pBS SK+ test-plasmids containing the target sequence with indicated sgRNAs. Plasmids were linearized with ScaI as in the control after Cas9 heat inactivation. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

(C) Denaturing acrylamide gel showing synthetic oligonucleotides before and after sgRNA/Cas9 mediated cleavage for 1h.

(D-G) Structure of sgRNAs. Nucleotides base-pairing in loop 1 are bold. Red lines in D indicate potential base-pairing with nucleotides in loop 2. Green nucleotides indicate mutations compared to sgRNA R13G.

Figure 3: Base-pairing of sgRNAs in the seed-region blocks Cas9 mediated cleavage of a test plasmid.

(A, B) Structure of sgRNAs. Horizontal red lines in A and B indicate artificially introduced base-pairing with the sgRNA scaffold, and vertical red lines in a indicate potential base-pairing with nucleotides in loop 2. Nucleotides base-pairing in loop 1 are bold.

(C) Denaturing acrylamide gel showing synthetic oligonucleotides before and after sgRNA/Cas9 mediated cleavage for 1h. Note that Cas9 cleavage is heterogeneous.

(D) Agarose gel showing Cas9 mediated cleavage after 6 h of 3 kb pBS SK+ test-plasmids containing the target sequence with indicated sgRNAs. Plasmids were linearized with ScaI as in the control after Cas9 heat inactivation. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8 kb.

Figure 4: *Drosophila* transformation vector for U6 promoter mediated expression of two sgRNAs.

(A) Forward and return primer sequences to incorporate sgRNA sequences. The first nucleotide of the gRNA is indicated by an asterisk.

(B) Cloning scheme for incorporating two sgRNAs into the destination vector.

(C) Plasmid map of the fly transformation vector *pUC 3GLA U6.1/U6.3* sgRNA expressing two sgRNAs under U6.1 and U6.3 promoters, respectively.

(D) Secondary structure of sgRNAs targeting *Ythdc1* and *Ythdf*.

Figure 5: Generation of gene deletions in *Drosophila* YTH protein genes using two sgRNAs/Cas9 and transposon markers.

(A) Schematic to the *Ythdc1* locus indicating transcripts (white boxes) and the ORF (black boxes) below the chromosome. Primers used are indicated on top and below the transcripts. The w+ marked transposon used for detecting deletions in the locus is indicated by a triangle and the deletion generated is indicated by a line.

(B) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdc1* transposon excision lines using primers flanking the deletion. Presence of a PCR product indicates the expected gene deletion. DNA markers are indicated on the left.

(C) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdc1* transposon excision lines using internal primers flanking an intron. Ectopic insertion in the opposite orientation is indicate by an asterisk (lane 4) as in this instance the transcript is not spliced. DNA markers are indicated on the left.

(D) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdc1*$^{\Delta 7}$ flies using internal primers and primers flanking the deletion. DNA markers are indicated on the left.

(E) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdc1*$^{\Delta 7}$ flies using internal primers in *Ythdf* and *ewg* genes. The PCR product of the *ewg* gene was used as loading control. DNA markers are indicated on the left.

(F) Schematic to the *Ythdf* locus indicating transcripts (white boxes) and the ORF (black boxes) below the chromosome. Primers used are indicated on top and below the transcripts. The RFP marked transposon used for detecting deletions in the locus is indicated by a triangle and the deletion generated is indicated by a line.

(G) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf* transposon excision lines using primers flanking the deletion. Presence of a PCR product indicates the expected gene deletion. DNA markers are indicated on the left.

(H) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf* transposon excision lines using primers flanking an intron. DNA markers are indicated on the left.

(I) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf*$^{\Delta B1}$ flies using internal primers and primers flanking the deletion. DNA markers are indicated on the left.

(J) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdf*$^{\Delta B1}$ flies using internal primers in *Ythdf* and *ewg* genes. The PCR product of the *ewg* gene was used as loading control. DNA markers are indicated on the left.

Supplementary Figures

Supplementary Figure S1. Secondary structures of sgRNAs used in Figs 1-3 predicted by RNAfold.

Supplementary Figure S2. Inactive sgRNAs targeted to the *Drosophila ythdf* (*CG6422*) gene.

(A) Sequences of sgRNAs targeting the CG6422 locus for generating a deletion.

(B, C) Secondary structures of sgRNAs. The scaffold is shown on the left indicating Watson-Crick base-pairing by lines and the predicted secondary structure by RNAfold is shown on the right.

Supplementary Figure S3. Secondary structure of sgRNAs targeted to the *Drosophila white* gene from Ren et al. (2014).

(A-AA) Secondary structures of sgRNAs predicted by RNAfold. Minimal energy and proximity base-pair structures are shown on the left and right, respectively. Red and blue indicated high and low probabilities for the adopted structural base-pairing assignment, respectively. The number on

the right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting DNA cleavage and green arrows point towards structural features supporting DNA cleavage.

Supplementary Figure S4. Secondary structure of sgRNAs targeted to the *Drosophila vermillion, ebony* and *yellow* gene from Ren et al. (2014).

(A-L) Secondary structures of sgRNAs predicted by RNAfold targeting the vermillion (A-D), the ebony (E-H) and the yellow gene (I-L) . Minimal energy and proximity structures are shown on the left and right, respectively. Red and blue indicated high and low probabilities for the adopted structural base-pairing assignment, respectively. The number on the right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting DNA cleavage.

Supplementary Figure S5. Secondary structure of sgRNAs targeted to human CD22 from Graf et al. (2019).

(A-V) Secondary structures of sgRNAs predicted by RNAfold. Minimal energy and proximity base-pair structures are shown on the left and right, respectively. Red and blue indicated high and low probabilities for the adopted structural base-pairing assignment, respectively. The number on the

right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting DNA cleavage.

Supplementary Figure S6. Key sgRNA secondary structural features directly interact with the Spy Cas9 endonuclease in the three way junction .

(A) Minimum Free Energy (MFE) predicted secondary structure of the sgRNA used for co-crystallization of Spy Cas9 (Nishimasu et al., 2014). Nucleotides involved in the three way junction interactions are indicated (A_{51}, A_{52}, G_{62} and U_{63})

(B) X-ray crystal structure of spyCas9 endonuclease in complex with chimeric sgRNA bound to genomic DNA target (PDB: 4OO8) (Nishimasu *et al* 2014). The gRNA portion is coloured in pink and the remainder of the sgRNA is coloured in orange. Genomic DNA is coloured yellow and the protein chain is coloured in green. Nucleotide residues situated at key structural features in the sgRNA are coloured cyan. Note that a small portion of the Cas9 protein chain depicting amino acid side chains to view the sgRNA three-way junction is transparent.

(C) Magnified view of the sgRNA three-way junction. Hydrogen bonds are indicated by black dashed lines and aromatic stacking interactions are shown by pink dashed lines. Here, the interaction of Phe1105 with A_{51}, A_{52} and U_{63} and the interaction of G_{62} with A_{51}, A_{52} and the phosphate backbone can be seen.

Supplementary Figure S7. Key sgRNA secondary structural features directly interact with the Spy Cas9 endonuclease in the tetraloop bulge.

(A) Minimum Free Energy (MFE) predicted secondary structure of the sgRNA used for co-crystallization of spyCas9 (Nishimasu et al., 2014). Nucleotides forming the bulge are indicated (A_{28}, A_{41}, A_{42} and G_{43}).

(B) X-ray crystal structure of Spy Cas9 endonuclease in complex with chimeric sgRNA bound to genomic DNA target (PDB: 4OO8) (Nishimasu *et al* 2014). The gRNA portion is coloured in pink and the remainder of the sgRNA is coloured in orange. Genomic DNA is coloured yellow and the protein chain is coloured in green. Nucleotide residues situated at key structural features in the sgRNA are coloured cyan. Note that the tetra loop after the bulge does not interact with Cas9 and sticks out of the structure.

(C) Magnified view of the sgRNA bulge present in the tetra loop. Hydrogen bonds are indicated by black dashed lines and aromatic stacking interactions are shown by pink dashed lines. Here, the interactions of Phe351, Tyr359 and Asp364 with A_{42} and G_{43} can be seen. Inlet on the right: 180º turn to visualize base-stacking of U_{44} with Tyr325 and His328. Through these interactions base-pairing with G_{27} is prevented.

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact Matthias Soller (msoller@bham.ac.uk).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Fly husbandry

All *Drosophila melanogaster* strains were reared at 25ºC and 40%–60% humidity on standard cornmeal-agar food in 12:12 h light:dark cycle. 14-18 h old embryos, wandering third instar larvae and 2-10 day old adult flies were used. Canton-S flies were used as wild-type.

Bacteria

E.coli DH5alpha was used to propagate plasmids.

