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.

INTRODUCTION

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

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

often defined by information gathered across empirical assays (21). One relevant example is CRISPRScan, a predictive-scoring system built from experimental zebrafish gene-editing data based on multiple factors such as nucleotide GC and AT content and nucleosome positioning (8,22). Bioinformatic tools also exist that define potential regions prone to off-target edits mainly based on sequence similarity and the type/amount of mismatches relative to the on-target region (22). More recent empirical approaches couple *in vitro* cleavage of genomic DNA (gDNA) with Cas9 ribonucleoprotein (RNP), such as CIRCLE-seq (23) or GUIDE-seq (24), to provide a blind assessment of editing sites but may not necessarily reflect the *in vivo* activity of the CRISPR/Cas9 complex.

Previous studies have characterized CRISPR off-target activity *in vivo* to be relatively low in zebrafish. A cross-generational study identified no inflation of transmitted *de novo* single nucleotide mutations due to CRISPR-editing using exome sequencing and a stringent bioinformatic pipeline (25) in a similar approach used to identify off-target mutations in mouse trios (26,27). Other studies have observed off-target mutation rates ranging from 1.1 to 2.5% (28) and 0.16% to 3.17% (12) in zebrafish by sequencing the top three or four off-target regions based on sequence homology, respectively. Although off-target mutations should not significantly impact studies of stable mutants, as unwanted mutations can be outcrossed out of studied lines relatively easily (11,17), they may be problematic in rapid genetic screens using G_0 mosaics that quickly test gene functions in a single generation.

The increasing number of tools available for the design and execution of CRISPR screens provide an important resource to the zebrafish community. Here, we assayed different available

CRISPR on- and off-target prediction methods using empirical data from Cas9-edited zebrafish embryos. We quantified CRISPR cleavage efficiencies *in vivo* employing a variety of experimental approaches and used these results to compare the accuracy of *in silico* and *in vitro* tools for predicting Cas9 on- and/or off-target activity. Finally, we assayed the impact of injecting G_0 negative 'mock' control embryos injected with a buffer containing either Cas9 enzyme or mRNA in the absence of gRNAs by performing RNA-seq. We obtained a list of genes as significantly differentially expressed versus uninjected wild-type siblings and carrying mosaic frameshift variants in protein-coding regions possibly due to spurious mutations incurred from Cas9 cleavage events. In all, these results will serve as a useful resource to the research community as larger-scale CRISPR screens become more common in assaying gene functions in zebrafish.

RESULTS

Identification of CRISPR-induced indels in zebrafish

We generated a dataset containing experimentally confirmed indels in injected G_0 zebrafish larvae with 14 protein-coding genes targeted by 50 gRNAs (2–4 different gRNAs/target gene) (Supplementary Tables 1 and 2). To obtain experimental *in vivo* editing efficiency values for each gRNA, DNA extracted from a pool of G_0 mutants—generated via microinjections of individual gRNAs at the one-cell stage and harvested at five days post-fertilization (dpf)—was subjected to PCR (~500 bp region) and Sanger sequencing of the gRNA predicted target site. From this, we inferred an *in vivo* 'efficiency score' measured as the percentage of DNA from injected embryos harboring indels compared to uninjected batch siblings. The percentage of indels was extracted using two different tools that deconvolve major mutations and their

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

A quicker and more affordable approach to quantify CRISPR cleavage efficiency is via PAGE, which takes advantage of the heteroduplexes produced from DNA harboring a mosaic mix of

