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

IRMGARD U. HAUSSMANN[1,2], THOMAS C. DIX[1], DAVID W. J. MCQUARRIE[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]College of Life Science, 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

Abstract

A 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. However, cleavage efficiency is below 50% for more than half of randomly selected sgRNA sequences in human cell culture screens or model organisms. Here, we used in vitro assays to determine intrinsic molecular parameters for maximal sgRNA activity including correct folding of sgRNAs and Cas9 structural information. From comparison of over 10 data sets, we find that major constraints in sgRNA design originate from maintaining the secondary structure of the sgRNA, sequence context of the seed region, GC context and detrimental motifs, but we also find considerable variation among different prediction tools when applied to different data sets. To aid selection of efficient sgRNAs, we developed web-based PlatinumCRISPr, a sgRNA design tool to evaluate base-pairing and known sequence composition parameters for optimal design of highly efficient sgRNAs for Cas9 genome editing. We applied this tool to select sgRNAs to efficiently generate gene deletions in *Drosophila Ythdc1* and *Ythdf*, that bind to N^6 methylated adenosines (m^6A) in mRNA. However, we discovered, that generating small deletions with sgRNAs and Cas9 leads to ectopic reinsertion of the deleted DNA fragment elsewhere in the genome. These insertions can be removed by standard genetic recombination and chromosome exchange. These new insights into sgRNA design and the mechanisms of CRISPR/Cas9 genome editing advances use of this technique for safer applications in humans.

45 ## Introduction

46 Bacterially derived Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-
47 associated protein 9 (Cas9) from *Streptococcus pyogenes* provides a powerful tool for precise
48 genome editing (Garcia-Doval and Jinek 2017; Jiang and Doudna 2017; Hille et al. 2018; Doudna
49 2020). To induce double stand-breaks in DNA at desired locations, the DNA scission enzyme Cas9
50 uses a guide RNA (gRNA) containing a 20 nucleotide complementary sequence to the genomic
51 target site (protospacer), which also requires the protospacer adjacent motif (PAM) at the 3'end,
52 comprised of NGG sequence (whereby N is any nucleotide and G is guanine). In addition to the
53 target complementary sequence (spacer) the gRNA also contains a constant crispr RNA (crRNA)
54 sequence that base-pairs with *trans*-activating crRNA (tracrRNA). Alternatively, a single gRNA
55 (sgRNA) can be used whereby the crRNA is fused to the tracrRNA through an artificial loop (Jinek
56 et al. 2012; Cong et al. 2013).

57 High efficiency CRISPR-Cas9 mediated DNA scission is essential to generate mutants at high
58 frequency in genetic screens and to provide the resource for efficient homologous recombination
59 directed gene replacements. Large scale analysis of sgRNA efficiencies revealed the whole
60 spectrum of on-target cleavage activities ranging from 0-100% arguing for a number of parameters
61 that need to be correct for high efficiency cleavage leading to models incorporating weighing of
62 features and/or thermodynamics of secondary structures (Hsu et al. 2013; Doench et al. 2014;
63 Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015;
64 Hart et al. 2015; Housden et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Xu et al.
65 2015; Doench et al. 2016; Liu et al. 2016; Abadi et al. 2017; Gandhi et al. 2017; Chuai et al. 2018;
66 Labuhn et al. 2018; Graf et al. 2019; Zhang et al. 2019; Michlits et al. 2020; Sledzinski et al. 2020;
67 Trivedi et al. 2020; Xiang et al. 2021; Riesenberg et al. 2023). These studies identified that

68 sequences with very low (≤35%) or very high (>80%) guanine-cytosine (GC) overall content were
69 less effective indicating a critical aspect for binding energy in target scission. In addition, purines in
70 the six nucleotides 5` to the PAM substantially increased Cas9 cleavage efficiency, while
71 pyrimidines and in particular uridine resulted in a lower efficiency (Ren et al. 2014; Wang et al.
72 2014; Housden et al. 2015; Graf et al. 2019). The lower efficiency of two uridine preceding the
73 PAM site was further associated with premature termination of RNA Pol III (Graf et al. 2019),
74 which terminates after a stretch of four to six uridines (Gao et al. 2018). Moreover, changes in
75 internal structure of sgRNA has been found associated with low activity (Moreno-Mateos et al.
76 2015; Thyme et al. 2016; Jensen et al. 2017). Recently, also bioinformatic approaches employing
77 machine-learning have been used to improve prediction of sgRNA cleavage efficiencies (Xiang et
78 al. 2021). These observations have helped to improve the design of sgRNAs to yield higher
79 efficiencies and incorporated into sgRNA design tools, but correlations between predictions and
80 guide activity vary considerably (Labun et al. 2016) (Haeussler et al. 2016; Sledzinski et al. 2020).
81 Accordingly, available rules to predict sgRNAs are currently not sufficient to guarantee high
82 cleavage efficiency and many sgRNA candidates scoring high fail to cleave efficiently (Labun et al.
83 2016) (Haeussler et al. 2016; Labuhn et al. 2018; Sledzinski et al. 2020). In particular, the impact of
84 sgRNA folding has not yet been analysed in detail and incorporated in web tools for sgRNA design.
85 The x-ray crystal structure of Cas9 bound to sgRNA has been determined and indicates four regions
86 of base-pairing termed tetraloop (tetraloop forms due to fusion of crRNA and tracrRNA) and stem
87 loops 1-3 that might be important for its function (Anders et al. 2014; Nishimasu et al. 2014). This
88 structure revealed many points of close interactions of the folded sgRNA with Cas9, but whether
89 disruptions in the sgRNA structure would impact on Cas9 cleavage efficiency has not

90 systematically been analyzed (Riesenberg et al. 2023). In particular, highly GC-rich guide RNAs
91 could disrupt the rather weak secondary structure of the sgRNA bound by Cas9.
92 The CRISPR-Cas9 genome editing tool is widely used to generate knock-out mutants by
93 introducing frameshifts. It has been recognized that introducing premature termination codons
94 (PTCs) can induce use of alternative translation initiation sites (Tuladhar et al. 2019). In addition, in
95 CRISPR-Cas9 engineered "knock-outs" of the m^6A mRNA methyltransferase METTL3, it has been
96 found that a functional ORF can be restored by altered splicing leaving considerable levels of m^6A
97 in mRNA (Poh et al. 2022). Likewise, compensatory responses have been observed involving
98 upregulating genes and as a consequence causing stronger phenotypes than compared to removing a
99 gene entirely (El-Brolosy et al. 2019; Ma et al. 2019). In addition, some genes have dual functions
100 as protein and RNA (Hachet and Ephrussi 2004). Hence, introducing a frameshift will only remove
101 the protein function. Likewise, many non-coding RNAs are present in introns suggesting that their
102 expression is connected to the expression of the host gene (Boivin et al. 2018), but such
103 relationships have not yet been explored comprehensively as they require more sophisticated
104 genome editing (Deveson et al. 2017). Thus, for generating gene knock-outs, removing the
105 transcription start site or the entire gene should be considered.
106 Initial concerns about CRISPR/Cas9 were about off target cleavage, but changing one nucleotide in
107 the spacer sequence complementary to the target efficiently abrogates activity (Ren et al. 2014).
108 However, from randomly chosen sgRNA sequences, more than half display cleavage efficiencies
109 below 50%, and we experienced complete inactivity in *Drosophila* mutagenesis or modification of
110 plasmids. From analyzing the causes of such inactivity, we discovered that maintaining the
111 secondary structure of sgRNAs essentially contributes to high-efficiency DNA scission of Cas9,
112 particularly in cold-blooded animals. Likewise, excessive base-pairing in the seed region also

113 impacts on Cas9 cleavage efficiency. Accordingly optimal design of sgRNA for high efficiency
114 DNA cleavage of Cas9 requires analysis of sgRNA secondary structure which is aberrant with
115 about 50% of PAM adjacent sequences in *Drosophila* and humans. To facilitate design of optimal
116 sgRNAs, we developed an online tool incorporating all currently known parameters for sgRNA
117 design including correct sgRNA folding (https://platinum-crispr.bham.ac.uk/predict.pl). However,
118 comparison of different sgRNA cleavage efficiency monitoring screens and various efficiency
119 prediction tools reveals considerable variation among different prediction tools when applied to
120 different data sets. We then applied the PlatinumCRISPr tool to identify high efficiency sgRNAs to
121 generate gene deletions in *Drosophila* using existing transposon marker lines. Using these novel
122 techniques we generated deletions of *Ythdc1* and *Ythdf*, that bind to N^6 methylated adenosines
123 (m^6A) in mRNA important in development (Dezi et al. 2016; Haussmann et al. 2016; Roignant and
124 Soller 2017; Balacco and Soller 2019; Anreiter et al. 2021). We further discovered, that generating
125 small deletions with sgRNAs and Cas9 leads to ectopic reinsertion of the deleted DNA fragment
126 elsewhere in the genome, but such inserts can be removed by standard genetic recombination and
127 chromosome exchange. Taken together, large scale analysis of sgRNA cleavage efficiencies in
128 screens together with new sgRNA design tools and insights into the mechanisms of CRISPR/Cas9
129 genome editing will help to develop this technique for safe application in humans.
130

131 ## Results

132 **RNA secondary structure constraints limit sgRNA/Cas9 activity**

133 Although sgRNA/Cas9 can cleave DNA efficiently, the first sgRNAs (L11GC and R13GC) we
134 designed according to previously published guidelines did not cut the *pUC 3GLA Dscam 3-5*
135 reporter we designed to study *Dscam* alternative splicing from introducing mutations by gap repair

136 recombineering (**Fig 1 A and B, Supplementary Fig S1, Supplementary Table S1**) (Hemani and
137 Soller 2012; Haussmann et al. 2019). Similarly, the first sgRNAs flanking the *Drosophila Ythdf*
138 gene, a reader for m^6A mRNA methylation did not result in a deletion of the locus based on
139 screening for loss of an RFP-marked transposon even though we had validated the target sequence
140 in the strain used (n=103, **Supplementary Fig S2, , Supplementary Table S1**)(Balacco and Soller
141 2019). To determine the intrinsic molecular parameters for maximal sgRNA activity, we devised an
142 in vitro assay to test the DNA scission efficiency of these two sgRNA based on in vitro transcribed
143 sgRNAs and commercially available Cas9 using oligonucleotide and plasmid substrates containing
144 matching protospacer sequences followed by a PAM.
145 It has previously been shown that extending the tetraloop in the constitutive component of sgRNAs
146 constituted by tracrRNA and crRNA enhances cleavage efficiency in vitro using oligonucleotide
147 substrates (**Fig 1C**) (Jinek et al. 2012). Introducing the extended sequence present in tracrRNA and
148 crRNA into sgRNA (L7GCext) did not increase efficiency of cleaving a plasmid at 37ºC, but
149 increasing the temperature to 42ºC enhanced cleavage by L7GCext (**Fig 1B**). Increasing the salt
150 concentration to 200 mM also did not result in enhanced cleavage by L7GC/R3G (**Fig 1D**). In
151 contrast, both L7GC and L7GCext could cleave an oligonucleotide, while R13GC did not (**Fig 1E**).
152 Using this assay also confirms the previous observation that sgRNAs of shorter length will lead to
153 cleavage. In addition, three guanosines introduced at the 5'end of sgRNAs required for efficient in
154 vitro transcription are tolerated (**Fig 1F and 1G**) (Jinek et al. 2012).
155 The sgRNA scaffold adopts a typical fold when bound to Cas9 consisting of the bulged tetraloop,
156 followed by small loop 1 and the more extended loops 2 and 3, which form a protective 3'end
157 structure (**Fig 1C**) (Nishimasu et al. 2014). The loop2/3 structure does not involve the uridines
158 incorporated for termination of RNA Pol III driven expression from plasmids (**Fig 1C**). When

159 comparing the secondary structures of the four sgRNAs, we noticed that the two well-cutting
160 sgRNAs L7GC or R3G maintained the secondary structure of the constitutive RNA part, while the
161 non-cutting sgRNA L11GC disrupted the structure of the tetraloop (**Fig 1H-K**). The effect of
162 R13GC seems more subtle as it could cut in the oligonucleotide assay suggesting that the repeated
163 bulge structure is the cause for its inefficiency, which is supported by x-ray crystal structure of the
164 Cas9-sgRNA-DNA complex. Here, the bulge structure is recognized by Cas9 where Tyr_{359} base-
165 stacks with G_{43}, that also forms hydrogen bonds with Asp_{364} and Phe_{351} , and Phe_{351} forms a
166 hydrogen bond with A_{42} (Nishimasu et al. 2014). In addition, the sgRNAs initially used for deleting
167 the *Ythdf* gene have a severely disrupted secondary structure (**Supplementary Fig S2**).
168 When we systematically analysed genome sequences from *Drosophila* or humans for correct
169 folding and activity of sgRNAs using the above parameters, about 50% of sgRNAs (241 from 481
170 and 503 from 973, respectively) did not fold properly. Also, only about 10-20 % of randomly
171 selected sgRNAs exert high cleavage efficiency suggesting that correct folding could essentially
172 contribute to high cleavage efficiency (Hsu et al. 2013; Doench et al. 2014; Gagnon et al. 2014; Ren
173 et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Moreno-Mateos et al.
174 2015; Doench et al. 2016; Liu et al. 2016; Abadi et al. 2017; Chuai et al. 2018; Labuhn et al. 2018;
175 Graf et al. 2019; Zhang et al. 2019; Michlits et al. 2020; Sledzinski et al. 2020).
176 When comparing the sequences of the cutting sgRNAs L7GC or R3G with the non-cutting L11GC
177 and R13GC sgRNAs, we further noticed that L11GC and R13GC sgRNAs contained more
178 guanosines, which in RNA can base-pair with C and U. To test if guanosines in the sgRNA limit
179 Cas9 activity we increased their number in R3G to 13 to make sgRNA 13G (**Fig 2A**). For the
180 design of the R13 sgRNA, care was taken not to disrupt the tetraloop, but we noticed the potential
181 to interfere with loop 2 (**Fig 2B**, see below). Intriguingly, sgRNA 13G is not capable of directing

182 Cas9 cleavage if the target sequence is present in a 3 kb plasmid, but is active with a short
183 oligonucleotide substrate (**Fig 2A-D**). Likewise, adding a restriction enzyme together with
184 sgRNA/Cas9 inhibited Cas9 in cleaving plasmid DNA suggesting that Cas9's ability to scan DNA
185 can be impaired separately from its ability to cleave DNA.

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

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

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

199 To further substantiate these findings, we analyzed the structures of sgRNAs from previous studies
200 in mammalian cells and *Drosophila* with regard to their cleavage efficiency of previous attempts to
201 define rules for sgRNA cleavage efficiency in vivo (Ren et al. 2014; Graf et al. 2019). Indeed, in 39
202 sgRNAs designed for use in *Drosophila* reduced cleavage efficiency in nine sgRNAs is associated
203 with disturbances of the sgRNA secondary structure resulting in a cleavage efficiency below 35%
204 (**Supplementary Figs S3 and S4, Supplementary Table S1**)(Ren et al. 2014). Similarly, from 22

205 sgRNAs designed for use in mammalian cells, 13 had a cleavage efficiency below 35% associated
206 with disturbances of the sgRNA secondary structure (**Supplementary Fig S5, Supplementary**
207 **Table S1**) (Graf et al. 2019). Similar result were also observed for the efficiency of sgRNAs in
208 honey bees (Roth et al. 2019).
209 Given the requirement for correct folding of the sgRNA for efficient Cas9 mediated DNA scission,
210 we further examined the x-ray crystal structure of the Cas9-sgRNA-DNA complex to see whether
211 this would provide additional instructions to design sgRNAs (Nishimasu et al. 2014). Indeed, the
212 first two nucleotides, adenosine 51 and 52 (A_{51} and A_{52}, **Supplementary Fig S6 and S7**) after the
213 tetraloop form an aromatic base-stacking interaction with phenylalanine 1105 (Phe_{1105}) of Cas9.
214 Furthermore, these interactions are stabilized by guanosine 62 (G_{62}) forming non-Watson Crick
215 hydrogen bonds with A_{51} A_{52} and Phe_{1105}, and uracil 63 (U_{63}) forms a base-stacking interaction with
216 A_{52}. These interactions indicate that base-pairing of the gRNA with these nucleotides of the
217 constant part of the sgRNA reduce Cas9 activity in Cas9 cleavage assays in vitro and mutagenesis
218 in vivo.
219

220 **Large scale evaluation of novel sgRNA design parameters**

221 Next, we incorporated all the features previously published and from this study into a bioinformatic
222 sgRNA design tool, https://platinum-crispr.bham.ac.uk/predict.pl (a stand-alone code is available
223 for non-commercial use upon request from the authors) (Hsu et al. 2013; Doench et al. 2014;
224 Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer 2015;
225 Moreno-Mateos et al. 2015; Doench et al. 2016; Liu et al. 2016; Abadi et al. 2017; Chuai et al.
226 2018; Labuhn et al. 2018; Graf et al. 2019; Zhang et al. 2019; Michlits et al. 2020; Sledzinski et al.
227 2020). The included features include validation of intact secondary structures (tetraloop, loop 2 and
228 3, **Figs 1-3**), presence of a tetraloop bulge mimic (**Fig 1K**), self-complementarity of the gRNA (nts

229 1-20, **Figs 2F and 2G**) (Moreno-Mateos et al. 2015; Thyme et al. 2016; Jensen et al. 2017), GC
230 content in the six nucleotide seed region of the gRNA (nts 15-20, **Supplementary Data 1**, (Ren et
231 al. 2014; Wang et al. 2014; Graf et al. 2019), GC content of the gRNA (nts 1-20, **Supplementary**
232 **Data 1**, (Ren et al. 2014; Wang et al. 2014; Graf et al. 2019), the UUYY motif (nts 16-20), which
233 results in complete base-pairing (**Fig 3**) and can act as a Pol III termination signal (Gao et al. 2018),
234 the UCYG and CYGR motifs (nts 16-20) associated with lower cleavage efficiency (Graf et al.
235 2019) and lack of base-pairing of nucleotides 40, 41, 51 and 52 that are engaged in contacts with
236 Cas9 (**Supplementary Fig S6 and S7**).
237 To identify additional parameters affecting sgRNA cleavage efficiency, we performed a motif
238 analysis among the 35% low scoring sgRNAs for a number of different data sets (Doench et al.
239 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer
240 2015; Hart et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Xu et al. 2015; Doench et
241 al. 2016; Gandhi et al. 2017; Xiang et al. 2021) , but we did not find motifs associated with low
242 performance in individual data sets.
243 We then analysed the performance of a *Drosophila* sgRNA data set (Ren et al. 2014) according to
244 sgRNA design parameters described above. Our design tool PlatinumCRISPr selected 13 from 39
245 sgRNAs and those showed a cleavage efficiency of 55 % or more (**Fig 4A**). We then analysed a
246 number of sgRNA prediction tools for this data set including Chariscore (Chari et al. 2015),
247 Crispron (Xiang et al. 2021), DeepSpCas9 (Kim et al. 2019), DoenchScore (Doench et al. 2014),
248 Azimuth (implemented in ChopChop)(Doench et al. 2016), Moreno-Mateos Score (Moreno-Mateos
249 et al. 2015), Wang Score (Wang et al. 2014), Wong Score (Wong et al. 2015) and Xu Score (Xu et
250 al. 2015). PlatinumCRISPr significantly outperformed all of these prediction tools with the
251 *Drosophila* data set (Ren et al. 2014)(**Fig 4B**).