METHODS DETAIL

sgRNA/Cas9 directed DNA cleavage

DNA templates for in vitro transcription were reconstituted from synthetic oligonucleotides. As only the T7 promoter needs to be double-stranded for in vitro transcription, a T7 promoter oligonucleotide (CCTGGCTAATACGACTCACTATAG) was annealed to an anti-sense Ultramer (IDT DNA) encoding the entire sgRNA in addition to the T7 promoter. Alternatively, a 60 nt T7 promoter oligonucleotide with a partial sgRNA was annealed to an anti-sense oligonucleotide encoding the tracrRNA (AAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTTAACTTGCTATTTCTAGCTCTAAAAC) for 15 min at 40º C (2 µM) and made double-stranded by extension with Klenow fragment of DNA polymerase I according to the manufacturer's instructions (NEB). Klenow was then heat-inactivated 10 min at 85º C and oligonucleotides were desalted with a G-50 Autoseq Sephadex spin column (GE) before using for in vitro transcription.

Then, sgRNAs were generated by in vitro transcription with T7 polymerase (T7 MEGAscript, Ambion) from synthetic oligonucleotides (0.2 µM) and trace-labeled with ^{32}P alpha-ATP (800 Ci/mmol, 12.5 µM, Perkin Elmer) in a 20 µl reaction according to the manufacturer's instructions.

After DNAse I digestion, free nucleotides were removed with a G-50 Probequant Sephadex spin column (GE). Then, sgRNAs were heated for 2 min to 95º C and left at room temperature to adopt folding. Then, sgRNAs were quantified by scintillation counting and analysed on 8 % denaturing polyacrylamide gels as described (Dix et al., 2020).

For synthetic substrate DNAs the sense oligonucleotide (1 µM, sgRNA flanking sequences are: TCGAGCATTATATGAAC-sgRNA-GGGTATTGGGGAATTCATTATGC) was labeled with ^{32}P gamma-ATP (6000 Ci/mmol, 25 µM, Perkin Elmer) with PNK (NEB). After heat-inactivation of PNK for 2 min at 95º C, sense and anti-sense (anti-sense sgRNA flanking sequences are: GGCCGCATAATGAATTCCCCAATACCC-as sgRNA-GTTCATATAATGC) oligonucleotides were annealed by letting cool down to room temperature. These annealed oligonucleotides were also cloned into a modified *pBS SK+* using a Xho I and Not I cut vector to assay sgRNA/Cas9 activity.

For sgRNA/Cas9 cleavage assays, DNA/sgRNA/Cas9 ratios of 1/10/10 were used in a 10 µl reaction using the buffer supplied (NEB) and DEPC-treated water. Typically Cas9 (100 nM final) was incubated with sgRNA (100 nM) for 10 min at 25º C before adding oligonucleotides (10 nM final) or plasmid DNA (10 nM, corresponds to ~25 ng/µl final concentration of a 3 kb plasmid). Plasmids were linearized after Cas9 digestion by first heat inactivating Cas9 for 2 min at 95ºC, and then adding 10 µl of a restriction enzyme (5 U) in NEB buffer 3. Adding a restriction enzyme together with Cas9 inhibited DNA scission by Cas9. Cleavage of oligonucleotides was analysed on 8 % denaturing polyacrylamide gels and plasmid DNA was analysed on ethidium bromide stained agarose gels.

RNA secondary structure was analyzed with RNAfold at http://rna.tbi.univie.ac.at (Gruber et al., 2008) using the following tracrRNA sequence:

GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAA AAGUGGCACCGAGUCGGUGCUUUUUU.

Criteria for optimal sgRNA design

Criteria to select sgRNAs that maintain the structure required for efficient Cas9 DNA scission were implemented in the server accessible at https://platinum-crispr.bham.ac.uk/predict.pl and are as follows. The first nucleotide of the sgRNA needs to be a G for efficient transcription initiation by RNA Pol III (Paule and White, 2000). For T7 mediated in vitro transcription, three Gs need to be added to sgRNAs of 23 nucleotides. Disruption of the tetraloop or loop 2 and 3 structures, and the sequence U A/G G C/U A/G of nucleotides 16-20, which will result in a tetraloop bulge mimic, were classified as low efficiency sgRNAs as these parts are recognized by Cas9 (Nishimasu et al., 2014). Similarly, a hairpin loop in the gRNA consisting of 4 or more base pairing nucleotides, or base pairing of nucleotides 17-20 of the gRNA (U/C U/C U U), or base-pairing of eight nucleotides within the seed region with looped-out nucleotides spaced by three base-pairing nucleotides (N11-N20), or a GC content below 15% (1 or 2 nts) or above 50% (11 or more nts) were also considered low efficiency. Medium efficiency was assigned for gRNAs with a GC content of 15-25 % (3-5 nts) or 40-50% (8-10 nts), or a low CG content in the seed region (less than 5 nts in nucleotides 11-20). In addition, two U's at position 19/20 of the gRNA reduce efficiency because of premature transcription termination (Graf et al., 2019). Further, we assigned a medium impact if both of the two nucleotides A_{51} A_{52} and G_{62} U_{63} in the three way junction of loop 1 were base paired or seven nucleotides within the seed region (N11-N20) base-paired with looped-out nucleotides spaced by three base-pairing nucleotides. Thus, we deemed an sgRNA optimal to allow Cas9 to cleave DNA with high efficiency, if the GC content is 30-35 % (6-7 nts) and none of the above criteria applied.

RNA extraction, RT-PCR and PCR on genomic DNA

Total RNA was extracted using Tri-reagent (SIGMA) and reverse transcription was done with Superscript II (Invitrogen) according to the manufacturer's instructions using an oligo dT primer. PCR was done for 40 cycles with 1 µl of cDNA, with 1 µg of genomic DNA or from a single fly after freezing and drying in 200 µl of isopropanol. Primers to detect the sgRNA/Cas9 induced deletion in *Ythdc1* were YT F1 (GCCGCTGTGACGCAGAATTTGTGTG) and YT R1 (GGCCGTGCATGTTGCGCATGTAGTCC), and in *Ythdf* were 64F1 (GCCGAGAAAGTGCACAAGGATACGGAG) and 64R1 (CAAGGAATGGCTGAAGCAGACTCCTTG). Primers to amplify parts of the body of the RNA also flanking an intron were for *Ythdc1* YT F2 (CCACGCTGCCGCAGAACGACGCCAATC) and YT R2 (GCGGCAGATCCAGTCAAGCTCGATGAC), and in *Ythdf* were 64F2 (GAGCTGCCTGTCGATTCCCAACTCGTG) and 64R2 (CCGCCCTCTTCGTGTCGCTCCTTGAAG). Primers to amplify parts of the *ewg* gene have been described elsewhere (Koushika et al., 1999).

Cloning of sgRNAs into *pUC 3GLA U6.1 BbsI*

To clone two sgRNAs expressed by U6 promoters, the "tracrRNA U6.3 promoter" fragment was amplified with left (AAGATATCCGGGTGAACTTC**G**N_{19}GTTTTAGAGCTAGAAATAGC) and right (GCTATTTCTAGCTCTAAAA**C**N_{19}CGACGTTAAATTGAAAATAGG) sgRNA primers from pUC 3GLA U6.1/3 sgRNA using Pwo polymerase (Roche) with initial 30 sec denaturation at 94ºC followed by two cycles 94ºC/30 sec, 49ºC/40 sec, 72ºC/45 sec, then two cycles 94ºC/30 sec, 51ºC/40 sec, 72ºC/45 sec and 22 cycles two cycles 94ºC/30 sec, 56ºC/40 sec, 72ºC/45 sec.

Transcription from the U6 promoter initiates with a G (bold, underlined). Although this G does not need to be present in the targeting sequence, it needs be included for folding of the sgRNA. The sequences for the sgRNAs in *Ythdc1* were gACAGGTATTCCCAAACTCAC and GACATGTAGCGTTCCCATGA, and for *Ythdf* were GTCCTGAAATACGAGCACAA and gATAACGAACATGTGGGATCT. The pUC 3GLA U6.1 BbsI vector was cut by BbsI and the "sgRNA1 tracrRNA U6.3 promoter sgRNA2" fragment was cloned by Gibson assembly according to the manufacturer's instructions (NEB). For sequencing, primer U6.1 Fseq (GCGCGTACGTCCTTCGCATCCTTATG) was used. The sequences for pUC 3GLA U6.1 BbsI and the pUC 3GLA U6.1/3 Ythdf sgRNA have been deposited (MK908408 and MK908407) and plasmids are available from Addgene.

