Full title: Evaluation of CRISPR gene-editing tools in zebrafish identifies spurious mutations in 'mock' control embryos

Short title: CRISPR gene-editing tools in zebrafish

José M. Uribe-Salazar[1,2], Aadithya Sekar[1], Gulhan Kaya[1], KaeChandra Weyenberg[1], Cole Ingamells[1], Megan Y. Dennis[1,2†]

[1]Genome Center, MIND Institute, and Department of Biochemistry & Molecular Medicine, University of California, Davis, CA, USA; [2]Integrative Genetics and Genomics Graduate Group, University of California, Davis, CA, USA

[†]Corresponding author:
Megan Y. Dennis, Ph.D.
University of California, Davis, School of Medicine
One Shields Avenue
Genome Center, 4303 GBSF
Davis, CA 95616
Email: mydennis@ucdavis.edu

Keywords: *Danio rerio*, zebrafish, CRISPR, Cas9, gene knockout, CIRCLE-seq, RNA-seq.

ABSTRACT

Zebrafish have practical features that make them a useful model for higher-throughput tests of gene function using CRISPR/Cas9 editing to create 'knockout' models. Due to the large number of available tools to design CRISPR assays and diversity of theories/model systems they were originally built on, we sought to systematically compare computational and empirical approaches for predicting gene-editing efficacy in zebrafish. We subjected zebrafish embryos to CRISPR/Cas9 with 50 different guide RNAs (gRNAs) targeting 14 genes and assayed individual editing efficiencies. We compared our experimental *in vivo* efficiencies in mosaic G_0 embryos with those predicted by seven commonly used gRNA design tools and found large discrepancies between methods. Assessing off-target mutations (predicted *in silico* and *in vitro*) found that the majority of tested loci had low *in vivo* frequencies (<1%). Moreover, understanding that recent segmental duplications in the zebrafish genome could exacerbate CRISPR targeting of individual genes, we cataloged these loci and have made them available as a resource. Lastly, we assessed the transcriptome of negative 'mock' control CRISPR larvae injected with Cas9 enzyme or mRNA with no gRNA using RNA-seq and identified differentially expressed genes with high variability between injections. Using these same data, we discovered on average ~60 putative somatic mosaic frameshift mutations impacting genes per pool of injected larvae, potentially due to background cutting of DNA with Cas9 in the absence of gRNA. To verify this previously unreported phenomenon in zebrafish, we validated seven of twelve genes tested carrying low frequency mosaic somatic mutations in the genomes of a separate batch of injected larvae. These results suggest that negative control embryos may carry mutations within genes leading to spurious phenotypes. Overall, our results provide a valuable resource for the zebrafish community for the design and execution of CRISPR/Cas9 experiments.

AUTHOR SUMMARY

Zebrafish have proven to be a powerful model organism for the functional characterization of genes. Development of new workflows targeting individual or multiple genes simultaneously require a thorough understanding of the advantages and limitations of current available methods for CRISPR-editing in zebrafish. Here, we systematically evaluated on- and off-target efficiencies prediction methods of 50 gRNAs by experimentally testing their CRISPR cutting efficiencies in embryos. Moreover, we performed a global assessment of duplicated portions of the zebrafish genome, providing a powerful resource for the design of future CRISPR assays. Lastly, we evaluated the possibility that spurious editing occurs in samples injected with the Cas9 nuclease without a gRNA, which are commonly used as a baseline control. This analysis revealed high variability in gene expression and the presence of frameshift variants in larvae injected solely with Cas9, suggesting that additional caution should be taken when using these samples as baseline controls in functional characterizations of genes.

67 **INTRODUCTION**

68 Zebrafish (*Danio rerio*) are increasingly used to rapidly and robustly characterize gene functions
69 (1–3). Features that make this model attractive over other classic vertebrate systems include
70 external fertilization, rapid development, a large number of progeny, embryonic transparency,
71 small size, and the availability of powerful gene-editing tools (4–8). Continuous improvements
72 of CRISPR-editing in zebrafish have allowed efficient targeting of multiple genes
73 simultaneously leading to rapid generation of either mosaic (G_0) or stable mutant lines and
74 subsequent characterizations of phenotypes (9–15), which have been used to test candidate genes
75 associated with a human diseases and developmental features (16). The trend towards more
76 affordable higher-throughput protocols using zebrafish requires a careful evaluation of methods
77 used for the design of CRISPR-based genetic screens.

78

79 New and creative CRISPR-based approaches in zebrafish address biological questions related to
80 developmental processes (e.g., via cell-lineage tracing) as well as gene functions (e.g., using
81 epigenome editing and targeted mutagenesis, review in (17)). In the latter application, important
82 factors in generating CRISPR gene knockouts include predicting/maximizing 'on target' Cas9
83 cleavage activity, predicting/minimizing unintended 'off-target' editing events, and efficiently
84 detecting small insertions or deletions (indels). Presence of indels at candidate loci can be
85 determined in an affordable manner via a number of approaches (reviewed in (18)), ranging from
86 simple identification of heteroduplexes visualized using a polyacrylamide gel electrophoresis
87 (PAGE) due to multiple alleles coexisting in the sampled DNA (19) to more sophisticated
88 sequencing methods (11,20) that precisely identify and quantify mutant alleles. On-target activity
89 of a particular guide RNA (gRNA) can be predicted using tools that provide efficiency scores,

90 often defined by information gathered across empirical assays (21). One relevant example is
91 CRISPRScan, a predictive-scoring system built from experimental zebrafish gene-editing data
92 based on multiple factors such as nucleotide GC and AT content and nucleosome positioning
93 (8,22). Bioinformatic tools also exist that define potential regions prone to off-target edits mainly
94 based on sequence similarity and the type/amount of mismatches relative to the on-target region
95 (22). More recent empirical approaches couple *in vitro* cleavage of genomic DNA (gDNA) with
96 Cas9 ribonucleoprotein (RNP), such as CIRCLE-seq (23) or GUIDE-seq (24), to provide a blind
97 assessment of editing sites but may not necessarily reflect the *in vivo* activity of the
98 CRISPR/Cas9 complex.
99
100 Previous studies have characterized CRISPR off-target activity *in vivo* to be relatively low in
101 zebrafish. A cross-generational study identified no inflation of transmitted *de novo* single
102 nucleotide mutations due to CRISPR-editing using exome sequencing and a stringent
103 bioinformatic pipeline (25) in a similar approach used to identify off-target mutations in mouse
104 trios (26,27). Other studies have observed off-target mutation rates ranging from 1.1 to 2.5% (28)
105 and 0.16% to 3.17% (12) in zebrafish by sequencing the top three or four off-target regions based
106 on sequence homology, respectively. Although off-target mutations should not significantly
107 impact studies of stable mutants, as unwanted mutations can be outcrossed out of studied lines
108 relatively easily (11,17), they may be problematic in rapid genetic screens using G_0 mosaics that
109 quickly test gene functions in a single generation.
110
111 The increasing number of tools available for the design and execution of CRISPR screens
112 provide an important resource to the zebrafish community. Here, we assayed different available

113 CRISPR on- and off-target prediction methods using empirical data from Cas9-edited zebrafish
114 embryos. We quantified CRISPR cleavage efficiencies *in vivo* employing a variety of
115 experimental approaches and used these results to compare the accuracy of *in silico* and *in vitro*
116 tools for predicting Cas9 on- and/or off-target activity. Finally, we assayed the impact of
117 injecting G_0 negative 'mock' control embryos injected with a buffer containing either Cas9
118 enzyme or mRNA in the absence of gRNAs by performing RNA-seq. We obtained a list of genes
119 as significantly differentially expressed versus uninjected wild-type siblings and carrying mosaic
120 frameshift variants in protein-coding regions possibly due to spurious mutations incurred from
121 Cas9 cleavage events. In all, these results will serve as a useful resource to the research
122 community as larger-scale CRISPR screens become more common in assaying gene functions in
123 zebrafish.
124
125 **RESULTS**
126 **Identification of CRISPR-induced indels in zebrafish**
127 We generated a dataset containing experimentally confirmed indels in injected G_0 zebrafish
128 larvae with 14 protein-coding genes targeted by 50 gRNAs (2–4 different gRNAs/target gene)
129 (Supplementary Tables 1 and 2). To obtain experimental *in vivo* editing efficiency values for
130 each gRNA, DNA extracted from a pool of G_0 mutants—generated via microinjections of
131 individual gRNAs at the one-cell stage and harvested at five days post-fertilization (dpf)—was
132 subjected to PCR (~500 bp region) and Sanger sequencing of the gRNA predicted target site.
133 From this, we inferred an *in vivo* 'efficiency score' measured as the percentage of DNA from
134 injected embryos harboring indels compared to uninjected batch siblings. The percentage of
135 indels was extracted using two different tools that deconvolve major mutations and their

