Targeted mutagenesis of specific genomic DNA sequences in animals for the *in vivo* generation of variant libraries.

Julia Falo-Sanjuan[1], Yuliana Diaz-Tirado[1], Meghan A. Turner[7,&], Olivia Rourke[1], Julian Davis[1], Claudia Medrano[1], Jenna Haines[1], Joey McKenna[1], Arman Karshenas[7], Michael B. Eisen[1,2,3,+], Hernan G. Garcia[1,4,5,6,7,+]

[1]Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA

[2]Department of Integrative Biology, University of California, Berkeley, CA, USA

[3]Howard Hughes Medical Institute, University of California, Berkeley, CA, USA

[4]Department of Physics, University of California, Berkeley, CA, USA

[5]Institute for Quantitative Biosciences-QB3, University of California, Berkeley, CA, USA

[6]Chan Zuckerberg Biohub – San Francisco, San Francisco, CA, USA

[7]Biophysics Graduate Group, University of California, Berkeley, CA, USA

[+]Corresponding authors: hggarcia@berkeley.edu, mbeisen@berkeley.edu

[&]Current affiliation: Allen Institute for Brain Science, Seattle, WA, USA

Abstract

Understanding how the number, placement and affinity of transcription factor binding sites dictates gene regulatory programs remains a major unsolved challenge in biology, particularly in the context of multicellular organisms. To uncover these rules, it is first necessary to find the binding sites within a regulatory region with high precision, and then to systematically modulate this binding site arrangement while simultaneously measuring the effect of this modulation on output gene expression. Massively parallel reporter assays (MPRAs), where the gene expression stemming from 10,000s of in vitro-generated regulatory sequences is measured, have made this feat possible in high-throughput in single cells in culture. However, because of lack of technologies to incorporate DNA libraries, MPRAs are limited in whole organisms. To enable MPRAs in multicellular organisms, we generated tools to create a high degree of mutagenesis in specific genomic loci *in vivo* using base editing. Targeting GFP integrated in the genome of *Drosophila* cell culture and whole animals as a case study, we show that the base editor AID^{evoCDA1} stemming from sea lamprey fused to nCas9 is highly mutagenic. Surprisingly, longer gRNAs increase mutation efficiency and expand the mutating window, which can allow the introduction of mutations in previously untargetable sequences. Finally, we demonstrate arrays of >20 gRNAs that can efficiently introduce mutations along a 200bp sequence, making it a promising tool to test enhancer function *in vivo* in a high throughput manner.

Introduction

The orchestrated regulation of gene expression—in time, space or in response to signals—is required for most aspects of life such as ensuring the correct development of body plans, maintaining homeostasis and regeneration (Howard and Davidson, 2004; Levine, 2010; Rubinstein and de Souza, 2013). Misregulation of gene expression has been increasingly linked to disease, even in cancers where mutated coding sequences were thought to be the main drivers (Cuykendall et al., 2017; Rheinbay et al., 2020). Among the many layers to gene regulation, short regulatory sequences termed enhancers play a major role in determining the levels and spatiotemporal pattern of the expression of a gene. Enhancers are bound by transcription factors (TFs) in 6-10 bp DNA motifs termed transcription factor binding sites. Differences in the TF binding site sequence, number, orientation and spacing, amongst other features, are thought to relate to TF binding strength and dynamics and, in turn, to recruitment of the transcriptional machinery (Barolo and Posakony, 2002; Lelli et al., 2012; Spitz and Furlong, 2012).

Despite the clear role of enhancers in driving gene expression programs, the specific rules by which the number, placement and affinity of binding sites within them dictates transcriptional dynamics remain unknown, particularly in the context of multicellular organisms. To uncover these rules it is therefore necessary to find the binding sites within an enhancer with high precision, and then to systematically modulate this binding site arrangement while simultaneously measuring the effect of this modulation on output gene expression. Massively Parallel Reporter Assays (MPRAs) have made it possible to perform this feat in the context of cells in culture and in some multicellular organisms (Arnold et al., 2013; Brown et al., 2022; Chan et al., 2023; Inoue and Ahituv, 2015; Ireland et al., 2020; Jores et al., 2020; Kheradpour et al., 2013; Lagunas et al., 2023; Lalanne et al., 2024; Qi et al., 2020). Here, a barcoded library of reporter genes, each driven by a randomly mutated regulatory region or genome fragments, is generated *in vitro* and transfected into cells. Next, RNA-Seq is used to correlate enhancer sequence and gene expression. These correlations yield an information footprint that reveals activator and repressor binding sites (Ireland et al., 2020). Further, by computationally designing these libraries, regulatory parameters such as the relative placement of binding sites can be systematically tested (de Almeida et al., 2022; Kircher et al., 2019; Qi et al., 2020; Sharon et al., 2012; Yu et al., 2021). The result has been a pipeline that allows for the rapid identification and experimental validation of binding sites within an enhancer, and for the engagement in a theory-experiment dialogue aimed at reaching a predictive understanding of how binding site architecture dictates gene expression.

Despite the great promise of MPRA approaches, for the most part, this tool can only be implemented in the context of cells in culture. Specifically, because MPRAs are based on the incorporation of *in vitro*-generated DNA libraries, this approach can only be deployed in systems where such transfection or viral infection is possible. Indeed, the incorporation of a large number of DNA variants is not possible in most multicellular organisms. As a result, with some notable exceptions in adult mouse brains, *C. elegans*, *Ciona intestinalis* and tobacco leaves (Brown et al., 2022; Chan et al., 2023; Farley et al., 2015; Jores et al., 2020; Lagunas et al., 2023; Stevenson et al., 2023), the field lacks a reliable pipeline to incorporate DNA diversity into animals and plants, and to deploy MPRAs to uncover the rules that dictate the development and physiology of multicellular organisms.

A second limitation of MPRAs—which applies to both cells in culture and multicellular organisms—is their reliance on reporter constructs: enhancers within the library are typically integrated into a plasmid that allows for the simultaneous measurement of enhancer sequence and reporter gene expression level (Arnold et al., 2013; Brown et al., 2022, 2022; Chan et al., 2023, 2023; Farley et al., 2015; Inoue and Ahituv, 2015; Ireland et al., 2020; Jores et al., 2020, 2020; Kheradpour et al., 2013; Lagunas et al., 2023, 2023; Lalanne et al., 2024; Qi et al., 2020). In most cases these reporter constructs remain episomal and lack chromatin such that they cannot capture information about the genomic context that might be at play such as histone modification landscape.

In this paper, we present new technology that makes significant progress towards circumventing the limitations that have held back MPRAs from being implemented in most multicellular organisms, and in the endogenous genomic context of enhancers. As an alternative to generating enhancer variability *in vitro*, we developed an approach for *in vivo* mutagenesis of specific genomic DNA sequences using the fruit fly *Drosophila melanogaster* as a case study. Specifically, as a means to test the activity of thousands of enhancer variants in multicellular organisms, we developed a system that can create mutations randomly *in vivo* in a region of interest. Here, mutator flies generate libraries of enhancer variants—Flybraries—in their germline cells. Each embryo laid by these mothers then has a unique random realization of the enhancer whose transcriptional activity can be assayed and correlated with the enhancer sequence.

To enable *in vivo* mutagenesis in the genome, we assayed multiple DNA editing tools such as base editors as well as error-prone DNA polymerases. Most notably, we deployed a series of deaminases derived from the somatic hypermutation enzyme AID fused to a nickase Cas9 (nCas9) (Doll et al., 2023; Kohli et al., 2021; Komor et al., 2016; Thuronyi et al., 2019). This enzyme produces cytosine to thymine (C→T) transitions by deamination of C to uracil (U). Through multiple rounds of optimization in *Drosophila* animals and cell culture, combined with mathematical modeling of the mutation process, we show that somatic hypermutation enzyme AID stemming from sea lamprey—and not the widely used version originating from humans, mouse and rats (Anzalone et al., 2020; Doll et al., 2023; Thuronyi et al., 2019)—fused to nickase Cas9 (nCase9) can lead to significant mutagenesis in the fly genome. Further, we optimize the spatial window of mutagenesis of the AID-nCas9 fusion by optimizing the length, number and overlap of guide RNAs (gRNA) that mediate the targeting of this fusion. Thus, we show the feasibility of our *in vivo* genomic mutagenesis approach based on targeting AID-nCas9 to a DNA locus with multiple gRNAs, which constitutes a first step towards realizing MPRAs in endogenous genomic loci of cultures cells as well as multicellular organisms.

Results

A well-established system to generate mutations based on AID-nCas9 performs well in *Drosophila* cell culture but not *in vivo*

In the past few years, multiple tools have been developed with the goal of *in vivo* targeted mutagenesis, such as prime and base editing, or EvolvR. Prime editing requires the incorporation of a template with the desired mutation to be introduced (Bosch et al., 2020). Thus, as a result of lack of technology to incorporate libraries into most multicellular organisms, this approach cannot be implemented in animals and plants in high-throughput. Deaminases used in base editing produce cytosine to thymine (C→T) or adenine to guanine (A→G) transitions by deamination of C to uracil (U) or A to inositol (I) respectively, resulting in a mutational spectrum of C→T and G→A for C deaminases and A→G and T→C for A deaminases. C and A deaminases have been engineered and selected for high efficiency and precision, with the goal of reverting disease-causing mutations when targeted with nCas9 (Anzalone et al., 2020; Fu et al., 2021; Koblan et al., 2021; Musunuru et al., 2021; Newby et al., 2021; Thuronyi et al., 2019; Villiger et al., 2018). To avoid the excision of U and the resulting introduction of indels by the base excision repair pathway, most base editing tools also incorporate uracil glycosylase inhibitor (UGI) fused to AID-nCas9, which inhibits the base excision repair machinery (Fig. 1A) (Anzalone et al., 2020; Komor et al., 2016). The EvolvR system relies on fusing enCas9 (enhanced-nickase Cas9) to an error prone DNA polymerase I from *E. coli* that was engineered to increase its error rate 12,000 fold compared to wild type (Fig. 1B). The mutational spectrum of EvolvR was broader than for base editing (between 18% and 33% for A, T, C or G mutations) (Tou et al., 2020).

We tested these two mutagenesis methods. First, we studied several variants of AID engineered to be about 120 times more mutagenic than wild-type human APOBEC3 (AID12*-UGI, AID12*, AID123*-UGI and AID123*), fused to nCas9, with or without UGI. Each of these variants harbors mutations that result in different mutation efficiencies in bacteria and mammalian cells (Berríos et al., 2024; Kohli et al., 2021, Rahul Kohli, personal communication). Second, we tested the previously described EvolvR system (Tou et al., 2020). Since both approaches rely on fusions to nCas9, it is possible to target them to a region of interest in the genome using guide RNAs (gRNAs) (Fig. 1A, B). Further, inspired by approaches to fluorescently label a genomic region by recruitment of multiple Cas9 proteins using multiple gRNAs (Gu et al., 2018), we expressed an array of gRNAs to target our genomic region region of interest with multiple Cas9-AID/EvolvR fusion proteins. As a result, we can simultaneously mutate large segments of an enhancer in a random manner (Fig. 1C).

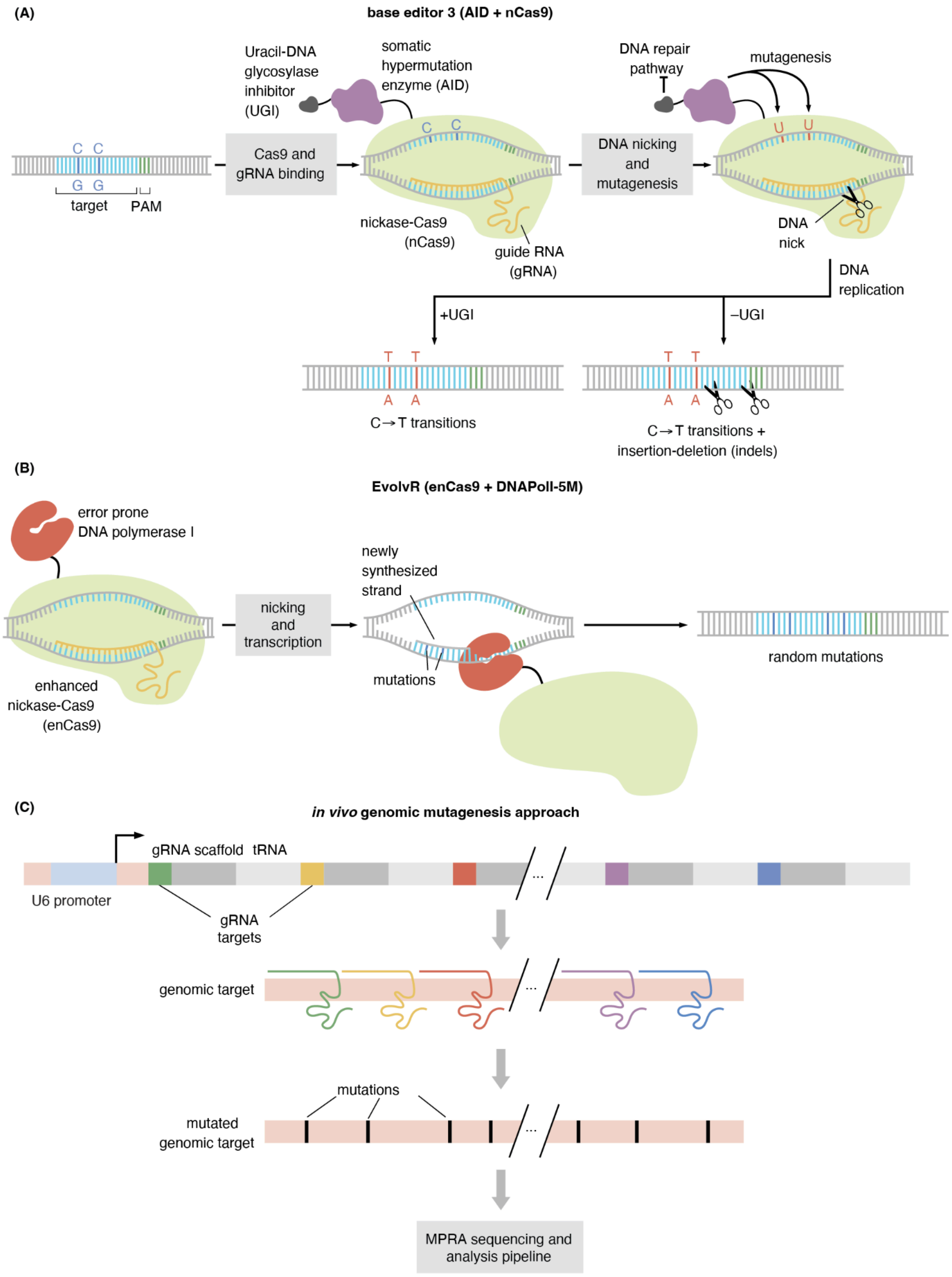
(A)
base editor 3 (AID + nCas9)
C C
G G
target
PAM
Cas9 and gRNA binding
Uracil-DNA glycosylase inhibitor (UGI)
somatic hypermutation enzyme (AID)
C C
nickase-Cas9 (nCas9)
guide RNA (gRNA)
DNA nicking and mutagenesis
DNA repair pathway
mutagenesis
U U
DNA nick
DNA replication
+UGI
−UGI
T T
A A
C→T transitions
T T
A A
C→T transitions + insertion-deletion (indels)
(B)
EvolvR (enCas9 + DNAPolI-5M)
error prone DNA polymerase I
enhanced nickase-Cas9 (enCas9)
nicking and transcription
newly synthesized strand
mutations
random mutations
(C)
in vivo genomic mutagenesis approach
gRNA scaffold tRNA
U6 promoter
gRNA targets
...
genomic target
...
mutations
mutated genomic target
...
MPRA sequencing and analysis pipeline

Figure 1. Tools to induce mutations *in vivo*. (A) Upon recognition of a specific DNA sequence targeted through the guide RNA (gRNA) and nickase Cas9 (nCas9), AID to create C→T transitions in the genome. The DNA nick introduced by nCas9 increases the efficiency of mutagenesis. Uracil glycosylase inhibitor (UGI) inhibits uracil excision and recruitment of the DNA repair machinery such indels are avoided**. (B)** In the EvolvR system, enhancer-nickase Cas9 (enCas9) targeted to a locus makes the DNA accessible for

an error prone DNA polymerase I from *E. coli* to transcribe the DNA. The DNA nick introduced by enCas9 is also required for mutagenesis. **(C)** Multiple gRNAs targeting an genomic region are expressed in a single RNA transcript flanked by tRNAs.

In order to achieve mutation rates throughout the target genomic site, achievable in the context of the in vitro-generated DNA libraries used in regular MPRAs, it is necessary for enough gRNAs to bind throughout the whole sequence. As a proof of concept, we cloned 4 gRNAs targeting GFP in a single construct, flanked by tRNA sequences to ensure that the gRNAs would be transcribed in one transcript and then cleaved. The 4xgRNA plasmid was transfected into *Drosophila* Kc cell lines expressing GFP, along with plasmids expressing the different AID-nCas9 variants driven by the *act5C* or the Cu^{+2} inducible *pMT* promoters (Fig. 2A, S1, see Methods) (Bunch et al., 1998). Following different amounts of time after transfection, genomic DNA extraction and amplicon sequencing of GFP was performed. Mismatches to the reference sequence could be detected to varying degrees in all samples and timepoints, with the variant AID12* at 4 weeks post-transfection producing the highest mismatch rate of up to ~20% in some positions along the segments of the GFP gene targeted by our gRNAs (Fig. 2B). These were C→T mutations on forward (FWD) gRNAs binding to the ‘minus’ strand, and G→A mutations for reverse (REV) gRNAs binding to ‘plus’ strand (Fig. 2B), consistent with AID only being able to mutate the strand of DNA not bound by the gRNA. Mutations only occurred on the gRNA sequence, with occasional mutations 1 to 3 bp 5’ of the gRNA, and exhibited a 5’ bias in the mutation rate, consistent with previous observations (Kohli et al., 2021; Marr and Potter, 2021). Notably, very few indels were detected even in samples without UGI (Fig. S2B), consistent with the recent realization that, unlike hemimetabolous insects, flies do not have the UNG gene required for the uracil excision repair pathway (Doll et al., 2023). No mutations were detected in EvolvR samples (Fig. S2B).

Having established that the AID system works in *Drosophila* cell culture, we moved on to determine its performance in the context of *Drosophila* embryos. In doing so, we started by restricting mutagenesis to oogenesis in order to avoid somatic mutations that might lead to toxicity as well as germline mutations that would get fixed in the germline and reduce the diversity of our library. To make this possible, it is necessary to ensure that AID gets only expressed after germline stem cell division throughout oogenesis or spermatogenesis. As a result, we expressed AID-nCas9 variants from the germline promoter *pBam.* This promoter is only active in the germline cyst stage following germ stem cell division, in both male and female germlines (H. M. Chen et al., 2020). The same 4 gRNAs as in our cell culture experiment were used to target genetically encoded GFP. F2 embryos from parents harboring GFP and expressing both AID-nCas9 and 4xgRNAs were collected (Fig. 2C) and GFP sequenced as described for the cell culture experiment above. In this case, only 4 positions on one of the gRNAs in the AID12* condition presented mutations, and these mutations occurred at a very low mutation rate of about 1% (Fig. 2D). Further, similarly to cell culture, no mutations were detected in the EvolvR samples. Whether mutations were actually occurring below our detection threshold or not, such low mutation rates are insufficient to recover information such as binding sites location in MPRAs, estimated at 10% in experiments mutating promoters in bacteria (Ireland et al., 2020; Pan et al., 2024).