Drosophila **genetics and phiC31 integrase-mediated transgenesis**

Flies were maintained on standard cornmeal agar food as described (Haussmann et al., 2013). *CantonS* was used as a wild type control. For the *Ythdc1*, the GFP marked *Mi(Michlits et al.)YT521-B*^{MI02006} transposon insert and the *w+* marked *Df(3L)Exel6094* deficiency were used, and for *Ythdf*, the RFP marked *PBac{SAstopDsRed}*^{LL04081} insert and the *w+* marked deficiency *Df(3R)ED6220* were used. For phiC31 mediated transformation, constructs were injected into *y*^{1} *w* M{vas-int.Dm}ZH-2A; PBac{y+-attP-3B}VK00013* with the landing site inserted at 76A as previously described (Haussmann et al., 2013). Prior to insertion of GFP marked constructs, the GFP and RFP markers had been removed from the *y1 w* M{vas-int.Dm}ZH-2A* landing site by Cre mediated recombination (Bischof et al., 2007; Zaharieva et al., 2015).

References

Abadi, S., Yan, W.X., Amar, D., and Mayrose, I. (2017). A machine learning approach for predicting CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. PLoS Comput Biol *13*, e1005807.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature *513*, 569-573.

Anreiter, I., Mir, Q., Simpson, J.T., Janga, S.C., and Soller, M. (2021). New Twists in Detecting mRNA Modification Dynamics. Trends Biotechnol *39*, 72-89.

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 *576*, 149-157.

Arimbasseri, A.G., and Maraia, R.J. (2015). Mechanism of Transcription Termination by RNA Polymerase III Utilizes a Non-template Strand Sequence-Specific Signal Element. Mol Cell *58*, 1124-1132.

Balacco, D.L., and Soller, M. (2019). The m(6)A Writer: Rise of a Machine for Growing Tasks. Biochemistry *58*, 363-378.

Bischof, J., Maeda, R.K., Hediger, M., Karch, F., and Basler, K. (2007). An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A *104*, 3312-3317.

Boivin, V., Deschamps-Francoeur, G., and Scott, M.S. (2018). Protein coding genes as hosts for noncoding RNA expression. Semin Cell Dev Biol *75*, 3-12.

Bosch, J.A., Birchak, G., and Perrimon, N. (2021). Precise genome engineering in Drosophila using prime editing. Proc Natl Acad Sci U S A *118*.

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.

Chuai, G., Ma, H., Yan, J., Chen, M., Hong, N., Xue, D., Zhou, C., Zhu, C., Chen, K., Duan, B.*, et al*. (2018). DeepCRISPR: optimized CRISPR guide RNA design by deep learning. Genome Biol *19*, 80.

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.

Deveson, I.W., Hardwick, S.A., Mercer, T.R., and Mattick, J.S. (2017). The Dimensions, Dynamics, and Relevance of the Mammalian Noncoding Transcriptome. Trends Genet *33*, 464-478.

Dezi, V., Ivanov, C., Haussmann, I.U., and Soller, M. (2016). Nucleotide modifications in messenger RNA and their role in development and disease. Biochem Soc Trans *44*, 1385-1393.

Dix, T.C., Brauer, U., Haussmann, I.U., and Soller, M. (2020). Gap-repair recombineering for manipulation of large high copy number plasmids. Current Protocols in Molecular Biology.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R., *et al*. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol *34*, 184-191.

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.

Doudna, J.A. (2020). The promise and challenge of therapeutic genome editing. Nature *578*, 229-236.

El-Brolosy, M.A., Kontarakis, Z., Rossi, A., Kuenne, C., Gunther, S., Fukuda, N., Kikhi, K., Boezio, G.L.M., Takacs, C.M., Lai, S.L., *et al*. (2019). Genetic compensation triggered by mutant mRNA degradation. Nature *568*, 193-197.

Farboud, B., and Meyer, B.J. (2015). Dramatic enhancement of genome editing by CRISPR/Cas9 through improved guide RNA design. Genetics *199*, 959-971.

Gagnon, J.A., Valen, E., Thyme, S.B., Huang, P., Akhmetova, L., Pauli, A., Montague, T.G., Zimmerman, S., Richter, C., and Schier, A.F. (2014). Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. PLoS One *9*, e98186.

Garcia-Doval, C., and Jinek, M. (2017). Molecular architectures and mechanisms of Class 2 CRISPR-associated nucleases. Curr Opin Struct Biol *47*, 157-166.

Graf, R., Li, X., Chu, V.T., and Rajewsky, K. (2019). sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. Cell Rep *26*, 1098-1103 e1093.

Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R., and Hofacker, I.L. (2008). The Vienna RNA websuite. Nucleic Acids Res *36*, W70-74.

Hachet, O., and Ephrussi, A. (2004). Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic localization. Nature *428*, 959-963.

Haussmann, I.U., Bodi, Z., Sanchez-Moran, E., Mongan, N.P., Archer, N., Fray, R.G., and Soller, M. (2016). m(6)A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. Nature *540*, 301-304.

Haussmann, I.U., Hemani, Y., Wijesekera, T., Dauwalder, B., and Soller, M. (2013). Multiple pathways mediate the sex-peptide-regulated switch in female Drosophila reproductive behaviours. Proc Biol Sci *280*, 20131938.

Haussmann, I.U., Li, M., and Soller, M. (2011). ELAV-mediated 3'-end processing of ewg transcripts is evolutionarily conserved despite sequence degeneration of the ELAV-binding site. Genetics *189*, 97-107.

Haussmann, I.U., Ustaoglu, P., Brauer, U., Hemani, Y., Dix, T.C., and Soller, M. (2019). Plasmid-based gap-repair recombineered transgenes reveal a central role for introns in mutually exclusive alternative splicing in Down Syndrome Cell Adhesion Molecule exon 4. Nucleic Acids Res *47*, 1389-1403.

Hille, F., Richter, H., Wong, S.P., Bratovic, M., Ressel, S., and Charpentier, E. (2018). The Biology of CRISPR-Cas: Backward and Forward. Cell *172*, 1239-1259.

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.

Jensen, K.T., Floe, 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.

Jiang, F., and Doudna, J.A. (2017). CRISPR-Cas9 Structures and Mechanisms. Annu Rev Biophys *46*, 505-529.

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.

Koushika, S.P., Soller, M., DeSimone, S.M., Daub, D.M., and White, K. (1999). Differential and inefficient splicing of a broadly expressed Drosophila erect wing transcript results in tissue-specific enrichment of the vital EWG protein isoform. Mol Cell Biol *19*, 3998-4007.

Labuhn, M., Adams, F.F., Ng, M., Knoess, S., Schambach, A., Charpentier, E.M., Schwarzer, A., Mateo, J.L., Klusmann, J.H., and Heckl, D. (2018). Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. Nucleic Acids Res *46*, 1375-1385.

Labun, K., Montague, T.G., Gagnon, J.A., Thyme, S.B., and Valen, E. (2016). CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. Nucleic Acids Res *44*, W272-276.

Lee, M.H., Siddoway, B., Kaeser, G.E., Segota, I., Rivera, R., Romanow, W.J., Liu, C.S., Park, C., Kennedy, G., Long, T., *et al*. (2018). Somatic APP gene recombination in Alzheimer's disease and normal neurons. Nature *563*, 639-645.

Liu, X., Homma, A., Sayadi, J., Yang, S., Ohashi, J., and Takumi, T. (2016). Sequence features associated with the cleavage efficiency of CRISPR/Cas9 system. Sci Rep *6*, 19675.

Ma, Z., Zhu, P., Shi, H., Guo, L., Zhang, Q., Chen, Y., Chen, S., Zhang, Z., Peng, J., and Chen, J. (2019). PTC-bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS components. Nature *568*, 259-263.

Michlits, G., Jude, J., Hinterndorfer, M., de Almeida, M., Vainorius, G., Hubmann, M., Neumann, T., Schleiffer, A., Burkard, T.R., Fellner, M., *et al*. (2020). Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. Nat Methods *17*, 708-716.

Moreno-Mateos, M.A., Vejnar, C.E., Beaudoin, J.D., Fernandez, J.P., Mis, E.K., Khokha, M.K., and Giraldez, A.J. (2015). CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat Methods *12*, 982-988.

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.

Paule, M.R., and White, R.J. (2000). Survey and summary: transcription by RNA polymerases I and III. Nucleic Acids Res *28*, 1283-1298.

Ren, X., Sun, J., Housden, B.E., Hu, Y., Roesel, C., Lin, S., Liu, L.P., Yang, Z., Mao, D., Sun, L., *et al*. (2013). Optimized gene editing technology for Drosophila melanogaster using germ line-specific Cas9. Proc Natl Acad Sci U S A *110*, 19012-19017.

Ren, X., Yang, Z., Xu, J., Sun, J., Mao, D., Hu, Y., Yang, S.J., Qiao, H.H., Wang, X., Hu, Q., *et al*. (2014). Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA parameters in Drosophila. Cell Rep *9*, 1151-1162.

