# CrispRVariants charts the mutation spectrum of genome engineering experiments

**To the Editor:**

CRISPR-Cas9 and related technologies enable the efficient alteration of genomic DNA at targeted positions and are widely used in applications ranging from individual gene knockouts to large-scale functional screening and therapeutic gene editing. Unlocking the full potential of these methods requires accurate evaluation of editing efficiencies. Here, we show that methodological decisions for analyzing sequencing data can substantially affect mutagenesis efficiency estimates. We provide a comprehensive R-based toolkit (CrispRVariants) and an accompanying web tool (CrispRVariantsLite) that resolve and localize individual mutant alleles with respect to the endonuclease cut site. CrispRVariants-enabled analyses of newly generated and existing genome editing data sets underscore how careful consideration of the full spectrum of variants resulting from genome engineering can not only inform effective guide RNA and amplicon design but also provide insights into the mutagenic process.

After the induction of a double-strand break by Cas9 or related enzymes[1,2] (coupled with a locus-specific guide RNA; sgRNA), typically, a random number of bases are inserted or deleted as the two DNA ends are reconnected by non-homologous end joining (NHEJ)[3]. Optionally, donor DNA can be introduced and integrated between the breakpoints by NHEJ or homology-directed repair[4,5]. The result is an 'edited' genome sequence at a chosen location, which can now be achieved in an increasing number of cell types and animal systems[5–7].

*In vivo* CRISPR applications, where multiple cells undergo independent rounds of mutagenesis and local NHEJ, generate particularly heterogenous sequencing data sets. Existing tools for the analysis of mutagenesis sequencing data report aggregated variant summaries (e.g., CRISPR-GA[6] and CRISPResso[7]) and are unsuited for applications that consider the entire, complex mutation spectrum, such as quantifying mosaicism[8] and allele-specific genome editing[9]. To facilitate such analyses, we have developed CrispRVariants, an R-based toolkit for quantifying and visualizing individual variant alleles from either traditional Sanger sequencing or high-throughput CRISPR-Cas9 mutagenesis sequencing experiments. CrispRVariants can be easily used to create a variant allele summary plot (**Fig. 1**) and accompanying table of counts. Individual variants can be removed, allowing allele-specific analysis and adjustment for heterozygosity. By localizing variant alleles with respect to the nuclease cut site instead of the PCR amplicon, CrispRVariants enables immediate comparison of variant spectra between target locations (**Supplementary Note 1**). This level of resolution enables users to directly relate variant genotypes to observed phenotypes and predict downstream effects of variants, such as protein structural changes or loss- or gain-of transcription factor binding sites when targeting non-coding sites (see examples in the CrispRVariants User Guide within the Bioconductor package). **Figure 1** summarizes the variant allele spectrum from several zebrafish embryos injected with an sgRNA targeting ENSDARG00000079624 (wtx); some variants reoccur independently in multiple embryos. Notably, visualization of variant alleles facilitates the detection of sequencing or alignment errors and previously unknown genetic variation. We designed CrispRVariants with interactivity