139

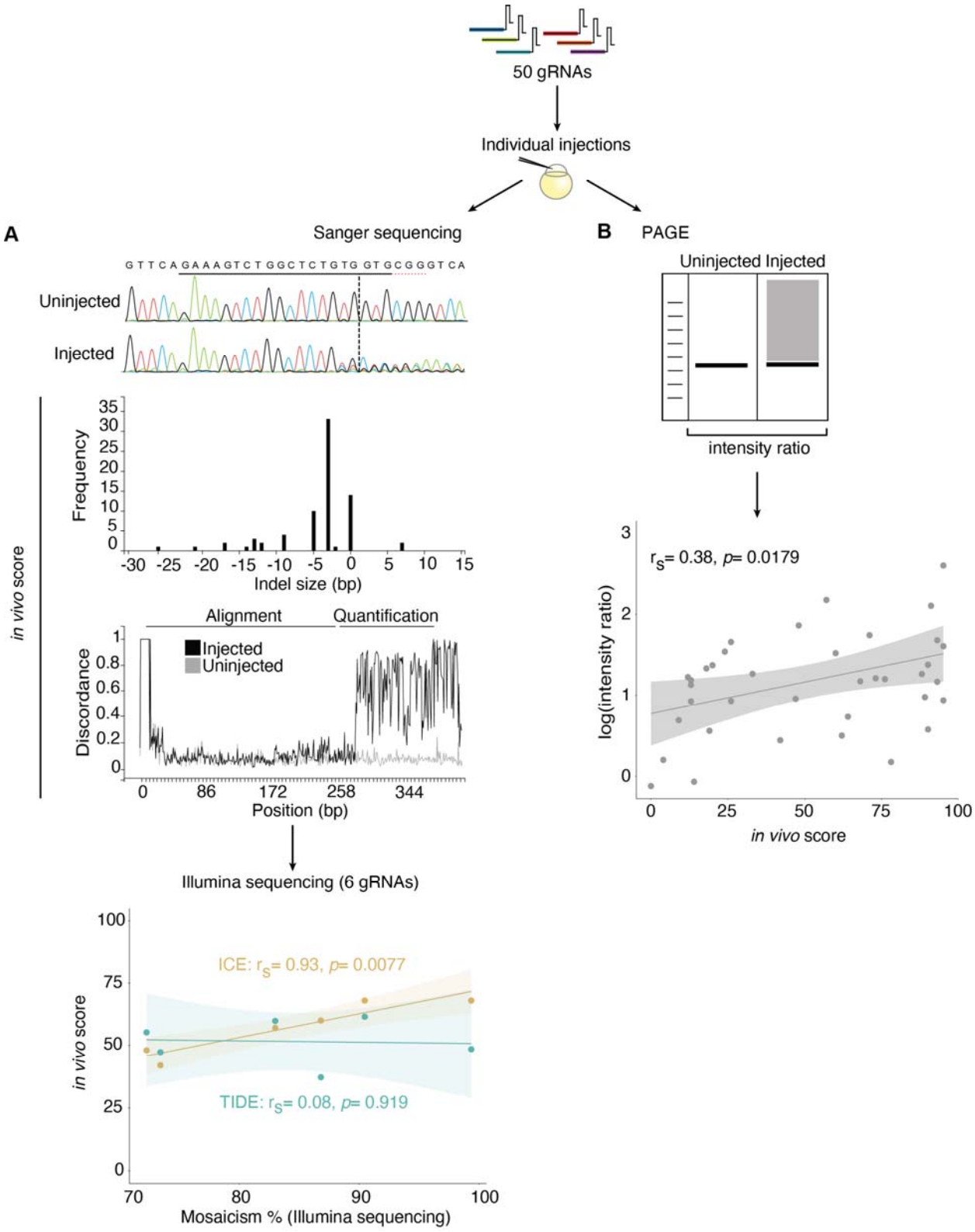


Figure 1. Workflow for the evaluation of CRISPR cleavages in zebrafish. 50 gRNAs were individually microinjected into one-cell stage embryos, DNA extracted, and Sanger sequencing and PAGE performed. (**A)** An *in vivo* score was obtained from the Sanger sequencing using the ICE and TIDE tools, and both 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.

predicted target site for each gRNA and quantified the PCR 'smear' intensity ratio of injected versus uninjected controls (see Methods). These intensity ratios were weakly correlated with our *in vivo* efficiency scores (Spearman r_s= 0.38, p= 0.0179; Figure 1B) indicating that accurate quantitative efficiencies cannot be directly deduced from PAGE but that the intensity of PCR 'smear' does qualitatively convey CRISPR-cleavage efficiency.

Accuracy of CRISPR on-target predictions by different methods

We next compared the accuracy of CRISPR on-target predictions computed by several published algorithms, including CRISPRScan (22), CHOPCHOP (32), E-CRISP (33), CRISPR-GE (34), CRISPR-RGEN (35), CCTop (36), as well as the design tool used for Integrated DNA Technologies (IDT, www.idtdna.com). Only five comparisons between tools exhibited significant correlations in their predicted scores and these relationships were weak (Spearman r_s <0.4; Figure 2). Thus, despite the research community broadly adapting all methods, there is little consensus in predicting activity of a particular gRNA among these tools. Only CRISPRScan, which we used to design our gRNAs, exhibited significant, albeit weak, correlation with *in vivo* experimentally determined values (Spearman r_s= 0.32, p= 0.03; Figure 2). Therefore, higher CRISPRScan scores predict higher *in vivo* efficiencies but with low accuracy.

Next, we evaluated the prediction power of an *in vitro* protocol, CIRCLE-seq (23), an approach designed to predict target sites of a given gRNA by subjecting naked genomic DNA to Cas9 enzyme/gRNA cleavage followed by Illumina sequencing. We tested individually the 50 gRNAs described above using CIRCLE-seq (37)and computed a log enrichment score normalized by the

library size, termed reads per million normalized (RPMN) (see Methods). We found that *in vitro* obtained enrichment scores were not correlated with *in vivo* efficiencies (Spearman r_s= -0.04, *p*= 0.81; Figure 2) or with *in silico* predictions (Figure 2), indicating that the CIRCLE-seq *in vitro* assay and *in silico* methods do not accurately predict on-target CRISPR cleavage activity, at least quantitatively. Previous work from *in vivo* CRISPR studies of zebrafish suggests that increased GC-content predicts increased activity of gRNAs (22). Examining our tested gRNAs, ranging in GC content from 31.8 to 77.3%, we observed a positive increase of CRISPRScan *in silico* scores (Linear model: beta= 68.2, *p*= 0.003, adjusted-r^2= 0.16) and CIRCLE-seq *in vitro* RPMN scores (Linear model: beta= 6.4, *p*= 0.006, adjusted-r^2= 0.14) with GC percentage (Supplementary Figures 2A,B). However, the experimentally determined *in vivo* scores were not correlated with GC content (Linear model: beta=12.4, *p*= 0.817, adjusted-r^2= -0.02; Supplementary Figure 2C). This result highlights that solely analyzing GC content cannot predict *in vivo* activity of a gRNA and more variables should also be considered (e.g., depletion of A nucleotide bases or nucleosome positioning (22)).