252 Next, we analysed 14 data sets from various organisms (*Drosophila*, zebrafish, sea squirt, worms
253 and cell culture cells,), which determined sgRNA cleavage efficiency for their performance using
254 the PlatinumCRISPr design tool (Doench et al. 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et
255 al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Hart et al. 2015; Moreno-Mateos et al. 2015;
256 Varshney et al. 2015; Xu et al. 2015; Doench et al. 2016; Gandhi et al. 2017; Xiang et al. 2021). For
257 overall performance (**Fig 5**), six data sets yielded significant ($p \leq 0.05$) enrichment of high efficiency
258 performing sgRNAs (Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Moreno-Mateos et al.
259 2015; Xiang et al. 2021), and five showed enrichment ($p \leq 0.25$) (Gagnon et al. 2014; Farboud and
260 Meyer 2015; Hart et al. 2015; Varshney et al. 2015; Xu et al. 2015), while two failed to show
261 enrichment for most of the parameters (Doench et al. 2014; Doench et al. 2016). In this analysis, we
262 noticed that structural constraints were significantly more important in cold-blooded organism,
263 where sgRNAs delivery is by injection in the absence of selection in contrast to cell culture cells,
264 where delivery is by transfection and selection for chronic exposure to sgRNAs for up to 10 days
265 before analysis.
266 When we analysed the performance of PlatinumCRISPr, Chariscore (Chari et al. 2015), Crispron
267 (Xiang et al. 2021), DeepSpCas9 (Kim et al. 2019), Doench Score (Doench et al. 2014), Azimuth
268 (implemented in ChopChop)(Doench et al. 2016), Moreno-Mateos Score (Moreno-Mateos et al.
269 2015), Wang Score (Wang et al. 2014), Wong Score (Wong et al. 2015) and Xu Score (Xu et al.
270 2015) prediction tools (**Supplementary Fig 8A-L**) with the different data sets determining sgRNA
271 cleavage efficiency we found that Wang Score performed best on the Doench data set and that
272 PlatinumCRISPr and Moreno-Mateos performed best with *Drosophila* and zebrafish generated data
273 sets, respectively, but we found no single prediction tool that stood out. In addition, in five out of

274 six cell culture generated data sets none of the prediction tools outperformed the others or
275 substantially increased prediction efficiency (**Supplementary Fig 8A-L**).
276 As part of this analysis, we noticed that the average cleavage efficiency in the analyzed data sets
277 varied substantially (from 20-75% average cleavage efficiency, Sup Fig 8A-L) pointing towards a
278 bias in outcome when testing various prediction tools with different data sets. To identify common
279 patterns among the different prediction tools applied to different data sets in the following analysis
280 we therefor excluded data sets with cleavage efficiencies below 30% for further analysis (Gagnon et
281 al. 2014; Chari et al. 2015; Farboud and Meyer 2015; Doench et al. 2016).
282 The default element of ChopChop is Azimuth (Labun et al. 2019), but also implements elements
283 from Doench Score, Chari Score, Xu Score and Moreno-Mateos Score. We therefore reasoned that
284 a combination of two prediction tools could result in more reliable selection of high efficiency
285 cleaving sgRNAs. When we combined two prediction tools the highest scoring combination was
286 PlatinumCRISPr together with Wong score with an average cleavage prediction of 61% (Fig 6A)
287 and this combination also outperformed in all the remaining data sets (Fig 6B), but this increased
288 the stringency and only very few sgRNAs were selected.
289

290 **A *Drosophila* transformation vector for expression of two sgRNAs**

291 Next, we applied these novel sgRNA design rules to generate *Drosophila* gene deletions. For this
292 purpose we generated a new fly transformation vector with a GFP marker (Solomon et al. 2018),
293 that is easier to select than previously generate *vermillion* marked vectors that require a *vermillion*
294 mutant background for transgene identification (Port et al. 2014; Trivedi et al. 2020). This vector
295 expresses two 20 nt sgRNAs from *U6.1* and *U6.3* promoters (Supplementary **Fig S9A-C**),
296 harboring a G as first nucleotide as a requirement for expression from the *U6* promoter (Paule and

297 White 2000; Ren et al. 2013). This plasmid can be generated by incorporating the two sgRNA
298 sequences in PCR primers for single step cloning into the plasmid, while previously published
299 vectors require two cloning steps or plasmid recombination (Port et al. 2014; Trivedi et al. 2020).
300 This sgRNA vector can then be injected into *Drosophila* expressing Cas9 in the germline for
301 CRISPR to induce mutations. Alternatively, this vector can be used to generate a transgenic line via
302 the attB site using phiC31 integrase mediated transformation. This fly strain is then crossed to a line
303 expressing Cas9 in the germline for generation of the desired genetic lesion. Transgenically
304 provided sgRNA/Cas9 generally results in a higher efficiency, because the sgRNAs are been
305 provided maternally.
306
307 **Efficient generation of gene deletions by sgRNA/Cas9 using transposon markers**
308 To generate deletions of the YTH protein genes *Ythdc1* and *Ythdf*, which are located on the third
309 chromosome, transposon inserts *Mi{MIC}YT521-B*^{MI02006} and *PBac{SAstopDsRed}*^{LL04081} marked
310 with *GFP* or *RFP*, respectively, were combined with an X-linked *vasCas9* or for *nosCas9* germline
311 expression of Cas9. To allow for detecting loss of the transposon in the YTH protein genes the GFP
312 and RFP markers of the *vasCas9* insert had been removed. These flies were then crossed to the GFP
313 marked sgRNA construct inserted on the 3rd chromosome (**Fig 7A and F**). The sgRNA insert has a
314 weak GFP marker and can generally be distinguished from Mi{MIC} inserts. Females from this
315 cross were mated with males containing *TM3 Sb/TM6 Tb* double-balancers in single crosses to
316 recover the sgRNA induced individual deletions and avoid analysis of clonal events. The male
317 progeny was then screened for loss of the GFP or RFP marker. Males which had lost the marker
318 were detected in 100 % (n=9) of the crosses for *Ythdc1* and in 88 % (n=9) for *Ythdf*, respectively.
319 This is a substantial increase of efficiency over imprecise P-element excision with a frequency of

320 0.01 to 1% (Soller et al. 2006; Haussmann et al. 2016; Haussmann et al. 2022). In those crosses
321 with loss of markers, all males for *Ythdc1* had lost the GFP marker, while for *Ythdf* the average
322 frequency of marker loss was 42 % (n=8). In the reverse cross for *Ythdf*, the frequency was 0%
323 (n=5) indicating that the low frequency in males is linked to the absence of recombination in males.
324 The identified single marker-less males were then crossed to *TM3 Sb/TM6 Tb* double-balancers to
325 establish a line and analysed with PCR using primers next to the deletion breakpoints yielding a
326 short PCR product (**Fig 7B and G**). A PCR product had been obtained in all lines were the marker
327 had been lost indicating that the expected deletion had indeed been generated. To generate *Ythdc1*
328 excision lines *nosCas9* was used for germline expression of Cas9 as *vasCas9* together with sgRNAs
329 targeting *Ythdc1* resulted in female sterility.
330

331 **sgRNA/Cas9 scissioned fragments insert elsewhere in the genome**

332 After establishing the lines, we noticed that all lines (n=10) established for the *Ythdc1* deletion did
333 not show the flightless phenotype previously reported (Haussmann et al. 2016). Therefore, we
334 selected four lines for further analysis by RT-PCR from RNA (*Ythdc1* excision lines, **Fig 7C**) or of
335 genomic DNA (*Ythdf* excision lines, **Fig 7H**) of homozygous flies with primers that were within the
336 deletion and also flanked an intron. Unexpectedly, a copy of the gene was still present in all *Ythdc1*
337 and *Ythdf* excision lines analyzed suggesting that the deleted fragment had been inserted elsewhere
338 in the genome.
339 To remove this ectopic insert(s), positive lines were crossed to *w*+ marked deficiencies and out-
340 crossed for two generations. The X and 2nd chromosomes were then exchanged to establish null-
341 mutant lines from single chromosomes for *Ythdc1* and *Ythdf* that were confirmed by RT-PCR to be
342 free of any ectopic inserts (**Fig 7D, E and Fig 7I, J**). To avoid such complications in the future we

343 generated a PBac w+ containing vector which can be efficiently inserted into a locus by cloning left
344 and right homology arms either to induce a partial deletion upon insertion. Alternatively, mutations
345 in sgRNA cleavage sites can be introduced into the homology arms to insert point mutations
346 followed by scar less removal of the PBac w+ by transposase.
347

348 **Discussion**

349 DNA scission by the sgRNA/Cas9 complex is highly specific and requires complete base-pairing
350 between the sgRNA and the target DNA generally not tolerating single miss-matches (Ren et al.
351 2014; Farboud and Meyer 2015). This feature makes the sgRNA/Cas9 complex an ideal tool for
352 genome editing, but its use is currently limited by the low predictability to cut its target in the
353 genome (Haeussler et al. 2016; Labuhn et al. 2018; Sledzinski et al. 2020).
354 Here, we discovered that the structure of the sgRNA is a key determinant for the scission efficiency
355 of the sgRNA/Cas9 complex in *Drosophila.* However, only about 50% of sequences adjacent to
356 PAM sites constitute sgRNAs that fold properly or are not compromised by unfavourable base-
357 pairing. In support of these two levels of interference, sgRNA R13GC correctly folds the tetraloop
358 and loop2/3, but did not cleave a short oligonucleotide substrate. This indicates that the structure of
359 this sgRNA blocks the catalytic activity of the sgRNA/Cas9 complex, likely by mimicking the
360 bulge structure of the tetraloop. Second, some sgRNAs allowed cleavage of short oligonucleotide
361 substrates (e.g. L7GC and R10ds6GC), but did not support efficient DNA scission of the target
362 sequence in the context of a 3 kb test-plasmid. Likely, these sgRNAs interfere with the ability of the
363 sgRNA/Cas9 complex to scan the DNA for target sites.
364 When we analyzed cleavage efficiencies of sgRNAs used in Drosophila (Ren et al. 2014), we
365 observed a good overlap with the ability of those sgRNAs to adopt the correct structure and having

366 a high cleavage efficiency. Further refinement to the design of sgRNAs comes from the recognition
367 that the GC content in the seed region is a major determinant to cleavage efficiency in addition to
368 general GC content. Analysis of the x-ray crystal structure of the Cas9-sgRNA-DNA complex also
369 revealed that the two As in the tracrRNA before the tetraloop engage with Cas9 through base
370 stacking and hydrogen bonds (Nishimasu et al. 2014). Base-pairing of these two As with Us at the
371 end of the sgRNA (N_{19} and N_{20}) before the start of the tracrRNA impact sgRNA/CAS9 complex
372 function and reduce cleavage efficiency (Graf et al. 2019). Likewise, if sgRNAs are made by RNA
373 Pol III, terminate occurs at the boundary of the tracrRNA if two UU precede the GUUUU of the
374 start of the tracrRNA (Arimbasseri and Maraia 2015; Graf et al. 2019). Motif searches in various
375 dataset to determine sgRNA cleavage efficiencies did not reveal any further motifs that impact on
376 cleavage efficiency. If such bias exists, this would like have been exploited by parasites of
377 prokaryotic hosts.
378 The bacterial CRISPR-Cas9 system consists of two RNAs (crRNA and trRNA) that assemble with
379 Cas9 to form the active complex. crRNA and trRNA base-pair through sequence complementarity
380 to form the tetra loop in the active Cas9 complex (Garcia-Doval and Jinek 2017; Jiang and Doudna
381 2017; Hille et al. 2018). In type II systems the tracrRNA is required for crRNA maturation
382 suggesting that base-pairing takes place while being assembled with Cas9. After the crRNA is
383 trimmed, the entire CRISPR-Cas9 complex can scan genomic DNA for DNA scission sites.
384 Alternatively, the crRNA could hybridize first to genomic DNA and recruit tracrRNA and Cas9 to
385 form a complex for DNA scission on site. In this scenario, the crRNA would not be able to interfere
386 with tracrRNA/Cas9 complex activity by forming an aberrant RNA secondary structure, but
387 whether this second scenario could be applied to more efficient genome editing with reduced off-
388 target cleavage needs to be tested. Of note, the effect of structural constraints are stronger in cold-

389 blooded animals likely reflecting that the optimal temperature for *E. coli* is 37ºC. In this context, it
390 would be worth exploring how much longer the sgRNA can be to still support Cas9 DNA scission
391 as longer RNAs would more stably hybridize to DNA. In either case, however, understanding
392 sgRNA/CAS9 complex assembly will inform how to prevent off-target DNA scission.
393 In this study, we also compared various sgRNA cleavage efficiency prediction tools with 12
394 datasets that have determined sgRNA cleavage efficiencies in human cells and various model
395 organisms including Drosophila, zebrafish, C. elegans, honey bees and seasquirt (Doench et al.
396 2014; Gagnon et al. 2014; Ren et al. 2014; Wang et al. 2014; Chari et al. 2015; Farboud and Meyer
397 2015; Hart et al. 2015; Moreno-Mateos et al. 2015; Varshney et al. 2015; Wong et al. 2015; Xu et
398 al. 2015; Doench et al. 2016; Gandhi et al. 2017; Kim et al. 2019; Roth et al. 2019; Xiang et al.
399 2021). Here, the PlatinumCRISPr tool deemed best for *Drosophila*, Moreno-Mateos Score for zebra
400 fish and Wang Score for one human cell culture screen, but surprisingly all prediction tools failed to
401 convince for the other five cell culture screens. Possibly, chronic exposure over several days in cell
402 culture systems could lead to a bias in determining sgRNA cleavage, compared to short exposure
403 when injected into early stage embryos like in insects and zebra fish.
404 Taken together, optimizing sequence composition and structural constraints in sgRNA design
405 essentially contributes to high DNA cleavage efficiency. Accordingly, optimized sgRNAs show
406 very little base-pairing with sequences adjacent to the loop 1 region, or are not complementary to
407 the tetraloop structure and/or the loop2/3 structure. In addition, avoiding base-pairing in the 10 nt
408 seed region prior to the PAM site predicts high efficiency of sgRNAs for DNA scission.
409 Furthermore, avoiding two Us before the PAM site prevents interference with the sgRNA/Cas9
410 structure. In any case, however, by introducing the target sequence into a plasmid the efficiency of a
411 particular sgRNA can be reliably determined in an in vitro cleavage assay. Although this assay will

412 determine the cleavage efficiency of an sgRNA, cellular features such as chromatin state can also
413 impact on Cas9 mediated DNA scission (Singh et al. 2015). In addition, whether genes are
414 expressed at the time of sgRNA cleavage likely also plays a role in the observed cleavage efficiency
415 as expression is associated with less compacted DNA.