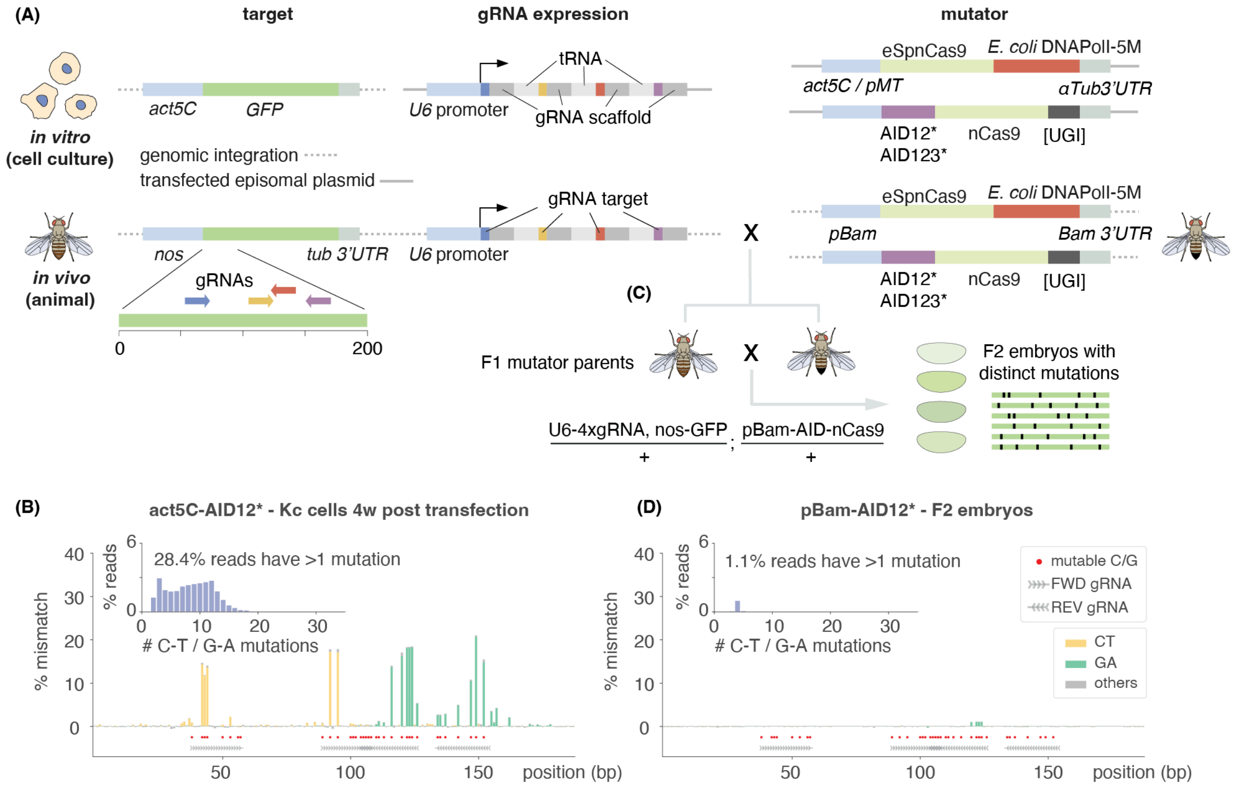

Figure 2. Human AID derived variants introduce C→T mutations at high rates at gRNA target sites in cell culture, but not in animals. (A) Plasmids used in experiments in cell culture and flies to mutagenize GFP in the genome of Kc cells or fly embryos using 4 gRNAs and nCas9 fused to mutagenic proteins AID or EvolvR. Different AID variants, promoters and 3'UTRs were tested. gRNAs were expressed off of U6 ubiquitous promoters. **(B)** Barplots showing mismatches to GFP classified by base substitution generated by AID12* in Kc-nls-GFP cells 4 weeks post-transfection. Most mismatches are C→T (yellow) on forward gRNAs or G→A (green) on reverse gRNAs. See Fig. S2 for other conditions **(C)** In embryos, the mutagenic enzymes were expressed in the germline from a *pBam* promoter and crossed to flies containing the same U6-4xgRNA construct used in cell culture as well as GFP. F2 embryos were collected, and GFP amplified to then be sequenced using Nanopore sequencing. **(E)** Barplots showing mismatches to GFP classified by base substitution generated by AID12* in F2 embryos following one generation of germline mutations. Few G→A (green) mutations were detected. In (C) and (E) red dots mark all "mutable bases" (Cs on FWD gRNAs and Gs on REV gRNAs). Insets in each plot show the distribution of the number of C→T or G→A mismatches per read on these mutable bases, after subtracting the number of mismatches observed in control samples (see methods).

Lamprey evoCDA1 is highly mutagenic *in vivo*

A few reasons could be causing the discrepancy in mutagenic activity between cell culture and flies. First, to ensure that the gRNAs were correctly expressed and that they were functional, we crossed the same 4xgRNA fly line with a line expressing catalytically active Cas9. We obtained a high mutation rate (up to 90% in some positions, Fig. S3A), indicating that the gRNAs were being expressed and could succesfully target the GFP gene. Second, it has been recently reported that some AIDs are sensitive to temperature (Doll et al., 2023). Therefore we tested the same constructs at 29C—towards the high range of fly viability—instead of room temperature but saw no significant difference in mutation rate (Fig. S3B). Third, a major difference between the cell culture, where mutations were observed at a high rate, and embryos, were mutations were hardly detectable, is the time the target gene is exposed to AID: cells could mutate for up to 4 weeks in culture while the *pBam* promoter is only active for 4 days during germline development in the animal (Rust et al., 2020). To determine whether more mutations could be accumulated over time we created stable stocks with each of the AID variants (Fig. 3A) and let them mutate for up to 10 generations, collecting embryos and sequencing GFP every 2 generations from generation 4. Although more mutations were detected over time (Fig. 3B, *pBam* AID12*-UGI, AID12* and AID123*-UGI samples, Fig. S4), the increase was very modest and still below the 10% that would be needed to successfully mutate a whole enhancer. Lastly, it is possible that the *pBam* promoter expressed AID at lower levels than the *actin5C* promoter used in cell culture. We therefore created flies expressing exactly the same *act5C* constructs that had been tested in cell culture and also replaced the promoter by a UASz (Fig. 3A), to achieve even higher levels (Deluca and Spradling, 2018). However, none of the *act5C* or *UASz* (with either *Bam-Gal4* or *osk-Gal4*) versions of our AIDs produced any significant increase in mutations (Fig. 3B). During this period, two papers were published using new AID-nCas9 constructs expressed from *act5C* promoters in *Drosophila* (Doll et al., 2023; Marr and Potter, 2021), so we tested these 3 variants together as well (Fig. 3A, rAPOEC1, evoAPOEC1, evoCDA1).

Surprisingly, one of the recently published AIDs, the *in vitro* evolved evoCDA1 variant from lamprey *Petromyzon marinus* pmCDA1 (Thuronyi et al., 2019), led to a remarkable amount of mutagenesis (Fig. 3B). In certain positions the percentage of mutated reads reached 60% and more than 80% of the reads had at least one C→T mutation (Fig. 3B). To measure the mutations produced in the male and female germline, we set up crosses that would allow us to distinguish, based on eye color, flies where mutations could only have occurred in the germline of its parents (see methods). Sequencing these F2 adults would also ensure we were detecting only germline mutations, since the mutations detected in F2 embryos above could be skewed by somatic mutations caused by *act5C* embryonic expression. This experiment revealed that mutations were indeed occurring in the germline, and that they were twice as likely to occur in the male germline, reaching 80% in some positions and 97% reads harboring at least one mutation, compared to 40% and and 72%, respectively, for mutations from the female germline (Fig. 3C, D). We hypothesize that the difference in mutation rate between males and females are due to sex-specific differences in AID efficiency, *act5C* expression, or due to the fact that male germline cells divide more times than in the female germline (Hempel et al., 2008). Regardless of the source of this sex-specific difference, our results suggest that choosing mutations generated in either germline can constitute a strategy for modulating mutation efficiency.

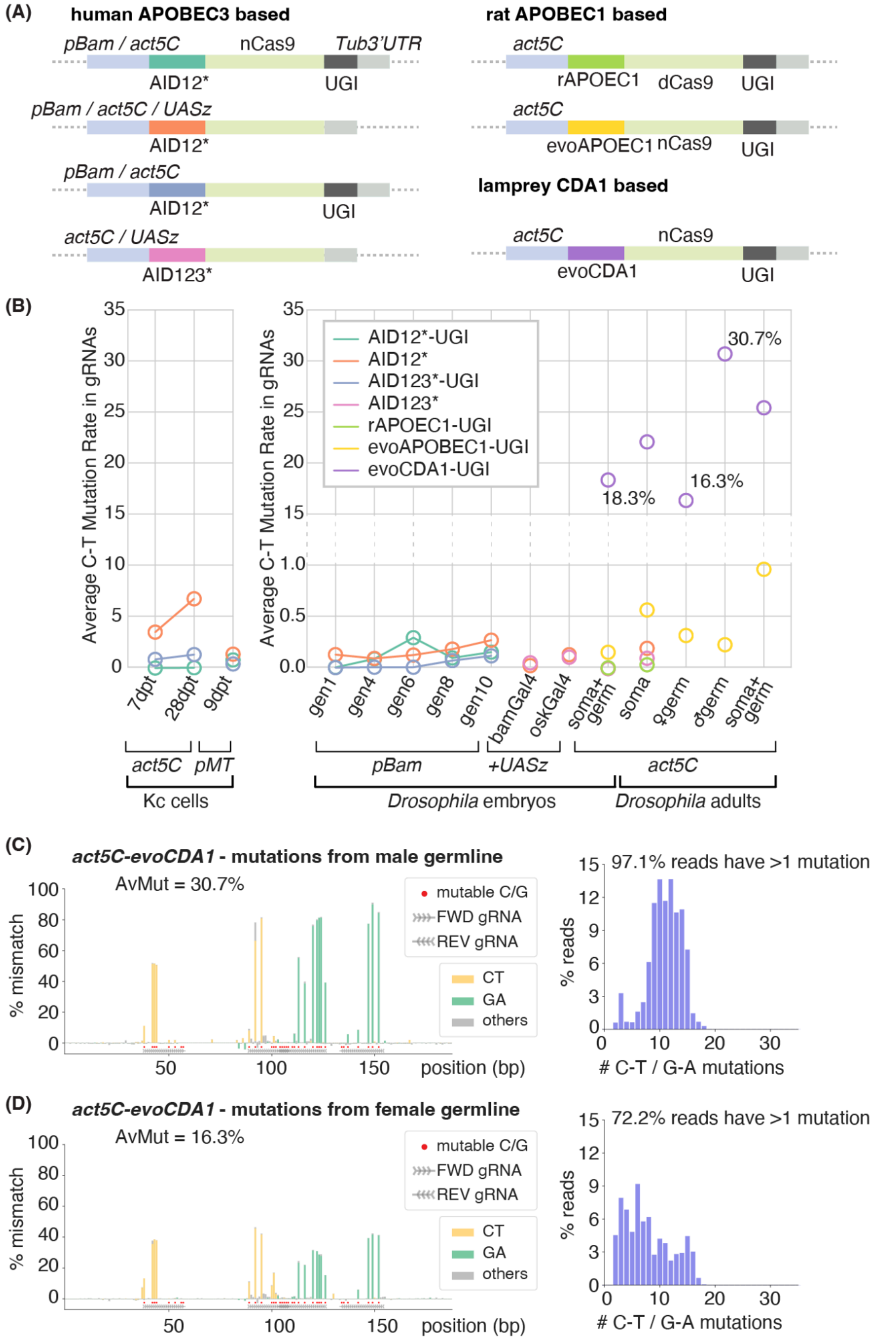

Figure 3. Sea lamprey AID^{evoCDA1} introduces C→T mutations at high rates in embryos. **(A)** Plasmids with different versions of AID used in experiments to mutagenize GFP in the genome fly embryos using 4 gRNAs. We tested AID stemming from different organisms under a plethora of different promoters. For the inducible UASz, we crossed to *bam-Gal4* and *osk-Gal4* to induce high germline expression. **(B)** Average mutation rate (defined as C→T or G→A mutations over all mutable positions) in conditions aiming to increase mutation rate in embryos. Mutation rates in the Kc experiment from Figure 2 are shown in the left panel for reference though relative levels of expression of AID between cell culture and embryos are likely not comparable. Propagating fly lines over multiple generations or replacing the *pBam* promoter for *act5C* of *UASz* had no or very little effect on the mutation rate. Only *act5C*-evoCDA1-nCas9

was successful at producing a high number of mutations in embryos. **(C-D)** Mismatches in GFP classified by base substitution in adults were *act5C*-evoCDA1-nCas9 introduced mutations with the 4 gRNAs from the male (C) or female (D) germline (left), and distribution of observed number of mutations per read after subtracting the number of mismatches observed in control samples (right). Red dots mark all “mutable bases” (Cs on FWD gRNAs and Gs on REV gRNAs).

Quantitative analysis of the variability in introduced mutations

evoCDA1 was remarkably efficient in introducing mutations in a short period of time. However, mutation rates between 40% and 80% at many positions could be too high for our purposes: if too many mutations are introduced in a given enhancer, even if they have a functional impact, it will not be possible to clearly discern which position(s) along the sequence encode for this functionality (Pan et al., 2024). Changing the *act5C* promoter back to the less active *pBam* promoter resulted in mutation rate of average 2% and 5.3% from the male germline, which could be too low for our purposes (Fig. S3).

In addition to overall mutation rates, our goal was to generate mutations that are as variable across embryos as possible. This would mean generating mutations along the sequence that are as uncorrelated from each other as possible, or even anticorrelated. Promoters active in different cell types could lead to different degrees of variability. For example, since the *act5C* promoter is active in the stem cell stage during germline development, it is possible that mutations could become fixed in the germline and, as a result, produce offspring with identical or very similar mutations. Such fixation would not occur if mutations are produced through the expression of AID using a promoter only active in later stages of germline development, such as *pBam*. To decide which construct was better suited for our purposes, we sought to develop a quantitative framework for understanding how mutations are introduced and what the resulting mutational variability between different molecules was.

Assessing the degree of correlation between sequences requires analysis of individual mutagenized DNA molecules. However, the experiments presented so far have relied on Nanopore sequencing, which can only provide mutational information averaged over multiple molecules (see Methods). To determine whether mutations within a given gRNA or between different gRNAs are correlated, we needed high quality data from individual DNA molecules. We re-sequenced our libraries using Illumina technology and quantified whether each mutable base and each gRNA was mutated for every DNA molecule. With this data in hand we calculated combinatorial probabilities and used mutual information theory to quantify the degree of mutational variability, first across gRNAs and then for mutations within each gRNA.

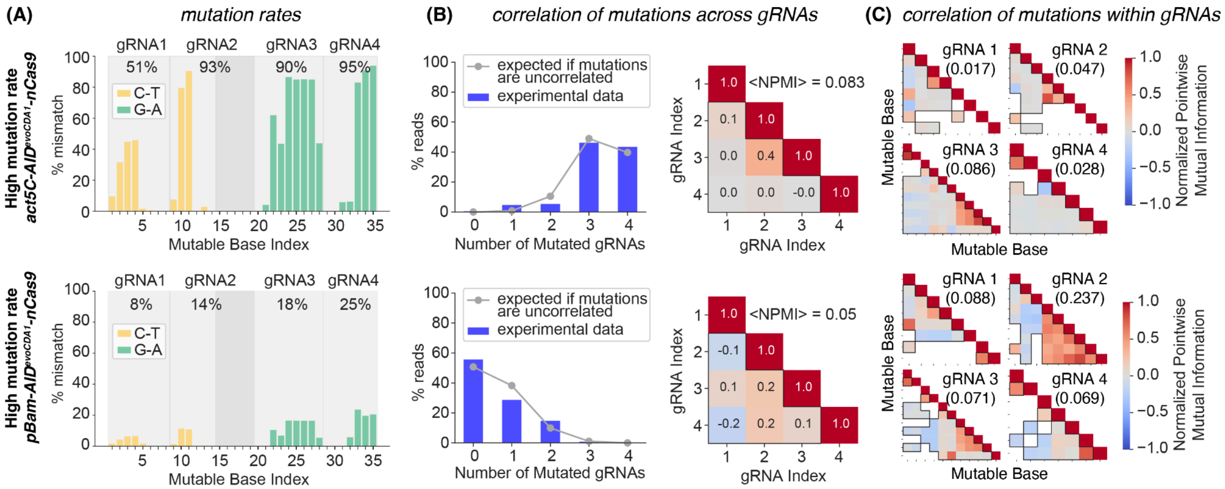

Figure 4. Quantifying the variability in mutations generated by AID^{evoCDA1} and 4 gRNAs. (A) Average mutation rates in each mutable base and gRNA. Bar plot shows the percentage of Illumina reads with mismatches to GFP in each of the mutable bases (marked in previous figures with red dots) produced by 4 gRNAs and *act5C-AID^{evoCDA1}-nCas9-UGI* or *pBam-AID^{evoCDA1}-nCas9-UGI*. Grey boxes mark each gRNA. Percentages on the top of the plot indicate the mutation rate for the whole gRNA, calculated as the percentage of reads with at least one mutation in each gRNA. **(B)** Quantification of the correlation in mutations across gRNAs for *act5C* (top) and *pBam* (bottom) constructs. Left plots show the expected distribution of the number of mutated gRNAs per read assuming mutations percentages from (A) and that mutations between different gRNAs are uncorrelated (Poisson Binomial Distribution, see Methods - Quantification mutational variability). Blue bars indicate the measured number of mutated gRNAs per read. Right heatmaps indicate the Normalized Pointwise Mutual Information (NPMI) between each pair of gRNAs, which measures the degree of co-occurrence between two events (0 if they co-occur by chance, 1 if completely correlated, -1 if completely uncorrelated). <NPMI> is the average correlation coefficient over non-diagonal matrix elements (marked by a black line). **(C)** Quantification of the correlation in mutations within each gRNA for *act5C* (top) and *pBam* (bottom) constructs. Heatmaps show NPMI across each pair of mutable bases for each gRNA. Numbers in parenthesis indicate the average NPMI over all combinations, corresponding to the non-diagonal elements of the matrix.

To analyze the correlation of mutations across the gRNAs, we quantified the percentage of reads with at least one mutation in each of the gRNAs (Fig. 4A). Using these percentages, we calculated how likely it would be to obtain reads containing 1, 2, 3 or 4 mutated gRNAs, i.e. the Poisson Binomial Distribution (see Methods - Quantification mutational variability) that calculates the probability of exactly k events occurring from a vector or probabilities assumed to be independent. We then compared the predicted distribution of the number of mutated gRNAs with the distribution from the experimental data (Fig. 4B, left). Distributions shifted towards a lower number of mutated gRNAs per read would indicate anticorrelated binding events, whereas distributions shifted towards a higher number of mutated gRNAs would indicate mutations across gRNA targets are correlated (see Methods). Using data from low (*pBam*) and high (*act5C*) mutation rates, we observed a similar distribution in the number of mutated gRNAs compared to when mutations are assumed to be independent (Fig. 4B, left).

To examine the degree of correlation in mutations between different gRNAs, we also calculated the normalized pointwise mutual information (NPMI) between each combination of pairs of gRNAs (see Methods - Quantification mutational variability). Events completely uncorrelated have a NPMI of 0, whereas events completely correlated or anticorrelated result in NPMIs of 1 and -1, respectively (Bouma, n.d.). NPMI values were close to 0 in most pairwise combinations (Fig. 4B, right). However, a positive correlation of 0.4 was found between gRNAs 2 and 3 in the *act5C* sample, whereas it was only 0.2 in the *pBam* sample (Fig. 4B, right). Since gRNAs 2 and 3 are overlapping, it is possible that binding of one increases the chances of another binding, for example by lowering the energy required to open the bubble of ssDNA (Corsi et al., 2022). Thus, our two quantification methods indicate that mutations across gRNAs are largely uncorrelated, with a potential positive correlation when gRNAs overlap.

We then used a similar approach to quantify the degree of correlation between each mutable base within a gRNA. Using information theory, we calculated the NPMI for every possible combination of 2 mutations within each gRNA, in the same samples with low (*pBam*) and high (*act5C*) mutation rates described above (Fig. 4C). Most combinations resulted in a NPMI close to 0 for both *act5C* and *pBam*. Some mutable bases within gRNAs 2 and 3 showed a NPMI ~ 0.5, indicating that those mutations were positively correlated (Fig. 4C). On average, quantified NPMIs for each gRNA were slightly positive but close to 0, < 0.1 for all gRNAs except for 0.237 in gRNA2 in the *pBam* sample (Fig. 4C).

From these quantifications we concluded that mutations across and within gRNAs are largely uncorrelated. Since the degree of correlation was similar in the context of the low (*pBam*) and high (*act5C*) mutation rates, we hypothesized that the difference in AID expression levels mainly affects overall mutation rates and not the degree of mutational variability within a gRNA. In this scenario, the construct producing average mutation rates closer to 10% along the whole enhancer would be our preferred choice.

Optimizing gRNA number and length to mutagenize a long stretches of genomic DNA

It is important to note that, while these pilot experiments were conducted using only four gRNAs, our ambition is to deploy our technology to mutagenize whole enhancers (or any other genomic sequence) in one experiment. To make this possible, it is necessary to carpet hundreds of base pairs of an enhancer by simultaneously expressing many gRNAs. Therefore the mutation rates quantified in the context of only 4 gRNAs might not be maintained when a higher number of gRNAs is present. We hypothesized that increasing the number of gRNAs will lead to reduced mutation rates due to the blocking effect of overlapping gRNAs, *i.e.* once a mutation has occurred on DNA it cannot be bound again by any gRNAs overlapping that DNA region, making that stretch of DNA refractory to further mutations. Other effects, such as lower levels of each gRNA produced in a single transcript from an individual promoter, could also result in lower overall mutation rates.