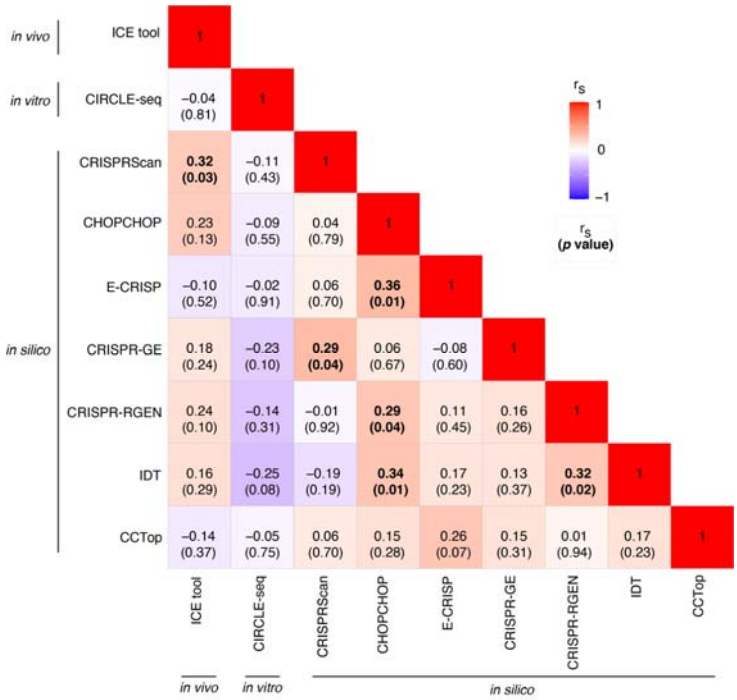


Figure 2. Correlation of on-target efficiencies calculated using different methods. Scores from *in silico* prediction tools, an *in vitro* protocol), and cutting cleavages obtained *in vivo* for 50 gRNAs were compared using Spearman correlations.

CRISPR off-target mutations prediction methods

To avoid spurious phenotypes, off-target mutations should be minimized when choosing gRNAs in CRISPR experiments. To characterize off-target mutations for our set of 50 gRNAs, we queried predictions from *in silico* (CRISPRScan) and *in vitro* (CIRCLE-seq) methods. CRISPRScan provided a list of the top 30 predicted off-target sites for each gRNA within the zebrafish reference genome (GRCz11/danRer11) based on a cutting frequency determination (CFD) score that primarily takes into account sequence similarity, location, and type of sequence mismatches (22,38). Alternatively, the CIRCLE-seq empirical approach produced variable numbers of sites (between 18 and 874 regions, median 142.6; Supplementary Table 3) per gRNA (defined as 'CIRCLE-seq sites') relative to the control library digested solely with Cas9. In general, none of the on-target scores previously obtained *in silico* or *in vitro* correlated with the number of CIRCLE-seq sites per gRNA (CRISPRScan score: Spearman r_s= -0.08, p= 0.559; RPMN: Spearman r_s= 0.09, p= 0.523). Moreover, *in vivo* scores also were not correlated with the number of CIRCLE-seq sites (Spearman r_s= 0.20, p= 0.204). To assess the concordance of *in silico* and *in vitro* off-target predictions, we intersected ±1 kbp regions flanking the predicted off-target sites and found an average overlap of 7.4±5.1% in regions between both prediction methods (overlap ranges from 0 to 26%).

Focusing on putatively impactful off-target predictions, an average of 19±12% CRISPRScan predicted and 60±15% CIRCLE-seq sites intersected at least one gene (Supplementary Table 3). To verify if predicted off-target sites were subject to *in vivo* Cas9 cleavage, we performed Sanger sequencing of predicted *in silico* (n=17) and *in vitro* (n=20) sites landing within genes for eight gRNAs with high *in vivo* efficiency scores (>90%), an average of six regions per gRNA (see

Supplementary Table 1 for description of sites). Using the ICE tool to quantify *in vivo* indel frequencies of predicted off-target sites, we found mosaic mutations at frequencies between 0 and 11%, with 23 out of the 37 sites evidencing indel frequencies below 1% (Supplementary Figure 3A). To validate the accuracy of ICE at these low indel frequencies, we again performed Illumina sequencing of predicted off-target sites for six of the eight evaluated gRNAs (see Supplementary Table 1 for description) and found significant concordance in results (0.29 - 7.62%; Spearman r_s= 0.83, p= 0.039; Supplementary Figure 3B). The average difference in mosaicism between ICE and Illumina was low (1.6±2.0) highlighting that ICE is more accurate at sites with lower indel frequencies and suitable to quickly and economically assess predicted off-targets regions.

No difference in indel frequencies was observed between sites predicted by CRISPRScan or CIRCLE-seq (Mann-Whitney U= 175.5, p= 0.873; Supplementary Figure 3C). Additionally, no significant correlation existed between the *in vivo* efficiency score of the gRNAs and the frequency of indels in their predicted off-targets (Spearman r_s= 0.27, p= 0.111), suggesting that higher on-target efficiencies do not necessarily translate into increased frequencies of off-target mutations at the top predicted sites (Supplementary Figure 3D). Moreover, we assessed if sites predicted with higher likelihoods of off-target cutting events resulted in higher mutation rates by comparing the indel frequencies among the different levels of prediction (top 1, 2, or 3 prediction scores by CRISPRScan or CIRCLE-seq). No difference was found between prediction groups, suggesting that the information used by the tools to assign probabilities of off-target activity (e.g. CFD scores in CRISPRScan or normalized read counts in CIRCLE-seq) do not necessarily predict the efficiency of cutting at off-target sites (Kruskal-Wallis: $H_{(2)}$= 2.26, p= 0.320;

Supplementary Figure 3E). Thus, off-target cutting mutations at the assessed sites exhibited low frequencies with no clear method performing best.