Samson, M.L. (2008). Rapid functional diversification in the structurally conserved ELAV family of neuronal RNA binding proteins. BMC Genomics *9*, 392.

Singh, R., Kuscu, C., Quinlan, A., Qi, Y., and Adli, M. (2015). Cas9-chromatin binding information enables more accurate CRISPR off-target prediction. Nucleic Acids Res *43*, e118.

Sledzinski, P., Nowaczyk, M., and Olejniczak, M. (2020). Computational Tools and Resources Supporting CRISPR-Cas Experiments. Cells *9*.

Soller, M. (2006). Pre-messenger RNA processing and its regulation: a genomic perspective. Cell Mol Life Sci *63*, 796-819.

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.

Tuladhar, R., Yeu, Y., Tyler Piazza, J., Tan, Z., Rene Clemenceau, J., Wu, X., Barrett, Q., Herbert, J., Mathews, D.H., Kim, J., *et al*. (2019). CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. Nat Commun *10*, 4056.

Wang, T., Wei, J.J., Sabatini, D.M., and Lander, E.S. (2014). Genetic screens in human cells using the CRISPR-Cas9 system. Science *343*, 80-84.

Wierson, W.A., Welker, J.M., Almeida, M.P., Mann, C.M., Webster, D.A., Torrie, M.E., Weiss, T.J., Kambakam, S., Vollbrecht, M.K., Lan, M., *et al*. (2020). Efficient targeted integration directed by short homology in zebrafish and mammalian cells. Elife *9*.

Zaharieva, E., Haussmann, I.U., Brauer, U., and Soller, M. (2015). Concentration and localization of co-expressed ELAV/Hu proteins control specificity of mRNA processing. Mol Cell Biol *35*.

Zhang, D., Hurst, T., Duan, D., and Chen, S.J. (2019). Unified energetics analysis unravels SpCas9 cleavage activity for optimal gRNA design. Proc Natl Acad Sci U S A *116*, 8693-8698.

Figure 1

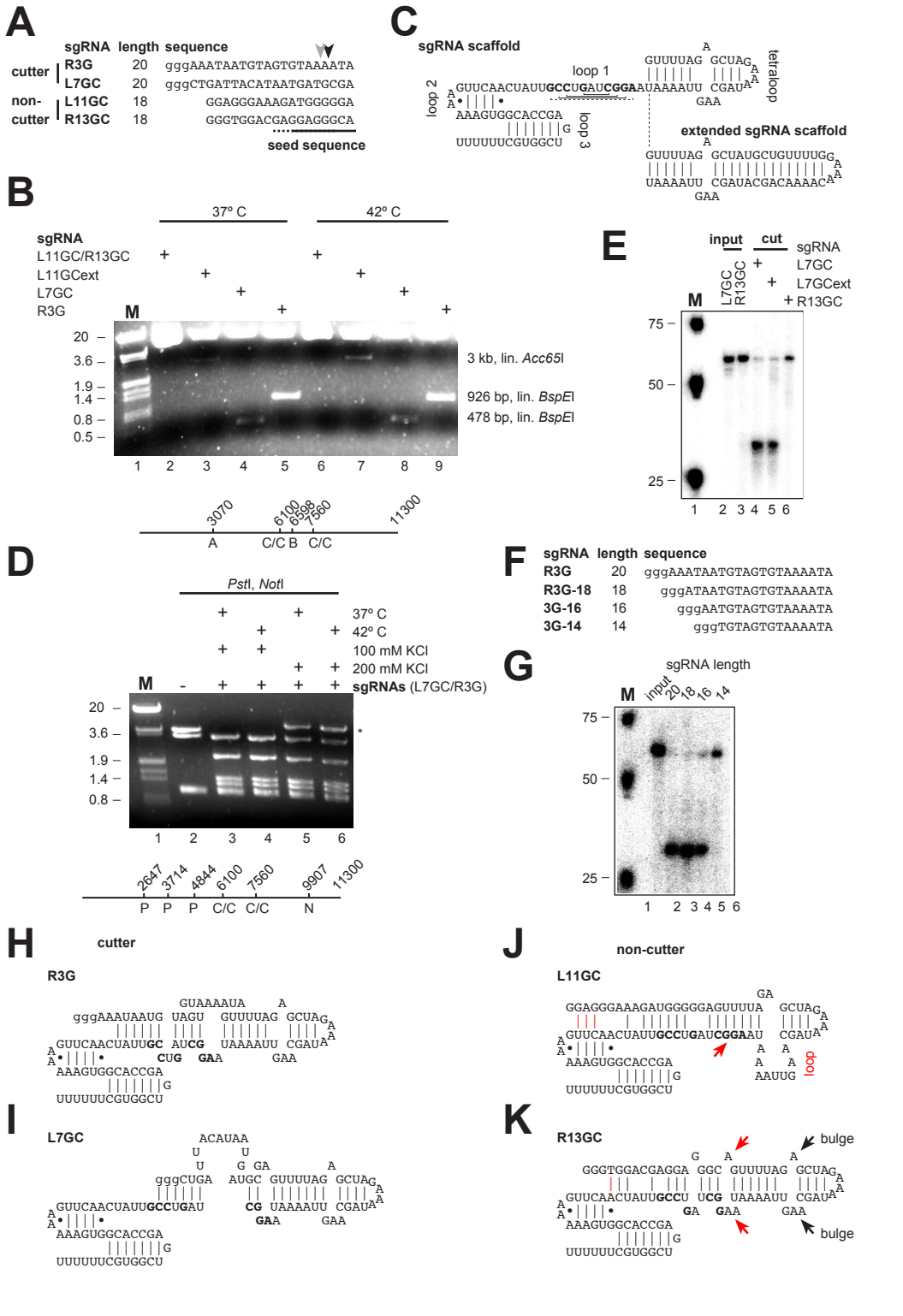
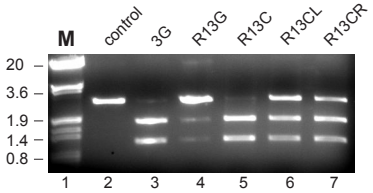

A
sgRNA length sequence
cutter R3G 20 gggAAATAATGTAGTGTAAAATA
L7GC 20 gggCTGATTACATAATGATGCGA
non-cutter L11GC 18 GGAGGGAAAGATGGGGGA
R13GC 18 GGGTGGACGAGGAGGGCA
seed sequence
B
37° C
42° C
sgRNA
L11GC/R13GC
L11GCext
L7GC
R3G
M
20
3.6
1.9
1.4
0.8
0.5
3 kb, lin. Acc65I
926 bp, lin. BspEI
478 bp, lin. BspEI
1 2 3 4 5 6 7 8 9
3070 6100 6398 7560 11300
A C/C B C/C
C
sgRNA scaffold
loop 1
loop 2
loop 3
tetraloop
extended sgRNA scaffold
D
PstI, NotI
37° C
42° C
100 mM KCl
200 mM KCl
sgRNAs (L7GC/R3G)
M
20
3.6
1.9
1.4
0.8
1 2 3 4 5 6
2647 3714 4844 6100 7560 9907 11300
P P P C/C C/C N
E
input cut
sgRNA
L7GC
L7GCext
R13GC
M
75
50
25
1 2 3 4 5 6
F
sgRNA length sequence
R3G 20 gggAAATAATGTAGTGTAAAATA
R3G-18 18 gggATAATGTAGTGTAAAATA
3G-16 16 gggAATGTAGTGTAAAATA
3G-14 14 gggTGTAGTGTAAAATA
G
sgRNA length
M input 20 18 16 14
75
50
25
1 2 3 4 5 6
H
cutter
R3G
I
L7GC
J
non-cutter
L11GC
loop
K
R13GC
bulge
bulge

Figure 2

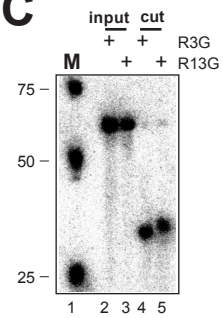
A

sgRNA	length	sequence
R3G	20	gggAAATAATGTAGTGTAAAATA
R13G	20	ggGAGGGAGGTGGTGTGAGAGG
R13C	20	ggGACCCACCTCCTCTCACACC
R13CL	20	ggGACCCACCTCCTCTGAGAGG
R13CR	20	ggGAGGGAGGTGGTGTCACACC