To quantitatively understand how DNA becomes refractory to mutations, we developed a simple theoretical model that predicts how mutations accumulate over time under different conditions. Briefly, our model posits that each nCas9 molecule binds each gRNA target with a probability p_{bound} (probability of binding in each fly generation or cell cycle). Only when bound will AID have the possibility of mutating each available C or G (depending on whether the gRNA is FWD or REV) with a probability p_{mut} (Fig. 5A). If p_{mut} = 1 each time nCas9 binds all possible bases within the reach of the gRNA will be mutated, leading to 100% mutation rate over time. If p_{mut} < 1, only some bases will be mutated. These new mutations will prevent gRNA binding such that p_{bound}=0 in further cell cycles. As a result, the mutated DNA sequence will become refractory to further mutations, leading to saturation of the mutation rate below 100%. In our simulations, each binding or mutation step was calculated as a random number, and depending on whether the obtained value was higher or lower than p_{bind} or p_{mut} binding events or mutations would occur or not (Fig. 5A). Simulating 4 gRNAs targeting GFP over an n number of generations, we observed that changing p_{bound} only changes the speed at which the saturation point is reached, whereas p_{mut} changes the maximum fraction of mutations at the saturation point (Fig. S6A). We used these simulations to test the effect of overlapping gRNAs and of increasing the number of gRNAs. As expected, adding overlaps between gRNAs led to lower mutation rates, and the effect was more pronounced at high p_{mut} (Fig. S6B).

To be able to compare simulations to experiments, we also considered uniform vs. 5' biased mutation rates along the gRNA sequence by modulating p_{mut} along the gRNA position. Adding the observed 5' bias from previous experiments to the simulations also decreased the overall mutation rate as expected (Fig. S6C) and allowed us to obtain predicted mutation profiles and histograms of the distribution of mutations that could be compared to the real experiment.

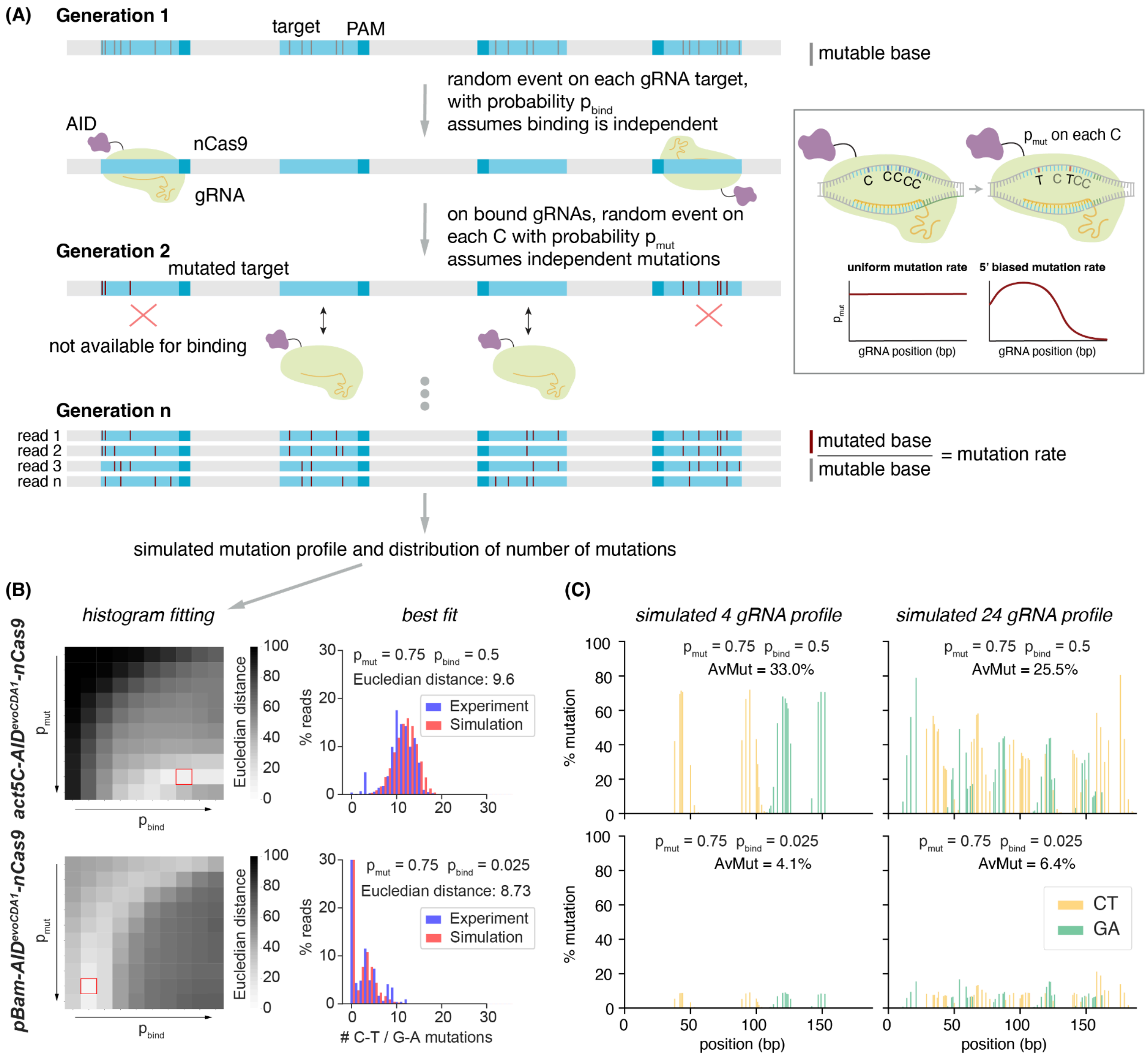

Figure 5. Simulating the expected number and distribution of mutations. **(A)** Simple model of mutations caused by AID-nCas9 in two probabilistic steps. First, each gRNA guides the binding of nCas9 to an intact target site with probability p_{bound}. Upon binding, AID mutates each of available C with a probability p_{mut}. Upon successful mutagenesis, the p_{bound} of any gRNA overlapping the mutated sequence becomes zero. **(B)** Finding the best fit parameters between simulations and experiments by comparing the histograms of the distribution of number of mutations per read. Left heatmaps show Euclidean distances between the two histograms for 100 combinations of p_{bound} and p_{mut} with the lowest value in red. Tested p_{bound} and p_{mut} values and fittings for all combinations are shown in Fig. S7, S8. Right plots show best fits for the *act5C* and *pBam* experiments, with experimental distribution of number of mutations shown in blue and best fit from the simulations in red. **(C)** Expected mutation profile when GFP is targeted with 4 gRNAs (left) or 24 gRNAs (right) with the best fit parameters to *act5C* samples (p_{bound} = 0.75 and p_{mut} = 0.5, top) and *pBam* samples (p_{bound} = 0.75 and p_{mut} = 0.025, bottom). Average of 1000 mutated simulated reads.

We generated 100 simulated datasets with the 5' biased mutation profile from a range of p_{bound} and p_{mut} values, and found the distribution of the number of mutations that better fit the experimental data (Fig. 5B, left, Fig. S7, S8). The best fit parameters for the experiment with a high mutation rate (*act5C*) was p_{mut} = 0.75 and p_{bound} = 0.5, whereas for the experiment with a low overall mutation rate (*pBam*) were p_{mut} = 0.75 and p_{bound} = 0.025 (Fig. 5B, right). It is interesting that the best fit p_{mut} values are the same in both experiments whereas p_{bound} is 20 times higher in the *act5C* experiment. This could be interpreted as different promoters producing different amounts of nCas9-AID, leading to different probabilities of each binding the DNA target. However, once bound, both mutate to similar rates. This suggests levels of nCas9-AID are limiting in this context and the main source of potential modulation of the mutagenic rate. Although finding an exact parameter that can explain the obtained results was not the goal of this approach, it was useful to generate hypotheses that could be tested with simulated data.

To more realistically model the mutagenesis of an enhancer, we designed gRNAs for all possible PAM sites in a 200 bp sequence of GFP, then removed those that were almost identical, ending up with 24 gRNAs that could cover the whole 200 bp sequence FWD and REV, potentially mutating all Cs and Gs in the region (Fig. S9A). Our simulations predicted that the mutation rates with these 24 gRNAs would be lower compared to the 4 gRNAs at high p_{mut} values, likely due to the effect of multiple overlaps, while they remained similar at low p_{mut} (Fig. S6C). Assuming the best fit parameters to *act5C-nCas9-evoCDA1* and *pBam-nCas9-evoCDA1*, our model predicts that mutating GFP with 24 gRNAs instead would result in average mutation rates of 25% and 6.4% respectively (Fig. 5C), assuming that all other parameters such as levels of AID-nCas9 and of each gRNA remain constant. Thus, our simulations provided useful context to inform our next experiments. Based on this result, we hypothesized *pBam* was the best promoter to achieve mutation rates close to 10%, but it remained to be tested how it performs *in vivo* when multiple gRNAs are used.

To experimentally test the mutagenic effect of a higher number of gRNAs, we cloned arrays of gRNAs containing the 24 gRNAs designed in the previous section that would cover the same 200 bp region of GFP in FWD and REV. Because of the challenge in cloning a repetitive array containing all 24 gRNAs, we cloned groups of these gRNAs into separate constructs (containing 6, 6 and 12 gRNAs) flanked by tRNA sequences in three different plasmids. In cell culture, we co-transfected these arrays of 24 gRNAs or 4 gRNAs together with 7 different AIDs expressed from *act5C* promoters: 4 versions from (Kohli et al., 2021), human APOEC/BE2 from (Marr and Potter, 2021), rat APOEC1 (Doll et al., 2023), sea lamprey evoCDA1 (Doll et al., 2023) (Fig. 3A), and extracted DNA at different timepoints (Fig. S9B, see methods). We found most mutations at the two week time point, which led us to further focus on these samples (Fig. S4C).

In the condition with 24gRNAs we detected mutations along the 200 bp sequence (compare top and bottom in Fig. 6A, B), resulting in a less 5’ biased and flatter average mutation profile than in the condition with 4 gRNAs (Fig. 6C). These results suggest that our approach of using an array of gRNAs can be effectively deployed to target a whole enhancer and obtain a relatively uniform mutation rate. Compared to experiments in flies and the expected mutation rates from simulations, we found average mutation rates of 12-14% for 4 gRNAs and 2-5% for 24 gRNAs were much lower than expected. This is likely due to incomplete transfection, leading to only a fraction of cells in culture being mutagenic. Assuming equal transfection rates in the 4 and 24 gRNA conditions in the cell culture experiment, we can estimate mutation rates with 24 gRNAs to be 6 times lower than with 4 gRNAs in cell culture (from 12% to 2%, Fig. 6B), as opposed to just a 20% reduction estimated from the simulations with 24 gRNAs (from 33% to 24%, Fig. 5C), suggesting other factors in addition to overlapping gRNAs are responsible for this reduction. In the previous section we concluded nCas9-AID is likely a limiting factor *in vivo*, and our simulations assume the same amount of nCas9-AID is available to bind on each gRNA independently of how many gRNAs exist. One hypothesis is that nCas9-AID becomes more limiting when a higher number of gRNAs is present, effectively leading to a lower probability of binding. Alternatively, it is possible gRNAs are also limiting. Since 24 gRNAs are produced in arrays of 6-12 gRNAs, it is possible each of them is produced at lower levels than the equivalent 4 gRNAs. Performing this test in embryos, where the same conditions can be ensured for the cases of low and high number of gRNAs, will help distinguish between these possibilities and to obtain the information necessary to estimate how mutation rate is modulated as the number of gRNAs increases.

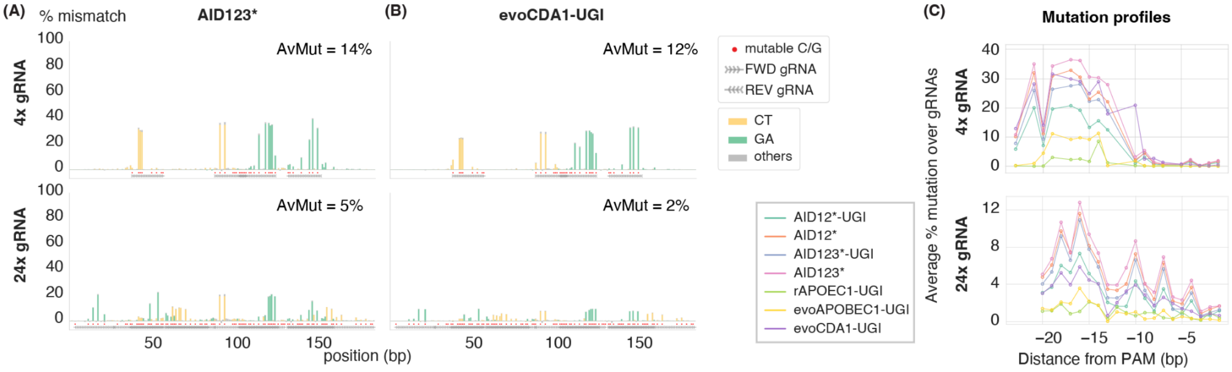

Figure 6. Testing the mutagenic activity of gRNA arrays in cell culture. **(A)** Mismatches to GFP classified by base substitution in Kc cells transfected with human act5C-AID123*-nCas9 and 4 (top) of 24 (bottom) gRNAs, 2 weeks post-transfection. **(B)** Mismatches to GFP classified by base substitution in Kc cells transfected with sea lamprey *act5C-AID*^{evoCDA1}*-nCas9* and 4 (top) of 24 (bottom) gRNAs, 2 weeks post-transfection. **(C)** Average mutation profiles over all gRNA used in the 4 (top) of 24 (bottom) gRNAs condition with different mutagenic enzymes. 24 gRNAs (bottom) mutate along the whole 200bp sequence and, when mutation rates are averaged over all gRNAs, the obtained mutation profiles exhibit less of a 5' bias than a smaller number of non overlapping gRNAs (top panels).

Finally we sought to optimize another parameter: the gRNA length. Although standard CRISPR uses gRNAs 20 bp in length (Jinek et al., 2013), we noticed that one of the four gRNAs initially tested was accidentally 23 bp long (Ma et al., 2016). This longer gRNA happened to be the most highly active one with respect to the remaining three gRNAs. These results suggested that 23 bp long gRNAs might be more efficient than the standard 20 bp ones in the context of base editing. As a result, to determine if an increase in gRNA length could lead to increased activity, we generated plasmids expressing the same gRNA in a 20 or 23 bp version. As explained above, we co-transfected these one 20bp gRNA or one 23bp gRNA plasmids, together with the 7 different *act5C*-AID variants (Fig. 3A, S4B). We found that human AID123* and sea lamprey evoCDA1 were the most efficient mutators in cell culture (Fig. S9C). As expected, and consistent with the differences observed between AIDs in cell culture and evoCDA1 in embryos, evoCDA1 had a wider 3' mutating window (Fig. 7, *) (Anzalone et al., 2020; Doll et al., 2023; Thuronyi et al., 2019). Confirming our hypothesis, the 23 bp gRNA was more efficient than the 20 bp gRNA. Further, the 23 bp gRNA extended the mutation window 5' with respect to the 20 bp gRNA (Fig. 7, **).

Having determined that a single 23 bp gRNA increases the mutagenesis window with respect to standard 20 bp long ones, we wondered if adopting even longer gRNAs could reduce the total number needed to target a whole enhancer. We determined the optimal gRNA length by increasing the length of the gRNA to target more C's on the 5' end to 23, 25, 30 or 45 bp. We found that increasing the gRNA length more than 23bp decreased the efficiency and returned to the narrower editing window of 20 bp in both AID123* and evoCDA1-UGI (Fig. S10). Thus, we concluded that 23 bp is the optimal gRNA length to mutate more Cs along the sequence and at higher efficiencies. Although the tested 24 gRNA array above contained gRNAs of 20bp in length, we expect increasing the length of all to 23bp will overall increase the efficiency and access more Cs available for mutation, leading to more uniform mutation profiles. When used to target an enhancer in whole animals, using these longer 23bp gRNAs in arrays we will be able to mutate over a long sequence and generate random uniform variability within it.

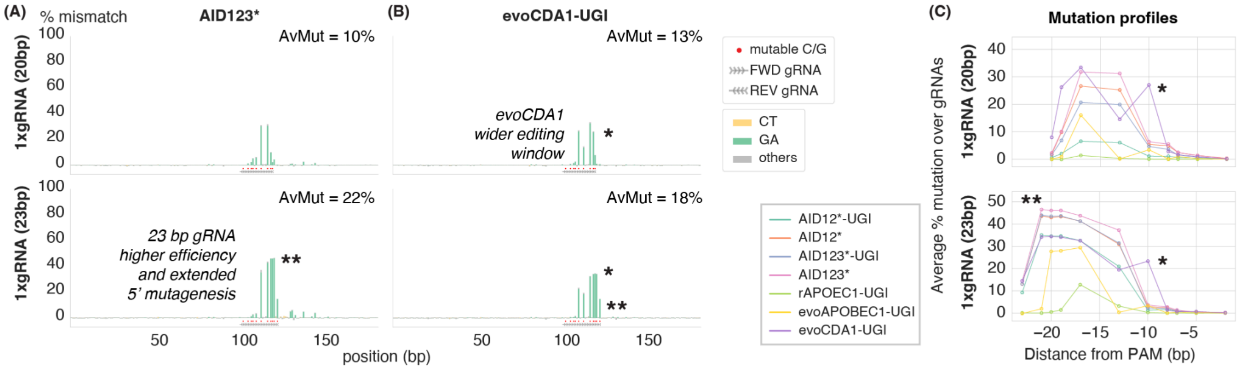

Figure 7. Testing gRNA length in cell culture. **(A)** Mismatches to GFP classified by base substitution in Kc cells transfected with act5C-AID123*-nCas9 and one 20 bp (top) or 23bp (bottom) gRNA, 2 weeks post-transfection. **(B)** Mismatches to GFP classified by base substitution in Kc cells transfected with *act5C-AID^{evoCDA1}-nCas9* and one 20 bp (top) or 23bp (bottom) gRNA, 2 weeks post-transfection. **(C)** Average mutation profiles over one 20 bp (top) or 23bp (bottom) gRNA with different mutagenic enzymes. * Lamprey AID^{evoCDA1} exhibits a wider editing window than APOBEC3 based AID123. ** A 23bp gRNA is more efficient than the same 20bp version, and extends the mutation window 5.

Discussion

The advent of MPRAs has revolutionized our ability to query biological phenomena, from dissecting gene regulatory regions to systematically studying protein function (Arnold et al., 2013; Brown et al., 2022; Chan et al., 2023; Inoue and Ahituv, 2015; Ireland et al., 2020; Jores et al., 2020; Kheradpour et al., 2013; Lagunas et al., 2023; Lalanne et al., 2024; Qi et al., 2020; Staller et al., 2022). Yet, these advances have been mostly relegated to cells in culture, where the transfection of massive libraries of DNA is feasible. Indeed, with some notable exceptions, multicellular organisms have remained on the sidelines of the great progress heralded by MPRAs. Further, because of their reliance on reporter constructs, MPRAs cannot be deployed in endogenous genomic *loci* neither in the context of single cells, nor in the context of multicellular organisms.

In this work, we introduced a key technology necessary for targeted *in vivo* genomic mutagenesis, a key step towards realizing MPRAs in multicellular organisms, as well as in endogenous *loci* both in multicellular organisms and single cells in culture. Specifically, we repurposed based editing tools (AID fused to nCas9) to generate random mutations within a 200 bp sequence by using multiple gRNAs expressed in an array. We found that lamprey AID (evoCDA1) was highly active in *Drosophila in vivo* and in cell culture, whereas other rat and human APOEC-derived AIDs, heavily used in mammalian cell culture, were active in *Drosophila* cell culture but not in animals. We showed that 24 gRNAs expressed in arrays can mutate along a 200 bp sequence, resulting in relatively uniform mutational profiles. Surprisingly, 23 bp gRNAs led to both higher mutation rates and wider editing windows than the canonical 20 bp CRISPR gRNAs.

Current limitations of our approach and further optimization

One clear aspect that our experiments revealed is that enzymes optimized in one system or species might not be well-suited to others. Specifically, even though human APOEC1 has been extensively engineered to be used in mouse and human cell culture (Thuronyi et al., 2019), it was not active in *Drosophila*. AID123*, which was highly active in bacteria (Kohli et al., 2021) was also not efficient in *Drosophila in vivo*, despite being highly active in *Drosophila* cell culture. In contrast, engineered lamprey AID, evoCDA1 (Thuronyi et al., 2019), was highly active both in *Drosophila* cell culture and *in vivo*. In previous works, temperature was identified as the main factor explaining this difference between deaminases (Doll et al., 2023). However, temperature does not explain why AID12* and AID123* worked well in cell culture (26C), but did not work in *Drosophila* at either 25C or 29C. Interestingly, because of their use in precision genome engineering (Koblan et al., 2021; Musunuru et al., 2021; Newby et al., 2021; Villiger et al., 2018), most AIDs have been selected to be highly efficient, accurate, and to have narrow editing windows (Anzalone et al., 2020; Thuronyi et al., 2019). However, for our purposes, enzymes with low but consistent mutation rates and with a wide editing window and random mutagenesis targets are needed. Thus, more optimization of these enzymes might be necessary to enable our *in vivo* genomic mutagenesis approach to MPRAs.