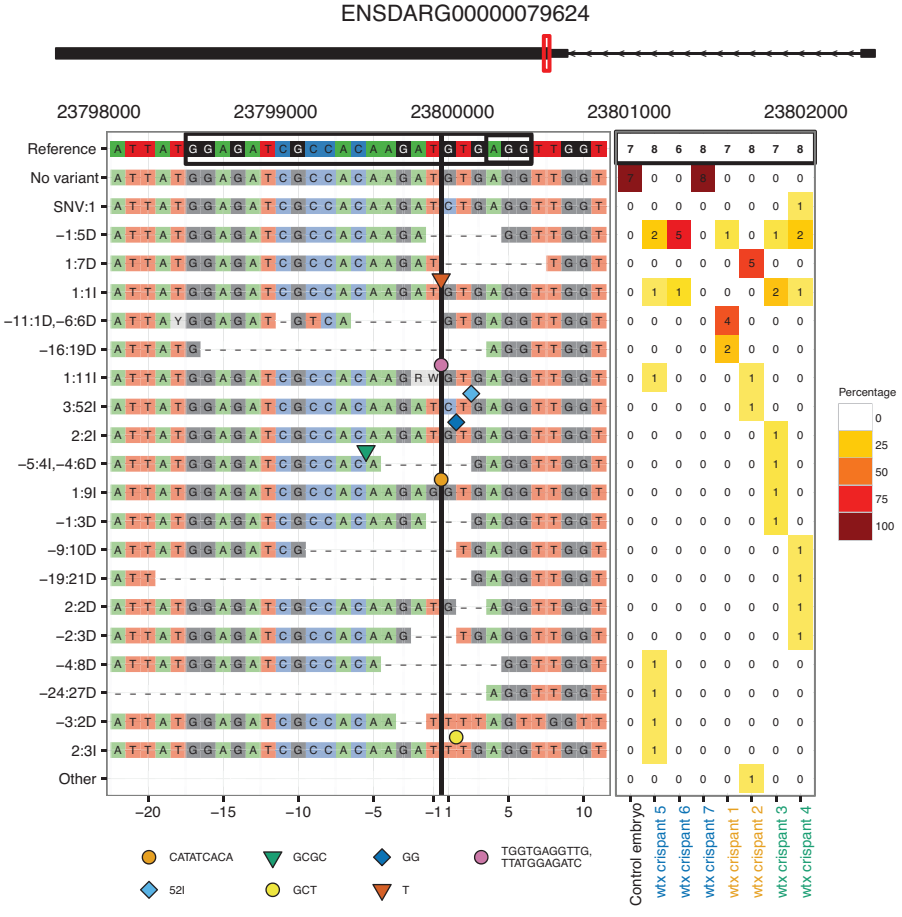


**Figure 1** The CrispRVariants plotVariants function summarizes variant types, locations and frequency across multiple clones from several injected animals. This function returns a ggplot2-based allele summary plot consisting of the following: first, a schematic of the target site location relative to the neighboring transcripts; second, an alignment of the consensus sequence for each variant combination to the reference sequence; and third, a heat map showing the frequency of the variants across samples (the heatmap can be plotted also with frequencies). Inserted sequences are shown below the alignments (key at bottom of figure), with large insertions indicated by the corresponding symbol. In this example, columns in the heat map represent sequences cloned from different embryos, with column labels colored by the embryonic phenotype: Black, uninjected; blue, wild-type-like; orange, developmental abnormalities or 'monsters'; green, heart phenotype.

in mind, explicitly allowing users to detect problems and filter sequences appropriately before estimating mutation efficiency. The accompanying web tool, CrispRVariantsLite, which is suitable for smaller-scale experiments, can be accessed through the website or downloaded and run locally, allowing users without bioinformatics expertise to examine and plot their data.

Distinguishing low-frequency mutation events from sequencing errors is challenging. In **Supplementary Note 2**, we show examples of sequencing errors and alignment uncertainty that affect the size, placement and ultimately variant classification (i.e., whether in-frame or not) of two germline mutant cohorts. Sequencing errors and genetic variation confound mutation efficiency estimation; for example, sequence polymorphisms in the targeted locus affect sgRNA binding and may lead to underestimation of the true editing efficiency. In **Supplementary Note 3**, we highlight unappreciated genetic variation in a recent study[10] as well as off-target sgRNAs that lack a canonical protospacer adjacent motif (PAM) sequence. We show through simulation that CrispRVariants matches or outperforms existing tools in estimating mutation efficiency (**Supplementary Note 4**). Notably, data processing decisions invisible to users contribute substantially to the differences between CrispRVariants and other available tools. We include with CrispRVariants a small synthetic benchmarking data set containing several types of commonly observed variants to facilitate transparent data processing.

Despite overwhelming evidence that data preprocessing choices affect variant calling in exome and whole-genome sequencing studies[11–13], their role in estimating mutagenesis efficiency has been largely neglected. Amplicon sequencing data may be aligned locally to the expected amplicon sequence (e.g., Gagnon *et al.*[5], CRISPR-GA[6], CRISPResso[7]), in which case pooled reads must first be separated, or globally to an entire reference genome (AmpliconDIVider[14] and CrispRVariants). Strategies that combine local and global alignment (CRISPResso (Pooled)[7]) or avoid separating reads by aligning to the set of all amplicons[15] are also possible. Inappropriate alignment and preprocessing settings can have a major impact on allele counts and efficiency estimates. In the most extreme case, tandem repeats and homology within an amplicon resulted in efficiency estimates that differed by 91% between methods (**Supplementary Note 5**). Local alignment strategies are vulnerable to miscounting off-target reads. For example, BLAT[16] local alignment (used in CRISPR-GA) can result in efficiency estimates that differ by >30% from estimates from global alignments (**Supplementary Note 5**).