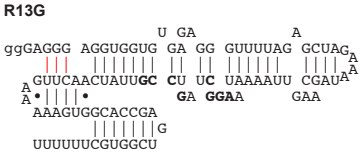
B



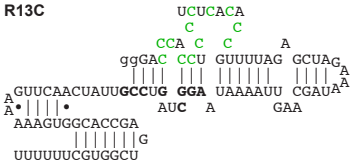
C



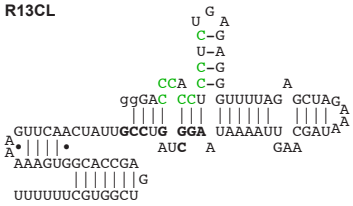
D



E



F



G

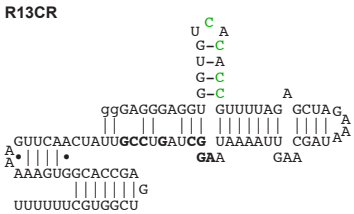


Figure 3

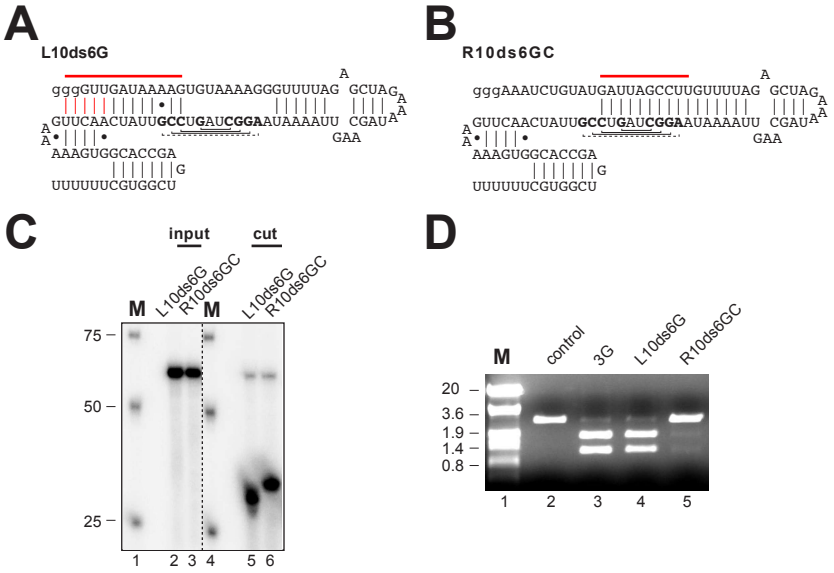

A
L10ds6G
B
R10ds6GC
C
input
cut
M
L10ds6G
R10ds6GC
75
50
25
1 2 3 4 5 6
D
M
control
3G
L10ds6G
R10ds6GC
20
3.6
1.9
1.4
0.8
1 2 3 4 5

Figure 4

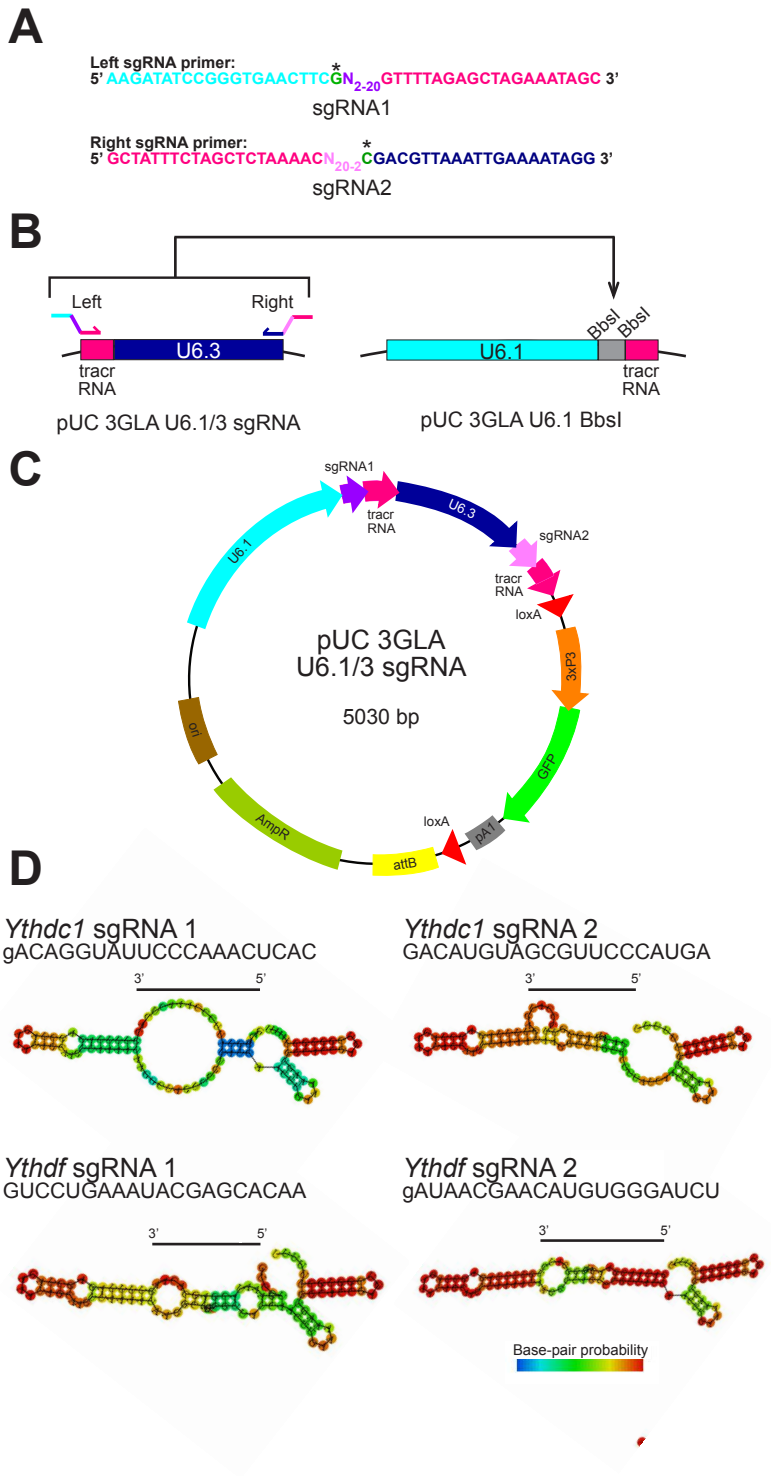
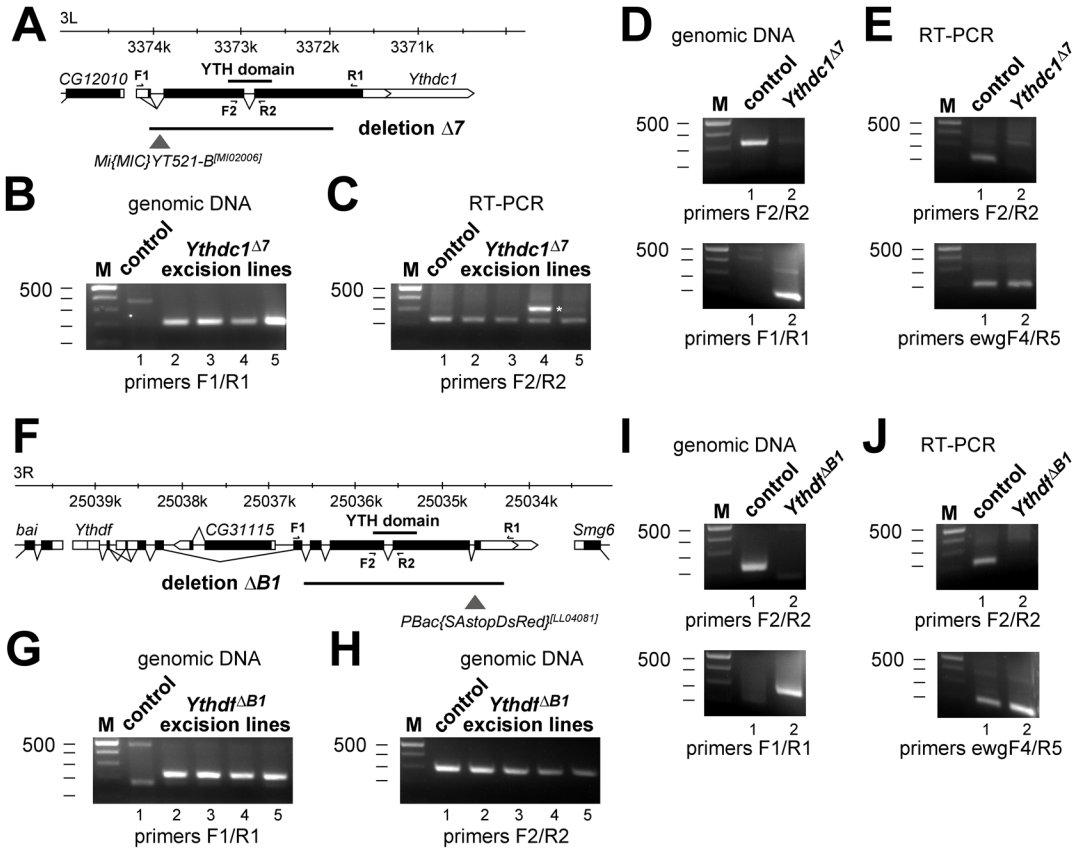
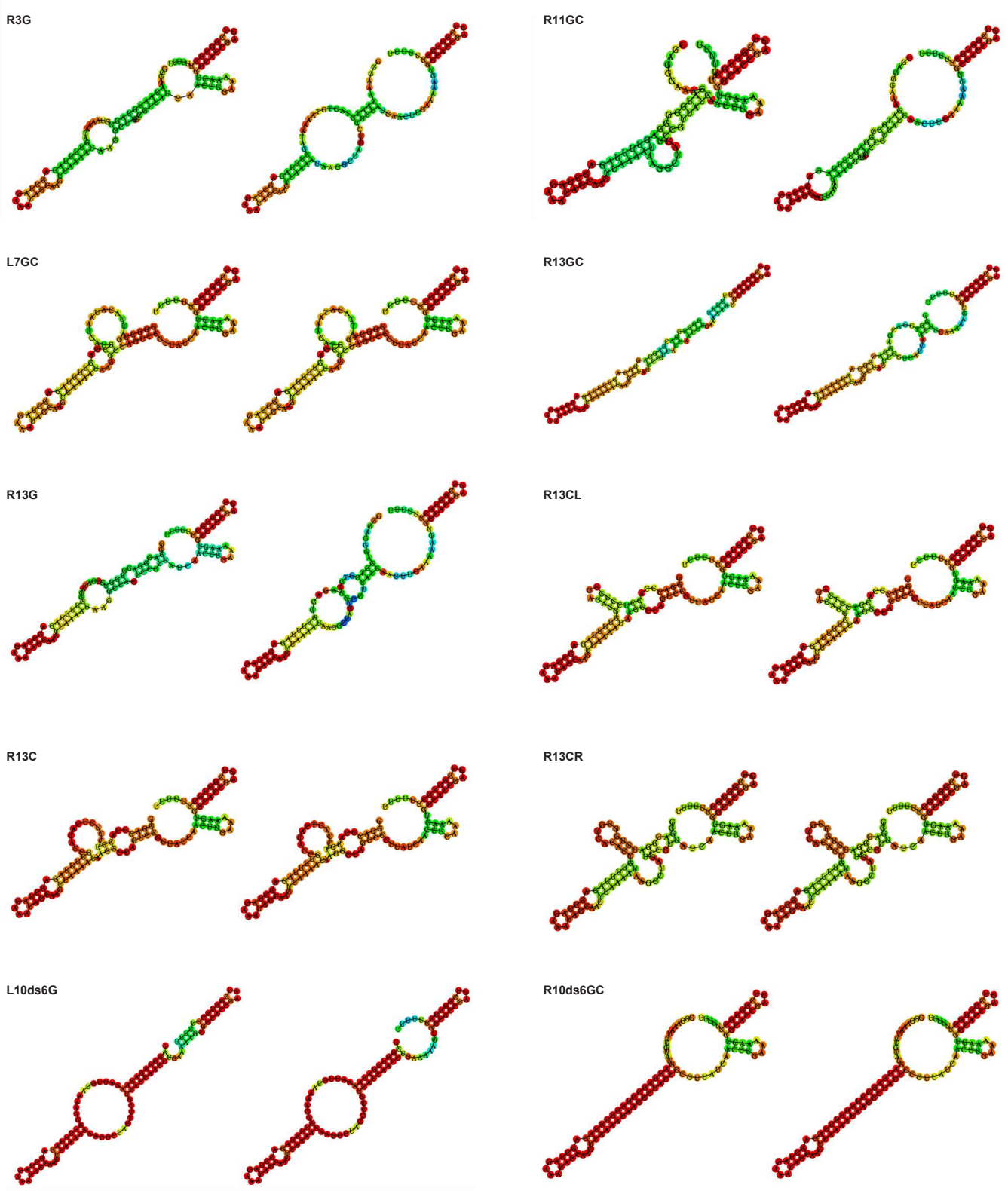