Recent studies have shown that the ability to tune the mutation rate in MPRAs can be critical. For example, using Reg-seq, an MPRA approach to find transcription factor binding sites and transcription factor binding energy matrices, it was determined that a 10% mutation rate is optimal for these purposes while minimizing noise (Ireland et al., 2020; Pan et al., 2024). While *in vitro*-generated libraries for MPRAs can be designed with a prescribed mutation rate, this is not the case in our approach to *in vivo* genomic mutagenesis. However, using different promoters to tune the timing and level of expression, and informed by simulations, we have demonstrated that we can also obtain average mutation rates between 10 and 20%.

Similarly, one significant difference between our approach and the generation of libraries *in vitro* is related to the mutational spectrum. *In vitro* DNA libraries can be designed to contain all possible base substitutions. However, our AID-nCas9 approach can only lead to C→T or G→A mutations. While it remains to be experimentally determined whether this limited mutational spectrum can provide the same information as an unbiased mutational spectrum, one recent study simulating similar experiments in bacteria (Pan et al., 2024) suggests that completely random substitutions can provide the same amount of information about where binding sites are in bacterial promoters as mutations produced by C and A deaminases (C→T and G→A by C deaminases, and A→G, and T→C by A deaminases). Regardless, we envision that in future work other base editors could be combined with AID to increase the mutational spectrum. For example, the adenine deaminases ABE8e (Fu et al., 2021) could be used to incorporate A→G and T→C mutations.

Finally, another limitation of targeting AID with nCas9 is that it can only be targeted to sequences that contain the protospacer adjacent motif (PAM) given by the NGG sequence. Although enough PAM sequences were present on the stretch of the GFP gene used for this study, it is possible that other sequences will have a more limited number of PAM sequences. In that case, other Cas9 versions, such as Cas12 (Chen et al., 2022) or the nearly PAMless SpRY-nCas9 (Walton et al., 2020) could be used. Alternatively, other targeting systems, such as those that rely on fusions to T7 polymerase (Álvarez et al., 2020; H. Chen et al., 2020; Cravens et al., 2021; Moore et al., 2018; Park and Kim, 2021) could also be deployed.

Future applications of MPRAs in animals and in endogenous loci

Regardless of the potential limitations outlined above, all of which we envision will become easier to solve with the continued development of base editors, our system offers the possibility of generating thousands of different mutations in DNA sequences in a single experiment, saving time and costs compared to designing DNA libraries and making transgenic animals individually or in pools. As a first step, our technology could be used to dissect the regulatory content of enhancers that are contained within a STARRseq (self-transcribing active regulatory region sequencing) construct. Here, enhancers are placed downstream of promoters such that the resulting transcript contains the enhancer sequence. As a result, enhancer sequence and resulting gene expression levels can be simultaneously measured without the need of barcoding (Arnold et al., 2013).

We envision that future developments in single cell DNA and single-cell RNAseq will make it possible to go beyond reporter platforms such as STARRseq (Olsen et al., 2023; Yu et al., 2023). Indeed, as these technologies become widely accessible, it will be possible to mutagenize an enhancer in its endogenous context while simultaneously measuring the transcriptional activity of the gene regulated by the enhancer. It is important to note that this ability to mutagenize endogenous regulatory regions would even be useful in contexts where reporter-based MPRAs have been traditionally feasible, such as bacteria, yeast or mammalian cell culture. Importantly, our technology is not limited to targeting regulatory sequences. We envision that our approach could also be used to systematically mutagenize small protein domains such as transcriptional transactivation domains to test the phenotypic outcome by measuring changes in gene expression in their endogenous context (Staller et al., 2022).

Overall, we have developed a versatile tool that can be used to mutagenize regulatory sequences, or virtually any other sequences, in *Drosophila*. We posit that this tool can be readily adapted to other multicellular and single-celled organisms. Building on recent previous works aimed at using MPRAs to find the number, placement and affinity of transcription factor binding sites in previously uncharted regulatory regions (de Almeida et al., 2022; Ireland et al., 2020), future work from our lab will harness this new technology to map the regulatory architecture of enhancers relevant to fly embryogenesis in a single experiment.

Supplementary figures

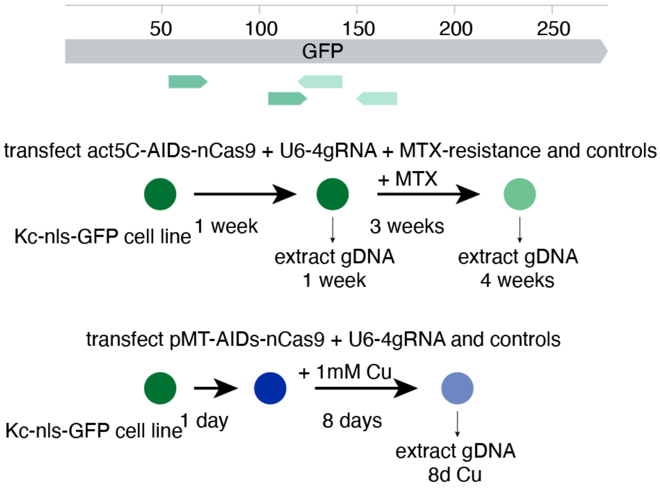

Figure S1. Experimental scheme for cell culture experiments. (A) Binding location of the 4 gRNAs targeting GFP. **(B)** Experimental protocol to induce and assay mutagenesis in cell culture. Timepoints of genomic DNA collection of Kc-NLS-GFP cells after transfection of gRNAs, AID-nCas9 and MTX resistance. Genomic DNA was collected after 1 and 4 weeks for *act5C* driven constructs, and after 9 days in the case of *pMT*. GFP was then amplified and sequenced using Nanopore sequencing.

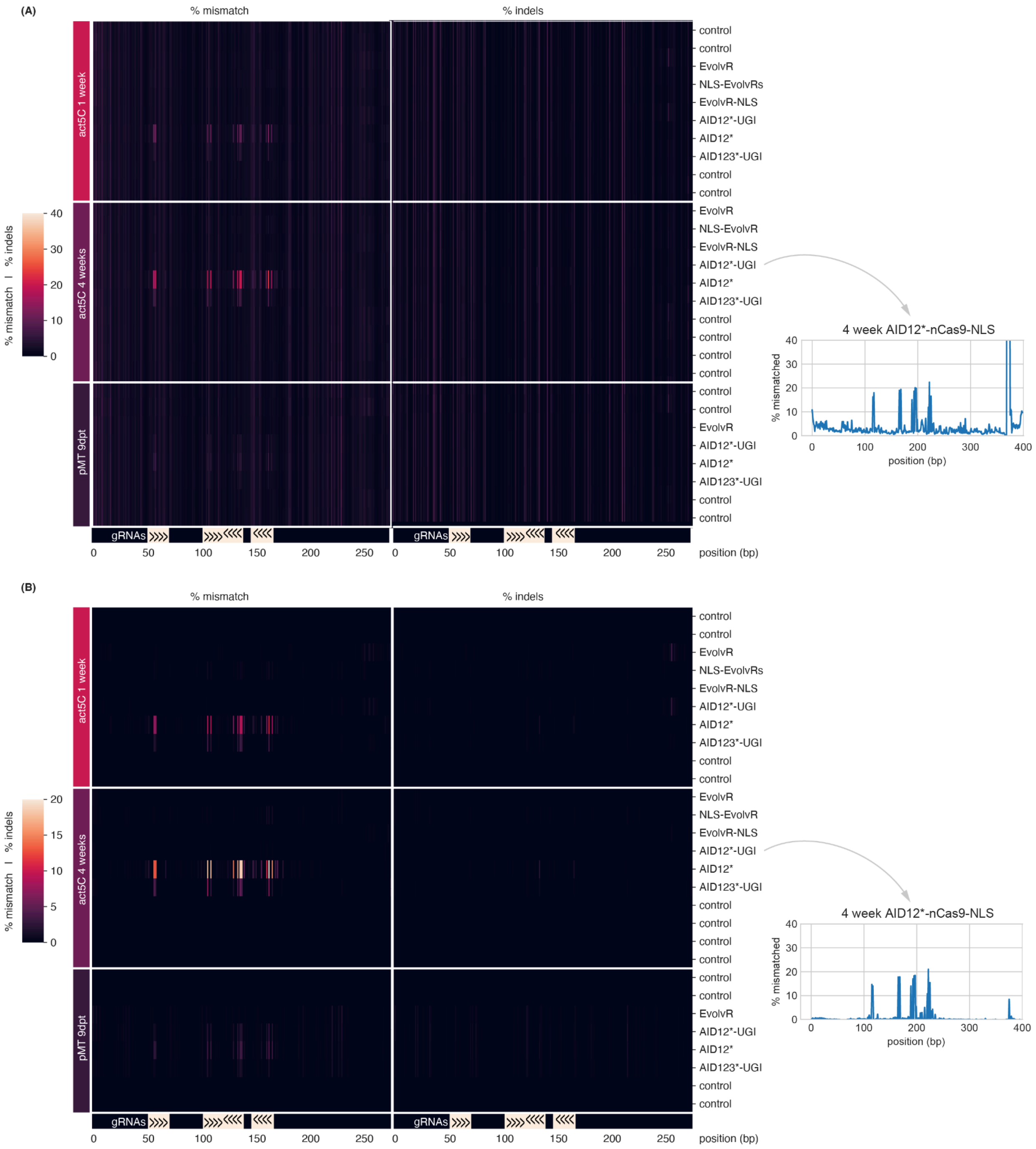

Figure S2. Nanopore sequencing results in cell culture. (A-B) Heatmap showing percentage of mismatch and indels to the GFP sequence before (A) and after (B) subtracting background mutations detected in controls (see methods) in the different timepoints and mutagenic enzyme conditions. AID samples exhibit a high percentage of mismatches aligning with the location of the gRNAs. Right plots show example profiles before and after background subtraction.

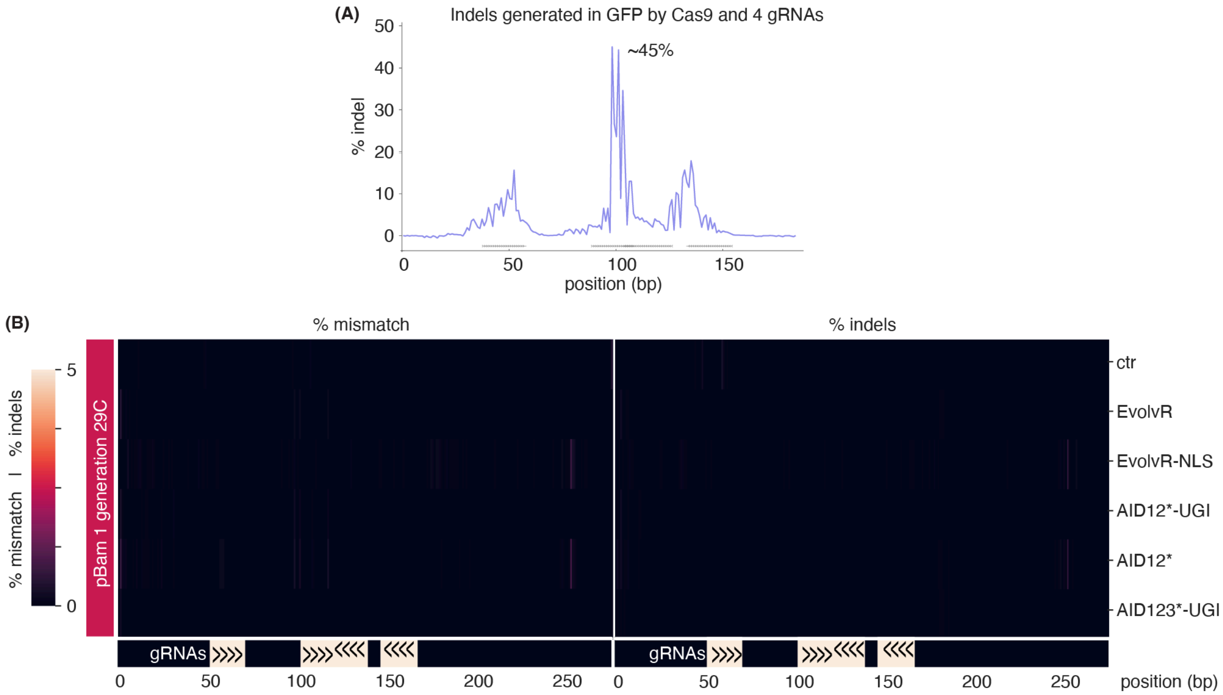

Figure S3. Testing gRNAs using catalytically active Cas9 and mutations in embryos at 29C (A) Catalytically active Cas9 introduces indels at target sites with variable efficiencies, with a maximum efficiency of approximately 45%. Since GFP could have been derived from males or females but *vasa*-Cas9 was only expressed in females, the actual efficiency would be approximately 90%. **(B)** Heatmap showing percentage of mismatch and indels to the GFP sequence after subtracting background mutations detected in controls in embryos where the indicated enzymes could have generated mutations for one generation at 29C. No mismatches or indels were detected.

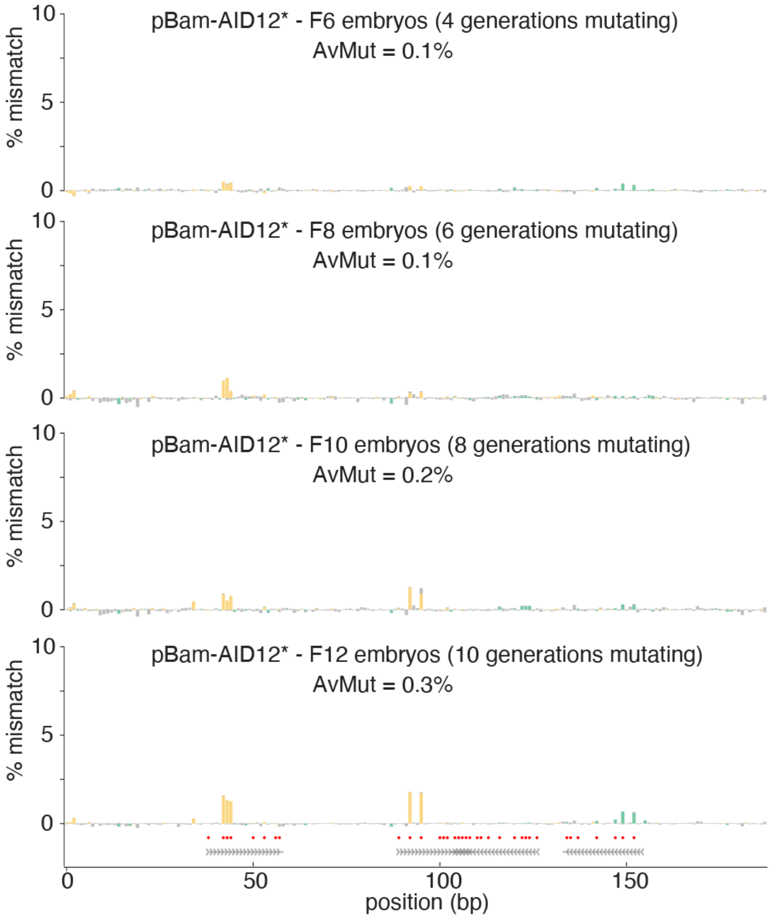

Figure S4. Mutagenesis in flies over multiple generations. Mismatches to GFP classified by base substitution in embryos from stable stocks expression pBam-AID12* and 4 gRNAs targeting GFP collected after increasing numbers of generations over which mutagenesis was carried out.

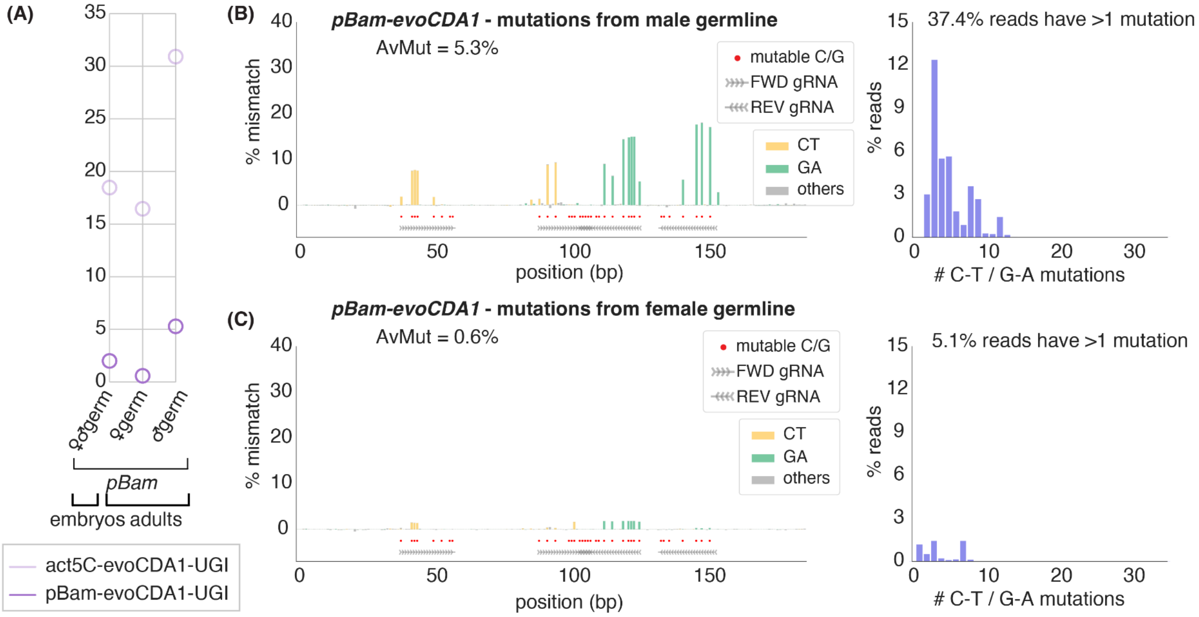

Figure S5. Mutagenesis by lamprey AID^{evoCDA1} expressed from a *bam* promoter. (A) Comparison of average mutation rates obtained by AID^{evoCDA1} expressed from *act5C* or *bam* promoters, in embryos and adults mutated from only the female or male germline. Mutations are significantly lower with *pBam* than *act5C.* **(B)** Mismatches in GFP classified by base substitution in adults were pBam-evoCDA1-nCas9 introduced mutations with the 4 gRNAs from the male (top) or female (bottom) germline (left), and distribution of the observed number of mutations per read after subtracting the number of mismatches observed in control samples (right). Red dots mark all "mutable bases" (Cs on FWD gRNAs and Gs on REV gRNAs).