416 Generating deletions of entire genes, or essential parts of them is the preferred way to generate a
417 null allele. This approach will avoid complications arising from introducing frameshifts at the
418 beginning of the ORF, as translation could reinitiate from later AUG or CUG start codons
419 (Koushika et al. 1999). In addition, in some genes the RNA has functions on its own, as shown for
420 *oscar* RNA that forms a large RNP particle with Oscar and other RNA binding proteins at the
421 posterior pole of a *Drosophila* oocyte (Hachet and Ephrussi 2004; Haussmann et al. 2011). Thus,
422 deletion of the entire gene region will discover such additional functions harbored in gene
423 transcripts.

424 When generating deletions of entire genes, we were very surprised to discover that the deleted DNA
425 fragment was inserted into the genome and transcribed, resulting in the expression of protein that
426 rescued the flightless phenotype in *Ythdc1* deletion allele. Although the mechanism for the
427 generation of these new inserts is not known, it seemed not to have led to complex chromosomal
428 aberrations, because the inserts could be removed by standard recombination and/or exchange of
429 chromosomes. Retro-transposition has been observed in the *elav* gene, which led to the loss of all
430 introns in *Drosophila* (Samson 2008). Likewise, holometabolous insects generally have three *elav*
431 genes, but honey bees have only one *elav* gene. The honey bee *elav* gene, however, carries features
432 of the other two genes present in *Drosophila*. Hence, *elav* in honey bees could have collapsed in an
433 ancestor from three to one gene by some form of recombination to include parts specific to the other
434 two *elav* genes (Ustaoglu et al. 2021). Likewise, the *amyloid-β precursor protein* (*APP*) gene

435 displays copy number variation in the human brain, which are increased by retro-transposition in
436 sporadic forms of Alzheimer's disease suggesting evolutionary conserved mechanisms for
437 reinsertion of genomic information (Lee et al. 2018). Re-insertion of fragments cut out by the
438 CRISPR-Cas9 system seems not to be specific to *Drosophila* as it has also been observed in human
439 cells (Geng et al. 2022). Since intron loss and gain occurs during evolution, the mechanism
440 underlying insertion of DNA fragments from sgRNA induced deletions might be responsible for
441 these changes. Thus, when generating sgRNA induced gene deletions, it is essential to test for the
442 absence of any transcripts by RT-PCR, but also to validate a deletion at the DNA level by PCR
443 using flanking primers or Southern blots. Alternatively, a GFP cassette with a polyA site can be
444 inserted to terminate the ORF in the beginning, but it needs to be evaluated whether the polyA site
445 in the beginning of the gene is used or whether the exon containing the GFP cassette is skipped
446 (Soller 2006; Wierson et al. 2020). For a more reliable way to generate gene knock-outs we have
447 now developed a PBac w+ marker that can be inserted when generating a deletion. Instead of
448 deleting the entire gene, however, deletion of a 5'part will render it non-functional. In addition, if
449 this 5'part inserts unwantedly, it will be non-functional. In any case, however, a marked locus will
450 allow for rigorous cleaning of the genetic background.
451 In essence, we have established the rules for designing highly efficient sgRNAs and established
452 methodology to efficiently generate gene deletions. These findings have implications for other RNA
453 based methodologies including prime editing (Anzalone et al. 2019; Bosch et al. 2021).
454

455 ## Materials & Methods

456 **sgRNA/Cas9 directed DNA cleavage**

457 DNA templates for in vitro transcription were reconstituted from synthetic oligonucleotides. As
458 only the T7 promoter needs to be double-stranded for in vitro transcription, a T7 promoter
459 oligonucleotide (CCTGGCTAATACGACTCACTATAG) was annealed to an anti-sense Ultramer
460 (IDT DNA) encoding the entire sgRNA in addition to the T7 promoter. Alternatively, a 60 nt T7
461 promoter oligonucleotide with a partial sgRNA was annealed to an anti-sense oligonucleotide
462 encoding the tracrRNA
463 (AAAAAAAGCACCGACTCGGTGCCACTTTTTCAAGTTGATAACGGACTAGCCTTATTTT
464 AACTTGCTATTTCTAGCTCTAAAAC) for 15 min at 40º C (2 µM) and made double-stranded by
465 extension with Klenow fragment of DNA polymerase I according to the manufacturer's instructions
466 (NEB). Klenow was then heat-inactivated 10 min at 85º C and oligonucleotides were desalted with
467 a G-50 Autoseq Sephadex spin column (GE) before using for in vitro transcription.
468 Then, sgRNAs were generated by in vitro transcription with T7 polymerase (T7 MEGAscript,
469 Ambion) from synthetic oligonucleotides (0.2 µM) and trace-labeled with ^{32}P alpha-ATP (800
470 Ci/mmol, 12.5 µM, Perkin Elmer) in a 20 µl reaction according to the manufacturer's instructions.
471 After DNAse I digestion, free nucleotides were removed with a G-50 Probequant Sephadex spin
472 column (GE). Then, sgRNAs were heated for 2 min to 95º C and left at room temperature to adopt
473 folding. Then, sgRNAs were quantified by scintillation counting and analysed on 8-20 %
474 denaturing polyacrylamide gels as described (Dix et al. 2022).
475 For synthetic substrate DNAs the sense oligonucleotide (1 µM, sgRNA flanking sequences are:
476 TCGAGCATTATATGAAC-sgRNA-GGGTATTGGGGAATTCATTATGC) was labeled with ^{32}P
477 gamma-ATP (6000 Ci/mmol, 25 µM, Perkin Elmer) with PNK (NEB). After heat-inactivation of
478 PNK for 2 min at 95º C, sense and anti-sense (anti-sense sgRNA flanking sequences are:
479 GGCCGCATAATGAATTCCCCAATACCC-as sgRNA-GTTCATATAATGC) oligonucleotides

480 were annealed by letting cool down to room temperature and used in sgRNA/Cas9 cleavage assays.
481 For plasmid sgRNA/Cas9 cleavage assays, these annealed oligonucleotides were cloned into a
482 modified *pBS SK+* using a Xho I and Not I cut vector to assay sgRNA/Cas9 activity.
483 For sgRNA/Cas9 cleavage assays, DNA/sgRNA/Cas9 ratios of 1/10/10 were used in a 10 µl
484 reaction using the buffer supplied (NEB) and DEPC-treated water (Haussmann et al. 2019).
485 Typically Cas9 (100 nM final) was incubated with sgRNA (100 nM) for 10 min at 25º C before
486 adding oligonucleotides (10 nM final) or plasmid DNA (10 nM, corresponds to ~25 ng/µl final
487 concentration of a 3 kb plasmid). Plasmids were linearized after Cas9 digestion by first heat
488 inactivating Cas9 for 2 min at 95ºC, and then adding 10 µl of a restriction enzyme (5 U) in NEB
489 buffer 3. Adding a restriction enzyme together with Cas9 inhibited DNA scission by Cas9.
490 Cleavage of oligonucleotides was analysed on 8 % denaturing polyacrylamide gels and plasmid
491 DNA was analysed on ethidium bromide stained agarose gels.
492 RNA secondary structure was analyzed with RNAfold at http://rna.tbi.univie.ac.at (Gruber et al.
493 2008) using the following tracrRNA sequence:
494 GUUUUAGAGCUAGAAAUAGCAAGUUAAAAUAAGGCUAGUCCGUUAUCAACUUGAAA
495 AAGUGGCACCGAGUCGGUGCUUUUUU.
496
497 **Criteria for optimal sgRNA design and Cas9 structural analysis**
498 Criteria to select sgRNAs that maintain the structure required for efficient Cas9 DNA scission were
499 implemented in the server accessible at https://platinum-crispr.bham.ac.uk/predict.pl and are as
500 follows. The first nucleotide of the sgRNA needs to be a G for efficient transcription initiation by
501 RNA Pol III (Paule and White 2000). For T7 mediated in vitro transcription, three Gs need to be
502 added to sgRNAs of 23 nucleotides. Disruption of the tetraloop or loop 2 and 3 structures, and the

503 sequence U A/G G C/U A/G of nucleotides 16-20, which will result in a tetraloop bulge mimic,
504 were classified as low efficiency sgRNAs as these parts are recognized by Cas9 (Nishimasu et al.
505 2014). Similarly, a hairpin loop in the gRNA consisting of 4 or more base pairing nucleotides, or
506 base pairing of nucleotides 17-20 of the gRNA (U/C U/C U U), or base-pairing of eight nucleotides
507 within the seed region with looped-out nucleotides spaced by three base-pairing nucleotides (N11-
508 N20), or a GC content below 15% (1 or 2 nts) or above 50% (11 or more nts) were also considered
509 low efficiency. Medium efficiency was assigned for gRNAs with a GC content of 15-25 % (3-5 nts)
510 or 40-50% (8-10 nts), or a low CG content in the seed region (less than 5 nts in nucleotides 11-20).
511 In addition, two U's at position 19/20 of the gRNA reduce efficiency because of premature
512 transcription termination (Gao et al. 2018; Graf et al. 2019). Further, we assigned a medium impact
513 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
514 paired or seven nucleotides within the seed region (N11-N20) base-paired with looped-out
515 nucleotides spaced by three base-pairing nucleotides. Thus, we deemed an sgRNA optimal to allow
516 Cas9 to cleave DNA with high efficiency, if the GC content is 30-35 % (6-7 nts) and none of the
517 above criteria applied.
518 Cas9/sgRNA structural complex analysis was done using Chimera as described (Dix et al. 2022).
519

520 **RNA extraction, RT-PCR and PCR on genomic DNA**

521 Total RNA was extracted using Tri-reagent (SIGMA) and reverse transcribed with Superscript II
522 (Invitrogen) according to the manufacturer's instructions using an oligo dT primer. PCR was done
523 for 40 cycles with 1 µl of cDNA, with 1 µg of genomic DNA or from a single fly after freezing and
524 drying in 200 µl of isopropanol. Primers to detect the sgRNA/Cas9 induced deletion in *Ythdc1* were
525 YT F1 (GCCGCTGTGACGCAGAATTTGTGTG) and YT R1

526 (GGCCGTGCATGTTGCGCATGTAGTCC), and in *Ythdf* were 64F1
527 (GCCGAGAAAGTGCACAAGGATACGGAG) and 64R1
528 (CAAGGAATGGCTGAAGCAGACTCCTTG). Primers to amplify parts of the body of the RNA
529 also flanking an intron were for *Ythdc1* YT F2 (CCACGCTGCCGCAGAACGACGCCAATC) and
530 YT R2 (GCGGCAGATCCAGTCAAGCTCGATGAC), and in *Ythdf* were 64F2
531 (GAGCTGCCTGTCGATTCCCAACTCGTG) and 64R2
532 (CCGCCCTCTTCGTGTCGCTCCTTGAAG). Primers to amplify parts of the *ewg* gene have been
533 described elsewhere (Koushika et al. 1999).
534
535 **Cloning of sgRNAs into *pUC 3GLA U6.1 BbsI***
536 To clone two sgRNAs expressed by U6 promoters, the "tracrRNA U6.3 promoter" fragment was
537 amplified with left (AAGATATCCGGGTGAACTTC**G**N_{19}GTTTTAGAGCTAGAAATAGC) and
538 right (GCTATTTCTAGCTCTAAAA**C**N_{19}CGACGTTAAATTGAAAATAGG) sgRNA primers
539 from pUC 3GLA U6.1/3 sgRNA using Pwo polymerase (Roche) with initial 30 sec denaturation at
540 94ºC followed by two cycles 94ºC/30 sec, 49ºC/40 sec, 72ºC/45 sec, then two cycles 94ºC/30 sec,
541 51ºC/40 sec, 72ºC/45 sec and 22 cycles two cycles 94ºC/30 sec, 56ºC/40 sec, 72ºC/45 sec.
542 Transcription from the U6 promoter initiates with a G (bold, underlined). Although this G does not
543 need to be present in the targeting sequence, it needs be included for folding of the sgRNA. The
544 sequences for the sgRNAs in *Ythdc1* were gACAGGTATTCCCAAACTCAC and
545 GACATGTAGCGTTCCCATGA, and for *Ythdf* were GTCCTGAAATACGAGCACAA and
546 gATAACGAACATGTGGGATCT. The pUC 3GLA U6.1 BbsI vector was cut by BbsI and the
547 "sgRNA1 tracrRNA U6.3 promoter sgRNA2" fragment was cloned by Gibson assembly according
548 to the manufacturer's instructions (NEB). For sequencing, primer U6.1 Fseq

549 (GCGCGTACGTCCTTCGCATCCTTATG) was used. The sequences for *pUC 3GLA HAi Dscam*
550 *3-5, pUC 3GLA U6.1 BbsI* and the *pUC 3GLA U6.1/3 Ythdf sgRNA* have been deposited
551 (MK908409, MK908408 and MK908407).
552

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

554 All *Drosophila melanogaster* strains were reared at 25ºC and 40%–60% humidity on standard
555 cornmeal-agar food in 12:12 h light:dark cycle as described (Haussmann et al. 2013). *CantonS* was
556 used as a wild type control. For the *Ythdc1*, the GFP marked *Mi{MIC}YT521-B*^{MI02006} transposon
557 insert and the *w+* marked *Df(3L)Exel6094* deficiency were used, and for *Ythdf*, the RFP marked
558 *PBac{SAstopDsRed}*^{LL04081} insert and the *w+* marked deficiency *Df(3R)ED6220* were used. For
559 phiC31 mediated transformation, constructs were injected into *y*^{1} *w* M{vas-int.Dm}ZH-2A;*
560 *PBac{y+-attP-3B}VK00013* with the landing site inserted at 76A as previously described
561 (Haussmann et al. 2013). Prior to insertion of GFP marked constructs, the GFP and RFP markers
562 had been removed from the *y1 w* M{vas-int.Dm}ZH-2A* landing site by Cre mediated
563 recombination (Bischof et al. 2007; Zaharieva et al. 2015).
564

565 **Implementation of PlatinumCRISPr**

566 PlatinumCRISPr is implemented as a Perl script based web-server iteratively evaluating the rule set
567 described in the main text. A guide is classified as "compromised" if any of the rules is violated.
568 For analysis of an sgRNA consisting the target complementary sequence (spacer) and the constant
569 crispr RNA (crRNA) fused to the tracrRNA through an artificial loop is used whereby the first
570 nucleotide of the 20 nt spacer sequence is a G, because a G is needed for transcription (Jinek et al.
571 2012; Cong et al. 2013). Folding of the sgRNA is computed using RNAFold (Version 2.4.17) and
572 further processed using bpRNA for subsequent interpretation of the dot-bracket code describing the

573 secondary structure by a custom-made Perl script. Notably, the sequence position is calculated from
574 the 3'end of the tracerRNA to allow for variable length of the spacer (between 18 and 23 nt) for
575 custom applications using synthesized RNA consisting of the spacer fused to the crRNA and
576 hybridized to the tracrRNA.

577 PlatinumCRISPr classifies guides by a binary outcome and typically reports around 70% of
578 sgRNAs as "compromised". Accordingly, only the top 30% were analysed for the distribution of
579 their reported cleavage efficiency in a given data set for each scoring application (**Supplementary**
580 **Fig S8).** Statistical significance was calculated using a one-sided Wilcoxon signed-rank test. We are
581 grateful to M. Haeussler for providing published guide sequencing and cleavage efficiency scores
582 (Haeussler et al. 2016). CrisperON and DeepSpCas9 guide sequencing and cleavage efficiency
583 scores were calculated using published web-interfaces (Wong et al. 2015; Xiang et al. 2021).

584

585 ## Acknowledgments

586 We thank Bloomington and Kyoto stock centers for fly lines, A. Handler for the PiggyBac clone,
587 BacPac for Bac clones and plasmids, the University of Cambridge Department of Genetics Fly
588 Facility for embryo injections for generating transgenic lines, M. Haeusler for datasets and pre-
589 calculated scores, Sung-Eun Yoon from the Korea Drosophila Resource Centre and Ying Tan from
590 Genetivision Corporation for sgRNA efficiency data, Yuan Tian for help with art work and Pawel
591 Grzechnik for comments on the manuscript. For this work we acknowledge funding from the
592 Biotechnology and Biological Science Research Council, the Leverhulme Trust, the MRC DTP to
593 DWJM and a Genetics Society summer studentship to AIH.

594

595 ## Author contributions

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

600

601 ## Declaration of interests

602 The authors declare no competing interests.

603

604 ## Accession codes

605 Genbank: MK908409 (*pUC 3GLA HAi Dscam 3-5*), MK908408 (*pUC 3GLA U6.1 BbsI*) and
606 MK908407 (*pUC 3GLA U6.1/3 YTHDF53 sgRNA*)

607

608 ## Supplementary Data

609 Supplementary Data accompany this paper.

610

611 ## References
612

613 Abadi S, Yan WX, Amar D, Mayrose I. 2017. A machine learning approach for predicting
614 CRISPR-Cas9 cleavage efficiencies and patterns underlying its mechanism of action. *PLoS*
615 *Comput Biol* **13**: e1005807.
616 Anders C, Niewoehner O, Duerst A, Jinek M. 2014. Structural basis of PAM-dependent target DNA
617 recognition by the Cas9 endonuclease. *Nature* **513**: 569-573.
618 Anreiter I, Mir Q, Simpson JT, Janga SC, Soller M. 2021. New Twists in Detecting mRNA
619 Modification Dynamics. *Trends Biotechnol* **39**: 72-89.
620 Anzalone AV, Randolph PB, Davis JR, Sousa AA, Koblan LW, Levy JM, Chen PJ, Wilson C,
621 Newby GA, Raguram A et al. 2019. Search-and-replace genome editing without double-
622 strand breaks or donor DNA. *Nature* **576**: 149-157.