A
Left sgRNA primer:
5' AAGATATCCGGGTGAACTTCGN2-20GTTTTAGAGCTAGAAATAGC 3'
sgRNA1
Right sgRNA primer:
5' GCTATTTCTAGCTCTAAAACN20-2CGACGTTAAATTGAAAATAGG 3'
sgRNA2
B
Left
Right
U6.3
tracr RNA
pUC 3GLA U6.1/3 sgRNA
U6.1
BbsI
BbsI
tracr RNA
pUC 3GLA U6.1 BbsI
C
sgRNA1
tracr RNA
U6.3
sgRNA2
tracr RNA
loxA
3xP3
GFP
pA1
loxA
attB
AmpR
ori
U6.1
pUC 3GLA
U6.1/3 sgRNA
5030 bp
D
Ythdc1 sgRNA 1
gACAGGUAUUCCCAAACUCAC
3'
5'
Ythdc1 sgRNA 2
GACAUGUAGCGUUCCCAUGA
3'
5'
Ythdf sgRNA 1
GUCCUGAAAUACGAGCACAA
3'
5'
Ythdf sgRNA 2
gAUAACGAACAUGUGGGAUCU
3'
5'
Base-pair probability

Figure 5


A
3L
3374k
3373k
3372k
3371k
CG12010
F1
YTH domain
R1
Ythdc1
F2
R2
deletion Δ7
Mi{MIC}YT521-B[MI02006]
B
genomic DNA
M
control
Ythdc1Δ7 excision lines
500
1 2 3 4 5
primers F1/R1
C
RT-PCR
M
control
Ythdc1Δ7 excision lines
500
1 2 3 4 5
primers F2/R2
D
genomic DNA
M
control
Ythdc1Δ7
500
1 2
primers F2/R2
500
1 2
primers F1/R1
E
RT-PCR
M
control
Ythdc1Δ7
500
1 2
primers F2/R2
500
1 2
primers ewgF4/R5
F
3R
25039k
25038k
25037k
25036k
25035k
25034k
bai
Ythdf
CG31115
F1
YTH domain
R1
Smg6
F2
R2
deletion ΔB1
PBac{SAstopDsRed}[LL04081]
G
genomic DNA
M
control
YthdfΔB1 excision lines
500
1 2 3 4 5
primers F1/R1
H
genomic DNA
M
control
YthdfΔB1 excision lines
500
1 2 3 4 5
primers F2/R2
I
genomic DNA
M
control
YthdfΔB1
500
1 2
primers F2/R2
500
1 2
primers F1/R1
J
RT-PCR
M
control
YthdfΔB1
500
1 2
primers F2/R2
500
1 2
primers ewgF4/R5

Supplementary Figure S1



Supplementary Figure S2

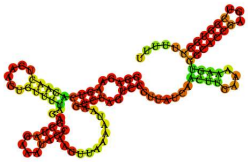
A

sgRNA	length	sequence
CG6422 5'	20	GGAGAAGUCAGAACUCAAGU
CG6422 3'	20	GAGGCAGAAACAAUGGAUCU

B

CG6422 5'

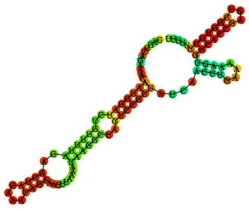
```
                          AA
                        C   G
                        U   U
                         C-G
                         A-U
                         A-U
                         G-U
                    GA   A-U
                GGA  AGUC  AGAGCUAG
                |||  ||||     ||||  A
  GUUCAACUAUUGCCU  UCGG     CGAU A A
A  ||||          GA    A   A
A AAAGUGGCACCGA        A   A
      ||||||| G        U   G
  UUUUUUCGUGGCU        A   U
                       A   U
                       A  A
                         A
```