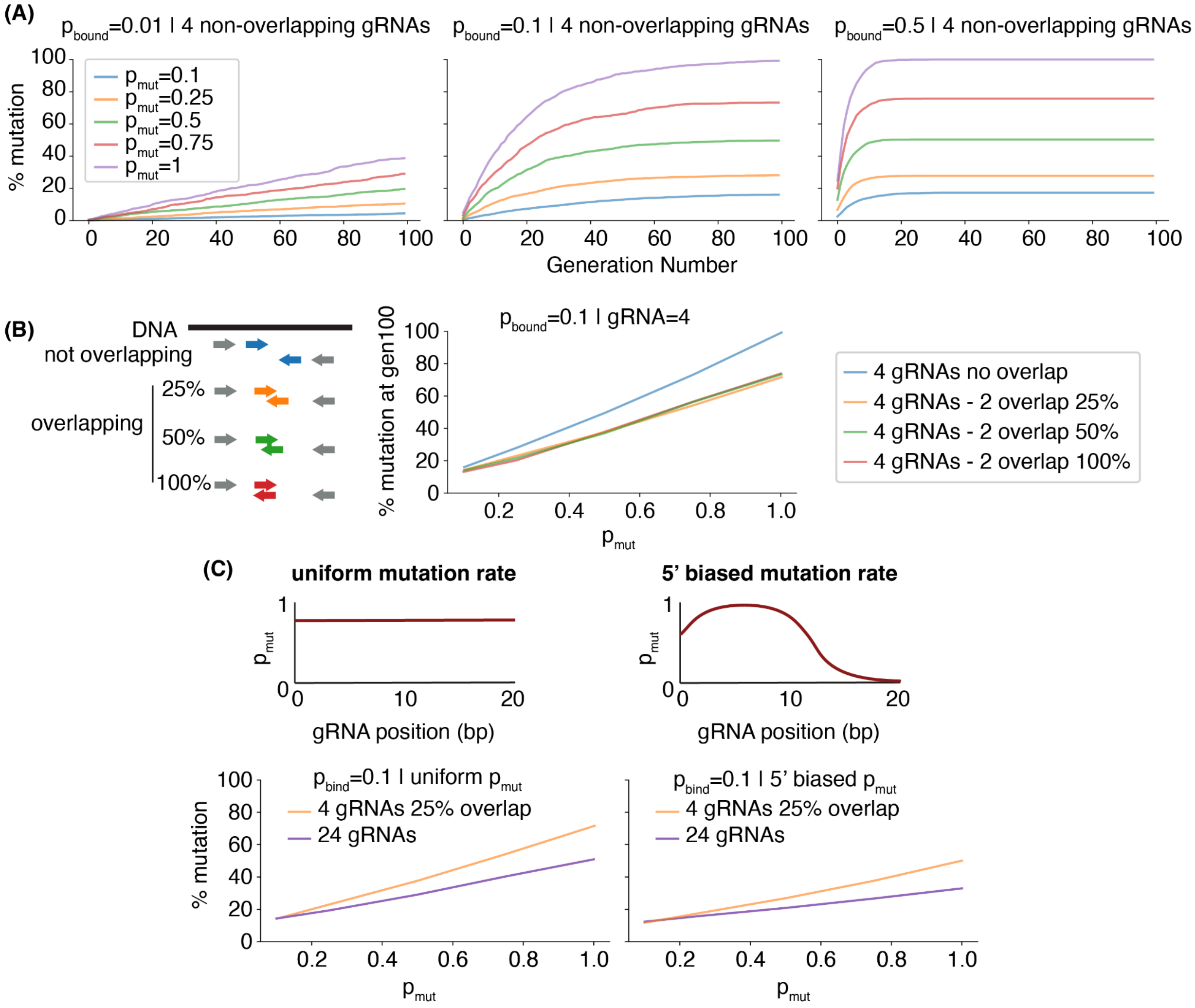

Figure S6. Simulating the effects of uniform and 5' biased mutation rates. (A) Average fraction of mutations at each generation when 4 gRNAs target GFP, for a range of p_{bound} (left to right, 0.01, 0.1 and 0.5) and p_{mut} (different colors) when mutation profiles were uniform along the gRNA (left) and when a 5' bias was incorporated (right). **(B)** Accumulated average fraction of mutations over time for different values of p_{bound} and p_{mut} for four non-overlapping gRNAs. Simulations of mutations accumulated in 200 cells over 100 generations. **(C)** Average fraction of mutations at the end of the simulation (generation 100) when different amounts of gRNA overlap were considered, for a range of p_{bound} (different colors) and p_{mut} (x-axis), in simulations using a uniform (left) or 5' biased (right) mutation rate along the gRNA.

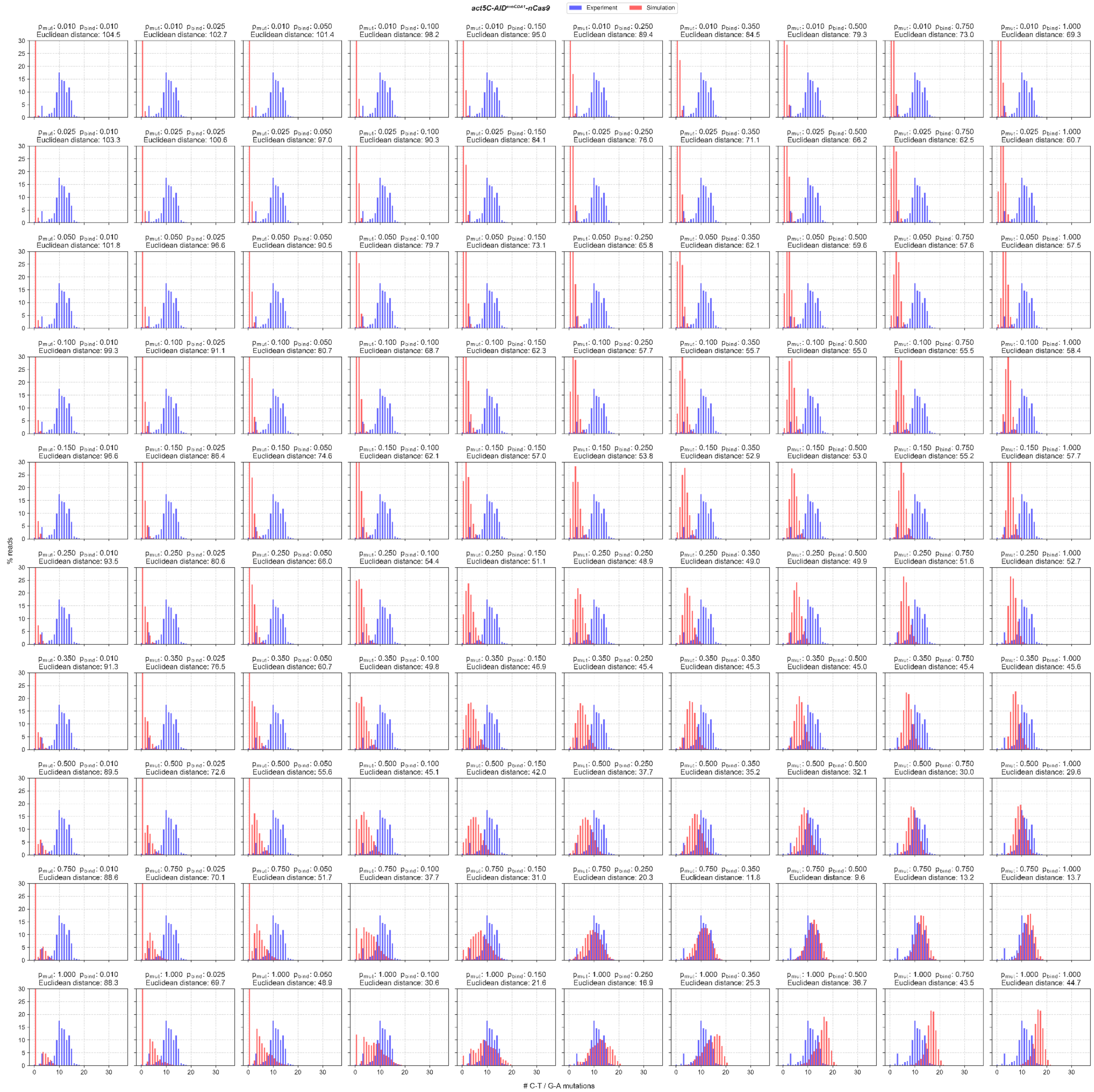

Figure S7. Fits between simulation and experiment for *act5C-AID^{evoCDA1}-nCas9-UGI*. Comparison of simulated histograms of number of introduced mutations per read (red) for a variety of p_{bound} (along columns) and p_{mut} values (along rows) with the histogram obtained from the high mutation rate experiment in embryos (*act5C-AID^{evoCDA1}-nCas9-UGI*). Euclidean distances between both histograms are shown on top of each plot.

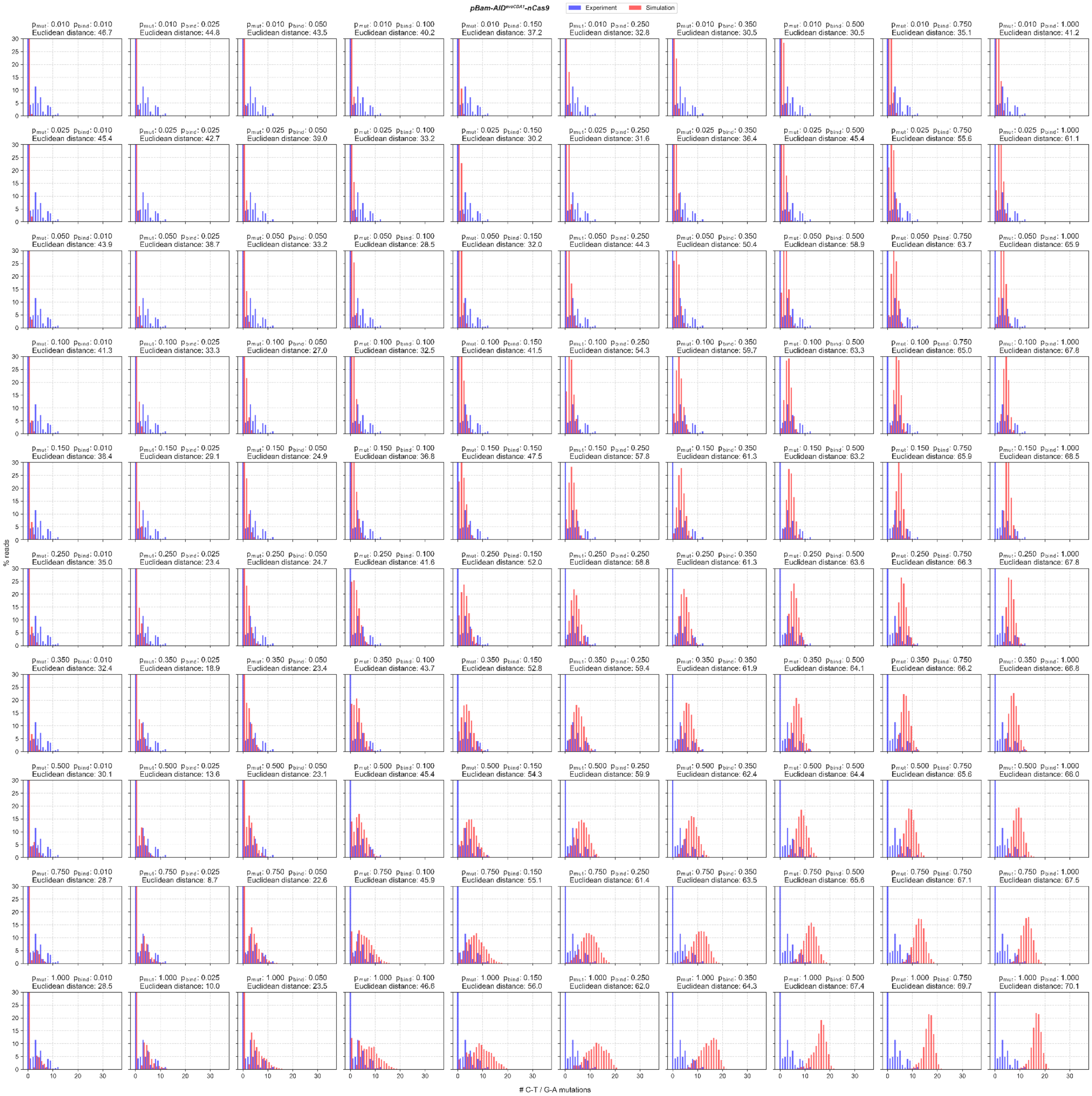

Figure S8. Fits between simulation and experiment for *pBam-AID^{evoCDA1}-nCas9-UGI*. Comparison of simulated histograms of number of introduced mutations per read (red) for a variety of p_{bound} (along columns) and p_{mut} values (along rows) with the histogram obtained from the low mutation rate experiment in embryos (*pBam-AID^{evoCDA1}-nCas9-UGI*). Euclidean distances between both histograms are shown on top of each plot.

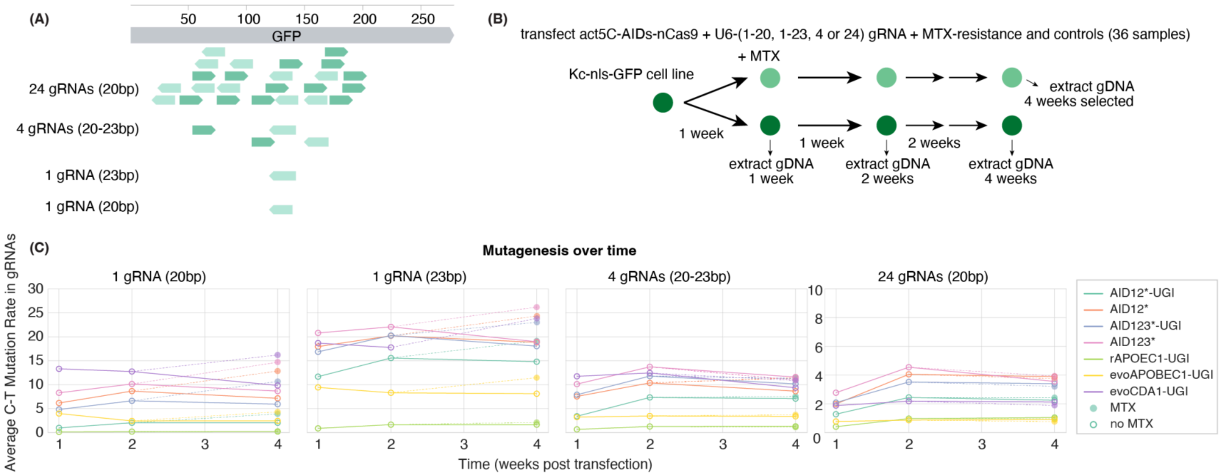

Figure S9. Overview of the experiment in cell culture testing different number and length of gRNA. (A) Scheme of the different gRNA array constructs used to target GFP. **(B)** Scheme of the timepoints where genomic DNA was collected from Kc-GFP lines after transfection of mutagenic enzymes and gRNAs. We attempted to make stable cell lines by co-transfecting with an MTX resistance plasmid and adding MTX (see Methods). **(C)** Average mutation rates obtained with each gRNA plasmid (left to right), mutagenic enzyme (lines of different colors) and timepoints (x-axis).

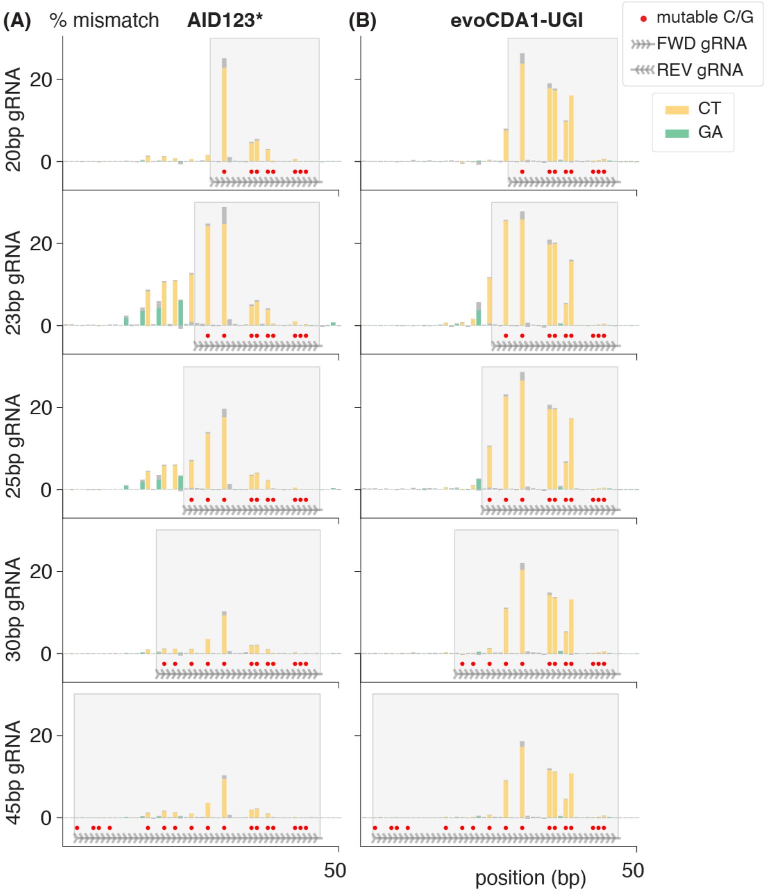

Figure S10. Increasing gRNA length beyond 23bp does not further increase efficiency. (A) Mismatches produced in GFP classified by base substitution in Kc cells transfected with act5C-AID123*-nCas9 and one gRNA of 20, 23, 25, 30 or 45bp in length, 1 week post-transfection. **(B)** Mismatches to GFP classified by base substitution in Kc cells transfected with act5C-AID^{evoCDA1}-nCas9 and one gRNA of 20, 23, 25, 30 or 45 bp in length, 1 week post-transfection.

Materials and Methods

Plasmids

Plasmids were generated using standard molecular biology techniques, either in house or by Genscript.

AID12* and AID123* or EvolvR (Addgene #83260) were kind gifts from Rahul Kohli (University of Pennsylvania) and John Dueber (UC Berkeley), respectively, and were cloned downstream of a *pBam* promoter and 5'UTR (Chen and McKearin, 2003) and upstream of nCas9 or enCas9 separated by a linker, and *bam 3'UTR* by Genscript, Inc (Piscataway, NJ) . These plasmids were further modified to replace *pBam* by *act5C* (González et al., 2011) or *UASz* (Deluca and Spradling, 2018), from the pUASz1.0 plasmid from DRSC #1431) and *bam 3'UTR* by *tub 3'UTR.*

Plasmids act5C-APOEC1-nCas-UGI and act5C-evoCDA1-nCas-UGI were kind gifts from Fillip Port (German Cancer ResearchCenter, Heidelberg University, Germany) (Doll et al., 2023).

pattB-Actin5C-BE2(pActin5c-CD-dCas9m4-UDI-attB) was obtained from Addgene# 104879 (Marr and Potter, 2021).

pCDF5-U6-[4xgRNA-tRNA]-GFP was generated by cloning the 4 gRNAs targeting GFP from (Ma et al., 2016) in pCDF5 (Addgene #Plasmid #73914) (Port and Bullock, 2016). pnos-nos5UTR-eGFP-tub3UTR was subsequently digested and cloned into pCDF5-U6-[4xgRNA-tRNA]-GFP to generate pCDF5-U6-[4xgRNA-tRNA]-GFP, nos-GFP.

Cloning of individual gRNAs in pCDF5 was performed by phosphorylating, annealing and ligating the product to *BbsI* digested pCDF5, following (Port et al., 2014).

Cloning of the 24xgRNAs in arrays (2 of 6 and one of 12), was achieved following (Port et al., 2014)as well. Briefly, primers containing gRNA sequences overgangs and sequence to amplify the gRNA scaffold and tRNA cassette were designed to assemble first 4 arrays of 6 gRNAs each. Each plasmid was assembled in a HiFi reaction (New England Biolabs) and transformed into NEB Stable bacteria (New England Biolabs) to avoid recombination. Recombination events were still observed so screening of multiple colonies was needed to find the 4 correctly assembled arrays. We then attempted to amplify by PCR and fuse the 4 arrays of 6 into 2 of 12, to subsequently create one of 24. Since these arrays were highly repetitive we were only successful in creating one of 12 gRNA, and therefore decided to transfect the 24 gRNAs in 3 plasmids (one array of 12 and 2 of 6).

All used primers can be found in Supplementary Table 1.

All plasmid sequences developed or used during this study and GFP sequences with annotated gRNAs can be found in this Benchling repository: benchling.com/juliafs/f_/NXjwqXma-optimizing-flybraries-paper/

Cell culture

In cell culture, Kc-nls-GFP (DGRC #227, Kc167-PP-93E (nls-GFP)) cells were maintained in cell media HyClone CCM3 and split twice a week to a density of 1.5 million cells/ml.

For the different experiments, Kc-nls-GFP cells were transfected with various nCas9 constructs driven by the *act5C* or *pMT* promoter and with plasmids expressing different gRNAs. We also co-transfected an MTX-resistance plasmid (p8HCO - DGRC #1003), aiming to select cells that had integrated the co-transfected plasmids with MTX and generate stable populations where mutating enzymes could be expressed long term (Hannon and Eisen, 2024; Rebay et al., 1991). Since integration rates are low and it can take 2 months to select stable populations, we anticipated not having generated completely stable populations by week 4 but expected to see increasing levels of mutations over time. If the mutagenic system was highly efficient, we could have also expected to see mutations saturating after one or two weeks in culture regardless of MTX selection. However very few cells survived MTX treatment, and mutation levels barely increased after MTX treatment or even decreased (Fig. S9). We therefore assumed MTX selection was not working but nevertheless sequenced samples at the different timepoints. Although the mutagenic enzymes were no longer active by week 4, we still expect to detect all the mutations generated in the genome while they were active.

In each condition, up to 750 ng of total DNA (containing the mutagenic enzyme and/or gRNA plasmid and the MTX resistance plasmid) was transfected in 0.75 million cells using the Effectene Transfection Reagent from Qiagen (cat. No 301425) in 24-well plates. The DNA was diluted in 75ul of EC buffer, mixed with 6ul of enhancer vortexed, and incubated at room temperature for 5 minutes. Afterwards, 8ul of Effectene reagent is added, vortexed, and incubated at room temperature for 15 minutes. The solution is mixed and added directly to the cells. The cells were incubated for several weeks and harvested for genomic DNA extraction. Methotrexate (MTX, Sigma-Aldrich M8407) was added at 0.1 µg/mL from week 1 post-transfection and half of the cells were kept in conditioned media with MTX. At the indicated time points (1, 2, or 4 weeks post-transfection) cells were harvested and genomic DNA extracted. In cells transfected with pMT plasmids, a low concentration of copper (1mM CuSO4) was added one day after transfection and cells were kept until harvesting at day 8 post-transfection.