136 frequencies within Sanger traces —Tracking of Indels by DEcomposition (TIDE) (29) and
137 Inference from CRISPR Edits (ICE) (30) (Figure 1) . Briefly, these tools use the gRNA sequence
138 to predict the cutting site in the control trace, map the sample trace to this reference, identify
139 indels by deconvolving all bases read at each position, and provide a frequency of the indel
140 spectrum (29,30). As previously reported (30), both tools provided positively correlated *in vivo*
141 scores across all gRNAs (Spearman r_s= 0.87, p= 6.78x10^{-15}) with an average score difference of
142 8.8±12.1 between tools (Supplementary Figure 1). We noted a higher correlation between tools
143 in scores below the median (Spearman r_s= 0.96, p= 1.23x10^{-13}) than above the median
144 (Spearman r_s= 0.65, p= 0.00072; Supplementary Figure 1). This result suggests that the
145 deconvolution process in both tools is more accurate when fewer molecules from the pool carry
146 indels. To determine the more accurate tool, we performed Illumina sequencing, which provides
147 more precise molecular estimates of mosaicism, of ~200 bp PCR fragments surrounding
148 predicted cut sites for a subset of gRNAs (n= 6, each targeting a different gene) with relatively
149 high *in vivo* efficiencies (>50%). Using *CrispRVariants* (31) with uninjected batch siblings DNA
150 as reference (Figure 1A), we extracted the proportion of reads carrying indel alleles and
151 identified a significant correlation with ICE (Spearman r_s= 0.93, p= 0.0077) but not with TIDE
152 (Spearman r_s= 0.08, p= 0.919). We therefore moved forward with *in vivo* efficiencies computed
153 via the ICE tool. Additionally, we observed mosaicisms quantified by Illumina to be 23.6±4.3
154 higher versus ICE editing scores suggesting the Sanger sequencing tool underestimates
155 mosaicism for gRNAs with higher editing efficiencies (Figure 1A, Supplementary Table 2).
156 A quicker and more affordable approach to quantify CRISPR cleavage efficiency is via PAGE,
157 which takes advantage of the heteroduplexes produced from DNA harboring a mosaic mix of
158 different types of indel mutations (19). We performed PAGE on ~200 bp regions surrounding the

159

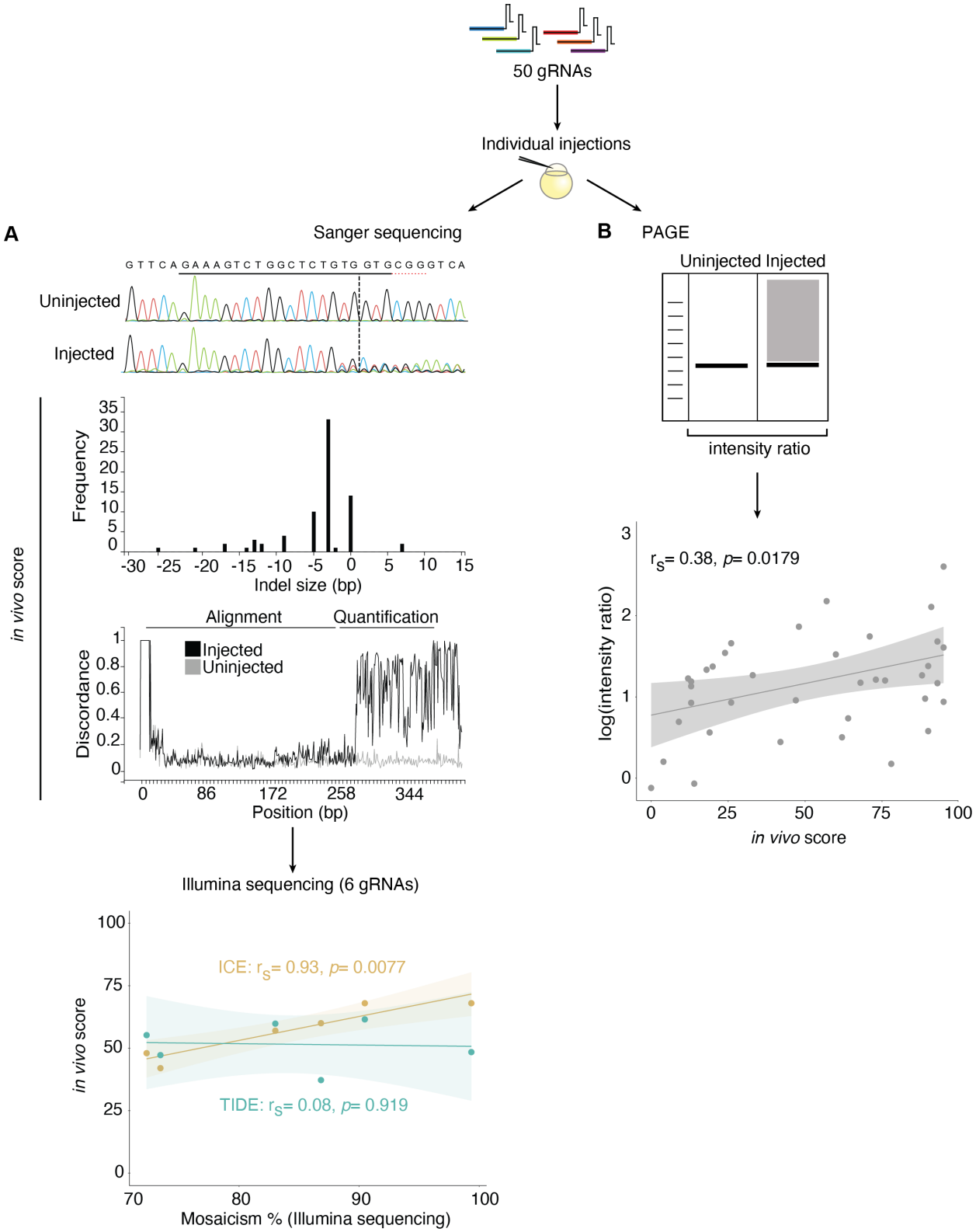


Figure 1. Workflow for the evaluation of CRISPR cleavages in zebrafish. 50 gRNAs were individually microinjected into one-cell stage embryos, DNA extracted, and Sanger sequencing and PAGE performed. (**A)** An *in vivo* score was obtained from the Sanger sequencing using the ICE and TIDE tools, and both scores for a subset of six gRNAs compared to mosaicism percentages from Illumina sequencing of the same region. **(B)** From the PAGE, an empirical intensity ratio was obtained and compared to the *in vivo* efficiency scores. Correlations results are shown in the scatter plots.

160 versus uninjected controls (see Methods). These intensity ratios were weakly correlated with our
161 *in vivo* efficiency scores (Spearman r_s= 0.38, p= 0.0179; Figure 1B) indicating that accurate
162 quantitative efficiencies cannot be directly deduced from PAGE but that the intensity of PCR
163 'smear' does qualitatively convey CRISPR-cleavage efficiency.
164
165 **Accuracy of CRISPR on-target predictions by different methods**
166 We next compared the accuracy of CRISPR on-target predictions computed by several published
167 algorithms, including CRISPRScan (22), CHOPCHOP (32), E-CRISP (33), CRISPR-GE (34),
168 CRISPR-RGEN (35), CCTop (36), as well as the design tool used for Integrated DNA
169 Technologies (IDT, www.idtdna.com). Only five comparisons between tools exhibited
170 significant correlations in their predicted scores and the se relationships were weak (Spearman r_s
171 <0.4; Figure 2). Thus, despite the research community broadly adapting all methods, there is
172 little consensus in predicting activity of a particular gRNA among these tools. Only
173 CRISPRScan, which we used to design our gRNAs, exhibited significant, albeit weak,
174 correlation with *in vivo* experimentally determined values (Spearman r_s= 0.32, p= 0.03; Figure
175 2). Therefore, higher CRISPRScan scores predict higher *in vivo* efficiencies but with low
176 accuracy.
177
178 Next, we evaluated the prediction power of an *in vitro* protocol, CIRCLE-seq (23), an approach
179 designed to predict target sites of a given gRNA by subjecting naked genomic DNA to Cas9
180 enzyme/gRNA cleavage followed by Illumina sequencing. We tested individually the 50 gRNAs
181 described above using CIRCLE-seq (37) and computed a log enrichment score normalized by the
182 library size, termed reads per million normalized (RPMN) (see Methods). We found that *in vitro*

183 obtained enrichment scores were

184 not correlated with *in vivo*

185 efficiencies (Spearman r_s= -0.04,

186 *p*= 0.81; Figure 2) or with *in silico*

187 predictions (Figure 2), indicating

188 that the CIRCLE-seq *in vitro* assay

189 and *in silico* methods do not

190 accurately predict on-target

191 CRISPR cleavage activity, at least

192 quantitatively. Previous work from

193 *in vivo* CRISPR studies of

194 zebrafish suggests that increased

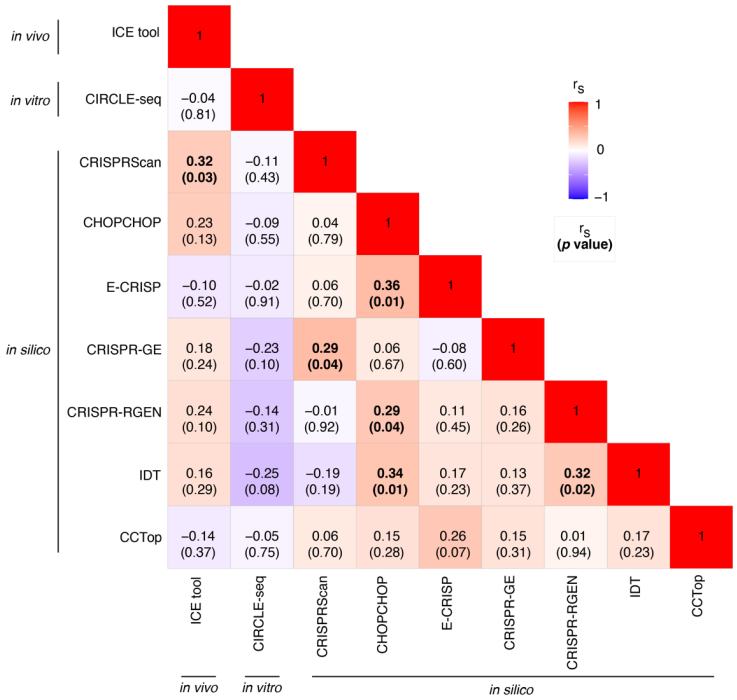


Figure 2. Correlation of on-target efficiencies calculated using different methods. Scores from *in silico* prediction tools, an *in vitro* protocol), and cutting cleavages obtained *in vivo* for 50 gRNAs were compared using Spearman correlations.

