Accurate analysis of genuine CRISPR editing events with ampliCan

Kornel Labun, Xiaoge Guo, Alejandro Chavez, George Church, James Gagnon, Eivind Valen

Abstract

We present ampliCan, an analysis tool that unites identification, quantification and visualization of genuine genome editing events from CRISPR amplicon sequencing data. ampliCan overcomes methodological challenges suffered by other tools to estimate the true mutational efficiency in a high-throughput automated fashion. ampliCan controls for biases at every step of the analysis and generates reports that allow users to quickly identify successful editing events or potential issues with their experiments. We benchmarked ampliCan against other leading tools demonstrating that it outperformed all in the face of common confounding factors.

Introduction

With the introduction of CRISPR[1,2], researchers obtained an inexpensive and effective tool for targeted mutagenesis. Despite its current limitations, CRISPR has been widely adopted within research settings and has begun to make inroads into medical applications[3]. The successful use of CRISPR relies on the ability to confidently identify CRISPR-induced mutations, generally insertions and deletions (indels). Indels are often identified by sequencing the targeted loci and comparing the sequenced reads to a reference sequence. Deep sequencing has the advantage of capturing the nature of the indel, readily identifying frameshift mutations or disrupted regulatory elements, and also characterizing the heterogeneity of the introduced mutations in a population. This is of particular importance when the aim is allele-specific editing or the experiment can result in mosaicism.

The reliability of a sequencing-based approach is dependent on the processing and interpretation of the sequenced reads and is contingent on several factors such as the inclusion of controls, the alignment algorithm and the filtering of experimental artifacts. To date, no tool considers and controls for the whole range of biases that can influence this interpretation and therefore, distorts the estimate of the mutation efficiency and leads to erroneous conclusions. Here we introduce a fully automated tool, ampliCan, designed to determine the true mutation frequencies of CRISPR experiments from high-throughput DNA amplicon sequencing. It scales to genome-wide experiments and can be used alone or integrated with the CHOPCHOP[4,5] guide RNA (gRNA) design tool.

Results

Estimation of the true mutation efficiency depends on multiple steps all subject to different biases[6]. Following sequencing, reads have to be aligned to the correct reference, filtered for artifacts, and then the mutation efficiency has to be quantified and normalized (**Fig. 1A**). In most existing tools, many of the choices made during these steps are typically hidden from the user leading to potential misinterpretation of the data. Furthermore, a subset of steps are often relegated to other tools not optimized for CRISPR experiments. Unlike other tools, ampliCan implements the complete pipeline from alignment to interpretation and can control for biases at all steps.