Fly husbandry

Drosophila melanogaster flies were grown and maintained on Fly Food B from LabExpress (Ann Arbor, MI), of the following ingredients: Agar 0.56%, Cornmeal 6.71%, Inactivated Yeast 1.59%, Soy Flour 0.92%, Corn Syrup 7.0%, Propionic Acid 0.44%, Tegosept 0.15%.

Embryos were collected on apple juice agar plates (75% water, 25% apple juice, 22.5 g/Ll Bacto agar, 25 g/L sucrose, 2.5%b Nipagin M mold inhibitor 20% diluted in Ethanol, CHSL protocol doi:10.1101/pdb.rec065672) with yeast paste. Animals of both sexes were used for this study.

Fly Strains and Genetics

The following transgenic lines were created by injection in phiC31 ; attP lines by BestGene (Chino Hills, CA) or RainbowTransgenic Flies, Inc. (Camarillo, CA):

yv ; pCFD5-GFP-4sgRNA-pNos-nos5UTR-eGFP-tub3UTR [VK22]
yw;+; pBam-bam5'UTR-enCas9-DNAPolI5M(EvolvR)-bam3'UTR [VK33]
yw;+; pBam-bam5'UTR-NLS-enCas9-DNAPolI5M(EvolvR)-bam3'UTR [VK33]
yw;+; pBam-bam5'UTR-enCas9-DNAPolI5M(EvolvR)-NLS-bam3'UTR [VK33]
yw;+; pBam-bam5'UTR-AID12*-nCas9-NLS-UGI-bam3'UTR [VK33]
yw;+; pBam-bam5'UTR-AID12*-nCas9-NLS-bam3'UTR [VK33]
yw;+; pBam-bam5'UTR-AID123*-nCas9-NLS-UGI-bam3'UTR [VK33]
yw;+; pBam-bam5'UTR-AID123*-nCas9-NLS-bam3'UTR [VK33]
yw;+; act5C-AID12*-nCas9-NLS-tub3'UTR [VK33]
yw;+; act5C-AID123*-nCas9-NLS-tub3'UTR [VK33]
yw;+; UASz-AID12*-nCas9-NLS-tub3'UTR [VK33]
yw;+; UASz-AID123*-nCas9-NLS-tub3'UTR [VK33]
yw;+; pBam-bam5'UTR-dCBEevoCDA1-nCas9-UGI-SV40 [VK33]

The following lines were obtained from Phillip Boutros (Doll et al., 2023) and BDSC:
act5C-act5C5’UTR-dCBEevoCDA1-nCas9-UGI-SV40 [attP40]
act5C-act5C5’UTR-APOEC1-nCas9-UGI-SV40 [attP40]
act5C-BE2(APOEC1-dCas9) (BDSC #92586)
osk-Gal4 (w[1118]; P{w[+mC]=osk-GAL4::VP16}F/TM3, Sb[1] - BDSC # 44242)
bam-Ga4 (y[1] w[*] P{w[+mC]=bam-GAL4:VP16}1 - BDSC # 80579)
vasa-Cas9 (w[*]; PBac{y[+mDint2] RFP[DsRed.OpIE2]=vas-Cas9.T2A.GFP}VK00033 - BDSC # 79006)

In all experiments where mutagenesis occurred for one generation, male or female nos-GFP, U6-4xgRNAs flies were crossed with male or female EvolvR-nCas9 or AID-nCas9 flies. All F1 males and females nos-GFP, U6-4xgRNAs / + ; EvolvR-nCas9 / + or nos-GFP, U6-4xgRNAs / + ; AID-nCas9 / + were placed in a cage and F2 embryos collected for DNA extraction.

In experiments where AID was left to mutate for multiple generations, stable stocks of nos-GFP, U6-4xgRNAs / (CyO) ; AID-nCas9 / (TM3) were created and maintained for multiple generations.

For the control with catalytically active Cas9, nos-GFP, U6-4xgRNAs was crossed with vasa-Cas9 and F1 males and females nos-GFP, U6-4xgRNAs / + ; vasa-Cas9 placed in a cage and F2 embryos collected for DNA extraction. Embryos could have received a GFP allele from either males or females, but only females expressed Cas9. Therefore obtained mutation rates to GFP are underestimated by half.

For the experiment to distinguish male and germline mutations, mutator parents were obtained as described above, by crossing males yv ; nos-GFP, U6-4xgRNAs with females yw ; act5C-evoCDA1-nCas9-UGI. To obtain mutations only generated in the male germline yw / Y ; nos-GFP, U6-4xgRNAs / act5C-evoCDA1-nCas9-UGI were crossed with yw females. white-eyed adult males and females (which could only be yw ; nos-GFP, U6-4xgRNAs-v; + / +) were selected and sequenced. To obtain mutations only generated in the female germline yw / yw ; nos-GFP, U6-4xgRNAs / act5C-evoCDA1-nCas9-UGI females were crossed with yw males. white-eyed adults (which could be yw ; nos-GFP,U6-4xgRNAs-v+ / + or yw ; + / + if recombination had occurred) were selected and sequenced. Although this was a mix of flies harboring nos-GFP,U6-4xgRNAs-v+ and not, GFP could only PCR-amplify from flies that harbored GFP, and which had been exposed to mutagenesis in the germline of its father.

gDNA extraction, amplicon sequencing and library preparation

Genomic DNA from Kc cells was extracted using Qiagen's DNeasy Blood & Tissue Kit (Cat. No. 69504) or the DNeasy Blood & Tissue QIAcube Kit (Cat. No. 69516). Genomic DNA from embryos was extracted by blending embryos in PBS using a pestle and pestle mixer and Qiagen's DNeasy Blood & Tissue Kit (Cat. No. 69504). Genomic DNA extraction from adult flies was performed by phenol:chloroform precipitation. Briefly flies were ground in 2.5ml of Tris HCl 0.1 M (pH 9.0) EDTA 0.1 M SDS 1% solution using Fisherbrand 15ml disposable tissue grinders (Cat no. 02-542-09) on ice. The mix was then incubated at 70C for 30 minutes. 10 ul of RNAse A were added and incubated at 37C for 15 minutes. 350 ul of KAc were added, inverted to mix and incubated on ice for 30 minutes. To remove debris, the mix was centrifuged for 15 minutes at 20G rpm in a table centrifuge at 4C. The supernatant was kept and then added 1:1 to phenol:chloroform (Sigma-Aldrich P1944). The mix was centrifuged for 5 minutes at 20G. The top layer was transferred to a new tube, 1:1 phenol:chloroform added again and centrifuged again. To precipitate DNA the top later was mixed with 0.7 volumes of isopropanol and centrifuged for 20 min at 20k G at 4C. The pellet was washed for 5 min in 70% Ethanol, left to dry and resuspended in 100 ul to 200 ul TE (10mM Tris-HCl 1mM EDTA).

Following DNA extraction, for Nanopore libraries, GFP was amplified using primers *CTCTTTCCCTACACGACGCTCTTCCGATCTtgcttcagccgctaccccgaccacatgaag* and *ACTGGAGTTCAGACGTGTGCTCTTCCGATCttgatgccgttcttctgcttgtcggccatg* that add sequences equivalent to NEB adaptor and NEB Ultra II Q5 polymerase. PCR products were purified during AMXPure beads. A second round of PCR was performed using single index Illumina primers from New England Biolabs (Cat no. E7335S, E7500S, E7710S and E7730S). The Nanopore 9.4 ligation kit (Cat no. 110) or 10.4 ligation kit (Cat no. 114) was used to ligate the nanopore adaptors and libraries were loaded on 9.4 (Cat no. FLO-MIN106D) or 10.4 (Cat no. FLO-MIN114) flow cells using the standard protocol and sequenced on MinION Mk1B devices .

Analysis of sequenced Nanopore libraries

Basecalling was performed using guppy with GPU settings and ‘super-accurate’ basecalling, in a desktop computer with an Intel(R) Xeon(R) Gold 6134 CPU @ 3.20GHz, 64GB RAM and NVIDIA Quadro P4000 graphic card. For 9.4 Nanopore flowcells the base calling command was:

```
"C:\Program Files\OxfordNanopore\ont-guppy\bin\guppy_basecaller.exe" --input_path
[INPUT_PATH] --save_path [SAVE_PATH] -r --config dna_r9.4.1_450bps_sup.cfg
--compress_fastq --verbose_logs --device cuda:0 --num_callers 8 --chunks_per_caller 1000
```

and for 10.4 flowcells:

```
"C:\Program Files\OxfordNanopore\ont-guppy\bin\guppy_basecaller.exe" --input_path
[INPUT_PATH]\ --save_path [SAVE_PATH] -r --config dna_r10.4.1_e8.2_260bps_sup.cfg
--compress_fastq --verbose_logs --device cuda:0 --num_callers 8 --chunks_per_caller 1000
```

Nanopore reads were first filtered for an average quality of 10, aligned to GFP using BioPython (with alignment scores of match = 3, mismatch = 0, gap_opening = -3, gap_extension = 0) and filtered again for an average alignment score >600. All samples contained a high degree of mismatches to GFP, consistent with the average 9.4 or 10.4 sequencing quality of Nanopore sequencing and simplex basecalling, so these parameters were adjusted to recover most alignments to GFP. Mismatch and gap_extension were set to 0 to not penalize reads with introduced mutations or Nanopore sequencing and basecalling errors. However, the type and location of the mismatches seemed to be consistent across samples once enough reads were averaged, suggesting that this “sequencing/base calling error background” is sequence dependent and reproducible across samples, and consistent with previous observations. We therefore used the average mismatch profiles from control samples to subtract from experimental samples and obtain mismatch and indel profiles specific to the mutagenic enzymes used (Fig. S2). Average mutation rates were calculated by counting the number of “mutable bases” (Cs on FWD gRNAs and Gs on REV gRNAs) that were within the 20 or 23bp of the gRNA target and dividing the average C-T or G-A mismatch rate by this number of “mutable bases”. Histograms showing the number of mutations per read were calculated by counting the number of C→T or G→A mutations on gRNA target sequences across all samples (Fig. 2BD, 3BD, 5AB, 6AB, S4, S5BC, S10). The histograms in control samples showed the expected number of mismatches due to sequencing and base calling errors, and were highly skewed towards a low (1-3) number of mismatches. Control histograms were subtracted from experimental histograms to obtain an estimation of the number of mutations per read caused by the mutagenic enzymes (Fig. 2BD, 3BD, S5BC). Mutation profiles along gRNAs were calculated by averaging C→T and G→A mutation rates over all gRNAs used in the experiment aligned by the PAM sequence (Fig. 5C, 6C).

Code used for sequence analysis can be found at github.com/juliafs93/CountMutations_Flybraries

Simulations

Simulations of AID-nCas9 mutagenesis were performed on Python using simple probability calculations. We started from a wild type sequence and a list of gRNAs targeting this sequence. A probability of each gRNA binding to each target was calculated by generating a random number between 0 and 1. If the obtained number was lower than p_{bound} , that gRNA would bind. If bound , AID would be allowed to mutate each C in that gRNA with a probability p_{mut}. Similarly, random numbers between 0 and 1 were calculated for each C on the gRNA sequence (or G in the reverse gRNA) to be mutated. Those with a value lower than p_{mut} resulted in mutations. In "uniform mutation profile" simulations, p_{mut} was constant along the gRNA position. In "5' biased mutation profile" simulations, p_{mut} was scaled by a smoothed distribution obtained from mutation profiles from experimental data, manually fitted from the experimental mutation profiles. This resulted in higher p_{mut} at the 5' end and lower p_{mut} at the 3' end, leading to a lower probability of mutating in the 3' half of the gRNA. The same process was repeated for every "cell" (n = 200) and for a set number of generations (100, although the parameters are unitless and refer to probability of binding/mutation per generation). Average mutation rates and distribution of the number of mutations was calculated and plotted similarly to the experiments.

Code used for simulations can be found at: github.com/juliafs93/SimulatingFlybraries

Analysis of sequenced Illumina libraries

Illumina libraries were sequenced in a MiSeq2 instrument with a MiSeq V2 Nano kit (Illumina, San Diego CA) by the Genomics Facility at the CZI Biohub (San Francisco, CA). Libraries were sequenced paired end (2x150 cycles) and de-multiplexed reads merged using BBMerge *(Bushnell et al., 2017)*. These were the inputs for similar scripts as for Nanopore data, where reads were filtered for an average quality of 30, filtered again by alignment score to GFP and each type of mismatch calculated. The code can be obtained here: github.com/juliafs93/CountMutations_Flybraries

Quantification mutational variability

From this point, as opposed to the analysis of Nanopore libraries, mutations were considered on a read by read basis. First we calculated whether gRNAs 1, 2, 3 or 4 were mutated on each read, considering them mutated if there was at least one C-to-T mismatch in forward gRNAs (1 and 2) or at least one G-to-A mismatch in reverse gRNAs (3 and 4). From these we calculated what was the average mutation rate of each gRNA across all reads and the distribution of the number of mutated gRNAs in each read. We used the mutation rate associated with each gRNA to calculate what the distribution of the number of mutated gRNAs would be if each event was assumed independent, which can be compared with the actual one measured in experiments.

As an analogy, in a scenario where all gRNAs had a 25% mutation rate, completely correlated mutation events would result in 25% of reads containing all 4 mutated gRNAs and 75% containing none. Completely anticorrelated binding would result in 100% reads containing only one mutated gRNA each. Uncorrelated events (expected by chance) would result in 1/256 reads containing all 4 mutations, 4/64-1/256 containing exactly 3 mutations and so on for probability of obtaining exactly 2 and 1 mutated gRNAs. We used the same principle but using different probabilities for each of the 4 events.

This is equivalent to calculating the probability that after 4 coin flips we obtain 0, 1, 2, 3 or 4 tails, but instead of 50:50 each coin has a distinct probability - the mutation rates calculated for each gRNA across all reads. This is equivalent to the Poisson binomial distribution, which is the probability distribution of the number of successes in a collection of n independent yes/no events with distinct success probabilities (Wadycki et al., 1973; Wikipedia contributors, 2024):

$$\Pr(K=k)=\begin{cases}\prod\limits_{i=1}^{n}(1-p_i) & k=0\\ \frac{1}{k}\sum\limits_{i=1}^{k}(-1)^{i-1}\Pr(K=k-i)T(i) & k>0\end{cases}$$

, where *k* is the exact number of successful events, and *p* is a vector of probabilities for each event, assumed Poissonian and independent, and:

$$T(i)=\sum_{j=1}^{n}\left(\frac{p_j}{1-p_j}\right)^{i}$$

The exact probability of exactly *k* gRNAs being mutated was then calculated by convolution of the vector *p*, using code similar to (Biscarri et al., 2018; Wikipedia contributors, 2024).

We also used Mutual Information theory to calculate if two events (each two gRNAs mutated on the same strand or each two Cs mutated on the same gRNA) were correlated, uncorrelated or anticorrelated. For each pair of events we calculated the normalized pointwise mutual information (pointwise mutual information normalized by the joint self-information), which results in -1 for 2 events never occurring together, 0 for independence, and +1 for complete co-occurrence (Bouma, n.d.).

$$npmi(x;y)=\frac{pmi\,(x;y)}{h\,(x,y)}=\frac{log_2\frac{p(x,y)}{p(x)p(y)}}{-log_2\,p(x,y)}$$

, where *p(x,y)* was the fraction of reads with the two events co-ocurring and *p(x)* and *p(y)* the fraction of events with either event occurring.

The code used for the analysis of correlations can be obtained here: github.com/oliviarourke/Between-gRNA-Mutation-Correlation and github.com/oliviarourke/Within-gRNA-Mutation-Correlation

Fitting of simulated data to experiments

To fit which combination of p_{bound} and p_{mut} values better fit the experiment, we carried simulations of 10,000 reads with 10 p_{bound} and p_{mut} values between 0 and 1, and calculated the number of mutations generated in each read, resulting in a grid of 100 histograms of the expected distribution of the number of mutations. We then compared these histograms to the ones obtained in the experiment. A visual comparison of the histograms made it clear what area of parameters was likely the best fit. The best fit was then obtained by calculating the Euclidean distances between both histograms and selecting the one with the lowest value. We tested other methods, such as chi-square or Kolmogorov–Smirnov tests, but we note these were less reliable because the histogram values are considered as part of a distribution and lose their positional information. In this situation, comparing the values of both histograms in each position was important to decide which was the best fit.

The code used for the fitting between simulations and experiments can be found here: github.com/oliviarourke/simulations_vs_data

Acknowledgements

We would like to thank Steve Quake for the initial discussion that launched this project. We also acknowledge Rahul Kohli**,** John Dueber, Fillip Port and Rob Phillips, as well as their lab members, for extensive discussions and reagents. We thank members of the Garcia and Eisen labs for helpful discussion and feedback. J. F-S was supported by EMBO (ALTF 1267-2020) and HFSP (LT000310/2021-L) Long Term Postdoctoral Fellowships. Y. D-T was supported by the PREP UC Berkeley/CND Post Baccalaureate Program (funded by the US National Institutes of Health via R25GM140276). M.A.T. was supported by the NSF Graduate Research Fellowship. J.D and O.R. were supported by the Cal-Bridge Summer Program. C. M was supported by a Chancellor's Fellowship from UC Berkeley's Office for Graduate Diversity. H.G.G. was supported by an NIH R01 Award (R01GM139913), by the Koret-UC Berkeley-Tel Aviv University Initiative in Computational Biology and Bioinformatics, and by a Winkler Scholar Faculty Award. H.G.G. is also a Chan Zuckerberg Biohub Investigator (Biohub – San Francisco). M.B.E. was supported by the Howard Hughes Medical Institute.

REFERENCES

Álvarez B, Mencía M, de Lorenzo V, Fernández LÁ. 2020. In vivo diversification of target genomic sites using processive base deaminase fusions blocked by dCas9. *Nat Commun* **11**:1–14. doi:10.1038/s41467-020-20230-z

Anzalone AV, Koblan LW, Liu DR. 2020. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat Biotechnol* **38**:824–844. doi:10.1038/s41587-020-0561-9

Arnold CD, Gerlach D, Stelzer C, Boryń ŁM, Rath M, Stark A. 2013. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**:1074–7. doi:10.1126/science.1232542

Barolo S, Posakony JW. 2002. Three habits of highly effective signaling pathways: principles of transcriptional control by developmental cell signaling. *Genes Dev* **16**:1167–1181. doi:10.1101/gad.976502

Berríos KN, Barka A, Gill J, Serrano JC, Bailer PF, Parker JB, Evitt NH, Gajula KS, Shi J, Kohli RM. 2024. Cooperativity between Cas9 and hyperactive AID establishes broad and diversifying mutational footprints in base editors. *Nucleic Acids Res* gkae024. doi:10.1093/nar/gkae024

Biscarri W, Zhao SD, Brunner RJ. 2018. A simple and fast method for computing the Poisson binomial distribution function. *Comput Stat Data Anal* **122**:92–100. doi:10.1016/j.csda.2018.01.007

Bosch JA, Birchak G, Perrimon N. 2020. Precise genome engineering in Drosophila using prime editing. *Proc Natl Acad Sci U S A* **118**:1–9. doi:10.1073/pnas.2021996118

Bouma G. n.d. Normalized (Pointwise) Mutual Information in Collocation Extraction.

Brown AR, Fox GA, Kaplow IM, Lawler AJ, Phan BN, Wirthlin ME, Ramamurthy E, May GE, Chen Z, Su Q, McManus CJ, Pfenning AR. 2022. An *in vivo* massively parallel platform for deciphering tissue-specific regulatory function (preprint). Genomics. doi:10.1101/2022.11.23.517755

Bunch TA, Grinblat Y, Goldstein LSB. 1998. Characterization and use of the Drosophila metaliothionein promoter in cultured Drosophia melanogaster cells. *Nucleic Acids Res* **16**:1043–1061.