Landscape of segmental duplications (SDs) in the zebrafish genome

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

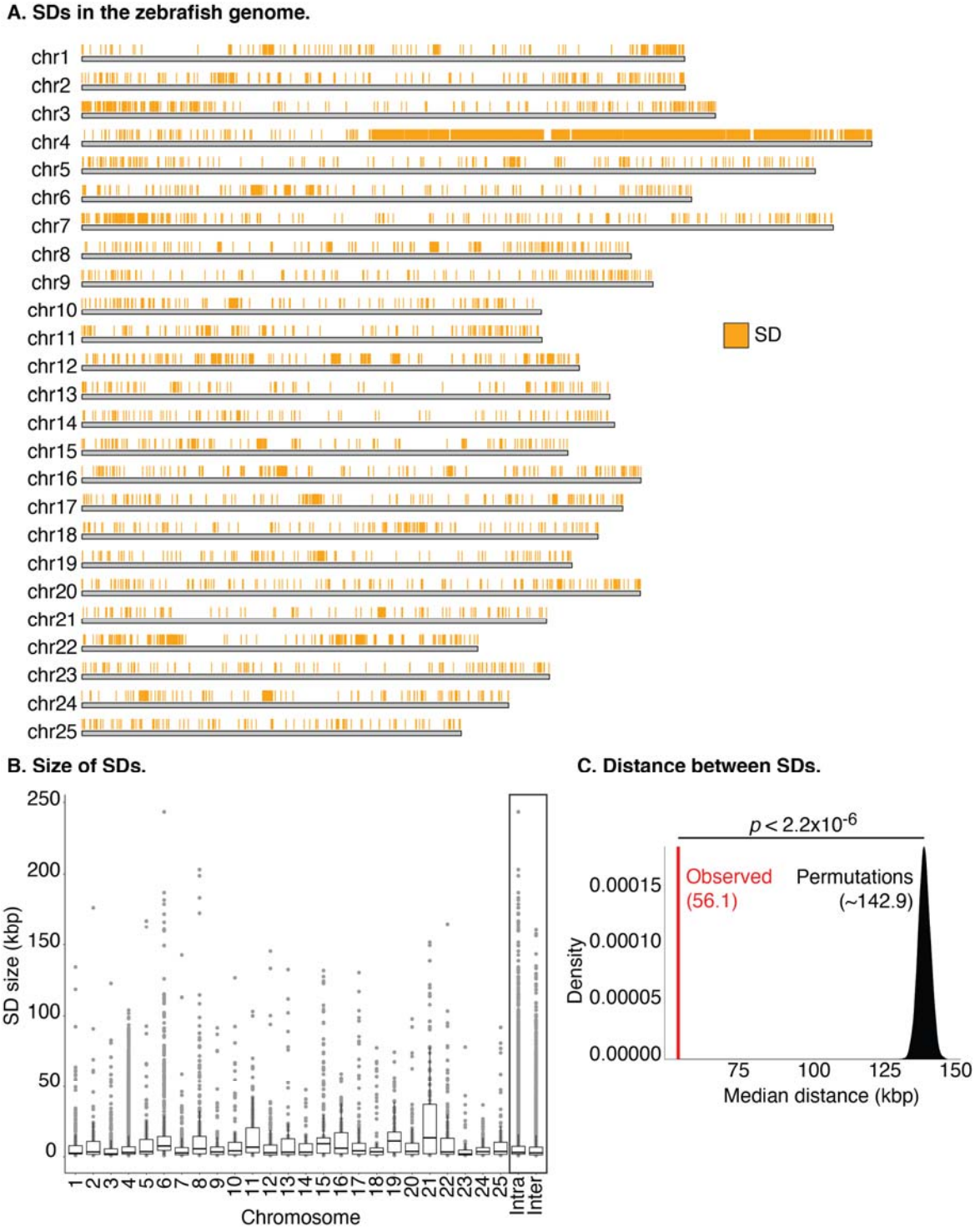


Figure 3. SDs in the zebrafish genome. (A) SDs (orange, upper section) highlighted above every zebrafish chromosome (gray blocks). **(B)** Size of the SDs identified at each chromosome. Intra: intrachromosomal, Inter: interchromosomal. **(C)** Median distance between SDs observed and estimated from a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

253

Evaluating CRISPR Cas9-injection controls

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

Recently, Sundaresan and colleagues found that Cas9 in the presence of Mn^{+2} ions can result in double-strand cleavage of genomic DNA in the absence of gRNA (44). Although their study did