195 GC-content predicts increased activity of gRNAs (22). Examining our tested gRNAs, ranging in

196 GC content from 31.8 to 77.3%, we observed a positive increase of CRISPRScan *in silico* scores

197 (Linear model: beta= 68.2, *p*= 0.003, adjusted-r^2= 0.16) and CIRCLE-seq *in vitro* RPMN scores

198 (Linear model: beta= 6.4, *p*= 0.006, adjusted-r^2= 0.14) with GC percentage (Supplementary

199 Figures 2A,B). However, the experimentally determined *in vivo* scores were not correlated with

200 GC content (Linear model: beta=12.4, *p*= 0.817, adjusted-r^2= -0.02; Supplementary Figure 2C).

201 This result highlights that solely analyzing GC content cannot predict *in vivo* activity of a gRNA

202 and more variables should also be considered (e.g., depletion of A nucleotide bases or

203 nucleosome positioning (22)).

204

205 **CRISPR off-target mutations prediction methods**

206 To avoid spurious phenotypes, off-target mutations should be minimized when choosing gRNAs
207 in CRISPR experiments. To characterize off-target mutations for our set of 50 gRNAs, we
208 queried predictions from *in silico* (CRISPRScan) and *in vitro* (CIRCLE-seq) methods.
209 CRISPRScan provided a list of the top 30 predicted off-target sites for each gRNA within the
210 zebrafish reference genome (GRCz11/danRer11) based on a cutting frequency determination
211 (CFD) score that primarily takes into account sequence similarity, location, and type of sequence
212 mismatches (22,38). Alternatively, the CIRCLE-seq empirical approach produced variable
213 numbers of sites (between 18 and 874 regions, median 142.6; Supplementary Table 3) per gRNA
214 (defined as 'CIRCLE-seq sites') relative to the control library digested solely with Cas9. In
215 general, none of the on-target scores previously obtained *in silico* or *in vitro* correlated with the
216 number of CIRCLE-seq sites per gRNA (CRISPRScan score: Spearman r_s= -0.08, p= 0.559;
217 RPMN: Spearman r_s= 0.09, p= 0.523). Moreover, *in vivo* scores also were not correlated with the
218 number of CIRCLE-seq sites (Spearman r_s= 0.20, p= 0.204). To assess the concordance of *in*
219 *silico* and *in vitro* off-target predictions, we intersected ±1 kbp regions flanking the predicted
220 off-target sites and found an average overlap of 7.4±5.1% in regions between both prediction
221 methods (overlap ranges from 0 to 26%).
222
223 Focusing on putatively impactful off-target predictions, an average of 19±12% CRISPRScan
224 predicted and 60±15% CIRCLE-seq sites intersected at least one gene (Supplementary Table 3).
225 To verify if predicted off-target sites were subject to *in vivo* Cas9 cleavage, we performed Sanger
226 sequencing of predicted *in silico* (n=17) and *in vitro* (n=20) sites landing within genes for eight
227 gRNAs with high *in vivo* efficiency scores (>90%), an average of six regions per gRNA (see
228 Supplementary Table 1 for description of sites). Using the ICE tool to quantify *in vivo* indel

229 frequencies of predicted off-target sites, we found mosaic mutations at frequencies between 0
230 and 11%, with 23 out of the 37 sites evidencing indel frequencies below 1% (Supplementary
231 Figure 3A). To validate the accuracy of ICE at these low indel frequencies, we again performed
232 Illumina sequencing of predicted off-target sites for six of the eight evaluated gRNAs (see
233 Supplementary Table 1 for description) and found significant concordance in results (0.29 -
234 7.62%; Spearman r_s= 0.83, *p*= 0.039; Supplementary Figure 3B). The average difference in
235 mosaicism between ICE and Illumina was low (1.6±2.0) highlighting that ICE is more accurate
236 at sites with lower indel frequencies and suitable to quickly and economically assess predicted
237 off-targets regions.
238
239 No difference in indel frequencies was observed between sites predicted by CRISPRScan or
240 CIRCLE-seq (Mann-Whitney U= 175.5, *p*= 0.873; Supplementary Figure 3C). Additionally, no
241 significant correlation existed between the *in vivo* efficiency score of the gRNAs and the
242 frequency of indels in their predicted off-targets (Spearman r_s= 0.27, *p*= 0.111), suggesting that
243 higher on-target efficiencies do not necessarily translate into increased frequencies of off-target
244 mutations at the top predicted sites (Supplementary Figure 3D). Moreover, we assessed if sites
245 predicted with higher likelihoods of off-target cutting events resulted in higher mutation rates by
246 comparing the indel frequencies among the different levels of prediction (top 1, 2, or 3 prediction
247 scores by CRISPRScan or CIRCLE-seq). No difference was found between prediction groups,
248 suggesting that the information used by the tools to assign probabilities of off-target activity (e.g.
249 CFD scores in CRISPRScan or normalized read counts in CIRCLE-seq) do not necessarily
250 predict the efficiency of cutting at off-target sites (Kruskal-Wallis: $H_{(2)}$= 2.26, *p*= 0.320;
251 Supplementary Figure 3E). Thus, off-target cutting mutations at the assessed sites exhibited low

252 frequencies with no clear method performing best.

253

254 **Landscape of segmental duplications (SDs) in the zebrafish genome**

255 Optimal CRISPR target sites should be both unique and invariant in an organism's genome.
256 Initial sequencing of the zebrafish genome revealed an average of 2.28 zebrafish genes for every
257 human gene (39) partly due to the teleost-specific whole-genome duplication that occurred 320–
258 350 million years ago (40); therefore, the evaluation of a particular gene using zebrafish often
259 requires targeting multiple paralogs simultaneously. More recent duplication events such as SDs
260 (duplicated regions >1 kbp in size and >90% sequence identity) (41,42). We used *sedef* (43) to
261 catalog SDs in the zebrafish reference build (GRCz11/danRer11) and identified 72,450 SDs
262 across all chromosomes comprising 5.5% of the genome and 4,572 (15.1%) annotated genes
263 (Figure 3A; Supplementary Table 4). From the total SDs identified, the majority (85.4%) were
264 intrachromosomal, with a median size ranging from 1.9 kbp on chromosome 23 to 13.5 kbp on
265 chromosome 21 (Figure 3B, Supplementary Table 4). Additionally, SDs tended to cluster
266 together in groups (56.1 kbp median distance between SDs; empirical $p < 0.0001$ from 10,000
267 permutations, Figure 3C). Detailing the largest SDs (>50 kbp, Supplementary Figure 4), revealed
268 notable links between chromosomes 6 and 17 and between a specific region in chromosome 3
269 with the SD-dense chromosome 4. Studies targeting genes on these chromosomes should be
270 cautious of off-target effects in duplicate paralogs, which we have made available as a UCSC
271 Genome Browser custom track hub (pending; data currently available on github).