623 Arimbasseri AG, Maraia RJ. 2015. Mechanism of Transcription Termination by RNA Polymerase
624 III Utilizes a Non-template Strand Sequence-Specific Signal Element. *Mol Cell* **58**: 1124-
625 1132.
626 Balacco DL, Soller M. 2019. The m(6)A Writer: Rise of a Machine for Growing Tasks.
627 *Biochemistry* **58**: 363-378.
628 Bischof J, Maeda RK, Hediger M, Karch F, Basler K. 2007. An optimized transgenesis system for
629 Drosophila using germ-line-specific phiC31 integrases. *Proc Natl Acad Sci U S A* **104**:
630 3312-3317.
631 Boivin V, Deschamps-Francoeur G, Scott MS. 2018. Protein coding genes as hosts for noncoding
632 RNA expression. *Semin Cell Dev Biol* **75**: 3-12.
633 Bosch JA, Birchak G, Perrimon N. 2021. Precise genome engineering in Drosophila using prime
634 editing. *Proc Natl Acad Sci U S A* **118**.
635 Chari R, Mali P, Moosburner M, Church GM. 2015. Unraveling CRISPR-Cas9 genome engineering
636 parameters via a library-on-library approach. *Nat Methods* **12**: 823-826.
637 Chuai G, Ma H, Yan J, Chen M, Hong N, Xue D, Zhou C, Zhu C, Chen K, Duan B et al. 2018.
638 DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* **19**: 80.
639 Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA et al.
640 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**: 819-823.
641 Deveson IW, Hardwick SA, Mercer TR, Mattick JS. 2017. The Dimensions, Dynamics, and
642 Relevance of the Mammalian Noncoding Transcriptome. *Trends Genet* **33**: 464-478.
643 Dezi V, Ivanov C, Haussmann IU, Soller M. 2016. Nucleotide modifications in messenger RNA
644 and their role in development and disease. *Biochem Soc Trans* **44**: 1385-1393.
645 Dix TC, Haussmann IU, Brivio S, Nallasivan MP, HadzHiev Y, Müller F, Müller B, Pettitt J, Soller
646 M. 2022. CMTr mediated 2'-O-ribose methylation status of cap-adjacent nucleotides across
647 animals. *RNA (New York, NY)* **28**: 1377-1390.
648 Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, Smith I, Tothova Z,
649 Wilen C, Orchard R et al. 2016. Optimized sgRNA design to maximize activity and
650 minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**: 184-191.
651 Doench JG, Hartenian E, Graham DB, Tothova Z, Hegde M, Smith I, Sullender M, Ebert BL,
652 Xavier RJ, Root DE. 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-
653 mediated gene inactivation. *Nat Biotechnol* **32**: 1262-1267.
654 Doudna JA. 2020. The promise and challenge of therapeutic genome editing. *Nature* **578**: 229-236.
655 El-Brolosy MA, Kontarakis Z, Rossi A, Kuenne C, Gunther S, Fukuda N, Kikhi K, Boezio GLM,
656 Takacs CM, Lai SL et al. 2019. Genetic compensation triggered by mutant mRNA
657 degradation. *Nature* **568**: 193-197.
658 Farboud B, Meyer BJ. 2015. Dramatic enhancement of genome editing by CRISPR/Cas9 through
659 improved guide RNA design. *Genetics* **199**: 959-971.
660 Gagnon JA, Valen E, Thyme SB, Huang P, Akhmetova L, Pauli A, Montague TG, Zimmerman S,
661 Richter C, Schier AF. 2014. Efficient mutagenesis by Cas9 protein-mediated
662 oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9**:
663 e98186.
664 Gandhi S, Haeussler M, Razy-Krajka F, Christiaen L, Stolfi A. 2017. Evaluation and rational design
665 of guide RNAs for efficient CRISPR/Cas9-mediated mutagenesis in Ciona. *Developmental*
666 *biology* **425**: 8-20.
667 Gao Z, Herrera-Carrillo E, Berkhout B. 2018. Delineation of the Exact Transcription Termination
668 Signal for Type 3 Polymerase III. *Molecular therapy Nucleic acids* **10**: 36-44.
669 Garcia-Doval C, Jinek M. 2017. Molecular architectures and mechanisms of Class 2 CRISPR-
670 associated nucleases. *Curr Opin Struct Biol* **47**: 157-166.
671 Geng K, Merino LG, Wedemann L, Martens A, Sobota M, Sanchez YP, Søndergaard JN, White RJ,
672 Kutter C. 2022. Target-enriched nanopore sequencing and de novo assembly reveals co-
673 occurrences of complex on-target genomic rearrangements induced by CRISPR-Cas9 in
674 human cells. *Genome research* **32**: 1876-1891.
675 Graf R, Li X, Chu VT, Rajewsky K. 2019. sgRNA Sequence Motifs Blocking Efficient
676 CRISPR/Cas9-Mediated Gene Editing. *Cell Rep* **26**: 1098-1103 e1093.

677 Gruber AR, Lorenz R, Bernhart SH, Neubock R, Hofacker IL. 2008. The Vienna RNA websuite.
678 *Nucleic Acids Res* **36**: W70-74.
679 Hachet O, Ephrussi A. 2004. Splicing of oskar RNA in the nucleus is coupled to its cytoplasmic
680 localization. *Nature* **428**: 959-963.
681 Haeussler M, Schönig K, Eckert H, Eschstruth A, Mianné J, Renaud JB, Schneider-Maunoury S,
682 Shkumatava A, Teboul L, Kent J et al. 2016. Evaluation of off-target and on-target scoring
683 algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol* **17**:
684 148.
685 Hart T, Chandrashekhar M, Aregger M, Steinhart Z, Brown KR, MacLeod G, Mis M, Zimmermann
686 M, Fradet-Turcotte A, Sun S et al. 2015. High-Resolution CRISPR Screens Reveal Fitness
687 Genes and Genotype-Specific Cancer Liabilities. *Cell* **163**: 1515-1526.
688 Haussmann IU, Bodi Z, Sanchez-Moran E, Mongan NP, Archer N, Fray RG, Soller M. 2016.
689 m(6)A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex
690 determination. *Nature* **540**: 301-304.
691 Haussmann IU, Hemani Y, Wijesekera T, Dauwalder B, Soller M. 2013. Multiple pathways
692 mediate the sex-peptide-regulated switch in female Drosophila reproductive behaviours.
693 *Proc Biol Sci* **280**: 20131938.
694 Haussmann IU, Li M, Soller M. 2011. ELAV-mediated 3'-end processing of ewg transcripts is
695 evolutionarily conserved despite sequence degeneration of the ELAV-binding site. *Genetics*
696 **189**: 97-107.
697 Haussmann IU, Ustaoglu P, Brauer U, Hemani Y, Dix TC, Soller M. 2019. Plasmid-based gap-
698 repair recombineered transgenes reveal a central role for introns in mutually exclusive
699 alternative splicing in Down Syndrome Cell Adhesion Molecule exon 4. *Nucleic Acids Res*
700 **47**: 1389-1403.
701 Haussmann IU, Wu Y, Nallasivan MP, Archer N, Bodi Z, Hebenstreit D, Waddell S, Fray R, Soller
702 M. 2022. CMTr cap-adjacent 2'-O-ribose mRNA methyltransferases are required for reward
703 learning and mRNA localization to synapses. *Nat Commun* **13**: 1209.
704 Hemani Y, Soller M. 2012. Mechanisms of Drosophila Dscam mutually exclusive splicing
705 regulation. *Biochem Soc Trans* **40**: 804-809.
706 Hille F, Richter H, Wong SP, Bratovic M, Ressel S, Charpentier E. 2018. The Biology of CRISPR-
707 Cas: Backward and Forward. *Cell* **172**: 1239-1259.
708 Housden BE, Valvezan AJ, Kelley C, Sopko R, Hu Y, Roesel C, Lin S, Buckner M, Tao R,
709 Yilmazel B et al. 2015. Identification of potential drug targets for tuberous sclerosis
710 complex by synthetic screens combining CRISPR-based knockouts with RNAi. *Science*
711 *signaling* **8**: rs9.
712 Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X,
713 Shalem O et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat*
714 *Biotechnol* **31**: 827-832.
715 Jensen KT, Floe L, Petersen TS, Huang J, Xu F, Bolund L, Luo Y, Lin L. 2017. Chromatin
716 accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing
717 efficiency. *FEBS Lett* **591**: 1892-1901.
718 Jiang F, Doudna JA. 2017. CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys* **46**: 505-
719 529.
720 Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-
721 RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**: 816-821.
722 Kim HK, Kim Y, Lee S, Min S, Bae JY, Choi JW, Park J, Jung D, Yoon S, Kim HH. 2019. SpCas9
723 activity prediction by DeepSpCas9, a deep learning-based model with high generalization
724 performance. *Science advances* **5**: eaax9249.
725 Koushika SP, Soller M, DeSimone SM, Daub DM, White K. 1999. Differential and inefficient
726 splicing of a broadly expressed Drosophila erect wing transcript results in tissue-specific
727 enrichment of the vital EWG protein isoform. *Mol Cell Biol* **19**: 3998-4007.
728 Labuhn M, Adams FF, Ng M, Knoess S, Schambach A, Charpentier EM, Schwarzer A, Mateo JL,
729 Klusmann JH, Heckl D. 2018. Refined sgRNA efficacy prediction improves large- and
730 small-scale CRISPR-Cas9 applications. *Nucleic Acids Res* **46**: 1375-1385.

731 Labun K, Montague TG, Gagnon JA, Thyme SB, Valen E. 2016. CHOPCHOP v2: a web tool for
732 the next generation of CRISPR genome engineering. *Nucleic Acids Res* **44**: W272-276.
733 Labun K, Montague TG, Krause M, Torres Cleuren YN, Tjeldnes H, Valen E. 2019. CHOPCHOP
734 v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res* **47**:
735 W171-w174.
736 Lee MH, Siddoway B, Kaeser GE, Segota I, Rivera R, Romanow WJ, Liu CS, Park C, Kennedy G,
737 Long T et al. 2018. Somatic APP gene recombination in Alzheimer's disease and normal
738 neurons. *Nature* **563**: 639-645.
739 Liu X, Homma A, Sayadi J, Yang S, Ohashi J, Takumi T. 2016. Sequence features associated with
740 the cleavage efficiency of CRISPR/Cas9 system. *Sci Rep* **6**: 19675.
741 Ma Z, Zhu P, Shi H, Guo L, Zhang Q, Chen Y, Chen S, Zhang Z, Peng J, Chen J. 2019. PTC-
742 bearing mRNA elicits a genetic compensation response via Upf3a and COMPASS
743 components. *Nature* **568**: 259-263.
744 Michlits G, Jude J, Hinterndorfer M, de Almeida M, Vainorius G, Hubmann M, Neumann T,
745 Schleiffer A, Burkard TR, Fellner M et al. 2020. Multilayered VBC score predicts sgRNAs
746 that efficiently generate loss-of-function alleles. *Nat Methods* **17**: 708-716.
747 Moreno-Mateos MA, Vejnar CE, Beaudoin JD, Fernandez JP, Mis EK, Khokha MK, Giraldez AJ.
748 2015. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo.
749 *Nat Methods* **12**: 982-988.
750 Nishimasu H, Ran FA, Hsu PD, Konermann S, Shehata SI, Dohmae N, Ishitani R, Zhang F, Nureki
751 O. 2014. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**:
752 935-949.
753 Paule MR, White RJ. 2000. Survey and summary: transcription by RNA polymerases I and III.
754 *Nucleic Acids Res* **28**: 1283-1298.
755 Poh HX, Mirza AH, Pickering BF, Jaffrey SR. 2022. Alternative splicing of METTL3 explains
756 apparently METTL3-independent m6A modifications in mRNA. *PLoS biology* **20**:
757 e3001683.
758 Port F, Chen HM, Lee T, Bullock SL. 2014. Optimized CRISPR/Cas tools for efficient germline
759 and somatic genome engineering in Drosophila. *Proc Natl Acad Sci U S A* **111**: E2967-2976.
760 Ren X, Sun J, Housden BE, Hu Y, Roesel C, Lin S, Liu LP, Yang Z, Mao D, Sun L et al. 2013.
761 Optimized gene editing technology for Drosophila melanogaster using germ line-specific
762 Cas9. *Proc Natl Acad Sci U S A* **110**: 19012-19017.
763 Ren X, Yang Z, Xu J, Sun J, Mao D, Hu Y, Yang SJ, Qiao HH, Wang X, Hu Q et al. 2014.
764 Enhanced specificity and efficiency of the CRISPR/Cas9 system with optimized sgRNA
765 parameters in Drosophila. *Cell Rep* **9**: 1151-1162.
766 Riesenberg S, Kanis P, Macak D, Wollny D, Düsterhöft D, Kowalewski J, Helmbrecht N, Maricic
767 T, Pääbo S. 2023. Efficient high-precision homology-directed repair-dependent genome
768 editing by HDRobust. *Nat Methods* **20**: 1388-1399.
769 Roignant JY, Soller M. 2017. m6A in mRNA: An Ancient Mechanism for Fine-Tuning Gene
770 Expression. *Trends Genet* **33**: 380-390.
771 Roth A, Vleurinck C, Netschitailo O, Bauer V, Otte M, Kaftanoglu O, Page RE, Beye M. 2019. A
772 genetic switch for worker nutrition-mediated traits in honeybees. *PLoS biology* **17**:
773 e3000171.
774 Samson ML. 2008. Rapid functional diversification in the structurally conserved ELAV family of
775 neuronal RNA binding proteins. *BMC Genomics* **9**: 392.
776 Singh R, Kuscu C, Quinlan A, Qi Y, Adli M. 2015. Cas9-chromatin binding information enables
777 more accurate CRISPR off-target prediction. *Nucleic Acids Res* **43**: e118.
778 Sledzinski P, Nowaczyk M, Olejniczak M. 2020. Computational Tools and Resources Supporting
779 CRISPR-Cas Experiments. *Cells* **9**.
780 Soller M. 2006. Pre-messenger RNA processing and its regulation: a genomic perspective. *Cell Mol*
781 *Life Sci* **63**: 796-819.
782 Soller M, Haussmann IU, Hollmann M, Choffat Y, White K, Kubli E, Schäfer MA. 2006. Sex-
783 peptide-regulated female sexual behavior requires a subset of ascending ventral nerve cord
784 neurons. *Current biology : CB* **16**: 1771-1782.