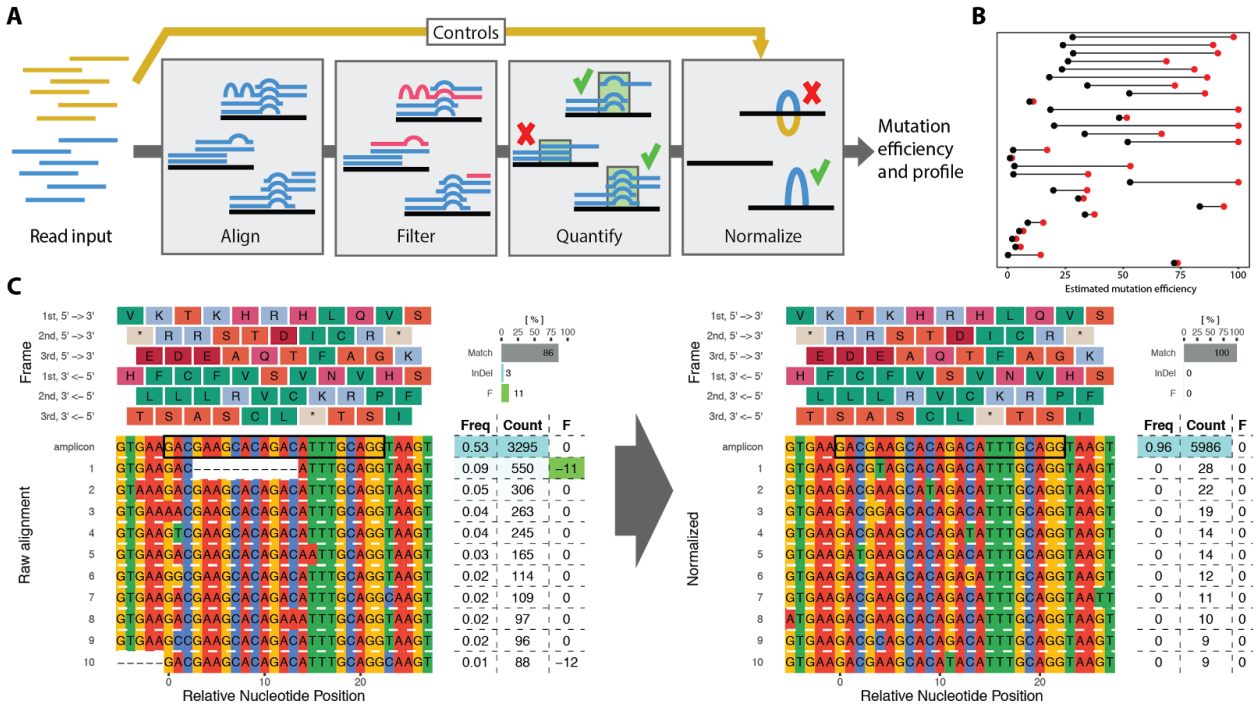

Fig 1. **A.** Estimation of mutation efficiency consists of multiple steps. At each of these steps biases can be introduced. Controls are processed identically to the main experiment and used for normalization. **B.** Overview of the change in estimated mutation efficiency when using controls that account for natural genetic variance in 28 experiments (mean change of 32%). Red dots show initial estimates based on unnormalized data, while black dots show the values after normalization. **C.** Alignment plot showing the top 10 most abundant reads in a real experiment. The table shows relative efficiency (Freq) of read, absolute number of reads (Count) and the summed size of the indel(s) (F), coloured green when inducing a frameshift. The bars (top right) shows the fraction of reads that contain no indels (Match), those having an indel without inducing frameshift (InDel) and frameshift inducing indels (F). The left panel shows the estimated

mutation efficiency from raw reads, which is 14% (11% with frameshift, 3% without). The right panel shows the same genomic loci after normalization with controls resulting in an mutation efficiency of 0%. The deletion of 11bp in 9% of the reads could not be found in GRCz10.88 Ensembl Variation database and would in the absence of controls give the impression of a real editing event.

Despite being arguably the most important step in any experiment, the use of controls is frequently overlooked in CRISPR assays. Discrepancies between a reference genome and the genetic variation in an organism of interest often lead to false positives and the false impression that mutations have been introduced (**Fig. 1B**, **Supplementary Fig. 1**)[7]. While the use of controls is (in principle) feasible with any tool, it commonly requires running the treated and control samples separately followed by a manual inspection and comparison of these. In ampliCan, controls are an integrated part of the pipeline and mutation frequencies are normalized and estimated automatically (**Fig. 1C**, Methods and **Supplementary Note 1**).

Estimating mutation efficiency starts with the alignment of the sequenced reads (**Fig. 1A**). A common strategy is to use standard genomic alignment tools. However, these tools do not align using knowledge about the known mechanisms of CRISPR-induced double stranded breaks and DNA repair. Genome editing typically results in a single deletion and/or insertion of variable length. Hence, correctly aligned reads will often have a low number of events (optimally 1 deletion and/or 1 insertion after normalization for controls) overlapping the cut site, while misaligned reads will result in a high number of events throughout the read due to discrepancies to the correct loci. Therefore an alignment strategy that penalizes multiple indel events (see Methods) is more consistent with DNA repair mechanisms and the CRISPR mode of action. ampliCan uses the Needleman-Wunsch algorithm with tuned parameters to ensure optimal alignments of the reads to their loci and models the number of indel and mismatch events to ensure that the reads originated from that loci (see Methods and **Supplementary Note 2**). In contrast, non-optimized aligners can create fragmented alignments resulting in misleading mutation profiles and possible distortion of downstream analyses and frameshift estimation (**Supplementary Fig. 2**). In assessments, ampliCan outperforms the tool CrispRVariants on its own synthetic dataset[6] contaminated with off-target reads matching the real on-target reads, but with a mismatch rate of 30% per bp (**Supplementary Fig. 3**). More problematic, the mapping strategy used in the pipelines of several tools (**Supplementary Table 1**) are not robust to small perturbations of this mismatch rate and altering it leads to a significant reduction in performance (**Fig. 2**, left). In contrast, ampliCan shows consistently high performance across a broad range of mismatch rates (**Fig. 2**, left and **Supplementary Figs. 3**, **4**).