272

273

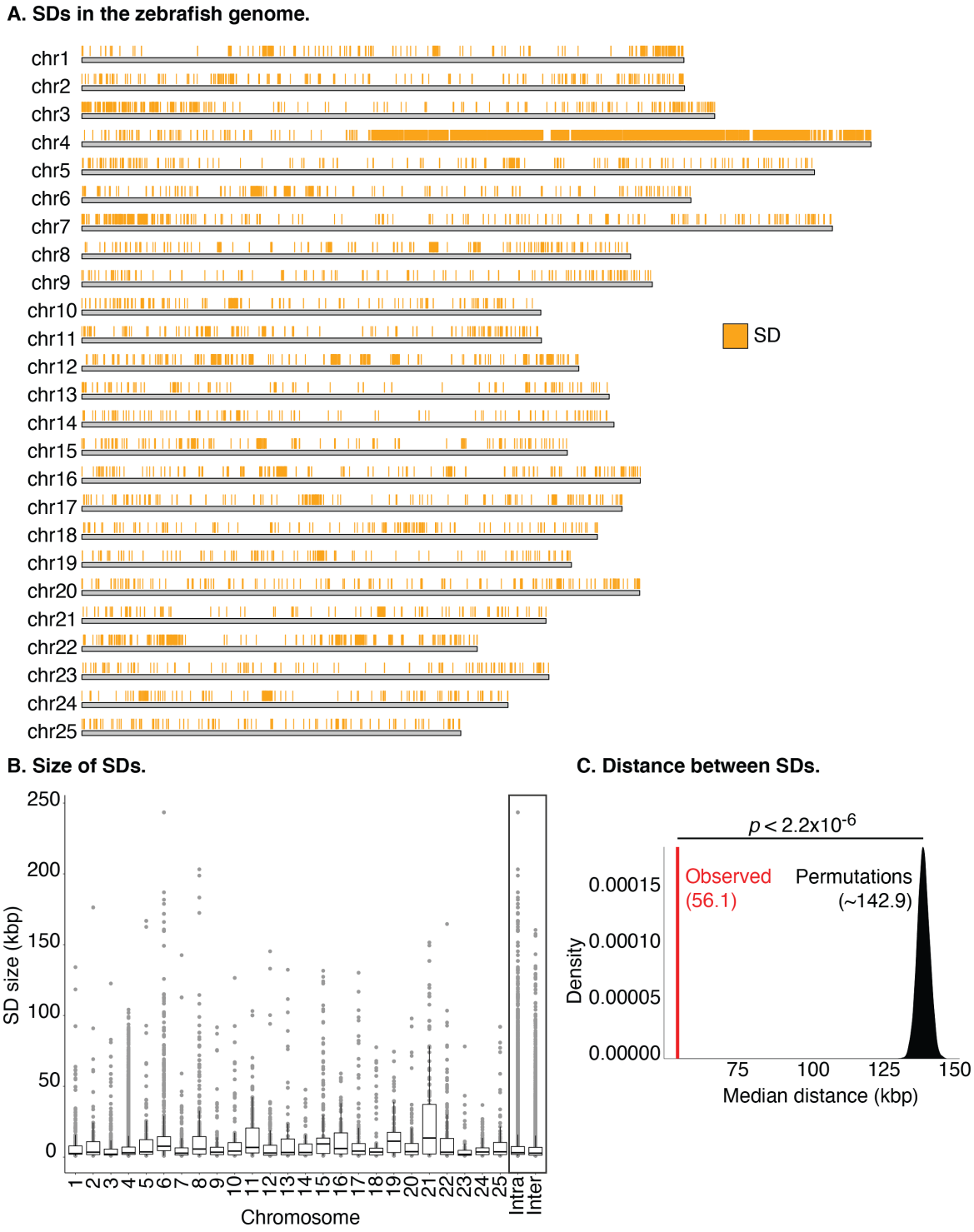


Figure 3. SDs in the zebrafish genome. (A) SDs (orange, upper section) highlighted above every zebrafish chromosome (gray blocks). **(B)** Size of the SDs identified at each chromosome. Intra: intrachromosomal, Inter: interchromosomal. **(C)** Median distance between SDs observed and estimated from a random sampling empirical test.

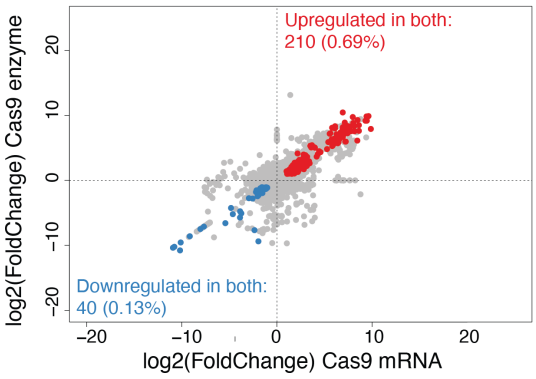
274

275 **Evaluating CRISPR Cas9-injection controls**

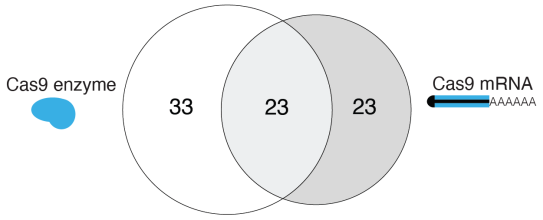
276 A commonly used 'mock' injection control for phenotypic screens of CRISPR-generated G_0
277 mosaic lines are embryos injected with buffer and Cas9 (enzyme or mRNA) in the absence of a
278 gRNA. We sought to determine if such control treatments could significantly impact the gene
279 expression profile of zebrafish larvae at the molecular and/or genomic level. We performed
280 RNA-seq of wild-type embryos injected with either Cas9 enzyme or Cas9 mRNA (three pools of
281 five injected larvae each) and their uninjected batch siblings (two pools of five larvae) as
282 controls. Compared with uninjected wild-type larvae, we identified hundreds of differentially
283 expressed (DE) genes in our Cas9-injected controls, with a greater number of upregulated genes
284 than downregulated genes (Supplementary Figure 5A, Supplementary Table 5). Specifically,
285 Cas9-enzyme injections resulted in more DE genes (total 971 genes (3.21% of genes assayed)
286 with 722 genes (2.39%) upregulated (Fold Change > 1) and 249 genes (0.82%) downregulated
287 (Fold Change < -1)) than Cas9-mRNA injections (total 462 genes, 1.53% of genes assays) with
288 351 genes (1.16%) upregulated and 111 genes (0.37%) downregulated. Two-hundred fifty DE
289 genes overlapped between the two treatments (Figure 4A), with genes involved in regeneration
290 processes upregulated (Figure 4C), suggesting a common response to damage incurred during
291 the injection process. We also noted significant variance across biological replicates when
292 quantifying transcript abundances in both Cas9 treatments (Supplementary Figure 5B-C). These
293 results suggest that considerable stochasticity may exist regarding the effects of Cas9 injections
294 in these controls.

295

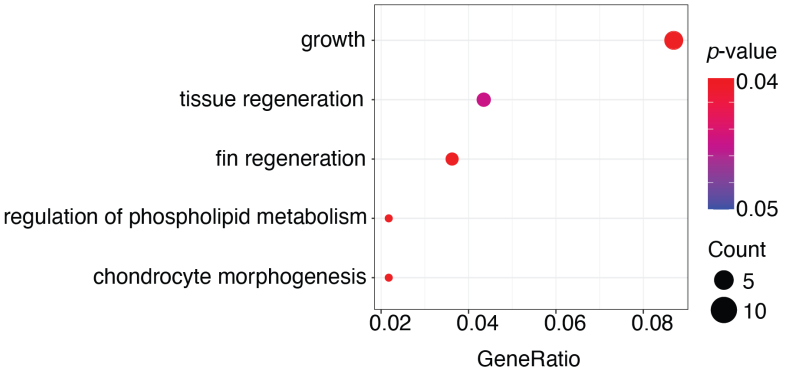
A. DE genes in Cas9-injected samples.



B. Genes with frameshift variants in Cas9-injected samples.



C. GO terms enriched in genes upregulated in both Cas9 treatments.



D. Validation of variants in Cas9-injected samples.

Gene	Alleles	Frequency (%) Enzyme	mRNA	Control
anp32b	Ref			
	3bp del	0.15	0.13	0
	1bp del	0.15	0	0
c3a1	Ref			
	1bp del	0.10	0.07	0
	1bp del	0.05	0.07	0
cyp3a65	Ref			
	2bp ins	0.02	0	0
	21bp del	0.02	0	0
pcnp	Ref			
	1bp del	0.05	0	0
	1bp del	0.02	0	0
sept3	Ref			
	2bp del	0.09	0	0
	1bp del	0	0.08	0
spry2	Ref			
	1bp del	0.05	0	0
	1bp del	0	0.04	0
tmem176	Ref			
	SNV: C/G	0.05	0	0
	1bp ins	0.02	0	0

- deletion AT G

Figure 4. Evaluation of Cas9 injection controls. **(A)** DE genes across samples injected with Cas9 enzyme or Cas9 mRNA relative to uninjected batch-siblings. Plot includes the number and percentage of genes downregulated (FoldChange < -1) and upregulated (FoldChange > 1) in both Cas9 treatments. **(B)** Protein-coding genes carrying frameshift variants for each Cas9-injected sample type. **(C)** Biological pathways enriched in genes upregulated in both Cas9 treatments using an enrichment test in *clusterProfiler (65)*. **(D)** Variants identified in larvae injected with Cas9 enzyme, Cas9 mRNA, or uninjected controls and their frequency. Two representative alleles showed for the seven genes with identified variants (see Supplementary Table 7 for complete results).

296

297 Most commonly we discovered only one variant per gene but did find a number of genes with
298 multiple variants (see Supplementary Table 6 for complete description of genes and alleles),
299 including *atp1a1a.1* with ten different alleles, *fga* with seven different alleles, and *anp32b* and
300 *slc25a4* with four different alleles each, highlighting the possibility that certain genes may be
301 consistently targeted by Cas9 in the absence of a gRNA. Variants identified were located closer
302 to a potential Cas9 PAM site (NGG) than by random chance (6-bp median distance to closest
303 PAM site; empirical $p = 0.0016$ from 10,000 permutations). To test if genes discovered as
304 mutated via RNA-seq data contain genomic indels, we performed targeted Illumina sequencing
305 of 12 genes in a separate set of Cas9-injection controls (enzyme or mRNA). We observed low
306 frequency (<1%) mosaic variants in 7 out of the 12 analyzed genes (Figure 4D, Supplementary
307 Table 7), but collectively observed mosaicism across 4.1±1.1% and 3.7±0.6% for Cas9 enzyme
308 and Cas9 mRNA injected, respectively, with no difference between injection type (Mann-
309 Whitney U= 31.5, p= 0.402) (Figure 4D, Supplementary Table 7). Considering the many
310 additional genes we identified as mutated in our RNA-seq data that remain to be validated, this
311 suggests that the collective combined effects of low frequency mutations at individual sites could
312 result in spurious phenotypes and impact subsequent characterizations.
313

314 **DISCUSSION**

315 Our study presents a comprehensive evaluation of empirical and predictive tools currently used
316 for CRISPR editing in zebrafish. Cleavage scores obtained by an *in vivo* assessment of 50
317 gRNAs via Sanger sequencing and deconvolution tools (ICE and TIDE) were highly congruent,
318 as previously reported (30), but only ICE scores were correlated with mosaicism percentages
319 from Illumina sequencing, the commonly used gold standard. ICE scores tended to underestimate