785 Solomon DA, Stepto A, Au WH, Adachi Y, Diaper DC, Hall R, Rekhi A, Boudi A, Tziortzouda P,
786 Lee YB et al. 2018. A feedback loop between dipeptide-repeat protein, TDP-43 and
787 karyopherin-α mediates C9orf72-related neurodegeneration. *Brain : a journal of neurology*
788 **141**: 2908-2924.
789 Thyme SB, Akhmetova L, Montague TG, Valen E, Schier AF. 2016. Internal guide RNA
790 interactions interfere with Cas9-mediated cleavage. *Nat Commun* **7**: 11750.
791 Trivedi D, Cm V, Bisht K, Janardan V, Pandit A, Basak B, H S, Ramesh N, Raghu P. 2020. A
792 genome engineering resource to uncover principles of cellular organization and tissue
793 architecture by lipid signaling. *Elife* **9**.
794 Tuladhar R, Yeu Y, Tyler Piazza J, Tan Z, Rene Clemenceau J, Wu X, Barrett Q, Herbert J,
795 Mathews DH, Kim J et al. 2019. CRISPR-Cas9-based mutagenesis frequently provokes on-
796 target mRNA misregulation. *Nat Commun* **10**: 4056.
797 Ustaoglu P, Gill JK, Doubovetzky N, Haussmann IU, Dix TC, Arnold R, Devaud JM, Soller M.
798 2021. Dynamically expressed single ELAV/Hu orthologue elavl2 of bees is required for
799 learning and memory. *Communications biology* **4**: 1234.
800 Varshney GK, Pei W, LaFave MC, Idol J, Xu L, Gallardo V, Carrington B, Bishop K, Jones M, Li
801 M et al. 2015. High-throughput gene targeting and phenotyping in zebrafish using
802 CRISPR/Cas9. *Genome research* **25**: 1030-1042.
803 Wang T, Wei JJ, Sabatini DM, Lander ES. 2014. Genetic screens in human cells using the CRISPR-
804 Cas9 system. *Science* **343**: 80-84.
805 Wierson WA, Welker JM, Almeida MP, Mann CM, Webster DA, Torrie ME, Weiss TJ, Kambakam
806 S, Vollbrecht MK, Lan M et al. 2020. Efficient targeted integration directed by short
807 homology in zebrafish and mammalian cells. *Elife* **9**.
808 Wong N, Liu W, Wang X. 2015. WU-CRISPR: characteristics of functional guide RNAs for the
809 CRISPR/Cas9 system. *Genome Biol* **16**: 218.
810 Xiang X, Corsi GI, Anthon C, Qu K, Pan X, Liang X, Han P, Dong Z, Liu L, Zhong J et al. 2021.
811 Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning.
812 *Nat Commun* **12**: 3238.
813 Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, Wu D, Cong L, Zhang F, Liu JS et al. 2015.
814 Sequence determinants of improved CRISPR sgRNA design. *Genome research* **25**: 1147-
815 1157.
816 Zaharieva E, Haussmann IU, Brauer U, Soller M. 2015. Concentration and localization of co-
817 expressed ELAV/Hu proteins control specificity of mRNA processing. *Mol Cell Biol* **35**.
818 Zhang D, Hurst T, Duan D, Chen SJ. 2019. Unified energetics analysis unravels SpCas9 cleavage
819 activity for optimal gRNA design. *Proc Natl Acad Sci U S A* **116**: 8693-8698.
820

821 # FIGURE LEGENDS

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

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

827 (B) Agarose gel showing Cas9 mediated cleavage of the 11.3 kb Dscam 3-5 plasmid for 24 h with
828 indicated sgRNAs. Plasmids were cut with either Acc65I (lanes 2, 3, 6 and 7) or BspEI (lanes 4, 5,
829 8 and 9) after Cas9 cleavage. The line at the bottom shows a map of the plasmid with restriction
830 sites indicated. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8
831 kb.
832 (C) Structure of the sgRNA scaffold from co-crystallization with Cas9 (Nishimasu et al., 2014).
833 Vertical or horizontal lines indicate Watson-Crick base-pairing, and dots or dashed lines indicate
834 non-Watson-Crick base-pairing. Nucleotides base-pairing in loop 1 are bold. Additional base-
835 pairing found in the tracrRNA-crRNA heterodimer is indicated in the extended scaffold (Jinek et
836 al., 2012).
837 (D) Agarose gel showing Cas9 mediated cleavage of the 11.3 kb Dscam 3-5 plasmid for 24 h with
838 indicated sgRNAs L7GC and R3G. Plasmids were cut with PstI and NotI. The star denotes
839 incomplete cleavage by NotI and the line at the bottom shows a map of the plasmid with restriction
840 sites indicated. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6 kb, 1.9 kb and 0.8
841 kb.
842 (E) Denaturing acrylamide gel showing Cas9 mediated cleavage of synthetic oligonucleotides with
843 indicated sgRNAs.
844 (F) Sequences of sgRNAs with variable length. Small letter guanosines used for in vitro
845 transcription are not present in the target DNA sequence.
846 (G) Denaturing acrylamide gel showing Cas9 mediated cleavage for 1 h of synthetic
847 oligonucleotides with indicated sgRNAs of variable length.
848 (H-K) Structure of sgRNAs. Nucleotides base-pairing in loop 1 are bold. Red lines in J and K
849 indicate potential base-pairing with nucleotides in loop 2. The red arrow in J indicates the sequence

850 complementarity leading to a bulge in the tetraloop. The red arrows in K indicate a duplication of
851 the bulge structure present in the tetraloop.
852
853 **Figure 2:** Secondary structure of sgRNAs affects Cas9 mediated cleavage efficiency.
854 (A) Sequences of sgRNAs. Small letter guanosines used for in vitro transcription are not present in
855 the target DNA sequence.
856 (B) Agarose gel showing Cas9 mediated cleavage after 6 h of 3 kb pBS SK+ test-plasmids
857 containing the target sequence with indicated sgRNAs. Plasmids were linearized with ScaI as in the
858 control after Cas9 heat inactivation. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6
859 kb, 1.9 kb and 0.8 kb.
860 (C) Denaturing acrylamide gel showing synthetic oligonucleotides before and after sgRNA/Cas9
861 mediated cleavage for 1h.
862 (D-G) Structure of sgRNAs. Nucleotides base-pairing in loop 1 are bold. Red lines in D indicate
863 potential base-pairing with nucleotides in loop 2. Green nucleotides indicate mutations compared to
864 sgRNA R13G.
865
866 **Figure 3:** Base-pairing of sgRNAs in the seed-region blocks Cas9 mediated cleavage of a test
867 plasmid.
868 (A, B) Structure of sgRNAs. Horizontal red lines in A and B indicate artificially introduced base-
869 pairing with the sgRNA scaffold, and vertical red lines in a indicate potential base-pairing with
870 nucleotides in loop 2. Nucleotides base-pairing in loop 1 are bold.
871 (C) Denaturing acrylamide gel showing synthetic oligonucleotides before and after sgRNA/Cas9
872 mediated cleavage for 1h. Note that Cas9 cleavage is heterogeneous.

873 (D) Agarose gel showing Cas9 mediated cleavage after 6 h of 3 kb pBS SK+ test-plasmids
874 containing the target sequence with indicated sgRNAs. Plasmids were linearized with ScaI as in the
875 control after Cas9 heat inactivation. Size markers are EcoRI/HinDIII digested λ DNA of 20 kb, 3.6
876 kb, 1.9 kb and 0.8 kb.
877
878 **Figure 4:** PlatinumCRISPr selects high efficiency sgRNAs and outperforms other sgRNA selection
879 tools for a *Drosophila* data set.
880 (A) Comparison of the in vivo efficiency of all sgRNAs from the Ren et al. (2014) data set with
881 PlatinumCRISPr selected sgRNAs.
882 B) Comparison of the sgRNA selection performance of PlatinumCRISPr with other sgRNA
883 selection tools.
884
885 **Figure 5:** Comparison of individual sgRNA selection criteria for performance with different data
886 sets. The different data sets with the number of sgRNAs tested are indicated on the left. The
887 different selection criteria are shown on top. Significant and enriched performances are indicated in
888 red and orange, respectively ($p<0.05$ and $p<0.25$). Criteria with numbers below 5% are indicated in
889 beige, and criteria already applied to the data set are shown in grey.
890
891 **Figure 6:** Combinations of two sgRNA selection tools select high efficiency cleaving sgRNAs for
892 several sgRNA efficiency screen data sets. Combinations of sgRNA selection tools in A are listed
893 according to overall cleavage efficiency of selected sgRNA significant ($p<0.05$) for both methods
894 (red) or one method (dark grey). Black indicates events with less than 5% of sgRNAs selected by at
895 least one method. Comparison of sgRNA selection by different sgRNA selection tools shown as

896 median of the cleavage efficiency for individual data sets (B). The distribution of cleavage
897 efficiencies for all sgRNAs is shown on the left (white box) and for PlatinumCRISPr-Wong Score
898 in red, Wond Score-Xu Score in purple and PlatinumCRISPr-Moreno-Mateos Score in blue.
899
900
901 **Figure 7:** Generation of gene deletions in *Drosophila* YTH protein genes using two sgRNAs/Cas9
902 and transposon markers.
903 (A) Schematic to the *Ythdc1* locus indicating transcripts (white boxes) and the ORF (black boxes)
904 below the chromosome. Primers used are indicated on top and below the transcripts. The w+
905 marked transposon used for detecting deletions in the locus is indicated by a triangle and the
906 deletion generated is indicated by a line.
907 (B) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdc1*
908 transposon excision lines using primers flanking the deletion. Presence of a PCR product indicates
909 the expected gene deletion. DNA markers are indicated on the left.
910 (C) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdc1*
911 transposon excision lines using internal primers flanking an intron. Ectopic insertion in the opposite
912 orientation is indicate by an asterisk (lane 4) as in this instance the transcript is not spliced. DNA
913 markers are indicated on the left.
914 (D) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdc1*$^{\Delta 7}$ flies
915 using internal primers and primers flanking the deletion. DNA markers are indicated on the left.
916 (E) Agarose gels showing RT-PCR products amplified from cDNA of control or *Ythdc1*$^{\Delta 7}$ flies
917 using internal primers in *Ythdf* and *ewg* genes. The PCR product of the *ewg* gene was used as
918 loading control. DNA markers are indicated on the left.

919 (F) Schematic to the *Ythdf* locus indicating transcripts (white boxes) and the ORF (black boxes)
920 below the chromosome. Primers used are indicated on top and below the transcripts. The RFP
921 marked transposon used for detecting deletions in the locus is indicated by a triangle and the
922 deletion generated is indicated by a line.
923 (G) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf*
924 transposon excision lines using primers flanking the deletion. Presence of a PCR product indicates
925 the expected gene deletion. DNA markers are indicated on the left.
926 (H) Agarose gels showing PCR products amplified from genomic DNA of control or *Ythdf*
927 transposon excision lines using primers flanking an intron. DNA markers are indicated on the left.
928 (I) Agarose gels showing PCR products amplified from genomic DNA of control or $Ythdf^{\Delta B1}$ flies
929 using internal primers and primers flanking the deletion. DNA markers are indicated on the left.
930 (J) Agarose gels showing RT-PCR products amplified from cDNA of control or $Ythdf^{\Delta B1}$ flies
931 using internal primers in *Ythdf* and *ewg* genes. The PCR product of the *ewg* gene was used as
932 loading control. DNA markers are indicated on the left.
933

934 ## Supplementary Figures

935

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

938

939 **Supplementary Figure S2.** Inactive sgRNAs targeted to the *Drosophila ythdf* (*CG6422*) gene.
940 (A) Sequences of sgRNAs targeting the CG6422 locus for generating a deletion.

941 (B, C) Secondary structures of sgRNAs. The scaffold is shown on the left indicating Watson-Crick
942 base-pairing by lines and the predicted secondary structure by RNAfold is shown on the right.
943
944 **Supplementary Figure S3.** Secondary structure of sgRNAs targeted to the *Drosophila white* gene
945 from Ren et al. (2014).
946 (A-AA) Secondary structures of sgRNAs predicted by RNAfold. Minimal energy and proximity
947 base-pair structures are shown on the left and right, respectively. Red and blue indicated high and
948 low probabilities for the adopted structural base-pairing assignment, respectively. The number on
949 the right indicates the effectiveness of inducing heritable mutations after injection into fly embryos
950 as determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too
951 high or too low compared to predicted DNA cleavage efficiency. Red arrows point towards
952 structural features likely limiting DNA cleavage and green arrows point towards structural features
953 supporting DNA cleavage.
954
955 **Supplementary Figure S4.** Secondary structure of sgRNAs targeted to the *Drosophila vermillion,*
956 *ebony* and *yellow* gene from Ren et al. (2014).
957 (A-L) Secondary structures of sgRNAs predicted by RNAfold targeting the vermillion (A-D), the
958 ebony (E-H) and the yellow gene (I-L) . Minimal energy and proximity structures are shown on the
959 left and right, respectively. Red and blue indicated high and low probabilities for the adopted
960 structural base-pairing assignment, respectively. The number on the right indicates the effectiveness
961 of inducing heritable mutations after injection into fly embryos as determined by Ren et al. (2014).
962 Red and green numbers indicate an effectiveness which is too high or too low compared to

963 predicted DNA cleavage efficiency. Red arrows point towards structural features likely limiting
964 DNA cleavage.
965
966 **Supplementary Figure S5.** Secondary structure of sgRNAs targeted to human CD22 from Graf et
967 al. (2019).
968 (A-V) Secondary structures of sgRNAs predicted by RNAfold. Minimal energy and proximity base-
969 pair structures are shown on the left and right, respectively. Red and blue indicated high and low
970 probabilities for the adopted structural base-pairing assignment, respectively. The number on the
971 right indicates the effectiveness of inducing heritable mutations after injection into fly embryos as
972 determined by Ren et al. (2014). Red and green numbers indicate an effectiveness which is too high
973 or too low compared to predicted DNA cleavage efficiency. Red arrows point towards structural
974 features likely limiting DNA cleavage.
975
976 **Supplementary Figure S6.** Key sgRNA secondary structural features directly interact with the Spy
977 Cas9 endonuclease in the three way junction .
978 (A) Minimum Free Energy (MFE) predicted secondary structure of the sgRNA used for co-
979 crystallization of Spy Cas9 (Nishimasu et al., 2014). Nucleotides involved in the three way junction
980 interactions are indicated (A_{51}, A_{52}, G_{62} and U_{63})
981 (B) X-ray crystal structure of spyCas9 endonuclease in complex with chimeric sgRNA bound to
982 genomic DNA target (PDB: 4OO8) (Nishimasu *et al* 2014). The gRNA portion is coloured in pink
983 and the remainder of the sgRNA is coloured in orange. Genomic DNA is coloured yellow and the
984 protein chain is coloured in green. Nucleotide residues situated at key structural features in the

985 sgRNA are coloured cyan. Note that a small portion of the Cas9 protein chain depicting amino acid
986 side chains to view the sgRNA three-way junction is transparent.
987 (C) Magnified view of the sgRNA three-way junction. Hydrogen bonds are indicated by black
988 dashed lines and aromatic stacking interactions are shown by pink dashed lines. Here, the
989 interaction of Phe1105 with A_{51}, A_{52} and U_{63} and the interaction of G_{62} with A_{51}, A_{52} and the
990 phosphate backbone can be seen.
991
992 **Supplementary Figure S7.** Key sgRNA secondary structural features directly interact with the Spy
993 Cas9 endonuclease in the tetraloop bulge.
994 (A) Minimum Free Energy (MFE) predicted secondary structure of the sgRNA used for co-
995 crystallization of spyCas9 (Nishimasu et al., 2014). Nucleotides forming the bulge are indicated
996 (A_{28}, A_{41}, A_{42} and G_{43}).
997 (B) X-ray crystal structure of Spy Cas9 endonuclease in complex with chimeric sgRNA bound to
998 genomic DNA target (PDB: 4OO8) (Nishimasu *et al* 2014). The gRNA portion is coloured in pink
999 and the remainder of the sgRNA is coloured in orange. Genomic DNA is coloured yellow and the
1000 protein chain is coloured in green. Nucleotide residues situated at key structural features in the
1001 sgRNA are coloured cyan. Note that the tetra loop after the bulge does not interact with Cas9 and
1002 sticks out of the structure.
1003 (C) Magnified view of the sgRNA bulge present in the tetra loop. Hydrogen bonds are indicated by
1004 black dashed lines and aromatic stacking interactions are shown by pink dashed lines. Here, the
1005 interactions of Phe351, Tyr359 and Asp364 with A_{42} and G_{43} can be seen. Inlet on the right: 180º
1006 turn to visualize base-stacking of U_{44} with Tyr325 and His328. Through these interactions base-
1007 pairing with G_{27} is prevented.

1008

1009 **Supplementary Figure S8.** Comparison of sgRNA selection by different sgRNA selection tools
1010 shown as median of the cleavage efficiency for individual data sets (A-L). The distribution of
1011 cleavage efficiencies for all sgRNAs is shown on the left (white box).

1012

1013 **Supplementary Figure S9:** *Drosophila* GFP-marked transformation vector for U6 promoter
1014 mediated expression of two sgRNAs.
1015 (A) Forward and return primer sequences to incorporate sgRNA sequences. The first nucleotide of
1016 the gRNA is indicated by an asterisk.
1017 (B) Cloning scheme for incorporating two sgRNAs into the destination vector.
1018 (C) Plasmid map of the fly transformation vector *pUC 3GLA U6.1/U6.3* sgRNA expressing two
1019 sgRNAs under U6.1 and U6.3 promoters, respectively.
1020 (D) Secondary structure of sgRNAs targeting *Ythdc1* and *Ythdf*.