Bushnell B, Rood J, Singer E. 2017. BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE* **12**:e0185056. doi:10.1371/journal.pone.0185056

Chan Y-C, Kienle E, Oti M, Di Liddo A, Mendez-Lago M, Aschauer DF, Peter M, Pagani M, Arnold C, Vonderheit A, Schön C, Kreuz S, Stark A, Rumpel S. 2023. An unbiased AAV-STARR-seq screen revealing the enhancer activity map of genomic regions in the mouse brain in vivo. *Sci Rep* **13**:6745. doi:10.1038/s41598-023-33448-w

Chen D, McKearin DM. 2003. A discrete transcriptional silencer in the bam gene determines asymmetric division of the Drosophila germline stem cell. *Development* **130**:1159–1170. doi:10.1242/dev.00325

Chen F, Lian M, Ma B, Gou S, Luo X, Yang K, Shi H, Xie J, Ge W, Ouyang Z, Lai C, Li N, Zhang Q, Jin Q, Liang Y, Chen T, Wang J, Zhao X, Li L, Yu M, Ye Y, Wang K, Wu H, Lai L. 2022. Multiplexed base editing through Cas12a variant-mediated cytosine and adenine base editors. *Commun Biol* **5**:1163. doi:10.1038/s42003-022-04152-8

Chen H, Liu S, Padula S, Lesman D, Griswold K, Lin A, Zhao T, Marshall JL, Chen F. 2020. Efficient, continuous mutagenesis in human cells using a pseudo-random DNA editor. *Nat Biotechnol* **38**:165–168. doi:10.1038/s41587-019-0331-8

Chen HM, Marques JG, Sugino K, Wei D, Miyares RL, Lee T. 2020. CAMIO: a transgenic CRISPR pipeline to create diverse targeted genome deletions in Drosophila. *Nucleic Acids Res* **48**:4344–4356. doi:10.1093/nar/gkaa177

Corsi GI, Qu K, Alkan F, Pan X, Luo Y, Gorodkin J. 2022. CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context. *Nat Commun* **13**:3006. doi:10.1038/s41467-022-30515-0

Cravens A, Jamil OK, Kong D, Sockolosky JT, Smolke CD. 2021. Polymerase-guided base editing enables in vivo mutagenesis and rapid protein engineering. *Nat Commun* **12**:1–12. doi:10.1038/s41467-021-21876-z

Cuykendall TN, Rubin MA, Khurana E. 2017. Non-coding genetic variation in cancer. *Curr Opin Syst Biol* **1**:9–15. doi:10.1016/j.coisb.2016.12.017

de Almeida BP, Reiter F, Pagani M, Stark A. 2022. DeepSTARR predicts enhancer activity from DNA sequence and enables the de novo design of synthetic enhancers. *Nat Genet* **54**:613–624. doi:10.1038/s41588-022-01048-5

Deluca SZ, Spradling AC. 2018. Efficient expression of genes in the drosophila germline using a uas promoter free of interference by hsp70 pirnas. *Genetics* **209**:381–387. doi:10.1534/genetics.118.300874

Doll RM, Boutros M, Port F. 2023. A temperature-tolerant CRISPR base editor mediates highly efficient and precise gene editing in *Drosophila*. *Sci Adv* **9**:eadj1568. doi:10.1126/sciadv.adj1568

Farley EK, Olson KM, Zhang W, Brandt AJ, Rokhsar DS, Levine MS. 2015. Suboptimization of developmental enhancers. *Science* **350**:325–328. doi:10.1126/science.aac6948

Fu J, Li Q, Liu X, Tu T, Lv X, Yin X, Lv J, Song Z, Qu J, Zhang J, Li J, Gu F. 2021. Human cell based directed evolution of adenine base editors with improved efficiency. *Nat Commun* **12**:1–11. doi:10.1038/s41467-021-26211-0

González M, Martín-Ruíz I, Jiménez S, Pirone L, Barrio R, Sutherland JD. 2011. Generation of stable Drosophila cell lines using multicistronic vectors. *Sci Rep* **1**. doi:10.1038/srep00075

Gu B, Swigut T, Spencley A, Bauer MR, Chung M, Meyer T, Wysocka J. 2018. Transcription-coupled changes in nuclear mobility of mammalian cis-regulatory elements. *Science* **3136**:eaao3136. doi:10.1126/science.aao3136

Hannon CE, Eisen MB. 2024. Intrinsic protein disorder is insufficient to drive subnuclear clustering in embryonic transcription factors. *eLife* **12**:RP88221. doi:10.7554/eLife.88221

Hempel LU, Kalamegham R, Smith JE, Oliver B. 2008. Chapter 4 Drosophila Germline Sex Determination: Integration of Germline Autonomous Cues and Somatic SignalsCurrent Topics in Developmental Biology. Elsevier. pp. 109–150. doi:10.1016/S0070-2153(08)00404-3

Howard ML, Davidson EH. 2004. cis-Regulatory control circuits in development. *Dev Biol* **271**:109–118. doi:10.1016/j.ydbio.2004.03.031

Inoue F, Ahituv N. 2015. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**:159–164. doi:10.1016/j.ygeno.2015.06.005

Ireland WT, Beeler SM, Flores-Bautista E, McCarty NS, Röschinger T, Belliveau NM, Sweredoski MJ, Moradian A, Kinney JB, Phillips R. 2020. Deciphering the regulatory genome of Escherichia coli, one hundred promoters at a time. *eLife* **9**:1–76. doi:10.7554/ELIFE.55308

Jinek M, East A, Cheng A, Lin S, Ma E, Doudna J. 2013. RNA-programmed genome editing in human cells. *eLife* **2**:e00471. doi:10.7554/eLife.00471

Jores T, Tonnies J, Dorrity MW, Cuperus J, Fields S, Queitsch C. 2020. Identification of Plant Enhancers and Their Constituent Elements by STARR-seq in Tobacco Leaves. *Plant Cell* tpc.00155.2020. doi:10.1105/tpc.20.00155

Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, Alston J, Mikkelsen TS, Kellis M. 2013. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res* **23**:800–811. doi:10.1101/gr.144899.112

Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, Costello JF, Shendure J, Ahituv N. 2019. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* **10**:3583. doi:10.1038/s41467-019-11526-w

Koblan LW, Erdos MR, Wilson C, Cabral WA, Levy JM, Xiong Z-M, Tavarez UL, Davison LM, Gete YG, Mao X, Newby GA, Doherty SP, Narisu N, Sheng Q, Krilow C, Lin CY, Gordon LB, Cao K, Collins FS, Brown JD, Liu DR. 2021. In vivo base editing rescues Hutchinson–Gilford progeria syndrome in mice. *Nature* **589**:608–614. doi:10.1038/s41586-020-03086-7

Kohli RM, Schutsky EK, Liu MY. 2021. United States Patent : US010961525B2 HYPERACTIVE AID / APOBEC AND HMC DOMINANT TET ENZYMES.

Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**:420–424. doi:10.1038/nature17946

Lagunas T, Plassmeyer SP, Fischer AD, Friedman RZ, Rieger MA, Selmanovic D, Sarafinovska S, Sol YK, Kasper MJ, Fass SB, Aguilar Lucero AF, An J-Y, Sanders SJ, Cohen BA, Dougherty JD. 2023. A Cre-dependent massively parallel reporter assay allows for cell-type specific assessment of the functional effects of non-coding elements in vivo. *Commun Biol* **6**:1151. doi:10.1038/s42003-023-05483-w

Lalanne J-B, Regalado SG, Domcke S, Calderon D, Martin BK, Li X, Li T, Suiter CC, Lee C, Trapnell C, Shendure J. 2024. Multiplex profiling of developmental cis-regulatory elements with quantitative single-cell expression reporters. *Nat Methods*. doi:10.1038/s41592-024-02260-3

Lelli KM, Slattery M, Mann RS. 2012. Disentangling the many layers of eukaryotic transcriptional regulation. *Annu Rev Genet* **46**:43–68. doi:10.1146/annurev-genet-110711-155437

Levine M. 2010. Transcriptional enhancers in animal development and evolution. *Curr Biol* **20**:R754–R763. doi:10.1016/j.cub.2010.06.070

Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. 2016. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat Methods* **13**:1029–1035. doi:10.1038/nmeth.4027

Marr E, Potter CJ. 2021. Base Editing of Somatic Cells Using CRISPR–Cas9 in Drosophila. *CRISPR J* **4**:1–10. doi:10.1089/crispr.2021.0062

Moore CL, Papa LJ, Shoulders MD. 2018. A Processive Protein Chimera Introduces Mutations across Defined DNA Regions *In Vivo*. *J Am Chem Soc* **140**:11560–11564. doi:10.1021/jacs.8b04001

Musunuru K, Chadwick AC, Mizoguchi T, Garcia SP, DeNizio JE, Reiss CW, Wang K, Iyer S, Dutta C, Clendaniel V, Amaonye M, Beach A, Berth K, Biswas S, Braun MC, Chen H-M, Colace TV, Ganey JD, Gangopadhyay SA, Garrity R, Kasiewicz LN, Lavoie J, Madsen JA, Matsumoto Y, Mazzola AM, Nasrullah YS, Nneji J, Ren H, Sanjeev A, Shay M, Stahley MR, Fan SHY, Tam YK, Gaudelli NM, Ciaramella G, Stolz LE, Malyala P, Cheng CJ, Rajeev KG, Rohde E, Bellinger AM, Kathiresan S. 2021. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. *Nature* **593**:429–434. doi:10.1038/s41586-021-03534-y

Newby GA, Yen JS, Woodard KJ, Mayuranathan T, Lazzarotto CR, Li Y, Sheppard-Tillman H, Porter SN, Yao Y, Mayberry K, Everette KA, Jang Y, Podracky CJ, Thaman E, Lechauve C, Sharma A, Henderson JM, Richter MF, Zhao KT, Miller SM, Wang T, Koblan LW, McCaffrey AP, Tisdale JF, Kalfa TA, Pruett-Miller SM, Tsai SQ, Weiss MJ, Liu DR. 2021. Base editing of haematopoietic stem cells rescues sickle cell disease in mice. *Nature* **595**:295–302. doi:10.1038/s41586-021-03609-w

Olsen TR, Talla P, Furnari J, Bruce JN, Canoll P, Zha S, Sims PA. 2023. Scalable co-sequencing of RNA and DNA from individual nuclei. doi:10.1101/2023.02.09.527940

Pan RW, Röschinger T, Faizi K, Phillips R. 2024. Dissecting endogeneous genetic circuits from first principles. doi:10.1101/2024.01.28.577658

Park H, Kim S. 2021. Gene-specific mutagenesis enables rapid continuous evolution of enzymes *in vivo*. *Nucleic Acids Res* **49**:e32–e32. doi:10.1093/nar/gkaa1231

Port F, Bullock SL. 2016. Augmenting CRISPR applications in Drosophila with tRNA-flanked sgRNAs. *Nat Methods* **13**:852–854. doi:10.1038/nmeth.3972

Port F, Chen H-M, Lee T, Bullock SL. 2014. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in *Drosophila*. *Proc Natl Acad Sci* **111**. doi:10.1073/pnas.1405500111

Qi Z, Jung C, Bandilla P, Ludwig C, Heron M, Kiesel AS, Philippou-Massier J, Nikolov M, Renna A, Schnepf M, Unnerstall U, Soeding J, Gaul U. 2020. Large-scale analysis of Drosophila core promoter function using synthetic promoters. *bioRxiv* **49**:2020.10.15.339325.

Rebay I, Fleming RJ, Fehon RG, Cherbas L, Cherbas P, Artavanis-Tsakonas S. 1991. Specific EGF repeats of Notch mediate interactions with Delta and serrate: Implications for notch as a multifunctional receptor. *Cell* **67**:687–699. doi:10.1016/0092-8674(91)90064-6

Rheinbay E, Nielsen MM, Abascal F, Wala JA, Shapira O, Tiao G, Hornshøj H, Hess JM, Juul RI, Lin Z, Feuerbach L, Sabarinathan R, Madsen T, Kim Jaegil, Mularoni L, Shuai S, Lanzós A, Herrmann C, Maruvka YE, Shen C, Amin SB, Bandopadhayay P, Bertl J, Boroevich KA, Busanovich J, Carlevaro-Fita J, Chakravarty D, Chan CWY, Craft D, Dhingra P, Diamanti K, Fonseca NA, Gonzalez-Perez A, Guo Q, Hamilton MP, Haradhvala NJ, Hong C, Isaev K, Johnson TA, Juul M, Kahles A, Kahraman A, Kim Y, Komorowski J, Kumar K, Kumar S, Lee Donghoon, Lehmann K-V, Li Y, Liu EM, Lochovsky L, Park Keunchil, Pich O, Roberts ND, Saksena G, Schumacher SE, Sidiropoulos N, Sieverling L, Sinnott-Armstrong N, Stewart C, Tamborero D, Tubio JMC, Umer HM, Uusküla-Reimand L, Wadelius C, Wadi L, Yao X, Zhang C-Z, Zhang Jing, Haber JE, Hobolth A, Imielinski M, Kellis M, Lawrence MS, Von Mering C, Nakagawa H, Raphael BJ, Rubin MA, Sander C, Stein LD, Stuart JM, Tsunoda T, Wheeler DA, Johnson R, Reimand J, Gerstein M, Khurana E, Campbell PJ, López-Bigas N, PCAWG Drivers and Functional Interpretation Working Group, Abascal F, Amin SB, Bader GD, Bandopadhayay P, Barenboim J, Beroukhim R, Bertl J, Boroevich KA, Brunak S, Campbell PJ, Carlevaro-Fita J, Chakravarty D, Chan CWY, Chen K, Choi JK, Deu-Pons J, Dhingra P, Diamanti K, Feuerbach L, Fink JL, Fonseca NA, Frigola J, Gambacorti-Passerini C, Garsed DW, Gerstein M, Getz G, Guo Q, Gut IG, Haan D, Hamilton MP, Haradhvala NJ, Harmanci AO, Helmy M, Herrmann C, Hess JM, Hobolth A, Hodzic E, Hong C, Hornshøj H, Isaev K, Izarzugaza JMG, Johnson R, Johnson TA, Juul M, Juul RI, Kahles A, Kahraman A, Kellis M, Khurana E, Kim Jaegil, Kim JK, Kim Y, Komorowski J, Korbel JO, Kumar S, Lanzós A, Larsson E, Lawrence MS, Lee Donghoon, Lehmann K-V, Li Shantao, Li Xiaotong, Lin Z, Liu EM, Lochovsky L, Lou S, Madsen T, Marchal K, Martincorena I, Martinez-Fundichely A, Maruvka YE, McGillivray PD, Meyerson W, Muiños F, Mularoni L, Nakagawa H, Nielsen MM, Paczkowska M, Park Keunchil, Park Kiejung, Pedersen JS, Pons T, Pulido-Tamayo S, Raphael BJ, Reimand J, Reyes-Salazar I, Reyna MA, Rheinbay E, Rubin MA, Rubio-Perez C, Sahinalp SC, Saksena G, Salichos L, Sander C, Schumacher SE, Shackleton M, Shapira O, Shen C, Shrestha R, Shuai S, Sidiropoulos N, Sieverling L, Sinnott-Armstrong N, Stein LD, Stuart JM, Tamborero D, Tiao G, Tsunoda T, Umer HM, Uusküla-Reimand L, Valencia A, Vazquez M, Verbeke LPC, Wadelius C, Wadi L, Wang Jiayin, Warrell J, Waszak SM, Weischenfeldt J, Wheeler DA, Wu G, Yu J, Zhang Jing, Zhang Xuanping, Zhang Y, Zhao Z, Zou L, Von Mering C, PCAWG Structural Variation Working Group, Akdemir KC, Alvarez EG, Baez-Ortega A, Beroukhim R, Boutros PC, Bowtell DDL, Brors B, Burns KH, Busanovich J, Campbell PJ, Chan K, Chen K, Cortés-Ciriano I, Dueso-Barroso A, Dunford AJ, Edwards PA, Estivill X, Etemadmoghadam D, Feuerbach L, Fink JL, Frenkel-Morgenstern M, Garsed DW, Gerstein M, Gordenin DA, Haan D, Haber JE, Hess JM, Hutter B, Imielinski M, Jones DTW, Ju YS, Kazanov MD, Klimczak LJ, Koh Y, Korbel JO, Kumar K, Lee EA, Lee JJ-K, Li Y, Lynch AG, Macintyre G, Markowetz F, Martincorena I, Martinez-Fundichely A, Meyerson M, Miyano S, Nakagawa H, Navarro FCP, Ossowski S, Park PJ, Pearson JV, Puiggròs M, Rippe K, Roberts ND, Roberts SA, Rodriguez-Martin B, Schumacher SE, Scully R, Shackleton M, Sidiropoulos N, Sieverling L, Stewart C, Torrents D, Tubio JMC, Villasante I, Waddell N, Wala JA, Weischenfeldt J, Yang Lixing, Yao X, Yoon S-S, Zamora J, Zhang C-Z, Weischenfeldt J, Beroukhim R, Martincorena I, Pedersen JS, Getz G, PCAWG Consortium, Aaltonen LA, Abascal F, Abeshouse A, Aburatani H, Adams DJ, Agrawal N, Ahn KS, Ahn S-M, Aikata H, Akbani R, Akdemir KC, Al-Ahmadie H, Al-Sedairy ST, Al-Shahrour F, Alawi M, Albert M, Aldape K, Alexandrov LB, Ally A, Alsop K, Alvarez EG, Amary F, Amin SB, Aminou B, Ammerpohl O, Anderson MJ, Ang Y, Antonello D, Anur P, Aparicio S, Appelbaum EL, Arai Y, Aretz A, Arihiro K, Ariizumi S, Armenia J, Arnould L, Asa S, Assenov Y, Atwal G, Aukema S, Auman JT, Aure MRR, Awadalla P, Aymerich M, Bader GD, Baez-Ortega A, Bailey MH, Bailey PJ, Balasundaram M, Balu S, Bandopadhayay P, Banks RE, Barbi S, Barbour AP, Barenboim J, Barnholtz-Sloan J, Barr H, Barrera E, Bartlett J, Bartolome J, Bassi C, Bathe OF, Baumhoer D, Bavi P, Baylin SB, Bazant W, Beardsmore D, Beck TA, Behjati S, Behren A, Niu B, Bell C, Beltran S, Benz C, Berchuck A, Bergmann AK, Bergstrom EN, Berman BP, Berney DM, Bernhart SH, Beroukhim R, Berrios M, Bersani S, Bertl J, Betancourt M, Bhandari V, Bhosle SG, Biankin AV, Bieg M, Bigner D, Binder H, Birney E, Birrer M, Biswas NK, Bjerkehagen B, Bodenheimer T, Boice L, Bonizzato G, De Bono JS, Boot A, Bootwalla MS, Borg A, Borkhardt A, Boroevich KA, Borozan I,