320 the presence of non-edited alleles by ~20% for gRNAs with high efficiencies (>50%), contrary
321 to previous comparisons of TIDE and Illumina sequencing in cell lines, where TIDE showed a
322 ~10–20% overestimation of non-edited alleles (44). For sites with lower indel frequencies, as we
323 observed for predicted off-target mutations, ICE scores were concordant with Illumina results
324 (~1–2% difference). Therefore, we suggest that Sanger sequencing deconvolution tools are
325 valuable for establishing relative gRNAs efficiencies but do not necessarily accurately predict
326 absolute cleavage efficiencies in zebrafish *in vivo*, except at sites with low indel frequencies. In
327 addition, we show that empirical 'smear' scores from PAGE can approximate cleavage
328 efficiencies, making it a more affordable and rapid approach to assay editing efficiency versus
329 sequencing.
330
331 On-target efficiency prediction tools showed large differences using the same set of gRNAs
332 sequences, highlighting the importance of understanding the features accounted for by each tool.
333 A recent review (21) provides a comprehensive overview of different design tools available and
334 the source of experimental data used to train each one. Compared with our *in vivo* efficiency
335 scores from zebrafish embryos, CRISPRScan was the only tool that could (somewhat) accurately
336 predict on-target efficiency in our set of gRNAs, while no other method provided scores that
337 were correlated with cleavage activities observed *in vivo*. Therefore, our results emphasize the
338 importance of utilizing a tool that has been trained using experimental data specifically from
339 zebrafish.
340
341 The *in silico* and *in vitro* off-target prediction tools sites predicted an overlapping set of genes
342 (>20% for our 50 gRNAs). Notably, we did not evaluate if these predicted sites included *cis*-

343 regulatory elements that could also potentially alter gene expression. Future assessments should
344 include tests targeting a diversity of loci for a more thorough understanding of the potential off-
345 target indels caused by unwanted CRISPR cleavage sites. We observed low off-target mutation
346 frequencies (most <1%), similar to those previously reported from using single (28) or multiple
347 (12) gRNAs. Notably, neither predictive method (CRISPRScan or CIRCLE-seq) nor intensity of
348 the prediction (using CFD or normalized read count) could accurately predict indel frequencies
349 at off-target sites. Typically, such low mutation frequencies should not be of high impact when
350 generating stable knockout zebrafish lines as these could be easily outcrossed. However, they
351 could potentially affect the phenotypic outcome when the injected G_0 mosaic populations were
352 analyzed directly.
353
354 The percentage of SDs observed in the zebrafish genome (~5.5%) is similar to reports of ~5%
355 observed in mammals, including humans (45–47). The distribution of these SDs across the
356 zebrafish genome is consistent with previous reports in mammalian genomes, where SDs tend to
357 be clustered (48–50), except for humans where these duplications are mostly interspersed (51).
358 Notably, it is unclear whether values observed in zebrafish are consistent across other teleost
359 species since few reports of SD quantification in other organisms are available. The higher
360 density of SDs present on chromosome 4, especially on its long arm, is consistent with previous
361 findings of extensive heterochromatic and protein-depleted regions found along this chromosome
362 (39). Importantly, we also observed a high percentage of SDs on chromosome 22 (6.70%),
363 making it a candidate region for subsequent analyses. All of these resources will aid future
364 CRISPR experiments targeting genes located in these duplicated regions for potential off-target
365 events in paralogous sequences.

366

367 Our data also showed that in the absence of gRNA, Cas9 alone may edit genomic DNA *in vivo*.
368 This could be problematic due to the difficulty in predicting where the cleavage sites could be
369 and the high impact these could have on functional assays. Studies *in vitro* have shown that
370 presence of Mn^{+2} ions can impact cleavage activity of Cas9 and Cas12a without a gRNA present
371 (52). Our injection buffer includes a functionally similar ion, Mg^{+2}, which has been shown to
372 compete with Mn^{+2} and activate common enzymes (53). Therefore, we investigated the effect of
373 Cas9 in genomic DNA of larvae solely injected with the nuclease, either as an enzyme or
374 mRNA. Our RNA-seq assay revealed that a set of genes involved in regeneration pathways are
375 differentially upregulated in the larvae injected with Cas9, highlighting the possibility that a
376 subset of gene expression can be perturbed simply due to the presence of Cas9 being present in
377 the developing embryo. Furthermore, we found an average of ~50 genes carrying frameshift
378 mutations per larvae injected with Cas9, with mutations validated in 7 of 12 genes through a
379 subsequent injection of Cas9 in a different batch of zebrafish embryos. A recent report showing
380 RNA targeting using the CRISPR/Cas9 system (54) suggests that we also cannot rule out the
381 possibility of Cas9 cutting mRNAs in our samples in the absence of a gRNA, which could
382 influence the presence of DE genes. The common genes carrying mutations across biological
383 replicates and the high confidence of the identified variants (mean read depth ~300x) highlight
384 the importance of future studies to assess the possibility of RNA-independent targeting by Cas9
385 in zebrafish and other organisms. In all, our results provide a list of genes that could be edited by
386 Cas9 alone in zebrafish embryos. Precaution should thus be taken when using Cas9 (no gRNA)-
387 injected embryos as a baseline control for comparing phenotypes, especially when performing
388 studies in the injected G_0 mosaic population directly.

389

390 Overall, our study is the first to perform a simultaneous assessment of gRNA activities predicted
391 by several commonly used *in silico* and *in vitro* methods with those determined experimentally
392 *in vivo* in injected zebrafish embryos. These results provide valuable information that can be
393 incorporated into the design and execution of CRISPR/Cas9 assays in zebrafish using available
394 workflows (10,11). Our aim was to provide information to aid in the decision-making process for
395 future projects using affordable and reliable gene-editing tools in zebrafish. As higher-
396 throughput methods are being developed for assaying multiple genes simultaneously, it will be
397 important to use optimal tools for predicting and assessing on- and off-target activity in zebrafish
398 larvae for accurate interpretation of phenotypic outcomes.

399

400 **METHODS**

401 **Zebrafish husbandry**

402 NHGRI-1 wild type zebrafish lines (55) were maintained through standard protocols (56) and
403 naturally spawned to obtain embryos. All animal use was approved by the Institutional Animal
404 Care and Use Committee from the Office of Animal Welfare Assurance, University of
405 California, Davis.

406

407 **Design and *in silico* predictions for gRNAs**

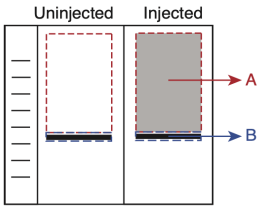
408 50 gRNAs targeting exons of 14 genes were designed using CRISPRScan (22) (scores ranging
409 between 24 and 83 with a mean value of 57.60) with zebrafish genome version
410 GRCz11/danRer11 as the reference (see description of gRNAs in Supplementary Tables 1 and
411 2). All targeted genes were protein coding. For each designed gRNA, we obtained the efficiency

412 scores predicted by CRISPRScan (22), CHOPCHOP (32), E-CRISP (33), CRISPR-GE (34),
413 CRISPR-RGEN (35), CCTop (36), and the IDT design tool (www.idtdna.com). From
414 CRISPRScan, we also gathered the top 30 predicted off-target sites for each gRNA defined by
415 the CFD score (22). Additionally, we utilized *bedtools* (57) to determine the GC percentage for
416 each gRNA.
417
418 **Microinjections to generate CRISPR G_0 mosaic mutants**
419 All gRNAs were individually injected into NHGRI-1 embryos to estimate the frequency of
420 indels. gRNAs were prepared following manufacturer's protocol (Integrated DNA
421 Technologies). Briefly, 2.5 µl of 100 µM crRNA, 2.5 µl of 100 µM tracrRNA, and 5 µl of
422 Nuclease-free Duplex Buffer using an annealing program consisting of 5 min at 95°C, a ramp
423 from 95°C to 50°C with a -0.1°C/s change, 10 minutes (min) at 50°C, and a ramp from 50°C to
424 4°C with a -1°C/s change. Ribonucleic protein (RNP) injection mix was prepared with 1.30 µl of
425 Cas9 enzyme (20 µM, New England BioLabs), 1.60 µl of prepared gRNAs, 2.5 µl of 4x Injection
426 Buffer (containing 0.2% phenol red, 800 mM KCl, 4 mM $MgCl_2$, 4 mM TCEP, 120 mM HEPES,
427 pH 7.0), and 4.6 µl of Nuclease-free water. Microinjections directly into the yolk of NHGRI-1
428 embryos at the one-cell stage were performed as described previously (58), using needles from a
429 micropipette puller (Model P-97, Sutter Instruments) and an air injector (Pneumatic MPPI-2
430 Pressure Injector). Embryos were collected and ~1 nl of RNP-mix was injected per embryo, after
431 previous calibration with a microruler. Twenty injected embryos per Petri dish were grown up to
432 5 dpf at 28°C.
433
434 **Sanger and Illumina amplicon sequencing**