1021

Figure 1

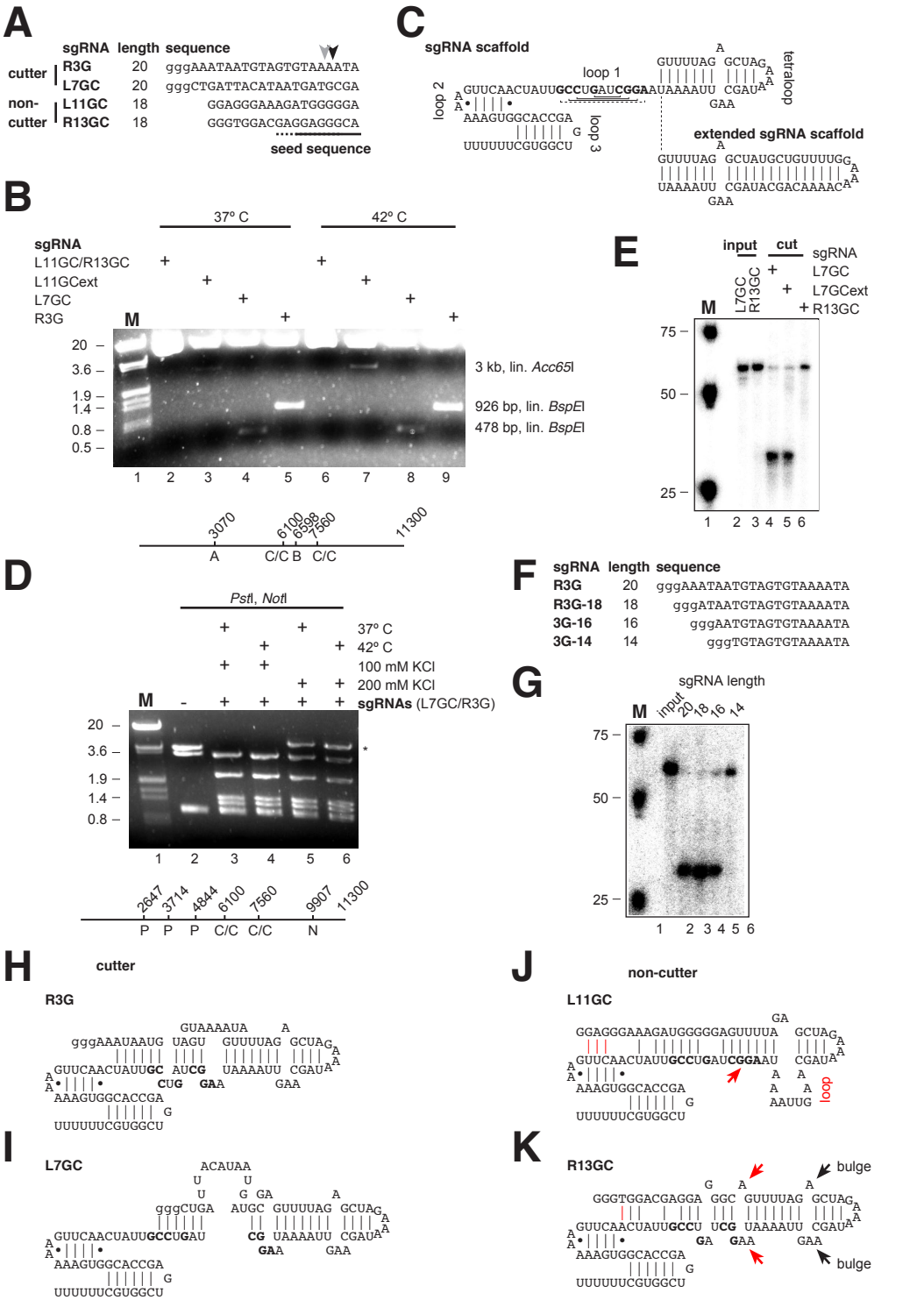
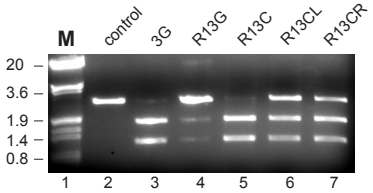

A
sgRNA length sequence
cutter R3G 20 gggAAATAATGTAGTGTAAAATA
L7GC 20 gggCTGATTACATAATGATGCGA
non-cutter L11GC 18 GGAGGGAAAGATGGGGGA
R13GC 18 GGGTGGACGAGGAGGGCA
seed sequence
B
37° C
42° C
sgRNA
L11GC/R13GC
L11GCext
L7GC
R3G
M
20
3.6
1.9
1.4
0.8
0.5
3 kb, lin. Acc65I
926 bp, lin. BspEI
478 bp, lin. BspEI
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

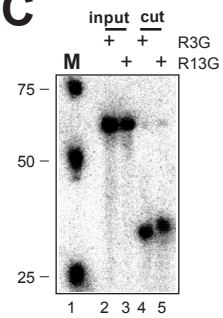
A

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

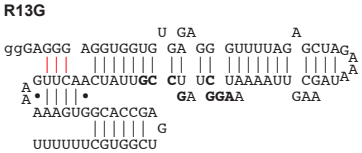
B



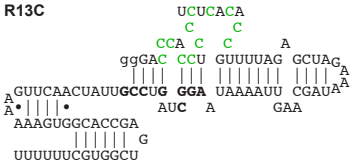
C



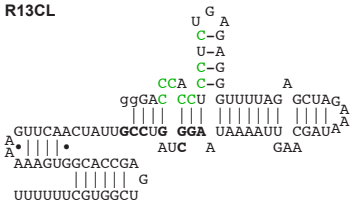
D



E



F



G

Figure 3

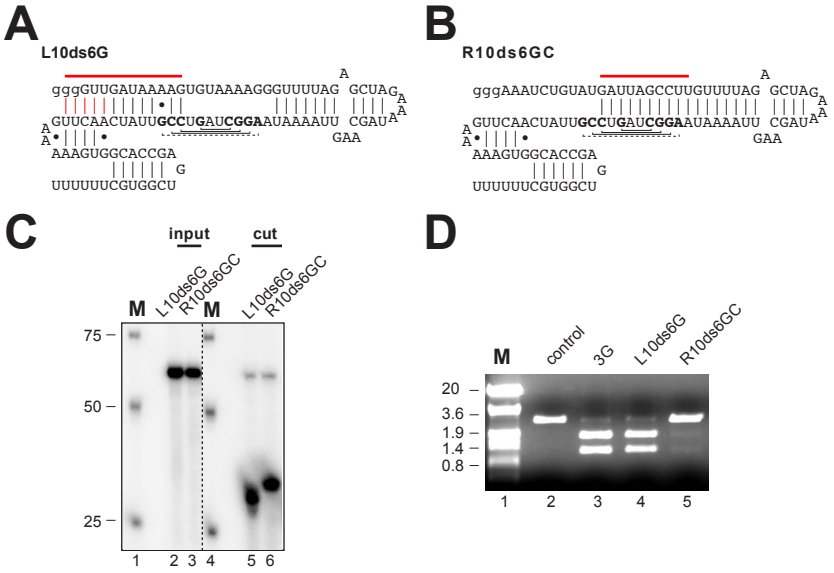

A
L10ds6G
B
R10ds6GC
C
input
cut
M
L10ds6G
R10ds6GC
75
50
25
1 2 3 4 5 6
D
M
control
3G
L10ds6G
R10ds6GC
20
3.6
1.9
1.4
0.8
1 2 3 4 5

Figure 4

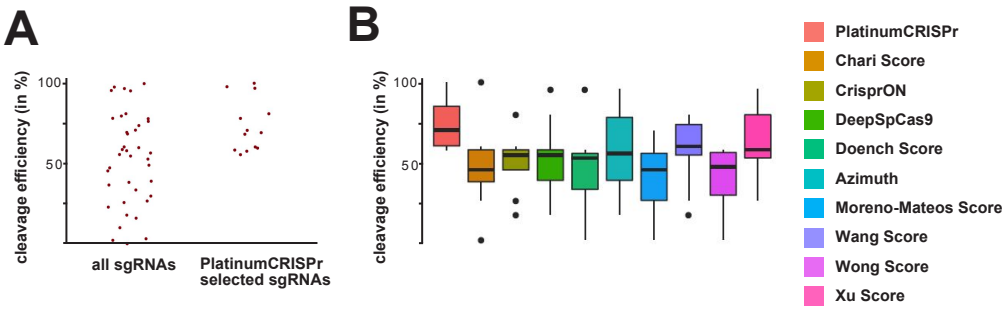

A
cleavage efficiency (in %)
100
50
all sgRNAs
PlatinumCRISPr selected sgRNAs
B
cleavage efficiency (in %)
100
50
PlatinumCRISPr
Chari Score
CrisprON
DeepSpCas9
Doench Score
Azimuth
Moreno-Mateos Score
Wang Score
Wong Score
Xu Score

Figure 5

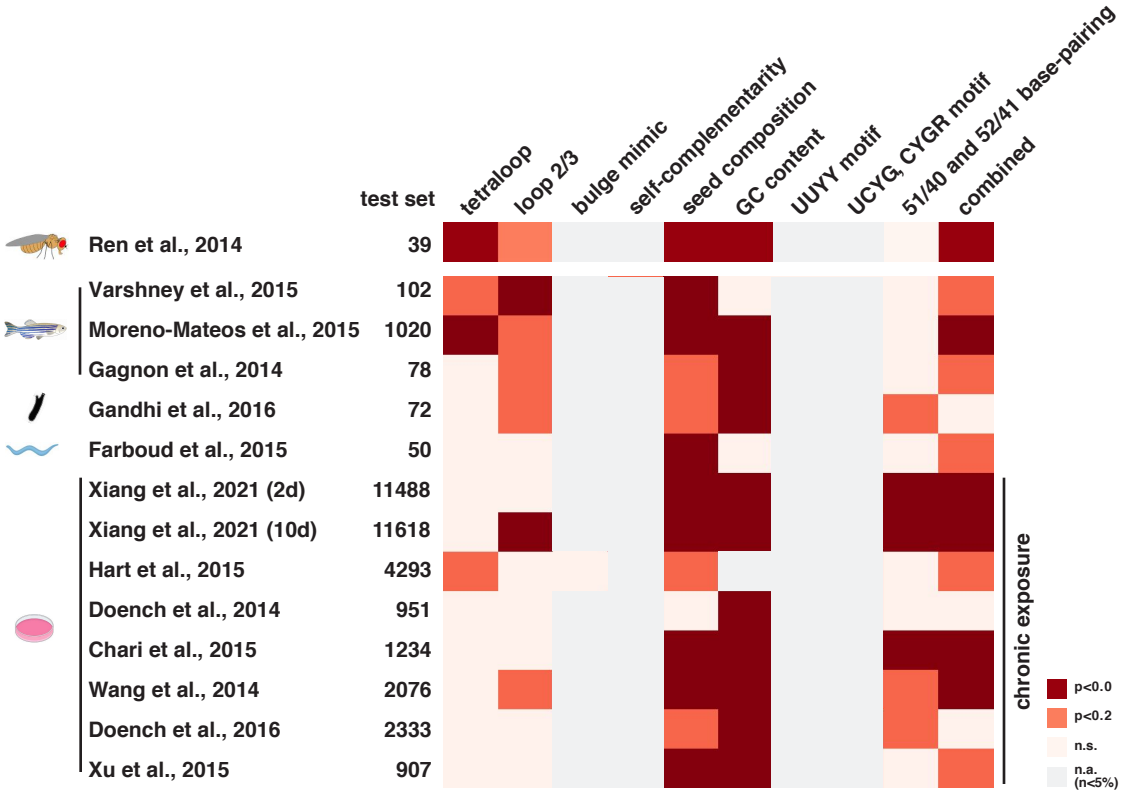

test set
tetraloop
loop 2/3
bulge mimic
self-complementarity
seed composition
GC content
UUYY motif
UCYG, CYGR motif
51/40 and 52/41 base-pairing
combined
Ren et al., 2014 39
Varshney et al., 2015 102
Moreno-Mateos et al., 2015 1020
Gagnon et al., 2014 78
Gandhi et al., 2016 72
Farboud et al., 2015 50
Xiang et al., 2021 (2d) 11488
Xiang et al., 2021 (10d) 11618
Hart et al., 2015 4293
Doench et al., 2014 951
Chari et al., 2015 1234
Wang et al., 2014 2076
Doench et al., 2016 2333
Xu et al., 2015 907
chronic exposure
p<0.0
p<0.2
n.s.
n.a. (n<5%)

Figure 6

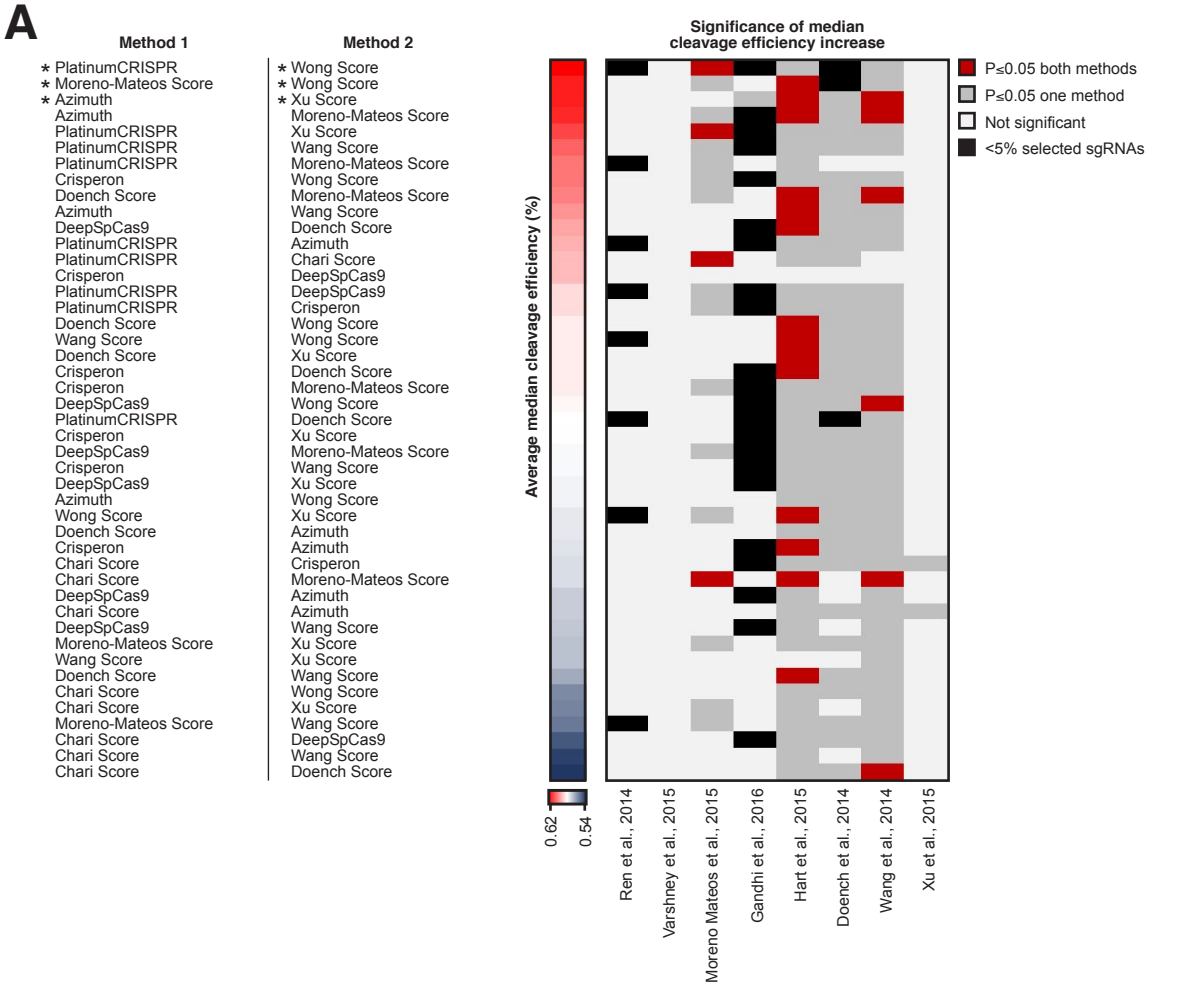
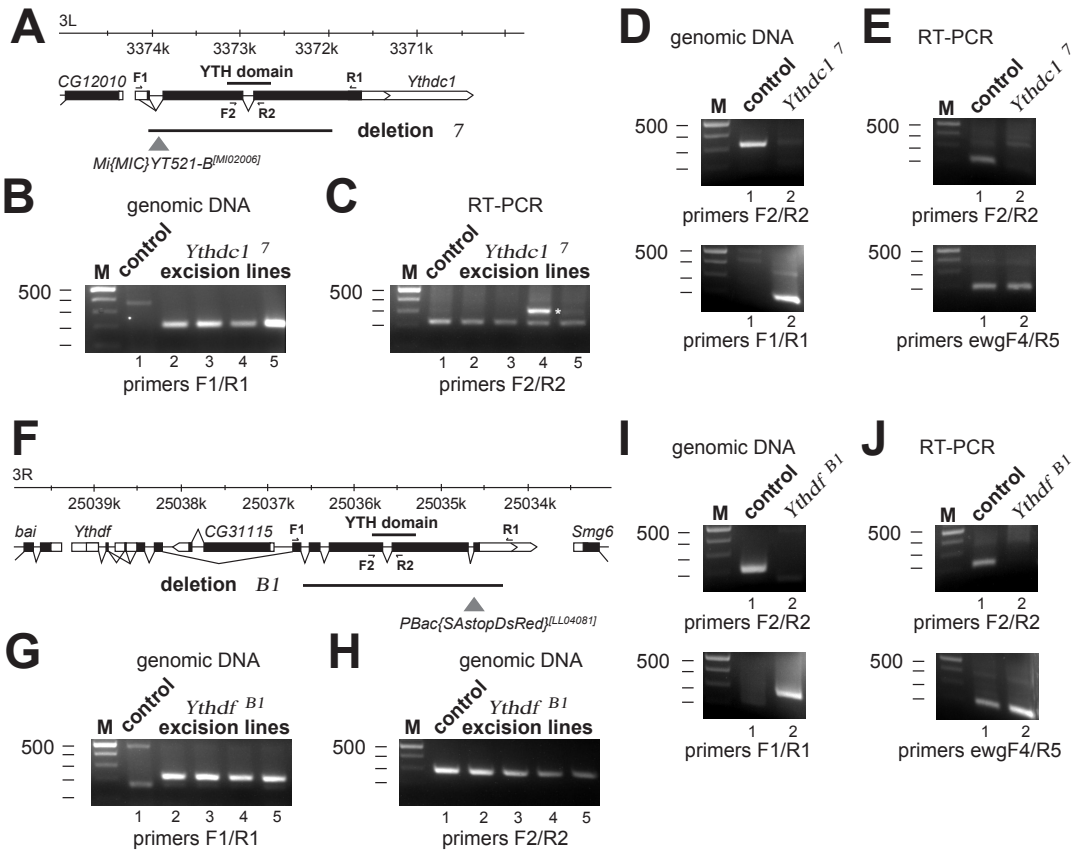
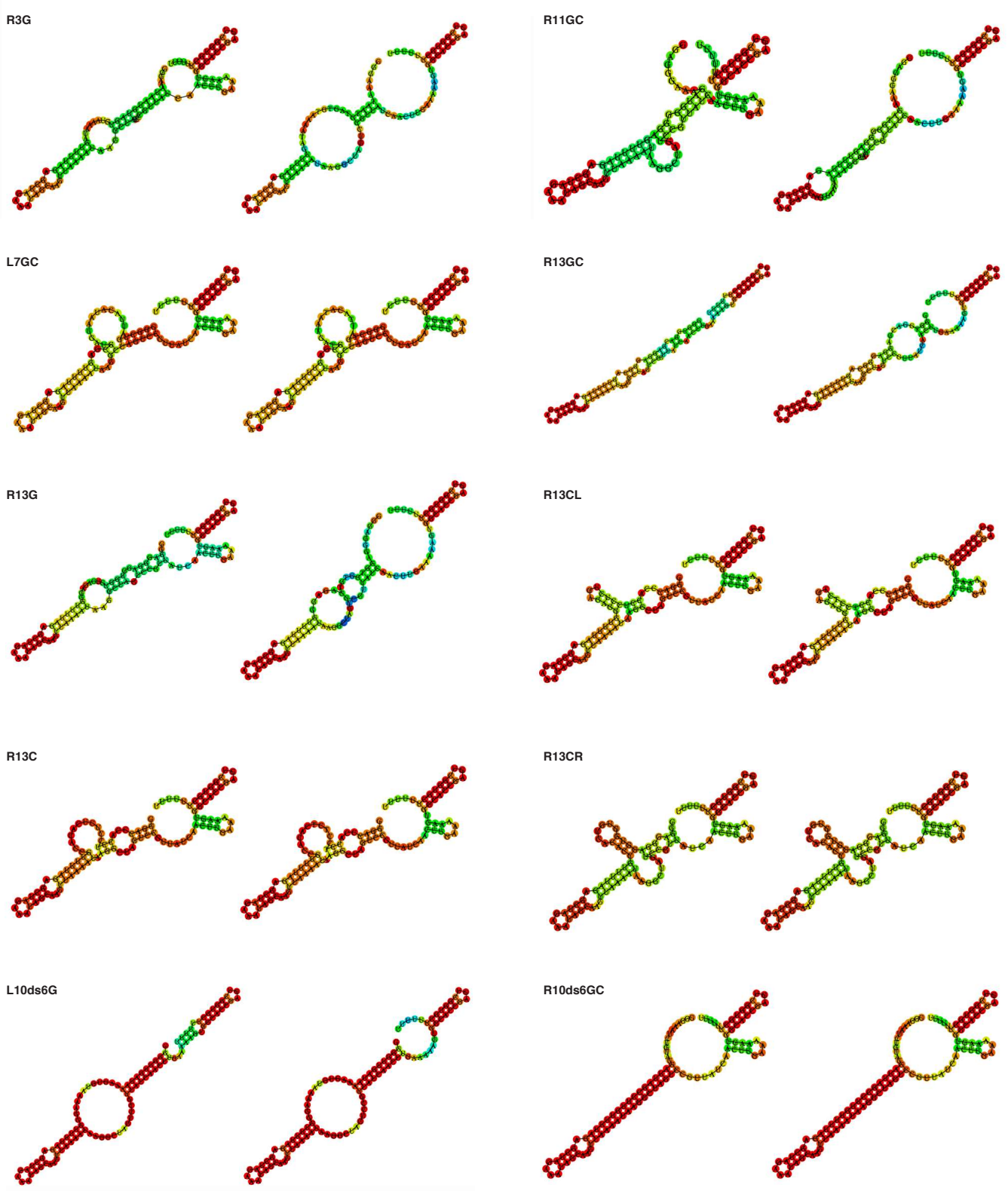