Stringency criteria when merging paired-end reads or dividing reads by PCR primers (as done in Shah *et al.*[10]) can also affect mutation efficiency estimates (**Supplementary Notes 6** and **7**). Specifically, altering the percentage overlap required for merging from 100% (as in Shah *et al.*[10]) to 90% changed the efficiency estimate for one sgRNA by 65%. CrispRVariants separates data preprocessing from mutation quantification, allowing critical parameters to be carefully selected and tailored to the experimental design (see Online Methods). By aggregating variant alleles instead of looking at and interpreting the full observed spectrum, existing tools make it difficult to assess whether appropriate bioinformatic decisions (alignment, merging and separation of reads) have been made; CrispRVariants facilitates this visual, interactive and iterative process.

In summary, the CrispRVariants package offers precise, transparent and reproducible preprocessing of low- and high-throughput amplicon sequencing experiments, provides easy visualizations of variant alleles across samples and allows careful calculation of efficiency, given all the complexities and confounders. The framework can also readily be applied to other mutagenesis systems. The software interfaces seamlessly with existing Bioconductor infrastructure and is available as **Supplementary Software**, but users are encouraged to download the most recent version that is available from Bioconductor[17] (https://www.bioconductor.org/packages/CrispRVariants).

## METHODS

Methods and any associated references are available in the online version of the paper.

*Editor's note: This article has been peer reviewed.*

*Note: Any Supplementary Information and Source Data files are available in the online version of the paper (doi:10.1038/nbt.3628).*


### ACKNOWLEDGMENTS

This work was supported by the Canton of Zürich, a Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (SNSF) professorship (PP00P3_139093), a Marie Curie Career Integration Grant from the European Commission, and a Schweizerische Herzstiftung grant to C.M.; European Research Council Starting Grant ANTIVIRNA (337284) and an SNSF Project Grant (31003A_149393) to M.J.; a European Commission 7th Framework Collaborative Project RADIANT grant (agreement number 305626) and an SNSF Project Grant (31003A_143883) to M.D.R.; a Universität Zürich (UZH) URPP Translational Cancer Research Seed Grant to A.B.; and a UZH Forschungskredit to C.H.


### COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.


***Helen Lindsay[1,2] Alexa Burger[1], Berthin Biyong[1,2], Anastasia Felker[1], Christopher Hess[1], Jonas Zaugg[1], Elena Chiavacci[1], Carolin Anders[3], Martin Jinek[3], Christian Mosimann[1] & Mark D Robinson[1,2]***

*[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland. [2]SIB Swiss Insitute of Bioinformatics, University of Zurich, Zurich, Switzerland. [3]Institute of Biochemistry, University of Zurich, Zurich, Switzerland.*
*e-mail: helen.lindsay@uzh.ch or christian.mosimann@imls.uzh.ch or mark.robinson@imls.uzh.ch*



1. Hsu, P.D., Lander, E.S. & Zhang, F. *Cell* **157**, 1262–1278 (2014).
2. Zetsche, B. *et al. Cell* **163**, 759–771 (2015).
3. Charpentier, E. & Doudna, J.A. *Nature* **495**, 50–51 (2013).
4. Ran, F.A. *et al. Cell* **154**, 1380–1389 (2013).
5. Gagnon, J.A. *et al. PLoS One* **9**, e98186 (2014).
6. Guell, M., Yang, L. & Church, G.M. *Bioinformatics* **30**, 2968–2970 (2014).
7. Pinello, L., Canver, M., Hoban, M. & Orkin, S. CRISPResso: sequencing analysis toolbox for CRISPR-Cas9 genome editing. Preprint at *bioRxiv* http://www.biorxiv.org/content/early/2015/11/10/031203 (2015).
8. Yen, S.T. *et al. Dev. Biol.* **393**, 3–9 (2014).
9. Yoshimi, K., Kaneko, T., Voigt, B. & Mashimo, T. *Nat. Commun.* **5**, 4240 (2014).
10. Shah, A.N., Davey, C.F., Whitebirch, A.C., Miller, A.C. & Moens, C.B. *Nat. Methods* **12**, 535–540 (2015).
11. McCarthy, D.J. *et al. Genome Med.* **6**, 26 (2014).
12. Li, H. *Bioinformatics* **30**, 2843–2851 (2014).
13. O'Rawe, J. *et al. Genome Med.* **5**, 28 (2013).
14. Varshney, G., Pei, W., LaFave, M. & Idol, J. *Genome Res.* **25**, 1030–1042 (2015).
15. Moreno-Mateos, M.A. *et al. Nat. Methods* **12**, 982–988 (2015).
16. Kent, W.J. *Genome Res.* **12**, 656–664 (2002).
17. Huber, W. *et al. Nat. Methods* **12**, 115–121 (2015).