435 DNA extractions were performed on 20 pooled embryos by adding 100 µl of 50 mM NaOH,
436 incubation at 95°C for 20 min, ramp from 95°C to 4°C at a 0.7°C/s decrease, followed by an
437 addition of 10 µl of 1 M Tris-HCl and a 15 min spin at 4680 rpm. We amplified a ~500 bp region
438 surrounding the targeted site of each gRNA (see Supplementary Table 1 for description of
439 primers). PCR amplifications were performed using 12.5 µl of 2X DreamTaq Green PCR Master
440 Mix (Thermo Fisher), 9.5 µl of Nuclease-Free water, 1 µl of 10 µM primers, and 1 µl extracted
441 DNA. Thermocycler program included 3 min at 95°C, followed by 35 cycles of 15 s at 95°C, 30
442 s at 60°C, and 20 s at 72°C, and a final 5 min incubation at 72°C. Reactions were purified using
443 Ampure XP magnetic beads (Beckman Coulter) and Sanger sequenced (Genewiz, San Diego,
444 CA). Raw ab1 files were used in the TIDE (29) and ICE (30) tools to predict the percentage of
445 indels, which we used as our *in vivo* editing score for each gRNA. To compare accuracy between
446 these tools, we PCR amplified ~200 bp around the targeted regions from the same extracted
447 DNA for six gRNAs and performed Illumina sequencing (Genewiz, San Diego, CA) to obtain
448 percent mosaicism of mutants by mapping paired-end fastq reads to the zebrafish reference
449 genome (GRCz11/danRer11) using *bwa* (59) and the R package *CrispRVariants* (31). For both
450 Sanger and Illumina sequencing, we used uninjected batch-sibling embryos as a control
451 reference.
452
453 **PAGE and intensity-ratio estimation**
454 An empirical cleavage analysis from each gRNA was performed using PAGE. Briefly, we
455 amplified a ~200 bp region in DNA around the targeted site from gRNA-injected and uninjected
456 embryos, as described above. Reactions were run on 7.5% polyacrylamide gels for 75 min at 110

457 V and revealed using GelRed (VWR International). Gel images were processed in the software

458 Fiji (60). For each sample, we defined areas **A** and **B** as follows:



459

460 For each gRNA, the mean-intensity value was obtained for the A and B areas in both the injected

461 and uninjected samples. The A and B areas were exactly the same size between samples. The

462 intensity ratio was calculated as: [injected B / injected A] / [uninjected B / uninjected A]. Log-

463 normalized intensity ratios followed a normal distribution (Shapiro-Wilk test: W= 0.96, p=

464 0.167) with an average value of 1.21±0.70.

465

466 **CIRCLE-seq**

467 CIRCLE-seq libraries were prepared for each gRNA (IDT) using gDNA extracted from NHGRI-

468 1 (DNA Blood & Tissue kit, Qiagen) following the described protocol (37). Libraries were

469 sequenced using one HiSeq XTen lane (Novogene, Sacramento, CA), providing an average of

470 7.3 million reads (range: 4.0 - 13.3 million reads) and >Q30 for 92% of reads per gRNA library.

471 Raw reads were processed using the bioinformatic pipeline described (37) to identify regions

472 with cutting events relative to a control sample (treated with Cas9 enzyme and no gRNA). To

473 obtain an on-target efficiency estimation from *in vitro* digestions, we calculated the reads per

474 million normalized (RPMN). For this purpose, we used *samtools* (59) to extract read coverage

475 from aligned bam files. For each gRNA, coverage was obtained for the third and fourth base

476 upstream of the PAM site as it is the region expected to be cut by Cas9 (61). RPMN for each

477 gRNA was calculated as the sum of coverage at these two sites divided by the total mapped reads

478 per sample and multiplied by one million to scale the values. RPMN scores ranged from 4.42 to
479 881 (median 99.3) so we decided to use a log normalization to reduce this range.
480
481 **SD discovery**
482 SDs in the zebrafish genome were identified using *sedef* (43) with a soft-masked reference
483 genome (GRCz11/danRer11). Filtering of these identified regions included a minimum
484 alignment size of 1 kbp and sequence similarity >90%. The median observed distance between
485 SDs was compared to the result of 10000 permutations of a random sampling across the genome.
486 In brief, sampling included the extraction of regions matching in number and size to the
487 identified SDs and the estimation of the median distance between them. One-tailed empirical *p*
488 values from this comparison were calculated as (M+N)/(N+1), where M is the number of
489 iterations with a median distance below the observed value and N is the total number of
490 iterations.
491
492 **RNA-seq**
493 We performed RNA-seq of Cas9 injected NHGRI-1 larvae to identify potential gRNA-
494 independent cleavage sites. One-cell stage NHGRI-1 embryos were injected with either Cas9
495 enzyme or Cas9 mRNA. Injection mix for Cas9 enzyme included Cas9 enzyme (20 µM, New
496 England BioLabs), 2.5 µl of 4x Injection Buffer (0.2% phenol red, 800 mM KCl, 4 mM $MgCl_2$, 4
497 mM TCEP, 120 mM HEPES, pH 7.0), and Nuclease-free water. Cas9 mRNA was obtained from
498 plasmid pT3TS-nCas9n (Addgene, plasmid #46757) (4), using the MEGAshotscript T3
499 transcription kit (Thermo Fisher) following manufacturer's guidelines of 3.5 h 56°C incubation
500 with T3. mRNA was purified with the MEGAclear transcription clean-up kit (Thermo Fisher)

501 and concentration of mRNA obtained using a NanoDrop (Thermo Fisher). The injection mix of
502 Cas9 mRNA contained 100 ng/µl of mRNA, 4x Injection Buffer (0.2% phenol red, 800 mM
503 KCl, 4 mM $MgCl_2$, 4 mM TCEP, 120 mM HEPES, pH 7.0), and Nuclease-free water.
504 Additionally, uninjected batch-siblings were used as controls. All embryos were grown at 28°C
505 in a density of <50 embryos per dish. At 5 dpf, three pools of five larvae were collected for each
506 group (Cas9 enzyme, Cas9 mRNA, and uninjected) for RNA extraction using the RNeasy kit
507 (Qiagen) with gDNA eliminator columns for DNA removal. Whole RNA samples were
508 subjected to RNA-seq using the poly-A selection method (Genewiz, San Diego, CA). Raw reads
509 were processed using the *elvers* (https://github.com/dib-lab/elvers; version 0.1, release DOI:
510 10.5281/zenodo.3345045) bioinformatic pipeline that utilizes *fastqc*, *trimmomatic* (62), and
511 *salmon (63)* to obtain the transcripts per kilobase million (TPM) for each gene. *DESeq2* (64) was
512 used to extract differentially expressed genes in the Cas9 enzyme or Cas9 mRNA injected
513 samples relative to the uninjected batch siblings. R package *cluterProfiler* (65) was used to
514 perform enrichment tests of differentially expressed genes in biological pathways.
515
516 **Variant identification from RNA-seq data**
517 We followed a previously described pipeline to identify somatic variants from RNA-seq data
518 (66). Briefly, we mapped reads with *STAR* (67) using the 2-pass mode and a genomic reference
519 created with GRCz11/danRer11 assembly and gtf files (release version 100). Variant calling was
520 performed with MuTect2 as part of *GATK* (68) using the tumor versus normal mode. 'Normal'
521 was defined by the two uninjected samples to identify all somatic mutations in our Cas9 injected
522 embryos. Variants inside SDs were removed. Variants were annotated using the Variant Effect
523 Predictor tool from ENSEMBL. High confidence variants (minimum sequencing depth of 20)

524 previously reported for the NHGRI-1 line (69) were removed. Only frameshift loss-of-function
525 variants with a minimum read depth of 20 in canonical protein-coding genes were considered.
526 We extracted the median distance between the identified variants and the nearest Cas9 PAM site
527 (NGG sequence) using coordinates from the CRISPRScan UCSC track. This median observed
528 distance was compared to the result of median distances of 10,000 permutations of random
529 sampling across the genome and their nearest PAM site. One-tailed empirical *p* values from this
530 comparison were calculated as (M+N)/(N+1), where M is the number of iterations with a median
531 distance below the observed value and N is the total number of iterations. We orthogonally
532 performed validations of variants in 12 genes via Illumina sequencing of a ~200 bp region
533 surrounding the identified variant location and the R package *CrispRVariants* (31)
534 (Supplementary Table 1 for primers description). For this purpose, we extracted DNA from a
535 pool of 20 embryos of each group (Cas9 enzyme, Cas9 mRNA, and uninjected). The 12 genes
536 assessed included three genes with most identified variants (*atp1a1a.1*, *fga*, and *anp32b*), four
537 DE genes carrying variants (*cyp3a65*, *galnt7*, *c3a.1*, *sept3*), two genes with variants found in
538 both Cas9 treatments (*actb1* and *slc25a4*), two genes with variants in all Cas9 enzyme injected
539 larvae (*spry2* and *pcnp*), and a gene with variants in all Cas9 mRNA injected samples (*tmem176*)
540 (see Supplementary Table 1 for description).
541
542 **Statistical analyses**
543 All analyses were performed in R version 4.0.2 (70) . Normality of variables was checked using
544 the Shapiro-Wilk test and parametric or nonparametric comparisons made accordingly.
545 Correlation tests (Pearson denoted as r and Spearman denoted as r_s) and linear regression models
546 were used to determine the relationship between variables. All analyses compared across

547 different experimental batches included *batch* as a factor in the model to prevent biases caused
548 by inter-batch differences. Averages include the standard deviation unless otherwise specified.
549 Alpha to determine significance across the different tests was set at 0.05 unless otherwise
550 specified. Additional R packages used for making figures included *clusterProfiler* (65),
551 *karyoploteR* (71), *ggcorrplot* (72), and *eulerr* (73).
552

553 **ACKNOWLEDGEMENTS**

554 We thank our team of UC Davis undergraduate students that maintain husbandry to keep our
555 zebrafish healthy and happy. Thank you to Dr. Li-En Jao for kindly providing the Cas9 mRNA
556 plasmid and comments on the manuscript, and Daniela C. Soto for bioinformatic support.
557

558 **DECLARATIONS**

559 **Abbreviations**

560 CRISPR: clustered regularly interspaced short palindromic repeats; gRNA: guide RNA; gDNA:
561 genomic DNA; RPMN: reads per million normalized. PAGE: polyacrylamide gel
562 electrophoresis; CFD: cutting frequency determination. SD: segmental duplication.
563

564 **Financial disclosure**

565 This work was supported, in part, by the U.S. National Institutes of Health (NIH) grants from the
566 National Institute of Neurological Disorder and Stroke (R00NS083627 to M.Y.D.), the Office of
567 the Director and National Institute of Mental Health (DP2 OD025824 to M.Y.D.), UC Davis
568 MIND Institute Intellectual and Developmental Disabilities Research Center pilot grant (U54
569 HD079125 to M.Y.D.), and a UC Davis Graduate Research Award (J.M.U-S.). M.Y.D. is also

570 supported by a Sloan Research Fellowship (FG-2016-6814). The funding sources did not play a
571 role in the research or publication process.
572

573 **Competing interests statement**

574 The authors declare that they have no competing interests.

575

576 **Ethics statement**

577 Animal use was properly approved by the Institutional Animal Care and Use Committee from
578 the Office of Animal Welfare Assurance, University of California, Davis.