A
Method 1
Method 2
* PlatinumCRISPR
* Moreno-Mateos Score
* Azimuth
Azimuth
PlatinumCRISPR
PlatinumCRISPR
PlatinumCRISPR
Crisperon
Doench Score
Azimuth
DeepSpCas9
PlatinumCRISPR
PlatinumCRISPR
Crisperon
PlatinumCRISPR
PlatinumCRISPR
Doench Score
Wang Score
Doench Score
Crisperon
Crisperon
DeepSpCas9
PlatinumCRISPR
Crisperon
DeepSpCas9
Crisperon
DeepSpCas9
Azimuth
Wong Score
Doench Score
Crisperon
Chari Score
Chari Score
DeepSpCas9
Chari Score
DeepSpCas9
Moreno-Mateos Score
Wang Score
Doench Score
Chari Score
Chari Score
Moreno-Mateos Score
Chari Score
Chari Score
Chari Score
* Wong Score
* Wong Score
* Xu Score
Moreno-Mateos Score
Xu Score
Wang Score
Moreno-Mateos Score
Wong Score
Moreno-Mateos Score
Wang Score
Doench Score
Azimuth
Chari Score
DeepSpCas9
DeepSpCas9
Crisperon
Wong Score
Wong Score
Xu Score
Doench Score
Moreno-Mateos Score
Wong Score
Doench Score
Xu Score
Moreno-Mateos Score
Wang Score
Xu Score
Wong Score
Xu Score
Azimuth
Azimuth
Crisperon
Moreno-Mateos Score
Azimuth
Azimuth
Wang Score
Xu Score
Xu Score
Wang Score
Wong Score
Xu Score
Wang Score
DeepSpCas9
Wang Score
Doench Score
Significance of median cleavage efficiency increase
Average median cleavage efficiency (%)
0.62
0.54
Ren et al., 2014
Varshney et al., 2015
Moreno Mateos et al., 2015
Gandhi et al., 2016
Hart et al., 2015
Doench et al., 2014
Wang et al., 2014
Xu et al., 2015
P≤0.05 both methods
P≤0.05 one method
Not significant
<5% selected sgRNAs
B
Doench et al., 2014
Moreno Mateos et al., 2015
Xu et al., 2015
Wang et al., 2014
Gandhi et al., 2016
Hart et al., 2015
Ren et al., 2014
Varshney et al., 2015
cleavage efficiency (in %)
100
50
All sgRNAs
PlatinumCRISPr - Wong Score
Moreno-Mateos Score - Wong Score
Azimuth - Xu Score

Figure 7


A
3L
3374k
3373k
3372k
3371k
CG12010
F1
YTH domain
R1
Ythdc1
F2
R2
deletion 7
Mi{MIC}YT521-B[MI02006]
B
genomic DNA
M
control
Ythdc1 7
excision lines
500
1 2 3 4 5
primers F1/R1
C
RT-PCR
M
control
Ythdc1 7
excision lines
500
1 2 3 4 5
primers F2/R2
D
genomic DNA
M
control
Ythdc1 7
500
1 2
primers F2/R2
500
1 2
primers F1/R1
E
RT-PCR
M
control
Ythdc1 7
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 B1
PBac{SAstopDsRed}[LL04081]
G
genomic DNA
M
control
Ythdf B1
excision lines
500
1 2 3 4 5
primers F1/R1
H
genomic DNA
M
control
Ythdf B1
excision lines
500
1 2 3 4 5
primers F2/R2
I
genomic DNA
M
control
Ythdf B1
500
1 2
primers F2/R2
500
1 2
primers F1/R1
J
RT-PCR
M
control
Ythdf B1
500
1 2
primers F2/R2
500
1 2
primers ewgF4/R5

Supplementary Figure S1



Supplementary Figure S2

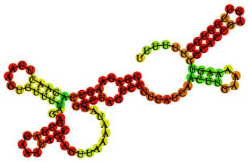
A

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

B

CG6422 5'

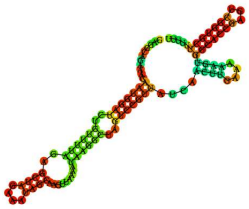
```
                          AA
                        C   G
                        U   U
                         C-G
                         A-U
                         A-U
                         G-U
                    GA   A-U
                GGA  AGUC  AGAGCUAG
                |||  ||||     ||||  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
```



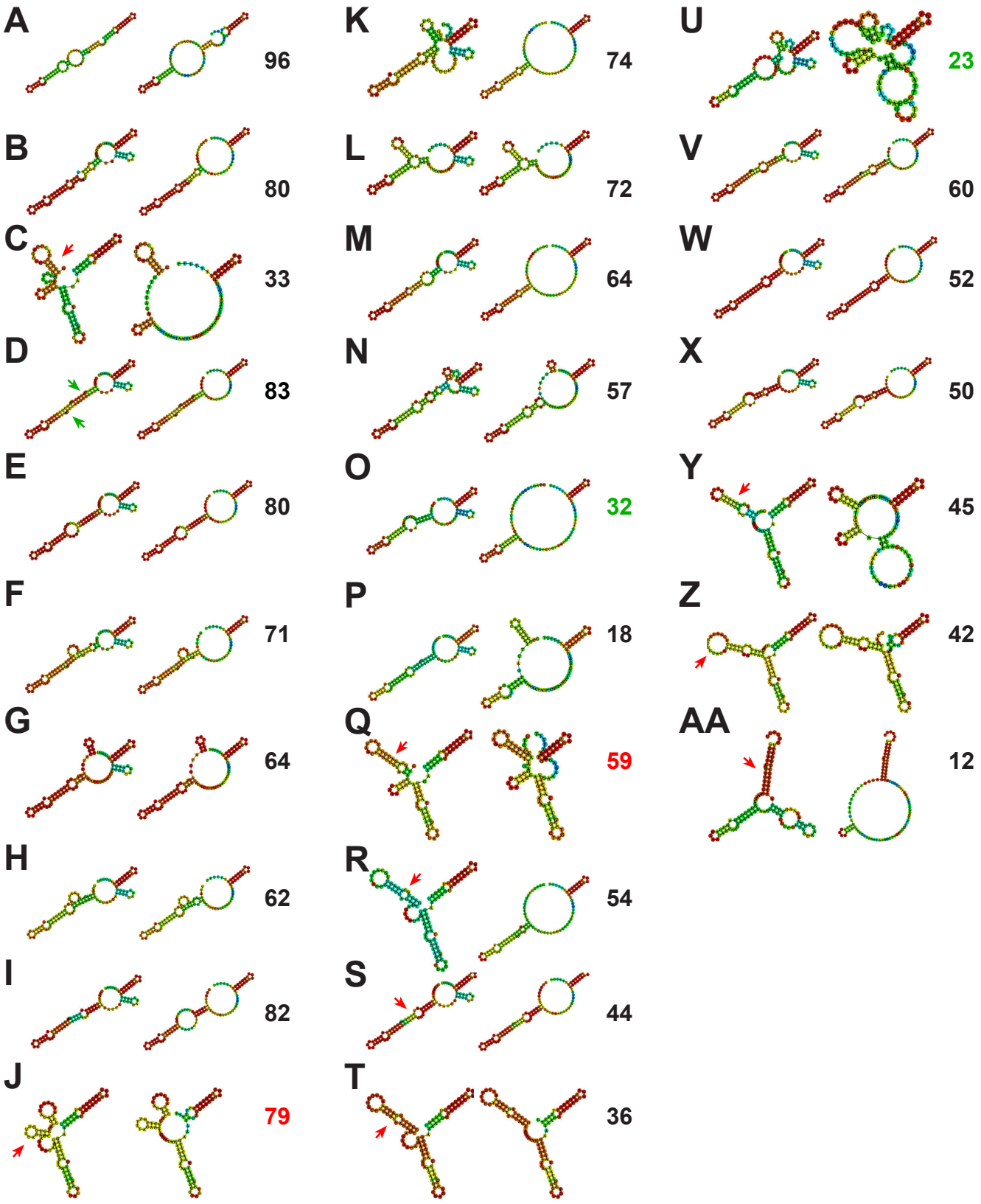
C

CG6422 3'