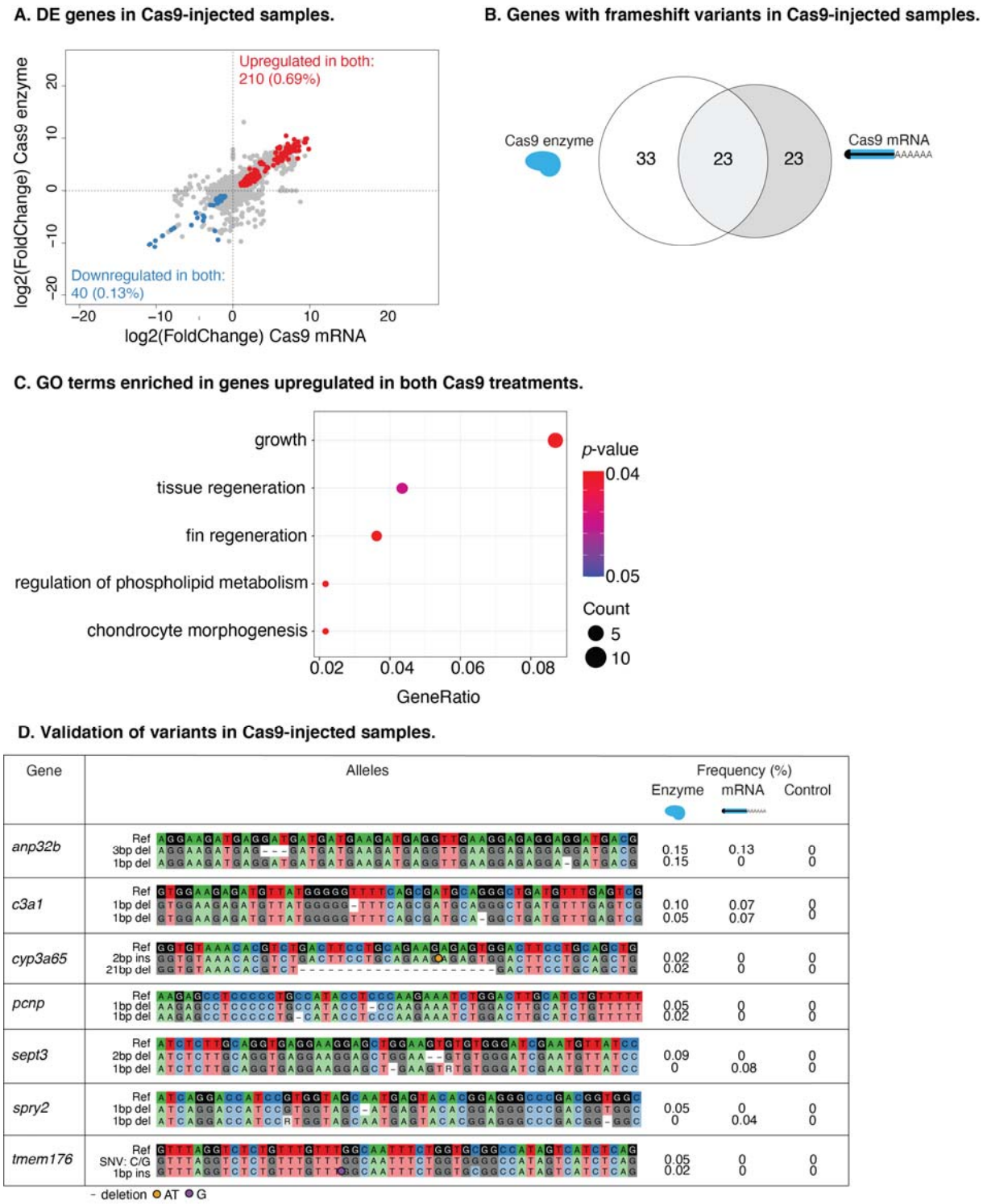


Figure 4. Evaluation of Cas9 injection controls. (**A)** DE genes across samples injected with Cas9 enzyme or Cas9 mRNA relative to uninjected batch-siblings. Plot includes the number and percentage of genes downregulated (FoldChange < -1) and upregulated (FoldChange > 1) in both Cas9 treatments. (**B)** 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).

not show this same off-target cleavage activity in the presence of Mg^{+2}, we hypothesized that some DE genes observed in our Cas9-injected controls could be the result of genomic mutations incurred by aberrant cutting of the Cas9 rather than a result of the injection process alone, due to the presence of $MgCl_2$ in the injection buffer. Using our RNA-seq data, we used an optimized pipeline (45) to identify somatic mosaic mutations with uninjected wild-type controls as a reference for common polymorphisms. Focusing only on high-confidence variants (minimum sequence read depth of 20), we filtered known variants in the NHGRI-1 zebrafish line (46), removed variants found in the SDs previously identified, and used the Variant Effect Predictor tool from ENSEMBL to obtain a list of frameshift mutations in protein-coding genes present in our Cas9-injected larvae. A total of 56 and 46 genes were identified with frameshifting variants in larvae injected with Cas9-enzyme and Cas9-mRNA, respectively, with 23 of these genes shared across both injection types (Figure 4B, Supplementary Table 6). Low allelic frequencies (Cas9-enzyme: average 0.053, range 0.003–0.022; Cas9-mRNA: average 0.072, range 0.002–0.357) agreed with what we expected for mosaic mutations as a result of Cas9 gRNA-independent cutting events (Supplementary Table 6). The observed read depth across all variants provided high confidence in their identification (Cas9-enzyme: average 307x, range 22x–2017x; Cas9-mRNA: 284x, range 20x–3453x) (Supplementary Table 6).

Most commonly we discovered only one variant per gene but did find a number of genes with multiple variants (see Supplementary Table 6 for complete description of genes and alleles), including *atp1a1a.1* with ten different alleles, *fga* with seven different alleles, and *anp32b* and *slc25a4* with four different alleles each, highlighting the possibility that certain genes may be consistently targeted by Cas9 in the absence of a gRNA. Variants identified were located closer

to a potential Cas9 PAM site (NGG) than by random chance (6-bp median distance to closest PAM site; empirical $p = 0.0016$ using the whole-genome and $p = 0.006$ using protein-coding regions only, from 10,000 permutations). To test if genes discovered as mutated via RNA-seq data contain genomic indels, we performed targeted Illumina sequencing of 12 genes in a separate set of Cas9-injection controls (enzyme or mRNA). We observed low frequency (<1%) mosaic variants in 7 out of the 12 analyzed genes (Figure 4D, Supplementary Table 7), but collectively observed mosaicism across 4.1±1.1% and 3.7±0.6% for Cas9 enzyme and Cas9 mRNA injected, respectively, with no difference between injection type (Mann-Whitney U= 31.5, $p = 0.402$) (Figure 4D, Supplementary Table 7). Considering the many additional genes we identified as mutated in our RNA-seq data that remain to be validated, this suggests that the collective combined effects of low frequency mutations at individual sites could result in spurious phenotypes and impact subsequent characterizations.

DISCUSSION

Our study presents a comprehensive evaluation of empirical and predictive tools currently used for CRISPR editing in zebrafish. Cleavage scores obtained by an *in vivo* assessment of 50 gRNAs via Sanger sequencing and deconvolution tools (ICE and TIDE) were highly congruent, as previously reported (30), but only ICE scores were correlated with mosaicism percentages from Illumina sequencing, the commonly used gold standard. ICE scores tended to underestimate the presence of non-edited alleles by ~20% for gRNAs with high efficiencies (>50%), contrary to previous comparisons of TIDE and Illumina sequencing in cell lines, where TIDE showed a ~10–20% overestimation of non-edited alleles (47). For sites with lower indel frequencies, as we observed for predicted off-target mutations, ICE scores were concordant with Illumina results

(~1–2% difference). Therefore, we suggest that Sanger sequencing deconvolution tools are valuable for establishing relative gRNAs efficiencies but do not necessarily accurately predict absolute cleavage efficiencies in zebrafish *in vivo*, except at sites with low indel frequencies. In addition, we show that empirical 'smear' scores from PAGE can approximate cleavage efficiencies, making it a more affordable and rapid approach to assay editing efficiency versus sequencing.