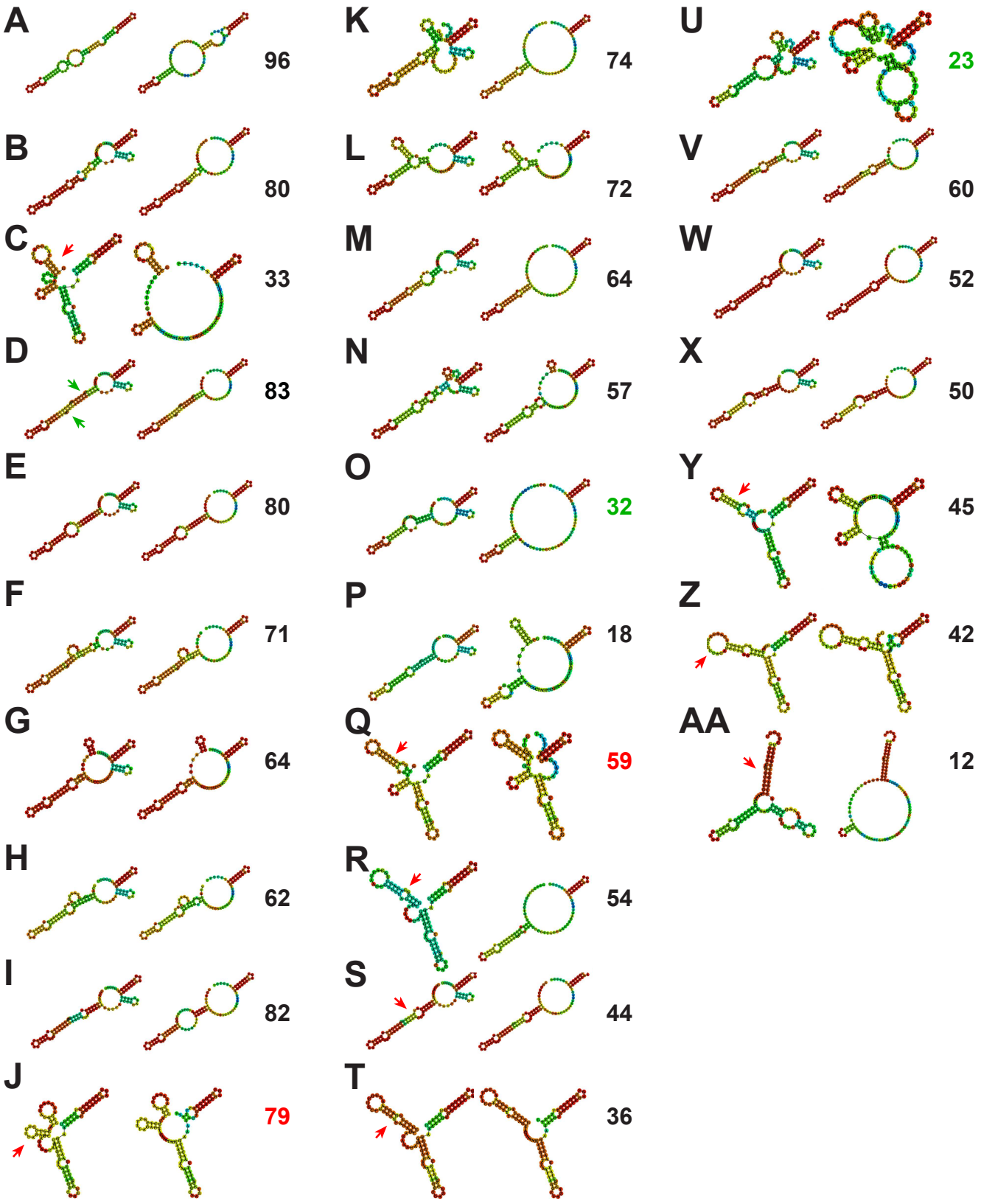
C

CG6422 3'

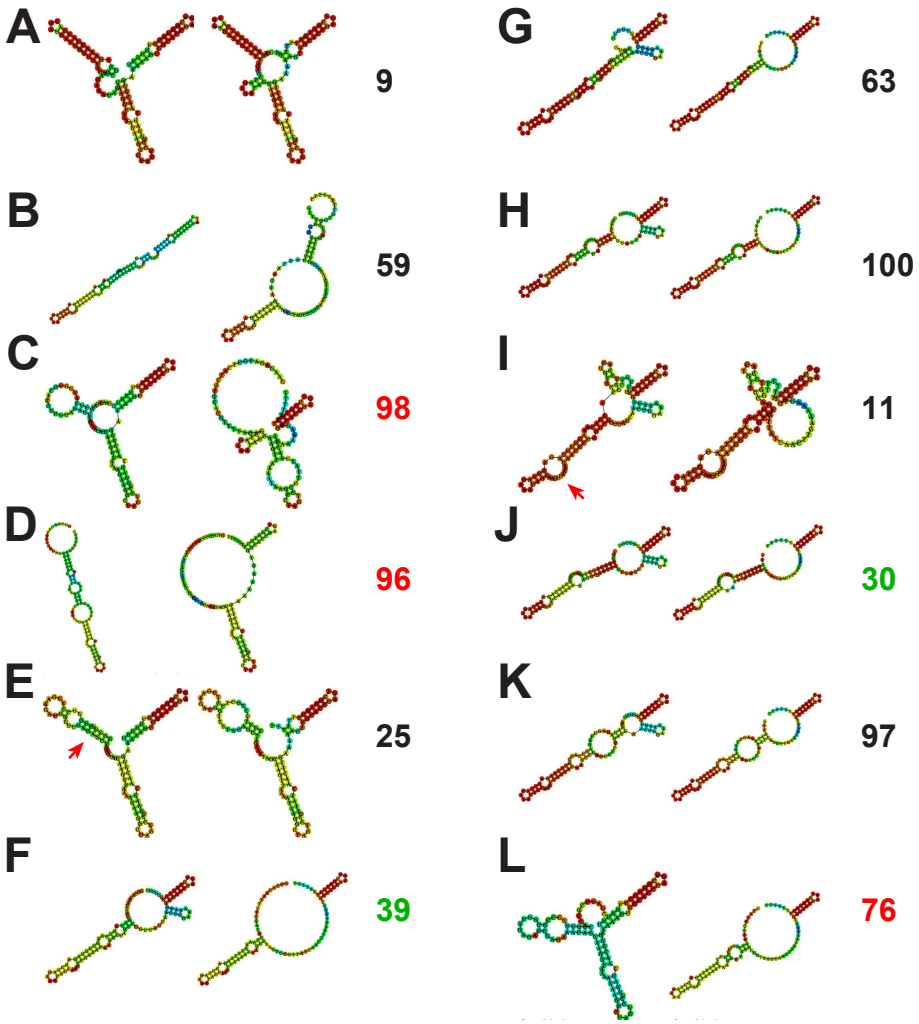
```
                        CU      G A
 GAGGCAGAAACAAUGGAU  GUUUUA   GCUA G
            |||||||  ||||||   |||| A
  AGUUCAACUAUUGCCUG  CGGAAU   CGAU A
  A ||||          AU      A  A
   AAAGUGGCACCGA          A  A
        ||||||| G         A  G
    UUUUUUCGUGGCU         A  U
                           U
```