579

580 **Data availability statement**

581 Original fastq files from the CIRCLE-seq and RNA-seq assays are deposited in the European
582 Nucleotide Archive repository under project PRJEB39643. Links to files are available in our
583 github repository (www.github.com/mydennislab/public_data).

584

585 **REFERENCES**

586 1. Meyers JR. Zebrafish: Development of a Vertebrate Model Organism: Zebrafish :
587 Development of a Vertebrate Model Organism. Current Protocols Essential Laboratory
588 Techniques. 2018 May 4;16(1):e19.

589 2. Holtzman NG, Kathryn Iovine M, Liang JO, Morris J. Learning to Fish with Genetics: A
590 Primer on the Vertebrate Model Danio rerio [Internet]. Vol. 203, Genetics. 2016. p. 1069–
591 89. Available from: http://dx.doi.org/10.1534/genetics.116.190843

592 3. Liu J, Zhou Y, Qi X, Chen J, Chen W, Qiu G, et al. CRISPR/Cas9 in zebrafish: an efficient
593 combination for human genetic diseases modeling. Hum Genet. 2017 Jan;136(1):1–12.

594 4. Jao L-E, Wente SR, Chen W. Efficient multiplex biallelic zebrafish genome editing using a
595 CRISPR nuclease system. Proc Natl Acad Sci U S A. 2013 Aug 20;110(34):13904–9.

596 5. Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, et al. Efficient genome
597 editing in zebrafish using a CRISPR-Cas system [Internet]. Vol. 31, Nature Biotechnology.
598 2013. p. 227–9. Available from: http://dx.doi.org/10.1038/nbt.2501

599 6. Irion U, Krauss J, Nüsslein-Volhard C. Precise and efficient genome editing in zebrafish
600 using the CRISPR/Cas9 system. Development. 2014 Dec;141(24):4827–30.

601 7. Varshney GK, Sood R, Burgess SM. Understanding and Editing the Zebrafish Genome
602 [Internet]. Advances in Genetics. 2015. p. 1–52. Available from:
603 http://dx.doi.org/10.1016/bs.adgen.2015.09.002

604 8. Vejnar CE, Moreno-Mateos MA, Cifuentes D, Bazzini AA, Giraldez AJ. Optimized
605 CRISPR-Cas9 System for Genome Editing in Zebrafish. Cold Spring Harb Protoc
606 [Internet]. 2016 Oct 3;2016(10). Available from: http://dx.doi.org/10.1101/pdb.prot086850

607 9. Gagnon JA, Valen E, Thyme SB, Huang P, Akhmetova L, Pauli A, et al. Efficient
608 mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment
609 of single-guide RNAs. PLoS One. 2014 May 29;9(5):e98186.

610 10. Varshney GK, Pei W, LaFave MC, Idol J, Xu L, Gallardo V, et al. High-throughput gene
611 targeting and phenotyping in zebrafish using CRISPR/Cas9. Genome Res. 2015
612 Jul;25(7):1030–42.

613 11. Varshney GK, Carrington B, Pei W, Bishop K, Chen Z, Fan C, et al. A high-throughput
614 functional genomics workflow based on CRISPR/Cas9-mediated targeted mutagenesis in
615 zebrafish. Nat Protoc. 2016 Dec;11(12):2357–75.

616 12. Shah AN, Davey CF, Whitebirch AC, Miller AC, Moens CB. Rapid reverse genetic
617 screening using CRISPR in zebrafish. Nat Methods. 2015 Jun;12(6):535–40.

618 13. Watson CJ, Monstad-Rios AT, Bhimani RM, Gistelinck C, Willaert A, Coucke P, et al.
619 Phenomics-Based Quantification of CRISPR-Induced Mosaicism in Zebrafish. Cell Syst.
620 2020 Mar 25;10(3):275–86.e5.

621 14. Wu RS, Lam II, Clay H, Duong DN, Deo RC, Coughlin SR. A Rapid Method for Directed
622 Gene Knockout for Screening in G0 Zebrafish. Dev Cell. 2018 Jul 2;46(1):112–25.e4.

623 15. Hoshijima K, Jurynec MJ, Klatt Shaw D, Jacobi AM, Behlke MA, Grunwald DJ. Highly
624 Efficient CRISPR-Cas9-Based Methods for Generating Deletion Mutations and F0
625 Embryos that Lack Gene Function in Zebrafish. Dev Cell. 2019 Dec 2;51(5):645–57.e4.

626 16. Thyme SB, Pieper LM, Li EH, Pandey S, Wang Y, Morris NS, et al. Phenotypic Landscape
627 of Schizophrenia-Associated Genes Defines Candidates and Their Shared Functions. Cell.
628 2019 Apr 4;177(2):478–91.e20.

629 17. Liu K, Petree C, Requena T, Varshney P, Varshney GK. Expanding the CRISPR Toolbox in
630 Zebrafish for Studying Development and Disease. Front Cell Dev Biol. 2019 Mar 4;7:13.

631 18. Zischewski J, Fischer R, Bortesi L. Detection of on-target and off-target mutations
632 generated by CRISPR/Cas9 and other sequence-specific nucleases. Biotechnol Adv. 2017
633 Jan;35(1):95–104.

634 19. Zhu X, Xu Y, Yu S, Lu L, Ding M, Cheng J, et al. An efficient genotyping method for
635 genome-modified animals and human cells generated with CRISPR/Cas9 system. Sci Rep.
636 2014 Sep 19;4:6420.

637 20. Brocal I, White RJ, Dooley CM, Carruthers SN, Clark R, Hall A, et al. Efficient
638 identification of CRISPR/Cas9-induced insertions/deletions by direct germline screening in
639 zebrafish. BMC Genomics. 2016 Mar 24;17:259.

640 21. Liu G, Zhang Y, Zhang T. Computational approaches for effective CRISPR guide RNA
641 design and evaluation. Comput Struct Biotechnol J. 2020;18:35–44.

642 22. Moreno-Mateos MA, Vejnar CE, Beaudoin J-D, Fernandez JP, Mis EK, Khokha MK, et al.
643 CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. Nat
644 Methods. 2015 Oct;12(10):982–8.

645 23. Tsai SQ, Nguyen NT, Malagon-Lopez J, Topkar VV, Aryee MJ, Keith Joung J. CIRCLE-
646 seq: a highly sensitive in vitro screen for genome-wide CRISPR–Cas9 nuclease off-targets
647 [Internet]. Vol. 14, Nature Methods. 2017. p. 607–14. Available from:
648 http://dx.doi.org/10.1038/nmeth.4278

649 24. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, et al. GUIDE-seq
650 enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat
651 Biotechnol. 2015 Feb;33(2):187–97.

652 25. Mooney MR, Davis EE, Katsanis N. Analysis of Single Nucleotide Variants in CRISPR-
653 Cas9 Edited Zebrafish Exomes Shows No Evidence of Off-Target Inflation. Front Genet.
654 2019 Oct 11;10:949.

655 26. Iyer V, Boroviak K, Thomas M, Doe B, Riva L, Ryder E, et al. No unexpected CRISPR-
656 Cas9 off-target activity revealed by trio sequencing of gene-edited mice. PLoS Genet. 2018
657 Jul;14(7):e1007503.

658 27. Dong Y, Li H, Zhao L, Koopman P, Zhang F, Huang JX. Genome-Wide Off-Target
659 Analysis in CRISPR-Cas9 Modified Mice and Their Offspring. G3 . 2019 Nov
660 5;9(11):3645–51.

661 28. Hruscha A, Krawitz P, Rechenberg A, Heinrich V, Hecht J, Haass C, et al. Efficient
662 CRISPR/Cas9 genome editing with low off-target effects in zebrafish. Development. 2013
663 Dec;140(24):4982–7.

664 29. Brinkman EK, Chen T, Amendola M, van Steensel B. Easy quantitative assessment of
665 genome editing by sequence trace decomposition. Nucleic Acids Res. 2014 Dec
666 16;42(22):e168.

667 30. Hsiau T, Conant D, Rossi N, Maures T, Waite K, Yang J, et al. Inference of CRISPR Edits
668 from Sanger Trace Data [Internet]. Available from: http://dx.doi.org/10.1101/251082

669 31. Lindsay H, Burger A, Biyong B, Felker A, Hess C, Zaugg J, et al. CrispRVariants charts the
670 mutation spectrum of genome engineering experiments. Nat Biotechnol. 2016 Jul
671 12;34(7):701–2.

672 32. Montague TG, Cruz JM, Gagnon JA, Church GM, Valen E. CHOPCHOP: a CRISPR/Cas9
673 and TALEN web tool for genome editing. Nucleic Acids Res. 2014 Jul;42(Web Server
674 issue):W401–7.

675 33. Heigwer F, Kerr G, Boutros M. E-CRISP: fast CRISPR target site identification. Nat
676 Methods. 2014 Feb;11(2):122–3.