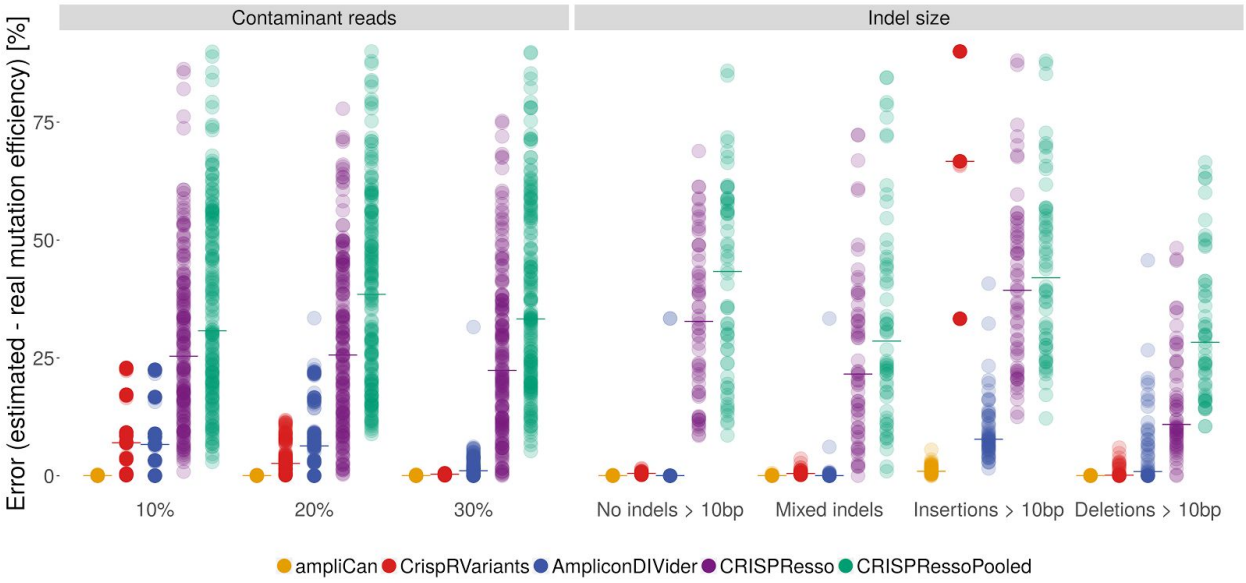

Fig 2. Benchmark of leading tools[6,16,17] when estimating mutation efficiency under different dataset conditions. Each dot shows the error of the estimate to the correct value for a single experiment. The median performance (Mixed indels) is indicated by the horizontal line. The left panel shows comparison of tools when datasets contain contaminant reads (see text and methods). The x-axis denotes how dissimilar the contaminant reads are to the correct reads. In cases where the contaminants are from homologous regions this may be low (10%), for other contaminants this is likely to be higher (30%). The right panel shows performance of tools as a function of the length of indel events. The sets in the left column contain no indels > 10bp, the right column contain only insertions > 10bp, the middle column (Mixed Indels) is a mix of shorter and longer events.

Targeted insertion of shorter fragments through co-opting of the homology directed repair (HDR) pathway is becoming increasingly popular[8,9]. This, together with long indels occurring in regular CRISPR experiments, presents a challenge for most CRISPR analysis tools. The inability of current pipelines to process these longer events (**Fig. 2**, right) typically stems from alignment strategies that are unable to assign reads with long indels to the correct loci. In previous assessments, simulated data has often been restricted to short indels where this weakness would not be apparent (**Supplementary Note 4**)[6]. Using a localized alignment strategy, based on primer matching (see Methods), ampliCan robustly handles these longer indels and on simulated data with indels >10 bp ampliCan outperforms all other tools (**Fig. 2**, right and **Supplementary Fig. 5**).

To aid in heterogeneous outcomes, ampliCan quantifies the heterogeneity of reads (**Supplementary Fig. 7**), the complete mutation efficiency for an experiment and the proportion of mutations resulting in a frameshift (**Fig. 1C**, top right). It also aggregates and quantifies mutation events of a specific type if a particular outcome is desired

(**Supplementary Fig. 6**). In addition, ampliCan provides overviews of the impact of all filtering steps (**Supplementary Fig. 8**). Reports can be generated in several formats (**Supplementary Tables 2 and 3**) and aggregated at multiple levels such as barcode, gRNA, gene, loci or any customized user specified grouping (**Supplementary Note 5**). This enables exploration of questions beyond mutation efficiency such as the rules of gRNA design, whether a particular researcher is better at designing gRNAs than others (**Supplementary Fig. 9**), whether a given barcode is not working or determining the stochasticity in the mutation outcome from a given gRNA (**Supplementary Fig. 19**).