On-target efficiency prediction tools showed large differences using the same set of gRNAs sequences, highlighting the importance of understanding the features accounted for by each tool. A recent review (21) provides a comprehensive overview of different design tools available and the source of experimental data used to train each one. Compared with our *in vivo* efficiency scores from zebrafish embryos, CRISPRScan was the only tool that could (somewhat) accurately predict on-target efficiency in our set of gRNAs, while no other method provided scores that were correlated with cleavage activities observed *in vivo*. Therefore, our results emphasize the importance of utilizing a tool that has been trained using experimental data specifically from zebrafish.

The *in silico* and *in vitro* off-target prediction tools sites predicted an overlapping set of genes (>20% for our 50 gRNAs). Notably, we did not evaluate if these predicted sites included *cis*-regulatory elements that could also potentially alter gene expression. Future assessments should include tests targeting a diversity of loci for a more thorough understanding of the potential off-target indels caused by unwanted CRISPR cleavage sites. We observed low off-target mutation frequencies (most <1%), similar to those previously reported from using single (28) or multiple

(12) gRNAs. Notably, neither predictive method (CRISPRScan or CIRCLE-seq) nor intensity of the prediction (using CFD or normalized read count) could accurately predict indel frequencies at off-target sites. Typically, such low mutation frequencies should not be of high impact when generating stable knockout zebrafish lines as these could be easily outcrossed. However, they could potentially affect the phenotypic outcome when the injected G_0 mosaic populations were analyzed directly.

The percentage of SDs observed in the zebrafish genome (~5.5%) is similar to reports of ~5% observed in mammals, including humans (48-50). The distribution of these SDs across the zebrafish genome is consistent with previous reports in mammalian genomes, where SDs tend to be clustered (51-53), except for humans where these duplications are mostly interspersed (54). Notably, it is unclear whether values observed in zebrafish are consistent across other teleost species since few reports of SD quantification in other organisms are available. The higher density of SDs present on chromosome 4, especially on its long arm, is consistent with previous findings of extensive heterochromatic and protein-depleted regions found along this chromosome (39). Importantly, we also observed a high percentage of SDs on chromosome 22 (6.70%), making it a candidate region for subsequent analyses. All of these resources will aid future CRISPR experiments targeting genes located in these duplicated regions for potential off-target events in paralogous sequences.

Our data also showed that in the absence of gRNA, Cas9 alone may edit genomic DNA *in vivo*. This could be problematic due to the difficulty in predicting where the cleavage sites could be and the high impact these could have on functional assays. Studies *in vitro* have shown that

presence of Mn^{+2} ions can impact cleavage activity of Cas9 and Cas12a without a gRNA present (44). Our injection buffer includes a functionally similar ion, Mg^{+2}, which has been shown to compete with Mn^{+2} and activate common enzymes (55). Therefore, we investigated the effect of Cas9 in genomic DNA of larvae solely injected with the nuclease, either as an enzyme or mRNA. Our RNA-seq assay revealed that a set of genes involved in regeneration pathways are differentially upregulated in the larvae injected with Cas9, highlighting the possibility that a subset of gene expression can be perturbed simply due to the presence of Cas9 being present in the developing embryo. Furthermore, we found an average of ~50 genes carrying frameshift mutations per larvae injected with Cas9, with mutations validated in 7 of 12 genes through a subsequent injection of Cas9 in a different batch of zebrafish embryos. A recent report showing RNA targeting using the CRISPR/Cas9 system (56) suggests that we also cannot rule out the possibility of Cas9 cutting mRNAs in our samples in the absence of a gRNA, which could influence the presence of DE genes. The common genes carrying mutations across biological replicates and the high confidence of the identified variants (mean read depth ~300x) highlight the importance of future studies to assess the possibility of RNA-independent targeting by Cas9 in zebrafish and other organisms. In all, our results provide a list of genes that could be edited by Cas9 alone in zebrafish embryos. Precaution should thus be taken when using Cas9 (no gRNA)-injected embryos as a baseline control for comparing phenotypes, especially when performing studies in the injected G_0 mosaic population directly.

Overall, our study is the first to perform a simultaneous assessment of gRNA activities predicted by several commonly used *in silico* and *in vitro* methods with those determined experimentally *in vivo* in injected zebrafish embryos. These results provide valuable information that can be

incorporated into the design and execution of CRISPR/Cas9 assays in zebrafish using available workflows (10,11). Our aim was to provide information to aid in the decision-making process for future projects using affordable and reliable gene-editing tools in zebrafish. As higher-throughput methods are being developed for assaying multiple genes simultaneously, it will be important to use optimal tools for predicting and assessing on- and off-target activity in zebrafish larvae for accurate interpretation of phenotypic outcomes.

METHODS

Zebrafish husbandry

NHGRI-1 wild type zebrafish lines (46) were maintained through standard protocols (57) and naturally spawned to obtain embryos. All animal use was approved by the Institutional Animal Care and Use Committee from the Office of Animal Welfare Assurance, University of California, Davis.

Design and *in silico* predictions for gRNAs

50 gRNAs targeting exons of 14 genes were designed using CRISPRScan (22) (scores ranging between 24 and 83 with a mean value of 57.60) with zebrafish genome version GRCz11/danRer11 as the reference (see description of gRNAs in Supplementary Tables 1 and 2). All targeted genes were protein coding. For each designed gRNA, we obtained the efficiency scores predicted by CRISPRScan (22), CHOPCHOP (32), E-CRISP (33), CRISPR-GE (34), CRISPR-RGEN (35), CCTop (36), and the IDT design tool (www.idtdna.com). From CRISPRScan, we also gathered the top 30 predicted off-target sites for each gRNA defined by

the CFD score (22). Additionally, we utilized *bedtools* (58) to determine the GC percentage for each gRNA.