## ONLINE METHODS

See **Supplementary Note** for the availability of CrispRVariants, CrispRVariantsLite and code for the analyses in this paper.

For both low- or high-throughput sequencing analysis, the main entry point to CrispRVariants is a set of sequences aligned to a reference genome in BAM (binary alignment) format. Reads that cannot be represented as a single linear alignment are instead represented by some alignment tools as multiple "chimeric" alignments. We find that some chimeric reads are genuine variants (**Supplementary Note 8**) and recommend the use of a chimera-aware aligner. In current pipelines, we use BWA MEM[18] with default parameters. The choice of aligner can substantially affect the mutation efficiency estimates (**Supplementary Note 5**). Applied Biosystems Sanger sequencing data, commonly available in AB1 file format, can be easily converted to FASTQ format for mapping; CrispRVariants uses the sangerseqR[19] package to perform this conversion. The entry points for CrispRVariantsLite include a ZIP file of BAM files (sets of already mapped reads), a ZIP file of directories with AB1 files or a ZIP file of FASTQ files (file size restrictions apply).

CrispRVariants can work directly with pooled amplicon sequencing data. Reads are assigned to the correct amplicon either by an alignment spanning the amplicon region almost exactly (strict), or by any base mapped to the unique portion of an amplicon (relaxed). Because of high error frequency, the endpoints of Illumina MiSeq data are often clipped by aligners. We extrapolate the mapped region to include clipped regions when matching amplicons. This dividing strategy is suitable either for paired-end reads where both reads span the entire amplicon or for merged paired-end reads. In cases where unique mapping to a single amplicon is insufficient to assign reads, alignments may be filtered in R and passed directly to CrispRVariants as a GenomicAlignments[20] object. CrispRVariants can collapse paired reads by checking for concordant variants in the vicinity of the cut site. However, if merging criteria are not overly strict, we find that merging reads before mapping improves speed without affecting efficiency estimates (**Supplementary Note 6**). Chimeric reads are assigned to all overlapping amplicons; however, to be counted as a variant, the mapped endpoint of one aligned segment must be close to the specified cut site. This criterion excludes PCR artifacts, such as primer dimers. Chimeric read sets are grouped into the 'Other' category. For amplicons with a non-trivial fraction of 'Other' reads, additional exploratory analyses are available within the software (see software vignette).

Once assigned, read alignments are narrowed to the target region (i.e., the user-specified local genomic region around the guide's target site). Reads that do not span the target region are discarded and reads that match the reference sequence are recorded as 'no variant'. Insertions and deletions (indels) are then localized in a strand-aware manner, labeled and counted; a 3 base pair deletion starting 2 bases upstream of the target location is designated '-2:3D'. Downstream variants are numbered similarly by their leftmost base. Reads that do not contain an indel can additionally be separated by the presence of single nucleotide variants (SNVs). By default, the zero point is at base 17 of a 23 bp sgRNA, i.e. the endonuclease cut site. The user is free to specify: i) the target region; ii) the corresponding zero point; and iii) a window within the target region for calling SNVs.

**Supplementary analyses**. The analyses presented in the **Supplementary Note** use data from Shah *et al.*[10], Burger *et al.*[21] and Cho *et al.*[22]. The performance of CrispRVariants (version 0.9.11), CRISPResso (version 0.9.1), CRISPR-GA and AmpliconDIVider was compared under a range of scenarios. Where not otherwise specified, the data are from Shah *et al.*[10]; CrispRVariants and AmpliconDIVider were run after BWA MEM alignment and CRISPResso was run in single amplicon mode. Further information about the data used is in **Supplementary Note 9**.


18. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 3 (2013). http://arxiv.org/abs/1303.3997.
19. Hill, J., Demarest, B. & Hill, M. Package 'sangerseqR' (2014).
20. Lawrence, M. *et al. PLOS Comput. Biol.* **9**, e1003118 (2013).
21. Burger, A. *et al.* Maximizing mutagenesis with solubilized CRISPR-Cas9 ribonucleoprotein complexes. *Development* **143**, 2025–2037 (2016).
22. Cho, S.W. *et al. Genome Res.* **24**, 132–141 (2014).






doi:10.1038/nbt.3628