677 34. Xie X, Ma X, Zhu Q, Zeng D, Li G, Liu Y-G. CRISPR-GE: A Convenient Software Toolkit
678 for CRISPR-Based Genome Editing. Mol Plant. 2017 Sep 12;10(9):1246–9.

679 35. Hwang G-H, Park J, Lim K, Kim S, Yu J, Yu E, et al. Web-based design and analysis tools
680 for CRISPR base editing. BMC Bioinformatics. 2018 Dec 27;19(1):542.

681 36. Stemmer M, Thumberger T, Del Sol Keyer M, Wittbrodt J, Mateo JL. Correction: CCTop:
682 An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. PLoS One. 2017
683 Apr 20;12(4):e0176619.

684 37. Lazzarotto CR, Nguyen NT, Tang X, Malagon-Lopez J, Guo JA, Aryee MJ, et al. Defining
685 CRISPR–Cas9 genome-wide nuclease activities with CIRCLE-seq [Internet]. Vol. 13,
686 Nature Protocols. 2018. p. 2615–42. Available from: http://dx.doi.org/10.1038/s41596-018-
687 0055-0

688 38. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized
689 sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat
690 Biotechnol. 2016 Feb;34(2):184–91.

691 39. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish
692 reference genome sequence and its relationship to the human genome. Nature. 2013 Apr
693 25;496(7446):498–503.

694 40. Meyer A, Schartl M. Gene and genome duplications in vertebrates: the one-to-four (-to-
695 eight in fish) rule and the evolution of novel gene functions. Curr Opin Cell Biol. 1999
696 Dec;11(6):699–704.

697 41. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, et al. Segmental
698 Duplications and Copy-Number Variation in the Human Genome [Internet]. Vol. 77, The
699 American Journal of Human Genetics. 2005. p. 78–88. Available from:
700 http://dx.doi.org/10.1086/431652

701 42. Eichler EE. Recent duplication, domain accretion and the dynamic mutation of the human
702 genome. Trends Genet. 2001 Nov;17(11):661–9.

703 43. Numanagic I, Gökkaya AS, Zhang L, Berger B, Alkan C, Hach F. Fast characterization of
704 segmental duplications in genome assemblies. Bioinformatics. 2018 Sep 1;34(17):i706–14.

705 44. Sentmanat MF, Peters ST, Florian CP, Connelly JP, Pruett-Miller SM. A Survey of
706 Validation Strategies for CRISPR-Cas9 Editing. Sci Rep. 2018 Jan 17;8(1):888.

707 45. Pu L, Lin Y, Pevzner PA. Detection and analysis of ancient segmental duplications in
708 mammalian genomes. Genome Res. 2018 Jun;28(6):901–9.

709 46. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental
710 duplications in the human genome. Science. 2002 Aug 9;297(5583):1003–7.

711 47. She X, Cheng Z, Zöllner S, Church DM, Eichler EE. Mouse segmental duplication and
712 copy number variation. Nat Genet. 2008 Jul;40(7):909–14.

713 48. Tuzun E, Bailey JA, Eichler EE. Recent segmental duplications in the working draft
714 assembly of the brown Norway rat. Genome Res. 2004 Apr;14(4):493–506.

715 49. Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC,
716 Gibbs RA, Muzny DM, et al. The genome sequence of taurine cattle: a window to ruminant
717 biology and evolution. Science. 2009 Apr 24;324(5926):522–8.

718 50. Nicholas TJ, Cheng Z, Ventura M, Mealey K, Eichler EE, Akey JM. The genomic
719 architecture of segmental duplications and associated copy number variants in dogs.
720 Genome Res. 2009 Mar;19(3):491–9.

721 51. Marques-Bonet T, Eichler EE. The evolution of human segmental duplications and the core
722 duplicon hypothesis. Cold Spring Harb Symp Quant Biol. 2009 Aug 28;74:355–62.

723 52. Sundaresan R, Parameshwaran HP, Yogesha SD, Keilbarth MW, Rajan R. RNA-
724 Independent DNA Cleavage Activities of Cas9 and Cas12a. Cell Rep. 2017 Dec
725 26;21(13):3728–39.

726 53. Wimhurst JM, Manchester KL. Comparison of ability of Mg and Mn to activate the key
727 enzymes of glycolysis [Internet]. Vol. 27, FEBS Letters. 1972. p. 321–6. Available from:
728 http://dx.doi.org/10.1016/0014-5793(72)80650-1

729 54. Strutt SC, Torrez RM, Kaya E, Negrete OA, Doudna JA. RNA-dependent RNA targeting by
730 CRISPR-Cas9. Elife [Internet]. 2018 Jan 5;7. Available from:
731 http://dx.doi.org/10.7554/eLife.32724

732 55. LaFave MC, Varshney GK, Vemulapalli M, Mullikin JC, Burgess SM. A Defined Zebrafish
733 Line for High-Throughput Genetics and Genomics: NHGRI-1. Genetics. 2014 Sep
734 1;198(1):167–70.

735 56. Westerfield M. The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio
736 Rerio). 1995.

737 57. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc
738 Bioinformatics. 2014 Sep 8;47:11.12.1–34.

739 58. Jao L-E, Appel B, Wente SR. A zebrafish model of lethal congenital contracture syndrome
740 1 reveals Gle1 function in spinal neural precursor survival and motor axon arborization.
741 Development. 2012 Apr;139(7):1316–26.

742 59. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
743 Bioinformatics. 2009 Jul 15;25(14):1754–60.

744 60. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an
745 open-source platform for biological-image analysis [Internet]. Vol. 9, Nature Methods.
746 2012. p. 676–82. Available from: http://dx.doi.org/10.1038/nmeth.2019

747 61. Wu X, Kriz AJ, Sharp PA. Target specificity of the CRISPR-Cas9 system [Internet]. Vol. 2,
748 Quantitative Biology. 2014. p. 59–70. Available from: http://dx.doi.org/10.1007/s40484-
749 014-0030-x

750 62. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence
751 data. Bioinformatics. 2014 Aug 1;30(15):2114–20.

752 63. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-
753 aware quantification of transcript expression. Nat Methods. 2017 Apr;14(4):417–9.

754 64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for
755 RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

756 65. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological
757 Themes Among Gene Clusters [Internet]. Vol. 16, OMICS: A Journal of Integrative
758 Biology. 2012. p. 284–7. Available from: http://dx.doi.org/10.1089/omi.2011.0118

759 66. Coudray A, Battenhouse AM, Bucher P, Iyer VR. Detection and benchmarking of somatic
760 mutations in cancer genomes using RNA-seq data. PeerJ. 2018 Jul 31;6:e5362.

761 67. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics.
762 2015 Sep 3;51:11.14.1–11.14.19.

763 68. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et
764 al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best
765 practices pipeline. Curr Protoc Bioinformatics. 2013;43:11.10.1–11.10.33.

766 69. LaFave MC, Varshney GK, Vemulapalli M, Mullikin JC, Burgess SM. A Defined Zebrafish
767 Line for High-Throughput Genetics and Genomics: NHGRI-1. Genetics. 2014 Sep
768 1;198(1):167–70.

769 70. Team RC. R: A language and environment for statistical com-puting. Vienna, Austria: R
770 Foundation for Statistical Computing. Retrieved from …; 2017.

71. Gel B, Serra E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. Bioinformatics. 2017 Oct 1;33(19):3088–90.

72. Kassambara A. ggcorrplot: Visualization of a Correlation Matrix using'ggplot2'. R package version 0 1. 2016;1.

73. Larsson J. eulerr: area-proportional Euler and Venn diagrams with ellipses. R package version. 2018;4(0).

FIGURE LEGENDS

Figure 1. Workflow for the evaluation of CRISPR cleavages in zebrafish. 50 gRNAs were individually microinjected into one-cell stage embryos, DNA extracted, and Sanger sequencing and PAGE performed. (**A)** An *in vivo* score was obtained from the Sanger sequencing using the ICE and TIDE tools, and both scores for a subset of six gRNAs compared to mosaicism percentages from Illumina sequencing of the same region. **(B)** From the PAGE, an empirical intensity ratio was obtained and compared to the *in vivo* efficiency scores. Correlations results are shown in the scatter plots.

Figure 2. Correlation of on-target efficiencies calculated using different methods. Scores from *in silico* prediction tools, an *in vitro* protocol), and cutting cleavages obtained *in vivo* for 50 gRNAs were compared using Spearman correlations.

Figure 3. SDs in the zebrafish genome. **(A)** SDs (orange, upper section) highlighted above every zebrafish chromosome (gray blocks). **(B)** Size of the SDs identified at each chromosome. Intra: intrachromosomal, Inter: interchromosomal. **(C)** Median distance between SDs observed and estimated from a random sampling empirical test.

796 **Figure 4.** Evaluation of Cas9 injection controls. (**A)** DE genes across samples injected with Cas9
797 enzyme or Cas9 mRNA relative to uninjected batch-siblings. Plot includes the number and
798 percentage of genes downregulated (FoldChange < -1) and upregulated (FoldChange > 1) in both
799 Cas9 treatments. (**B)** Protein-coding genes carrying frameshift variants for each Cas9-injected
800 sample type. (**C)** Biological pathways enriched in genes upregulated in both Cas9 treatments
801 using an enrichment test in *clusterProfiler (65)*. **(D)** Variants identified in larvae injected with
802 Cas9 enzyme, Cas9 mRNA, or uninjected controls and their frequency. Two representative
803 alleles showed for the seven genes with identified variants (see Supplementary Table 7 for
804 complete results).