Microinjections to generate CRISPR G_0 mosaic mutants

All gRNAs were individually injected into NHGRI-1 embryos to estimate the frequency of indels. gRNAs were prepared following manufacturer's protocol (Integrated DNA Technologies). Briefly, 2.5 µl of 100 µM crRNA, 2.5 µl of 100 µM tracrRNA, and 5 µl of Nuclease-free Duplex Buffer using an annealing program consisting of 5 min at 95°C, a ramp 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 4°C with a -1°C/s change. Ribonucleic protein (RNP) injection mix was prepared with 1.30 µl of Cas9 enzyme (20 µM, New England BioLabs), 1.60 µl of prepared gRNAs, 2.5 µl of 4x Injection Buffer (containing 0.2% phenol red, 800 mM KCl, 4 mM $MgCl_2$, 4 mM TCEP, 120 mM HEPES, pH 7.0), and 4.6 µl of Nuclease-free water. Microinjections directly into the yolk of NHGRI-1 embryos at the one-cell stage were performed as described previously (59), using needles from a micropipette puller (Model P-97, Sutter Instruments) and an air injector (Pneumatic MPPI-2 Pressure Injector). Embryos were collected and ~1 nl of RNP-mix was injected per embryo, after previous calibration with a microruler. Twenty injected embryos per Petri dish were grown up to 5 dpf at 28°C.

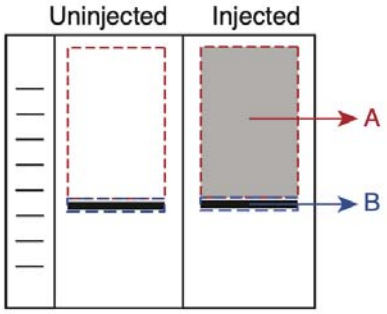
Sanger and Illumina amplicon sequencing

DNA extractions were performed on 20 pooled embryos by adding 100 µl of 50 mM NaOH, incubation at 95°C for 20 min, ramp from 95°C to 4°C at a 0.7°C/s decrease, followed by an addition of 10 µl of 1 M Tris-HCl and a 15 min spin at 4680 rpm. We amplified a ~500 bp region

surrounding the targeted site of each gRNA (see Supplementary Table 1 for description of primers). PCR amplifications were performed using 12.5 µl of 2X DreamTaq Green PCR Master Mix (Thermo Fisher), 9.5 µl of Nuclease-Free water, 1 µl of 10 µM primers, and 1 µl extracted DNA. Thermocycler program included 3 min at 95°C, followed by 35 cycles of 15 s at 95°C, 30 s at 60°C, and 20 s at 72°C, and a final 5 min incubation at 72°C. Reactions were purified using Ampure XP magnetic beads (Beckman Coulter) and Sanger sequenced (Genewiz, San Diego, CA). Raw ab1 files were used in the TIDE (29) and ICE (30) tools to predict the percentage of indels, which we used as our *in vivo* editing score for each gRNA. To compare accuracy between these tools, we PCR amplified ~200 bp around the targeted regions from the same extracted DNA for six gRNAs and performed Illumina sequencing (Genewiz, San Diego, CA) to obtain percent mosaicism of mutants by mapping paired-end fastq reads to the zebrafish reference genome (GRCz11/danRer11) using *bwa* (60) and the R package *CrispRVariants* (31). For both Sanger and Illumina sequencing, we used uninjected batch-sibling embryos as a control reference.

PAGE and intensity-ratio estimation

An empirical cleavage analysis from each gRNA was performed using PAGE. Briefly, we amplified a ~200 bp region in DNA around the targeted site from gRNA-injected and uninjected embryos, as described above. Reactions were run on 7.5% polyacrylamide gels for 75 min at 110 V and revealed using GelRed (VWR International). Gel images were processed in the software Fiji (61). For each sample, we defined areas **A** and **B** as follows:



For each gRNA, the mean-intensity value was obtained for the A and B areas in both the injected and uninjected samples. The A and B areas were exactly the same size between samples. The intensity ratio was calculated as: [injected B / injected A] / [uninjected B / uninjected A]. Log-normalized intensity ratios followed a normal distribution (Shapiro-Wilk test: W= 0.96, *p*= 0.167) with an average value of 1.21±0.70.

CIRCLE-seq

CIRCLE-seq libraries were prepared for each gRNA (IDT) using gDNA extracted from NHGRI-1 (DNA Blood & Tissue kit, Qiagen) following the described protocol (37). Libraries were sequenced using one HiSeq XTen lane (Novogene, Sacramento, CA), providing an average of 7.3 million reads (range: 4.0 - 13.3 million reads) and >Q30 for 92% of reads per gRNA library. Raw reads were processed using the bioinformatic pipeline described (37) to identify regions with cutting events relative to a control sample (treated with Cas9 enzyme and no gRNA). To obtain an on-target efficiency estimation from *in vitro* digestions, we calculated the reads per million normalized (RPMN). For this purpose, we used *samtools* (60) to extract read coverage from aligned bam files. For each gRNA, coverage was obtained for the third and fourth base upstream of the PAM site as it is the region expected to be cut by Cas9 (62). RPMN for each gRNA was calculated as the sum of coverage at these two sites divided by the total mapped reads

per sample and multiplied by one million to scale the values. RPMN scores ranged from 4.42 to 881 (median 99.3) so we decided to use a log normalization to reduce this range.