ampliCan offers a complete pipeline controlling for biases at every step of evaluation. It can be integrated with the CHOPCHOP tool for gRNA design to incorporate all computational steps necessary for a CRISPR experiment. It scales from a single gRNA to genome-wide screens and can be run with a single command. For more advanced users, it provides a complete and adaptable framework, implemented in R and bioconductor, enabling further exploration of the data. Collectively, these advances will minimize misinterpretation of genome editing experiments and allow effective analysis of the outcome in an automated fashion.

Methods

ampliCan pipeline.

ampliCan is a multi-step pipeline (**Supplementary Fig. 10**) accepting a configuration file describing the experiment(s) and FASTQ files of sequenced reads as input. The configuration file contains information about barcodes, gRNAs, forward and reverse primers, amplicons and paths to corresponding FASTQ files (**Supplementary Table 4**). From here, the pipeline generates reports summarizing the key features of the experiments.

In the first step, ampliCan filters low quality reads which either have ambiguous nucleotides, an average quality or individual base quality under a default or user-specified threshold (**Supplementary Note 6**). After quality filtering, ampliCan assigns reads to the particular experiment by searching for matching primers (default up to two mismatches, but ampliCan supports different stringency, **Supplementary Note 7**). Unassigned reads are summarized and reported separately for troubleshooting. After read assignment ampliCan uses the Biostrings[10] implementation of the Needleman-Wunsch algorithm with optimized parameters (gap opening = -25, gap extension = 0, match = 5, mismatch = -4, no end gap penalty) to align all assigned reads to the loci/amplicon sequence. Subsequently, primer dimer reads are removed by detecting deletions larger than the size of the amplicon subtracting the length of the two primers and a short buffer. Additionally, sequences that contain a high number of indels or mismatch events compared to the remainder of the reads are filtered as these are potential sequencing artifacts or originate from off-target amplification (**Supplementary Note 6** and **Supplementary Fig. 11**). Mutation frequencies are

calculated from the remaining reads using the frequency of indels that (**Supplementary Fig. 6**) overlap a region (+/- 5bp) around the expected cut site and, if paired-end sequencing is used, follows consensus rules for the paired forward and reverse read (**Supplementary Fig. 12**). The expected cut site can be specified as a larger region for nickase or TALEN experiments. Any indel or mismatch also observed above a 1% threshold in the control are removed. Frameshifts are identified by summing the impact of deletions and insertions on the amplicon.

A series of automated reports is prepared in form of ".Rmd" files which can be converted to multiple formats, but also immediately transformed into html reports with knitr[11] for convenience. There are six different default reports prepared by ampliCan with statistics grouped at the corresponding level: id, barcode, gRNA, amplicon, summary and group (user specified, typically person conducting experiments, treatment or other grouping of interest). In addition to alignments of top reads (**Fig. 1C, Supplementary Fig. 1**) reports contain plots summarized over all deletions, insertions and variants (**Supplementary Fig. 6**). In addition a number of plots showing the general state of the experiments are shown including the heterogeneity of reads to investigate mosaicism or sequencing issues (**Supplementary Figs. 7**, **13**, **14**) and overviews of how many reads were filtered/assigned at each step (**Supplementary Fig. 15**). In addition to the default plots ampliCan produces R objects that contain all alignments and read information, these can be manipulated, extended and visualized through the R statistical package.

ampliCan provides a simple tool that can be used out-of-the-box, yet equipped with a flexible framework that can be exploited and extended by advanced R users. The default pipeline is compressed into a single convenient wrapper, amplicanPipeline, which generates all default reports. More advanced users can gain complete control over all processing steps and produce novel plots for more specialized use cases. Compatibility with the most popular plotting packages (ggplot2[12] and ggbio[13]) as well as the most popular data processing packages (dplyr[14] and data.table) provides a full fledged and elastic framework. Output files are encoded as GenomicRanges[15] tables of aligned read events for easy parsing (**Supplementary Table 3**) and human readable alignment results (**Supplementary Table 2**) and fasta. We would like to encourage users to communicate their needs and give us feedback, for future development.

Data Availability.