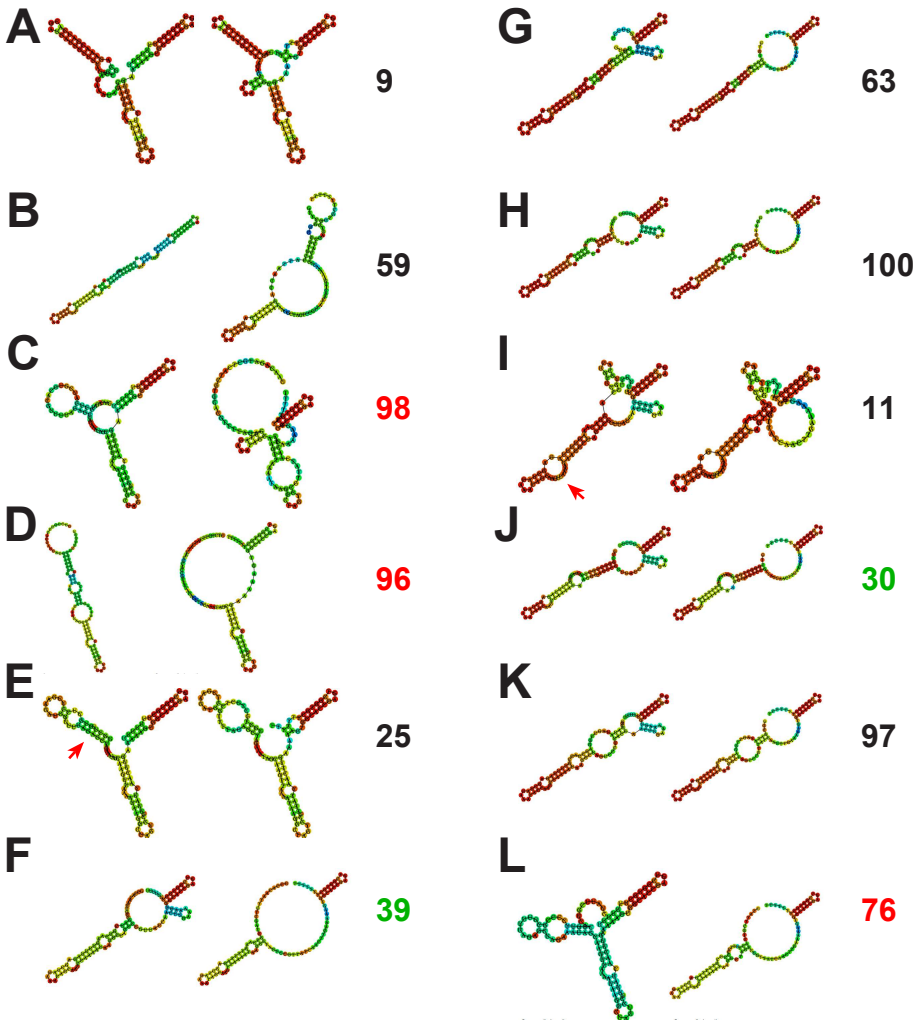
```
                        CU      G A
 GAGGCAGAAACAAUGGAU  GUUUUA   GCUA G
            |||||||  ||||||   |||| A
  AGUUCAACUAUUGCCUG  CGGAAU   CGAU A
  A ||||          AU      A  A
   AAAGUGGCACCGA          A  A
        ||||||| G         A  G
    UUUUUUCGUGGCU         A  U
                         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



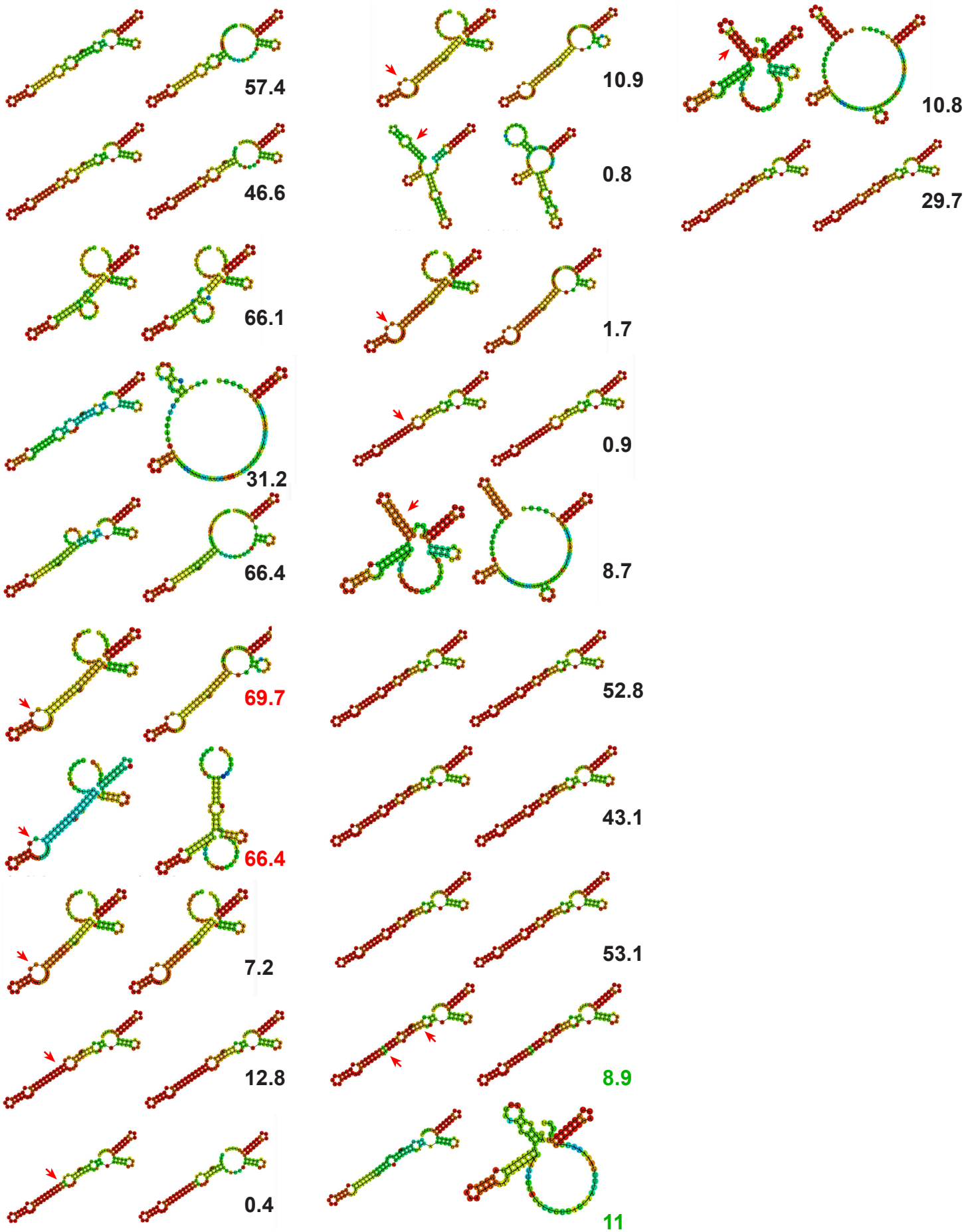
Supplementary Figure S3



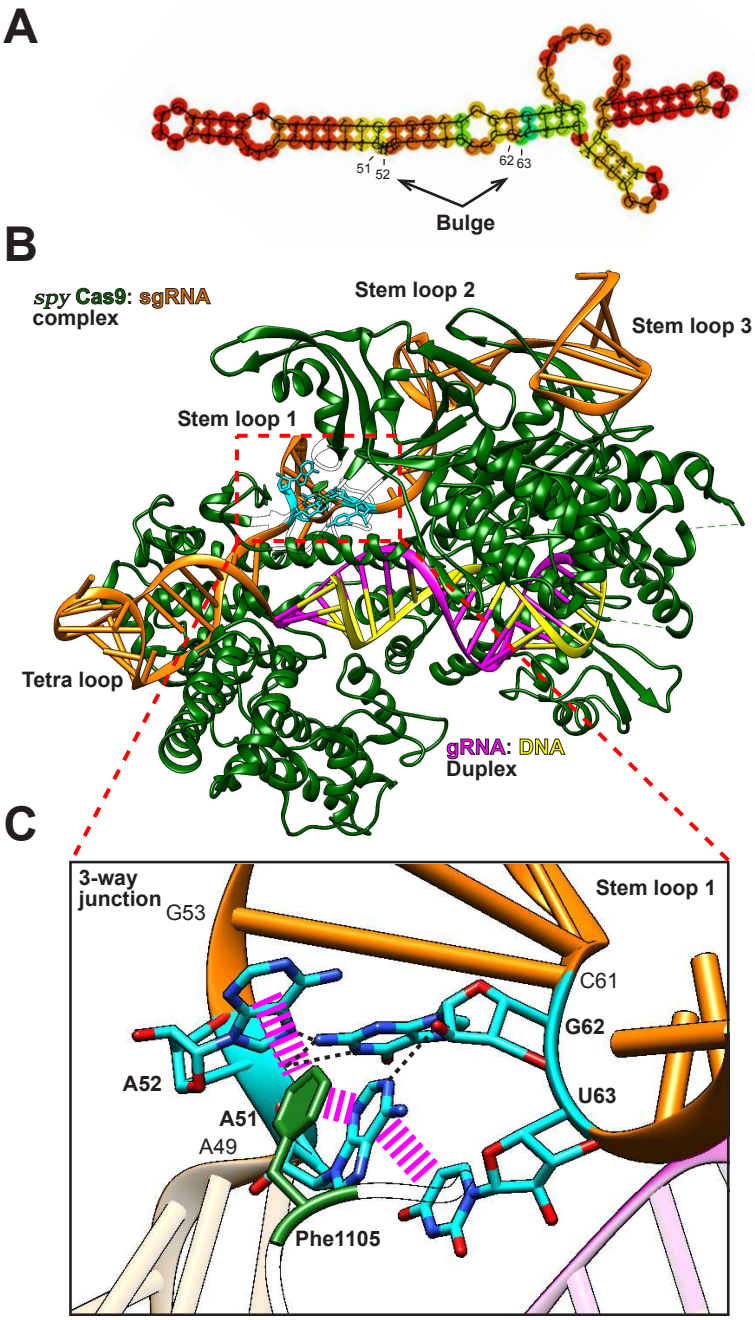
Supplementary Figure S4



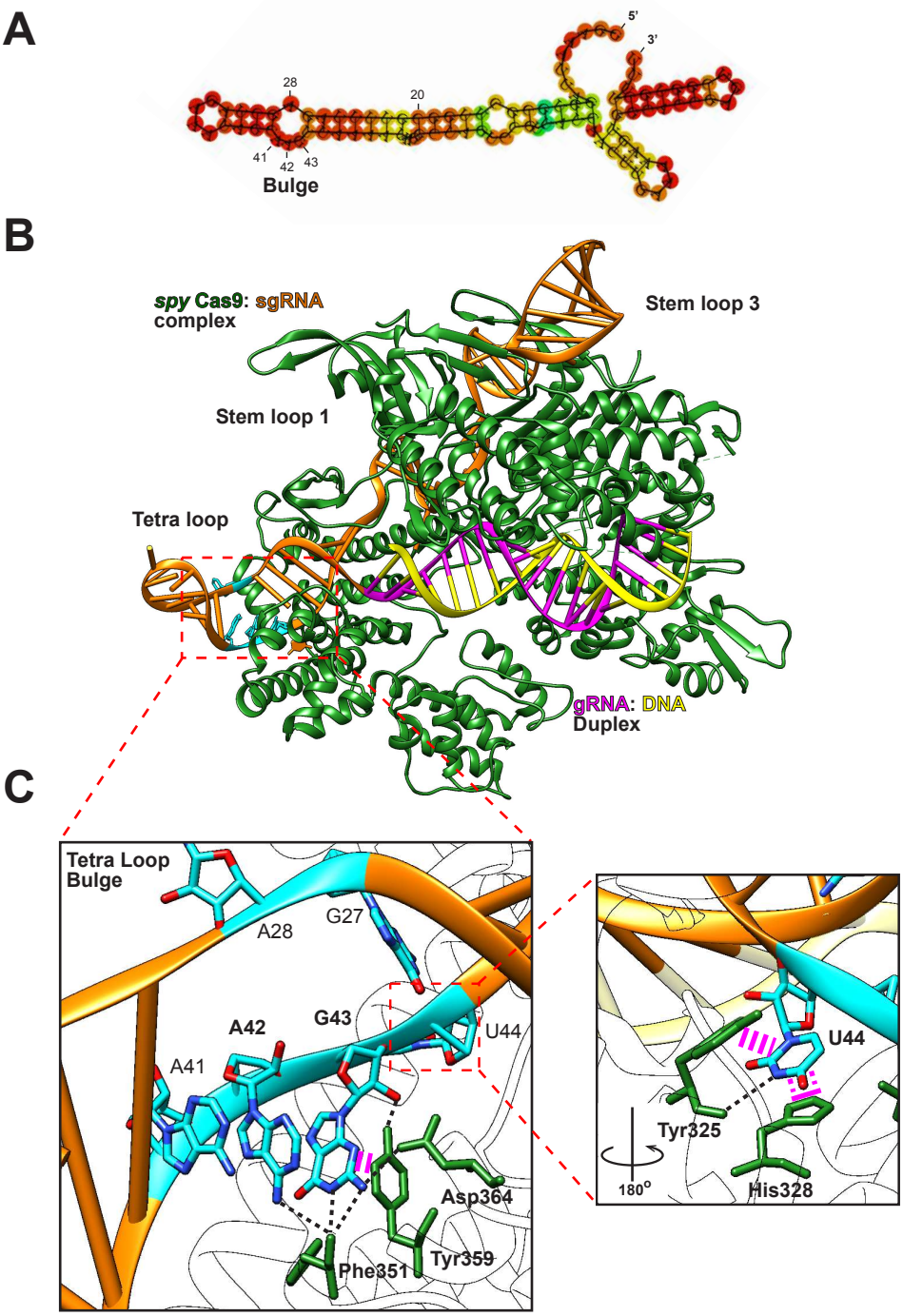
Supplementary Figure S5



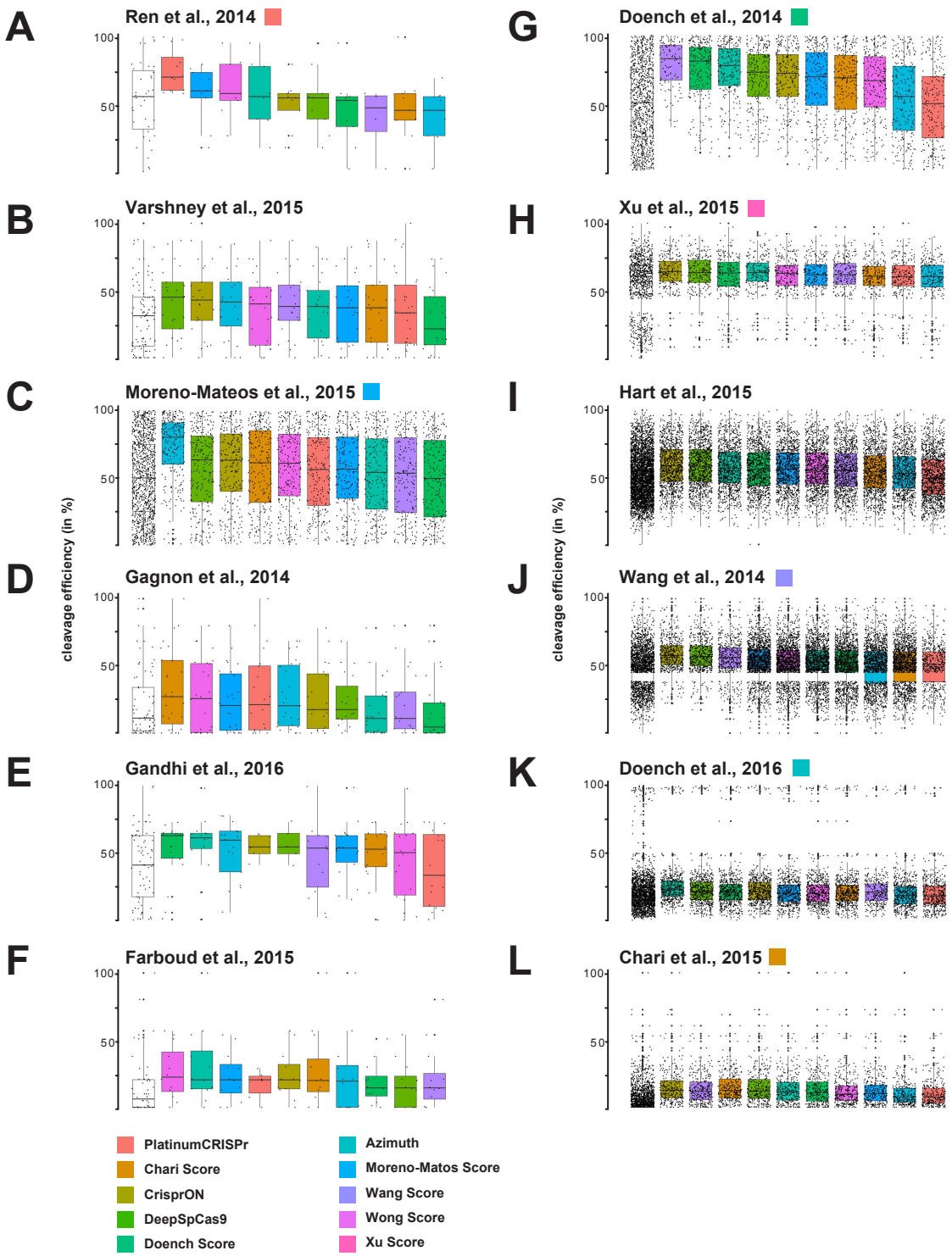
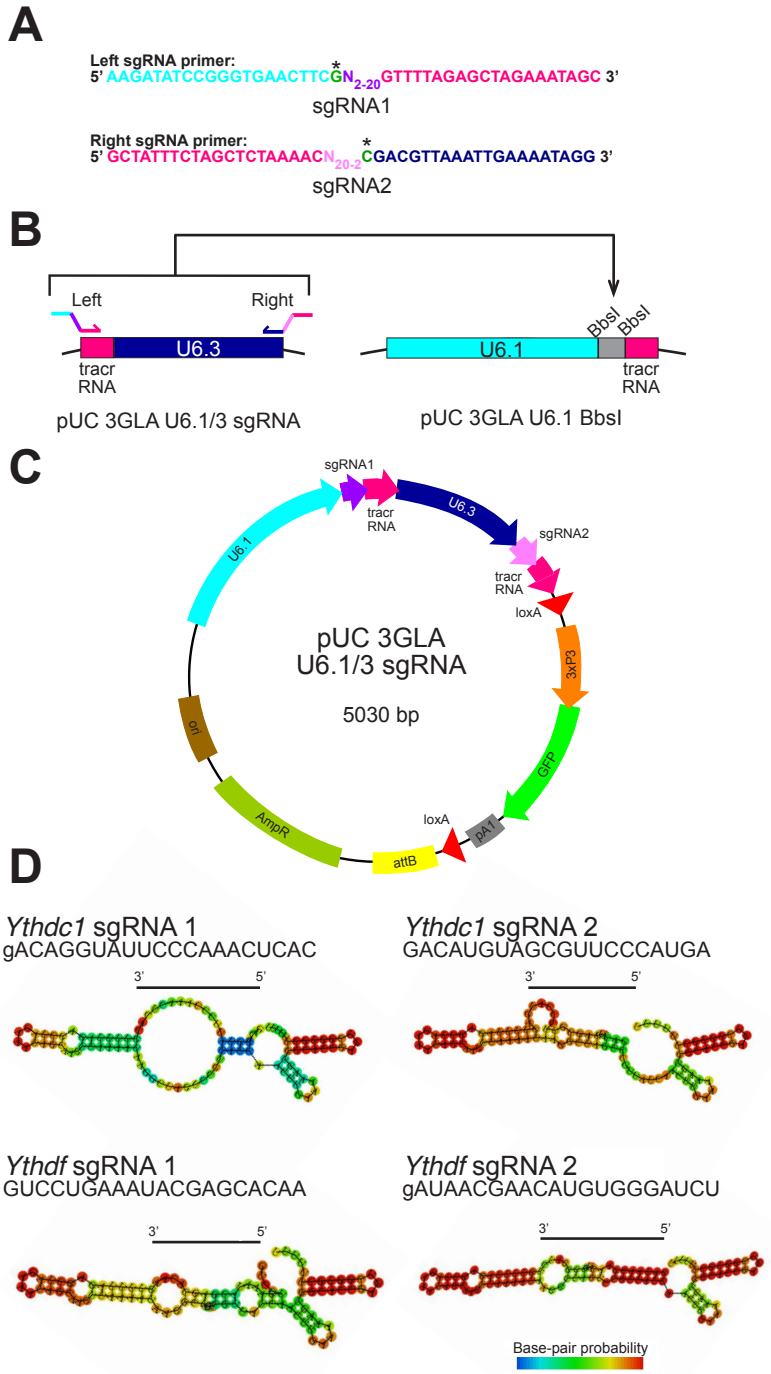
Supplementary Figure S6



Supplementary Figure S7



Supplementary Figure S8



Supplementary Figure S9