Borst C, Bosenberg M, Bosio M, Boultwood J, Bourque G, Boutros PC, Bova GS, Bowen DT, Bowlby R, Bowtell DDL, Boyault S, Boyce R, Boyd J, Brazma A, Brennan P, Brewer DS, Brinkman AB, Bristow RG, Broaddus RR, Brock JE, Brock M, Broeks A, Brooks AN, Brooks D, Brors B, Brunak S, Bruxner TJC, Bruzos AL, Buchanan A, Buchhalter I, Buchholz C, Bullman S, Burke H, Burkhardt B, Burns KH, Busanovich J, Bustamante CD, Butler AP, Butte AJ, Byrne NJ, Børresen-Dale A-L, Caesar-Johnson SJ, Cafferkey A, Cahill D, Calabrese C, Caldas C, Calvo F, Camacho N, Campbell PJ, Campo E, Cantù C, Cao S, Carey TE, Carlevaro-Fita J, Carlsen R, Cataldo I, Cazzola M, Cebon J, Cerfolio R, Chadwick DE, Chakravarty D, Chalmers D, Chan CWY, Chan K, Chan-Seng-Yue M, Chandan VS, Chang DK, Chanock SJ, Chantrill LA, Chateigner A, Chatterjee N, Chayama K, Chen H-W, Chen J, Chen K, Chen Y, Chen Z, Cherniack AD, Chien J, Chiew Y-E, Chin S-F, Cho J, Cho S, Choi JK, Choi W, Chomienne C, Chong Z, Choo SP, Chou A, Christ AN, Christie EL, Chuah E, Cibulskis C, Cibulskis K, Cingarlini S, Clapham P, Claviez A, Cleary S, Cloonan N, Cmero M, Collins CC, Connor AA, Cooke SL, Cooper CS, Cope L, Corbo V, Cordes MG, Cordner SM, Cortés-Ciriano I, Covington K, Cowin PA, Craft B, Craft D, Creighton CJ, Cun Y, Curley E, Cutcutache I, Czajka K, Czerniak B, Dagg RA, Danilova L, Davi MV, Davidson NR, Davies H, Davis IJ, Davis-Dusenbery BN, Dawson KJ, De La Vega FM, De Paoli-Iseppi R, Defreitas T, Tos APD, Delaneau O, Demchok JA, Demeulemeester J, Demidov GM, Demircioğlu D, Dennis NM, Denroche RE, Dentro SC, Desai N, Deshpande V, Deshwar AG, Desmedt C, Deu-Pons J, Dhalla N, Dhani NC, Dhingra P, Dhir R, DiBiase A, Diamanti K, Ding L, Ding S, Dinh HQ, Dirix L, Doddapaneni H, Donmez N, Dow MT, Drapkin R, Drechsel O, Drews RM, Serge S, Dudderidge T, Dueso-Barroso A, Dunford AJ, Dunn M, Dursi LJ, Duthie FR, Dutton-Regester K, Eagles J, Easton DF, Edmonds S, Edwards PA, Edwards SE, Eeles RA, Ehinger A, Eils J, Eils R, El-Naggar A, Eldridge M, Ellrott K, Erkek S, Escaramis G, Espiritu SMG, Estivill X, Etemadmoghadam D, Eyfjord JE, Faltas BM, Fan D, Fan Y, Faquin WC, Farcas C, Fassan M, Fatima A, Favero F, Fayzullaev N, Felau I, Fereday S, Ferguson ML, Ferretti V, Feuerbach L, Field MA, Fink JL, Finocchiaro G, Fisher C, Fittall MW, Fitzgerald A, Fitzgerald RC, Flanagan AM, Fleshner NE, Flicek P, Foekens JA, Fong KM, Fonseca NA, Foster CS, Fox NS, Fraser M, Frazer S, Frenkel-Morgenstern M, Friedman W, Frigola J, Fronick CC, Fujimoto A, Fujita M, Fukayama M, Fulton LA, Fulton RS, Furuta M, Futreal PA, Füllgrabe A, Gabriel SB, Gallinger S, Gambacorti-Passerini C, Gao J, Gao S, Garraway L, Garred Ø, Garrison E, Garsed DW, Gehlenborg N, Gelpi JLL, George J, Gerhard DS, Gerhauser C, Gershenwald JE, Gerstein M, Gerstung M, Getz G, Ghori M, Ghossein R, Giama NH, Gibbs RA, Gibson B, Gill AJ, Gill P, Giri DD, Glodzik D, Gnanapragasam VJ, Goebler ME, Goldman MJ, Gomez C, Gonzalez S, Gonzalez-Perez A, Gordenin DA, Gossage J, Gotoh K, Govindan R, Grabau D, Graham JS, Grant RC, Green AR, Green E, Greger L, Grehan N, Grimaldi S, Grimmond SM, Grossman RL, Grundhoff A, Gundem G, Guo Q, Gupta M, Gupta S, Gut IG, Gut M, Göke J, Ha G, Haake A, Haan D, Haas S, Haase K, Haber JE, Habermann N, Hach F, Haider S, Hama N, Hamdy FC, Hamilton A, Hamilton MP, Han L, Hanna GB, Hansmann M, Haradhvala NJ, Harismendy O, Harliwong I, Harmanci AO, Harrington E, Hasegawa T, Haussler D, Hawkins S, Hayami S, Hayashi S, Hayes DN, Hayes SJ, Hayward NK, Hazell S, He Y, Heath AP, Heath SC, Hedley D, Hegde AM, Heiman DI, Heinold MC, Heins Z, Heisler LE, Hellstrom-Lindberg E, Helmy M, Heo SG, Hepperla AJ, Heredia-Genestar JM, Herrmann C, Hersey P, Hess JM, Hilmarsdottir H, Hinton J, Hirano S, Hiraoka N, Hoadley KA, Hobolth A, Hodzic E, Hoell JI, Hoffmann S, Hofmann O, Holbrook A, Holik AZ, Hollingsworth MA, Holmes O, Holt RA, Hong C, Hong EP, Hong JH, Hooijer GK, Hornshøj H, Hosoda F, Hou Y, Hovestadt V, Howat W, Hoyle AP, Hruban RH, Hu J, Hu T, Hua X, Huang K, Huang M, Huang MN, Huang V, Huang Y, Huber W, Hudson TJ, Hummel M, Hung JA, Huntsman D, Hupp TR, Huse J, Huska MR, Hutter B, Hutter CM, Hübschmann D, Iacobuzio-Donahue CA, Imbusch CD, Imielinski M, Imoto S, Isaacs WB, Isaev K, Ishikawa S, Iskar M, Islam SMA, Ittmann M, Ivkovic S, Izarzugaza JMG, Jacquemier J, Jakrot V, Jamieson NB, Jang GH, Jang SJ, Jayaseelan JC, Jayasinghe R, Jefferys SR, Jegalian K, Jennings JL, Jeon S-H, Jerman L, Ji Y, Jiao W, Johansson PA, Johns AL, Johns J, Johnson R, Johnson TA, Jolly C, Joly Y, Jonasson JG, Jones CD, Jones DR, Jones DTW, Jones N, Jones SJM, Jonkers J, Ju YS, Juhl H, Jung J, Juul M, Juul RI, Juul S, Jäger N, Kabbe R, Kahles A, Kahraman A, Kaiser VB, Kakavand H, Kalimuthu S, Von Kalle C, Kang KJ, Karaszi K, Karlan B, Karlić R, Karsch D, Kasaian K, Kassahn KS, Katai H, Kato M, Katoh H, Kawakami Y, Kay JD, Kazakoff SH, Kazanov MD, Keays M, Kebebew E, Kefford RF, Kellis M, Kench JG, Kennedy CJ, Kerssemakers JNA, Khoo D, Khoo V, Khuntikeo N, Khurana E, Kilpinen H, Kim HK, Kim H-L, Kim H-Y, Kim H, Kim Jaegil, Kim Jihoon, Kim JK, Kim Y, King TA, Klapper W, Kleinheinz K, Klimczak LJ, Knappskog S, Kneba M, Knoppers BM, Koh Y, Komorowski J, Komura D, Komura M, Kong G, Kool M, Korbel JO, Korchina V, Korshunov A, Koscher M, Koster R, Kote-Jarai Z, Koures A, Kovacevic M, Kremeyer B, Kretzmer H, Kreuz M, Krishnamurthy S, Kube D, Kumar K, Kumar P, Kumar S, Kumar Y, Kundra R, Kübler K, Küppers R, Lagergren J, Lai PH, Laird PW, Lakhani SR, Lalansingh CM, Lalonde E, Lamaze FC, Lambert A, Lander E, Landgraf P, Landoni L, Langerød A, Lanzós A, Larsimont D, Larsson E, Lathrop M, Lau LMS, Lawerenz C, Lawlor RT, Lawrence MS, Lazar AJ, Lazic AM, Le X, Lee Darlene, Lee Donghoon, Lee EA, Lee HJ, Lee JJ-K, Lee J-Y, Lee J, Lee MTM, Lee-Six H, Lehmann K-V, Lehrach H, Lenze D, Leonard CR, Leongamornlert DA, Leshchiner I, Letourneau L, Letunic I, Levine DA, Lewis L, Ley T, Li C, Li CH, Li HI, Li J, Li L, Li Shantao, Li Siliang, Li Xiaobo, Li Xiaotong, Li Xinyue, Li Y, Liang H, Liang S-B, Lichter P, Lin P, Lin Z, Linehan WM, Lingjærde OC, Liu D, Liu EM, Liu F-FF, Liu F, Liu J, Liu X, Livingstone J, Livitz D, Livni N, Lochovsky L, Loeffler M, Long GV, Lopez-Guillermo A, Lou S, Louis DN, Lovat LB, Lu Yiling, Lu Y-J, Lu Youyong, Luchini C, Lungu I, Luo X, Luxton HJ, Lynch AG, Lype L, López C, López-Otín C, Ma EZ, Ma Y, MacGrogan G, MacRae S, Macintyre G, Madsen T, Maejima K, Mafficini A, Maglinte DT, Maitra A, Majumder PP, Malcovati L, Malikic S, Malleo G, Mann GJ, Mantovani-Löffler L, Marchal K, Marchegiani G, Mardis ER, Margolin AA, Marin MG, Markowetz F, Markowski J, Marks J, Marques-Bonet T, Marra MA, Marsden L, Martens JWM, Martin S, Martin-Subero JI, Martincorena I, Martinez-Fundichely A, Maruvka YE, Mashl RJ, Massie CE, Matthew TJ, Matthews L, Mayer E, Mayes S, Mayo M, Mbabaali F, McCune K, McDermott U, McGillivray PD, McLellan MD, McPherson JD, McPherson JR, McPherson TA, Meier SR, Meng A, Meng S, Menzies A, Merrett ND, Merson S, Meyerson M, Meyerson W, Mieczkowski PA, Mihaiescu GL, Mijalkovic S, Mikkelsen T, Milella M, Mileshkin L, Miller CA, Miller DK, Miller JK, Mills GB, Milovanovic A, Minner S, Miotto M, Arnau GM, Mirabello L, Mitchell C, Mitchell TJ, Miyano S, Miyoshi N, Mizuno S, Molnár-Gábor F, Moore MJ, Moore RA, Morganella S, Morris QD, Morrison C, Mose LE, Moser CD, Muiños F, Mularoni L, Mungall AJ, Mungall K, Musgrove EA, Mustonen V, Mutch D, Muyas F, Muzny DM, Muñoz A, Myers J, Myklebost O, Möller P, Nagae G, Nagrial AM, Nahal-Bose HK, Nakagama H, Nakagawa H, Nakamura H, Nakamura T, Nakano K, Nandi T, Nangalia J, Nastic M, Navarro A, Navarro FCP, Neal DE, Nettekoven G, Newell F, Newhouse SJ, Newton Y, Ng AWT, Ng A, Nicholson J, Nicol D, Nie Y, Nielsen GP, Nielsen MM, Nik-Zainal S, Noble MS, Nones K, Northcott PA, Notta F, O'Connor BD, O'Donnell P, O'Donovan M, O'Meara S, O'Neill BP, O'Neill JR, Ocana D, Ochoa A, Oesper L, Ogden C, Ohdan H, Ohi K, Ohno-Machado L, Oien KA, Ojesina AI, Ojima H, Okusaka T, Omberg L, Ong CK, Ossowski S, Ott G, Ouellette BFF, P'ng C, Paczkowska M, Paiella S, Pairojkul C, Pajic M, Pan-Hammarström Q, Papaemmanuil E, Papatheodorou I, Paramasivam N, Park JW, Park J-W, Park Keunchil, Park Kiejung, Park PJ, Parker JS,

Parsons SL, Pass H, Pasternack D, Pastore A, Patch A-M, Pauporté I, Pea A, Pearson JV, Pedamallu CS, Pedersen JS, Pederzoli P, Peifer M, Pennell NA, Perou CM, Perry MD, Petersen GM, Peto M, Petrelli N, Petryszak R, Pfister SM, Phillips M, Pich O, Pickett HA, Pihl TD, Pillay N, Pinder S, Pinese M, Pinho AV, Pitkänen E, Pivot X, Piñeiro-Yáñez E, Planko L, Plass C, Polak P, Pons T, Popescu I, Potapova O, Prasad A, Preston SR, Prinz M, Pritchard AL, Prokopec SD, Provenzano E, Puente XS, Puig S, Puiggròs M, Pulido-Tamayo S, Pupo GM, Purdie CA, Quinn MC, Rabionet R, Rader JS, Radlwimmer B, Radovic P, Raeder B, Raine KM, Ramakrishna M, Ramakrishnan K, Ramalingam S, Raphael BJ, Rathmell WK, Rausch T, Reifenberger G, Reimand J, Reis-Filho J, Reuter V, Reyes-Salazar I, Reyna MA, Reynolds SM, Rheinbay E, Riazalhosseini Y, Richardson AL, Richter J, Ringel M, Ringnér M, Rino Y, Rippe K, Roach J, Roberts LR, Roberts ND, Roberts SA, Robertson AG, Robertson AJ, Rodriguez JB, Rodriguez-Martin B, Rodríguez-González FG, Roehrl MHA, Rohde M, Rokutan H, Romieu G, Rooman I, Roques T, Rosebrock D, Rosenberg M, Rosenstiel PC, Rosenwald A, Rowe EW, Royo R, Rozen SG, Rubanova Y, Rubin MA, Rubio-Perez C, Rudneva VA, Rusev BC, Ruzzenente A, Rätsch G, Sabarinathan R, Sabelnykova VY, Sadeghi S, Sahinalp SC, Saini N, Saito-Adachi M, Saksena G, Salcedo A, Salgado R, Salichos L, Sallari R, Saller C, Salvia R, Sam M, Samra JS, Sanchez-Vega F, Sander C, Sanders G, Sarin R, Sarrafi I, Sasaki-Oku A, Sauer T, Sauter G, Saw RPM, Scardoni M, Scarlett CJ, Scarpa A, Scelo G, Schadendorf D, Schein JE, Schilhabel MB, Schlesner M, Schlomm T, Schmidt HK, Schramm S-J, Schreiber S, Schultz N, Schumacher SE, Schwarz RF, Scolyer RA, Scott D, Scully R, Seethala R, Segre AV, Selander I, Semple CA, Senbabaoglu Y, Sengupta S, Sereni E, Serra S, Sgroi DC, Shackleton M, Shah NC, Shahabi S, Shang CA, Shang P, Shapira O, Shelton T, Shen C, Shen H, Shepherd R, Shi R, Shi Y, Shiah Y-J, Shibata T, Shih J, Shimizu E, Shimizu K, Shin SJ, Shiraishi Y, Shmaya T, Shmulevich I, Shorser SI, Short C, Shrestha R, Shringarpure SS, Shriver C, Shuai S, Sidiropoulos N, Siebert R, Sieuwerts AM, Sieverling L, Signoretti S, Sikora KO, Simbolo M, Simon R, Simons JV, Simpson JT, Simpson PT, Singer S, Sinnott-Armstrong N, Sipahimalani P, Skelly TJ, Smid M, Smith J, Smith-McCune K, Socci ND, Sofia HJ, Soloway MG, Song L, Sood AK, Sothi S, Sotiriou C, Soulette CM, Span PN, Spellman PT, Sperandio N, Spillane AJ, Spiro O, Spring J, Staaf J, Stadler PF, Staib P, Stark SG, Stebbings L, Stefánsson ÓA, Stegle O, Stein LD, Stenhouse A, Stewart C, Stilgenbauer S, Stobbe MD, Stratton MR, Stretch JR, Struck AJ, Stuart JM, Stunnenberg HG, Su H, Su X, Sun RX, Sungalee S, Susak H, Suzuki A, Sweep F, Szczepanowski M, Sültmann H, Yugawa T, Tam A, Tamborero D, Tan BKT, Tan D, Tan P, Tanaka H, Taniguchi H, Tanskanen TJ, Tarabichi M, Tarnuzzer R, Tarpey P, Taschuk ML, Tatsuno K, Tavaré S, Taylor DF, Taylor-Weiner A, Teague JW, Teh BT, Tembe V, Temes J, Thai K, Thayer SP, Thiessen N, Thomas G, Thomas S, Thompson A, Thompson AM, Thompson JFF, Thompson RH, Thorne H, Thorne LB, Thorogood A, Tiao G, Tijanic N, Timms LE, Tirabosco R, Tojo M, Tommasi S, Toon CW, Toprak UH, Torrents D, Tortora G, Tost J, Totoki Y, Townend D, Traficante N, Treilleux I, Trotta J-R, Trümper LHP, Tsao M, Tsunoda T, Tubio JMC, Tucker O, Turkington R, Turner DJ, Tutt A, Ueno M, Ueno NT, Umbricht C, Umer HM, Underwood TJ, Urban L, Urushidate T, Ushiku T, Uusküla-Reimand L, Valencia A, Van Den Berg DJ, Van Laere S, Van Loo P, Van Meir EG, Van Den Eynden GG, Van Der Kwast T, Vasudev N, Vazquez M, Vedururu R, Veluvolu U, Vembu S, Verbeke LPC, Vermeulen P, Verrill C, Viari A, Vicente D, Vicentini C, VijayRaghavan K, Viksna J, Vilain RE, Villasante I, Vincent-Salomon A, Visakorpi T, Voet D, Vyas P, Vázquez-García I, Waddell NM, Waddell N, Wadelius C, Wadi L, Wagener R, Wala JA, Wang Jian, Wang Jiayin, Wang L, Wang Q, Wang W, Wang Y, Wang Z, Waring PM, Warnatz H-J, Warrell J, Warren AY, Waszak SM, Wedge DC, Weichenhan D, Weinberger P, Weinstein JN, Weischenfeldt J, Weisenberger DJ, Welch I, Wendl MC, Werner J, Whalley JP, Wheeler DA, Whitaker HC, Wigle D, Wilkerson MD, Williams A, Wilmott JS, Wilson GW, Wilson JM, Wilson RK, Winterhoff B, Wintersinger JA, Wiznerowicz M, Wolf S, Wong BH, Wong T, Wong W, Woo Y, Wood S, Wouters BG, Wright AJ, Wright DW, Wright MH, Wu C-L, Wu D-Y, Wu G, Wu J, Wu K, Wu Y, Wu Z, Xi L, Xia T, Xiang Q, Xiao X, Xing R, Xiong H, Xu Q, Xu Y, Xue H, Yachida S, Yakneen S, Yamaguchi R, Yamaguchi TN, Yamamoto M, Yamamoto S, Yamaue H, Yang F, Yang H, Yang JY, Yang Liming, Yang Lixing, Yang S, Yang T-P, Yang Y, Yao X, Yaspo M-L, Yates L, Yau C, Ye C, Ye K, Yellapantula VD, Yoon CJ, Yoon S-S, Yousif F, Yu J, Yu K, Yu W, Yu Y, Yuan K, Yuan Y, Yuen D, Yung CK, Zaikova O, Zamora J, Zapatka M, Zenklusen JC, Zenz T, Zeps N, Zhang C-Z, Zhang F, Zhang Hailei, Zhang Hongwei, Zhang Hongxin, Zhang Jiashan, Zhang Jing, Zhang Junjun, Zhang Xiuqing, Zhang Xuanping, Zhang Y, Zhang Z, Zhao Z, Zheng L, Zheng X, Zhou W, Zhou Y, Zhu B, Zhu H, Zhu J, Zhu S, Zou L, Zou X, deFazio A, Van As N, Van Deurzen CHM, Van De Vijver MJ, Van'T Veer L, Von Mering C. 2020. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**:102–111. doi:10.1038/s41586-020-1965-x

Rubinstein M, de Souza FSJ. 2013. Evolution of transcriptional enhancers and animal diversity. *Philos Trans R Soc B Biol Sci* **368**:20130017. doi:10.1098/rstb.2013.0017

Rust K, Byrnes LE, Yu KS, Park JS, Sneddon JB, Tward AD, Nystul TG. 2020. A single-cell atlas and lineage analysis of the adult Drosophila ovary. *Nat Commun* **11**:5628. doi:10.1038/s41467-020-19361-0

Sharon E, Kalma Y, Sharp A, Raveh-Sadka T, Levo M, Zeevi D, Keren L, Yakhini Z, Weinberger A, Segal E. 2012. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* **30**:521–530. doi:10.1038/nbt.2205

Spitz F, Furlong EEM. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet* **13**:613–626. doi:10.1038/nrg3207

Staller MV, Ramirez E, Kotha SR, Holehouse AS, Pappu RV, Cohen BA. 2022. Directed mutational scanning reveals a balance between acidic and hydrophobic residues in strong human activation domains. *Cell Syst* **13**:334-345.e5. doi:10.1016/j.cels.2022.01.002

Stevenson ZC, Moerdyk-Schauwecker MJ, Banse SA, Patel DS, Lu H, Phillips PC. 2023. High-Throughput Library Transgenesis in <em>Caenorhabditis elegans</em> via Transgenic Arrays Resulting in Diversity of Integrated Sequences (TARDIS). *bioRxiv* 2022.10.30.514301. doi:10.1101/2022.10.30.514301

Thuronyi BW, Koblan LW, Levy JM, Yeh W-H, Zheng C, Newby GA, Wilson C, Bhaumik M, Shubina-Oleinik O, Holt JR, Liu DR. 2019. Continuous evolution of base editors with expanded target compatibility and improved activity. *Nat Biotechnol* **37**:1070–1079. doi:10.1038/s41587-019-0193-0

Tou CJ, Schaffer DV, Dueber JE. 2020. Targeted Diversification in the S. cerevisiae Genome with CRISPR-Guided DNA Polymerase i. *ACS Synth Biol* **9**:1911–1916. doi:10.1021/acssynbio.0c00149

Villiger L, Grisch-Chan HM, Lindsay H, Ringnalda F, Pogliano CB, Allegri G, Fingerhut R, Häberle J, Matos J, Robinson MD, Thöny B, Schwank G. 2018. Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. *Nat Med* **24**:1519–1525. doi:10.1038/s41591-018-0209-1

Wadycki WJ, Shah BK, Ghangurde PD, Dudewicz EJ, Mantel N, Brown CC, Larson HJ, Barr DR, Frane JW, Saperstein B, Good IJ, Jones HL. 1973. A Step-Wise Clustering Procedure. *Am Stat* **27**:123–127.

Walton RT, Christie KA, Whittaker MN, Kleinstiver BP. 2020. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**:290–296. doi:10.1126/science.aba8853

Wikipedia contributors. 2024. Poisson binomial distribution — Wikipedia, The Free Encyclopedia.

Yu L, Wang X, Mu Q, Tam SST, Loi DSC, Chan AKY, Poon WS, Ng H-K, Chan DTM, Wang J, Wu AR. 2023. scONE-seq: A single-cell multi-omics method enables simultaneous dissection of phenotype and genotype heterogeneity from frozen tumors. *Sci Adv* **9**:eabp8901. doi:10.1126/sciadv.abp8901

Yu TC, Liu WL, Brinck MS, Davis JE, Shek J, Bower G, Einav T, Insigne KD, Phillips R, Kosuri S, Urtecho G. 2021. Multiplexed characterization of rationally designed promoter architectures deconstructs combinatorial logic for IPTG-inducible systems. *Nat Commun* **12**. doi:10.1038/s41467-020-20094-3