SD discovery

SDs in the zebrafish genome were identified using *sedef* (43) with a soft-masked reference genome (GRCz11/danRer11). Filtering of these identified regions included a minimum alignment size of 1 kbp and sequence similarity >90%. The median observed distance between SDs was compared to the result of 10000 permutations of a random sampling across the genome. In brief, sampling included the extraction of regions matching in number and size to the identified SDs and the estimation of the median distance between them. One-tailed empirical *p* values from this comparison were calculated as (M+N)/(N+1), where M is the number of iterations with a median distance below the observed value and N is the total number of iterations.

RNA-seq

We performed RNA-seq of Cas9 injected NHGRI-1 larvae to identify potential gRNA-independent cleavage sites. One-cell stage NHGRI-1 embryos were injected with either Cas9 enzyme or Cas9 mRNA. Injection mix for Cas9 enzyme included Cas9 enzyme (20 μM, New England BioLabs), 2.5 μl of 4x Injection Buffer (0.2% phenol red, 800 mM KCl, 4 mM $MgCl_2$, 4 mM TCEP, 120 mM HEPES, pH 7.0), and Nuclease-free water. Cas9 mRNA was obtained from plasmid pT3TS-nCas9n (Addgene, plasmid #46757) (4), using the MEGAshotscript T3 transcription kit (Thermo Fisher) following manufacturer's guidelines of 3.5 h 56°C incubation with T3. mRNA was purified with the MEGAclear transcription clean-up kit (Thermo Fisher)

and concentration of mRNA obtained using a NanoDrop (Thermo Fisher). The injection mix of Cas9 mRNA contained 100 ng/µl of mRNA, 4x Injection Buffer (0.2% phenol red, 800 mM KCl, 4 mM $MgCl_2$, 4 mM TCEP, 120 mM HEPES, pH 7.0), and Nuclease-free water. Additionally, uninjected batch-siblings were used as controls. All embryos were grown at 28°C in a density of <50 embryos per dish. At 5 dpf, three pools of five larvae were collected for each group (Cas9 enzyme, Cas9 mRNA, and uninjected) for RNA extraction using the RNeasy kit (Qiagen) with gDNA eliminator columns for DNA removal. Whole RNA samples were subjected to RNA-seq using the poly-A selection method (Genewiz, San Diego, CA). Raw reads were processed using the *elvers* (https://github.com/dib-lab/elvers; version 0.1, release DOI: 10.5281/zenodo.3345045) bioinformatic pipeline that utilizes *fastqc*, *trimmomatic* (63), and *salmon* (64) to obtain the transcripts per kilobase million (TPM) for each gene. *DESeq2* (65) was used to extract differentially expressed genes in the Cas9 enzyme or Cas9 mRNA injected samples relative to the uninjected batch siblings. R package *cluterProfiler* (66) was used to perform enrichment tests of differentially expressed genes in biological pathways.

Variant identification from RNA-seq data

We followed a previously described pipeline to identify somatic variants from RNA-seq data (45). Briefly, we mapped reads with *STAR* (67) using the 2-pass mode and a genomic reference created with GRCz11/danRer11 assembly and gtf files (release version 100). Variant calling was performed with MuTect2 as part of *GATK* (68) using the tumor versus normal mode. 'Normal' was defined by the two uninjected samples to identify all somatic mutations in our Cas9 injected embryos. Variants inside SDs were removed. Variants were annotated using the Variant Effect Predictor tool from ENSEMBL. High confidence variants (minimum sequencing depth of 20)

previously reported for the NHGRI-1 line (46) were removed. Only frameshift loss-of-function variants with a minimum read depth of 20 in canonical protein-coding genes were considered. We extracted the median distance between the identified variants and the nearest Cas9 PAM site (NGG sequence) using the coordinates in the CRISPRScan UCSC track. This median observed distance was compared to the result of median distances of 10,000 permutations of random sampling across the genome and their nearest PAM site. One-tailed empirical *p* values from this comparison were calculated as (M+N)/(N+1), where M is the number of iterations with a median distance below the observed value and N is the total number of iterations. We orthogonally performed validations of variants in 12 genes via Illumina sequencing of a ~200 bp region surrounding the identified variant location and the R package *CrispRVariants* (31) (Supplementary Table 1 for primers description). For this purpose, we extracted DNA from a pool of 20 embryos of each group (Cas9 enzyme, Cas9 mRNA, and uninjected). The 12 genes assessed included three genes with most identified variants (*atp1a1a.1*, *fga*, and *anp32b*), four DE genes carrying variants (*cyp3a65*, *galnt7*, *c3a.1*, *sept3*), two genes with variants found in both Cas9 treatments (*actb1* and *slc25a4*), two genes with variants in all Cas9 enzyme injected larvae (*spry2* and *pcnp*), and a gene with variants in all Cas9 mRNA injected samples (*tmem176*) (see Supplementary Table 1 for description).

Statistical analyses

All analyses were performed in R version 4.0.2 (69) . Normality of variables was checked using the Shapiro-Wilk test and parametric or nonparametric comparisons made accordingly. Correlation tests (Pearson denoted as r and Spearman denoted as r_s) and linear regression models were used to determine the relationship between variables. All analyses compared across

different experimental batches included *batch* as a factor in the model to prevent biases caused by inter-batch differences. Averages include the standard deviation unless otherwise specified. Alpha to determine significance across the different tests was set at 0.05 unless otherwise specified. Additional R packages used for making figures included *clusterProfiler* (66), *karyoploteR* (70), *ggcorrplot* (71), and *eulerr* (72).

ACKNOWLEDGEMENTS

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

DECLARATIONS

Abbreviations

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

Financial disclosure

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

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

Competing interests' statement

The authors declare that they have no competing interests.

Ethics statement

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

Data availability statement

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

71. Kassambara A. ggcorrplot: Visualization of a Correlation Matrix using'ggplot2'. R package

version 0 1. 2016;1.

72. 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.

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).