All data is available online under accession numbers: PRJNA245510 (BioProject, run 1 and run 5) and E-MTAB-6310, E-MTAB-6355, E-MTAB-6356, E-MTAB-6357, E-MTAB-6358, (run 6-10). Synthetic datasets can be reconstructed with the use of code from https://github.com/valenlab/amplican_manuscript.

Code availability.
ampliCan is developed as an R package under GNU General Public License version 3 and available through Bioconductor under http://bioconductor.org/packages/amplican or https://github.com/valenlab/amplican.

Additional information

References

1. Jinek, M. *et al.* A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337,** 816–821 (2012).
2. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339,** 819–823 (2013).
3. Courtney, D. G. *et al.* CRISPR/Cas9 DNA cleavage at SNP-derived PAM enables both in vitro and in vivo KRT12 mutation-specific targeting. *Gene Ther.* **23,** 108–112 (2016).
4. Montague, T. G., Cruz, J. M., Gagnon, J. A., Church, G. M. & Valen, E. CHOPCHOP: a CRISPR/Cas9 and TALEN web tool for genome editing. *Nucleic Acids Res.* **42,** W401–7 (2014).
5. Labun, K., Montague, T. G., Gagnon, J. A., Thyme, S. B. & Valen, E. CHOPCHOP v2: a web tool for the next generation of CRISPR genome engineering. *Nucleic Acids Res.* **44,** W272–6 (2016).
6. Lindsay, H. *et al.* CrispRVariants charts the mutation spectrum of genome engineering experiments. *Nat. Biotechnol.* **34,** 701–702 (2016).
7. Gagnon, J. A. *et al.* Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. *PLoS One* **9,** e98186 (2014).
8. Kuscu, C. *et al.* CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nat. Methods* **14,** 710–712 (2017).
9. Lackner, D. H. *et al.* A generic strategy for CRISPR-Cas9-mediated gene tagging. *Nat.*

Commun. **6,** 10237 (2015).

10. Pages, H., Gentleman, R., Aboyoun, P. & DebRoy, S. Biostrings: String objects representing biological sequences, and matching algorithms, 2008. *R package version* **2,** 160
11. Xie, Y. knitr: A general-purpose package for dynamic report generation in R. *R package version* **1,** 1 (2013).
12. Wickham, H. & Wickham, M. H. The ggplot package. (2007).
13. Yin, T., Cook, D. & Lawrence, M. ggbio: an R package for extending the grammar of graphics for genomic data. *Genome Biol.* **13,** R77 (2012).
14. Wickham, H. & Francois, R. dplyr: A grammar of data manipulation. *R package version 0. 4* **1,** 20 (2015).
15. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9,** e1003118 (2013).
16. Pinello, L. *et al.* Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat. Biotechnol.* **34,** 695–697 (2016).
17. Varshney, G. K. *et al.* High-throughput gene targeting and phenotyping in zebrafish using CRISPR/Cas9. *Genome Res.* **25,** 1030–1042 (2015).

Acknowledgements

We would like to thank Jason Rihel, Tessa Montague and Alex Schier for support and many useful comments, and members of the Schier lab for their contributions.

The project was supported by the Bergen Research Foundation and the Norwegian Research Council (FRIMEDBIO #250049) (E.V.), University of Bergen core funding (K.L.) and the American Cancer Society and University of Utah startup funding (J.A.G.).

Author information

Affiliations

Computational Biology Unit, Department of Informatics, University of Bergen, 5008 Bergen, Norway.

Kornel Labun & Eivind Valen

Department of Biology, University of Utah, Salt Lake City, UT 84112, USA.
James A Gagnon

Sars International Centre for Marine Molecular Biology, University of Bergen, 5008 Bergen, Norway.
Eivind Valen

Department of Pathology and Cell Biology, Columbia University, New York, NY 10032, USA.
Alejandro Chavez

Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA 02115, USA.
Department of Genetics, Harvard Medical School, Boston, Massachusetts, 02115, USA.
Xiaoge Guo

Wyss Institute for Biologically Inspired Engineering, Harvard University, Cambridge, MA 02115, USA.
Department of Genetics, Harvard Medical School, Boston, Massachusetts, 02115, USA.
George Church

Contributions

E.V. conceived and supervised the project. J.A.G. performed wet-lab experiments and prepared datasets. X.G., G.C. and A.C. assisted in data interpretation and writing the manuscript. K.L. developed the R package and performed all computational work.

Competing financial interests

The authors declare no competing financial interests.

Corresponding authors

Correspondence to: Eivind Valen; Tel: +47 55584074; Email: eivind.valen@gmail.com