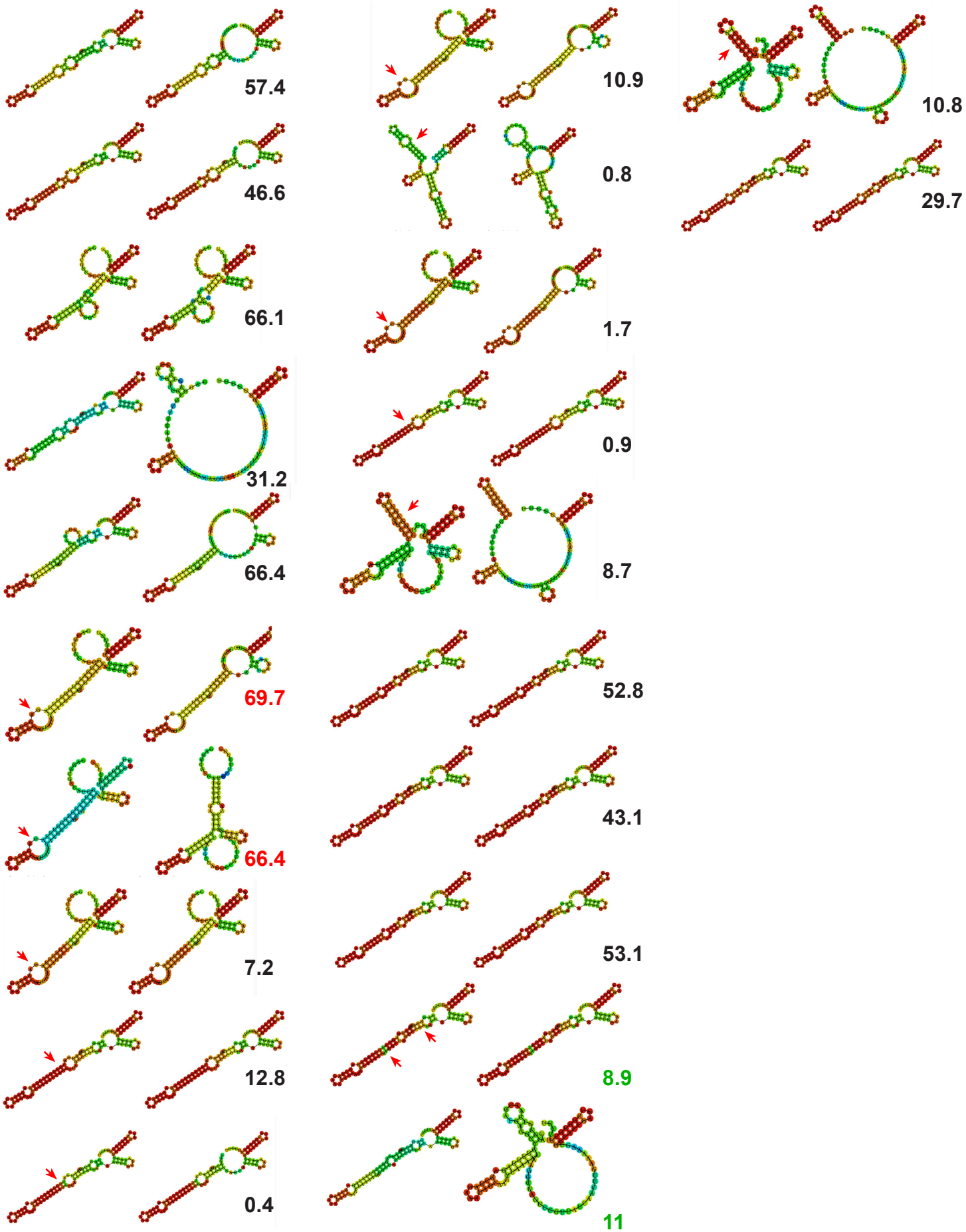
Supplementary Figure S3



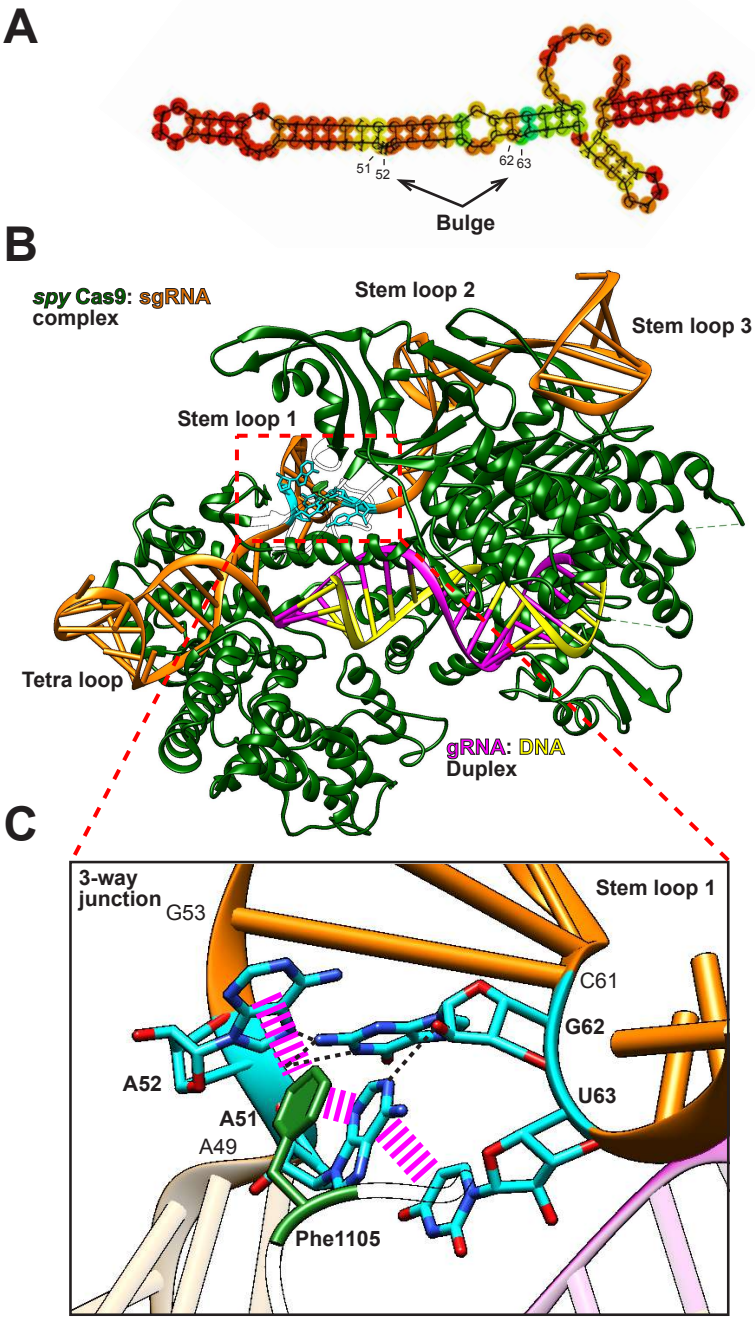
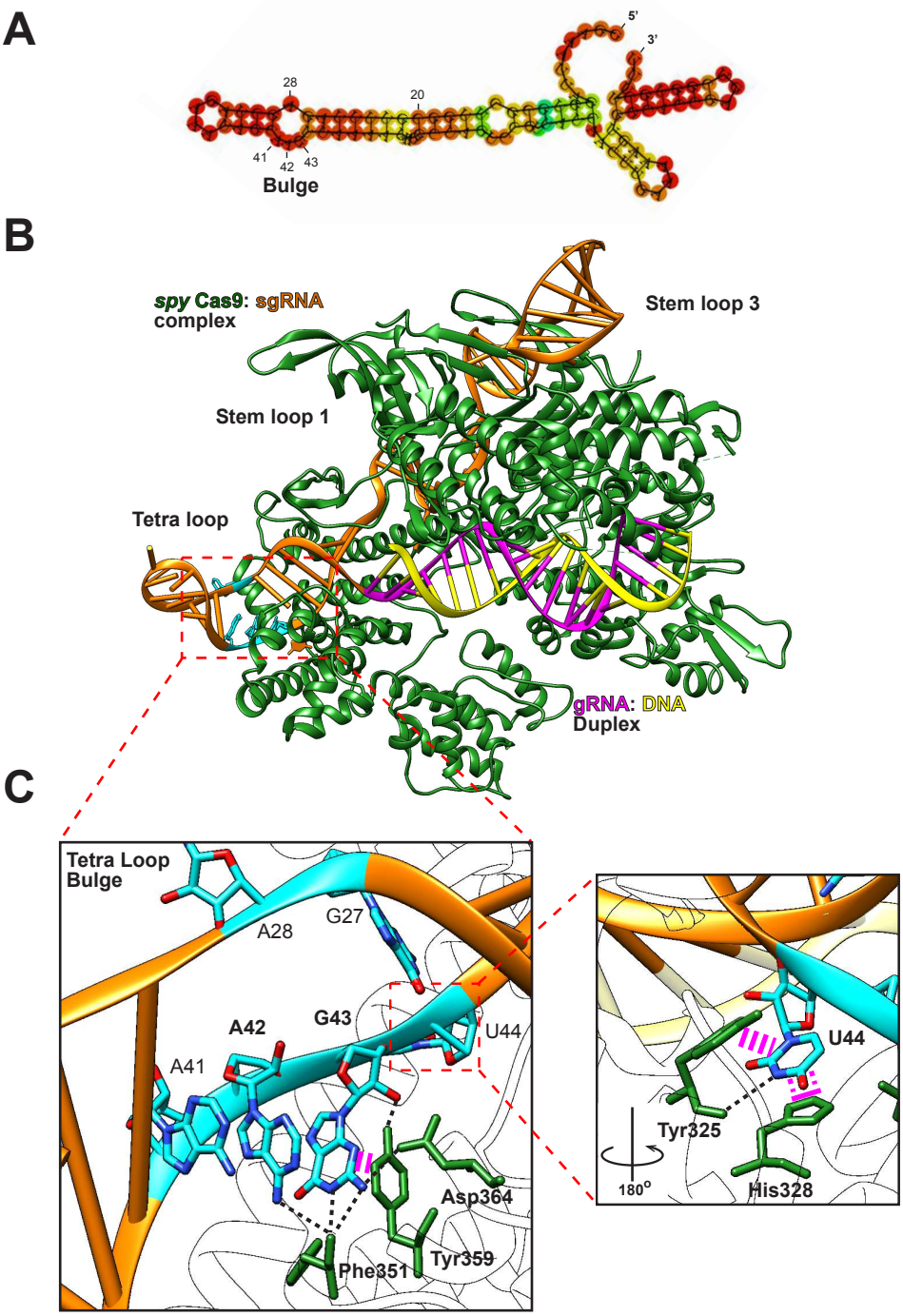
Supplementary Figure S4



Supplementary Figure S5



Supplementary Figure S6



Supplementary Figure S7