

10. Leinonen, R., Sugawara, H. & Shumway, M. & International Nucleotide Sequence Database Collaboration. *Nucleic Acids Res.* **39**, D19–D21 (2011).
11. Dyke, S. O. M. et al. *Eur. J. Hum. Genet.* **24**, 1676–1680 (2016).
12. Erlich, Y. & Narayanan, A. *Nat. Rev. Genet.* **15**, 409–421 (2014).
13. Homer, N. et al. *PLoS Genet.* **4**, e1000167 (2008).
14. Shringarpure, S. S. & Bustamante, C. D. *Am. J. Hum. Genet.* **97**, 631–646 (2015).
15. Dyke, S. O. M. et al. *PLoS Genet.* **12**, e1005772 (2016).
16. Dyke, S. O. M., Dove, E. S. & Knoppers, B. M. *Genomic Med.* **1**, 16024 (2016).
17. Raisaro, J. L. et al. *J. Am. Med. Inform. Assoc.* **24**, 799–805 (2017).
18. Philippakis, A. A. et al. *Hum. Mutat.* **36**, 915–921 (2015).
19. Buske, O. J. et al. *Hum. Mutat.* **36**, 922–927 (2015).
20. Peplow, M. *Br. Med. J.* **353**, i1757 (2016).
21. Ashley, E. A. *J. Am. Med. Assoc.* **313**, 2119–2120 (2015).
22. Mailman, M. D. et al. *Nat. Genet.* **39**, 1181–1186 (2007).

Acknowledgements

J. Ostell conceived the project; Global Alliance for Genomics & Health provided substantial guidance and support. The Beacon Project team designed and developed the Beacon API. Members of various organizations implemented Beacons and contributed to its APIs. We are thankful for data contributors who elect to share their data. M.F. and S.O.M.D. are supported by Genome Quebec, Genome Canada, the Government of Canada, and

the Ministère de l'Économie, Innovation et Exportation du Québec (Can-SHARE grant 141210); S.O.M.D. is supported by the Canadian Institutes of Health Research (grants EP1-120608; EP2-120609) and the Canada Research Chair in Law and Medicine; M.H. is supported by BD2K NIH/NCI 5U54HG007990-02; S. Scollen, S.V., M.B., I.L., J.T., S.U.-R., S.d.I.T., M.L., H.S. and the EGA are supported by ELIXIR, the research infrastructure for life-science data. This work was supported by ELIXIR-EXCELERATE, funded by the European Commission within the Research Infrastructures programme of Horizon 2020, grant agreement number 676559 (J.D.S., I.L.), the Wellcome Trust grant numbers WT201535/Z/16/Z (P.F.) and WT098051 (S.K., D.L., P.F.), and the European Molecular Biology Laboratory (P.F., S.K., J.D.S., I.L.); A.J.B. is supported by the European Union FP7 Programme 'EMIF' IMI-JU grant no. 115372, and H2020 Programme 'GCOF' grant no. 643439.

Author contributions

M.F., S. Scollen, G.S., S.V., S.K., D.L., P.G., S. Sherry, M.B., I.L. and D.H. provided project leadership and management; M.C., J.R., S.d.I.T., J.T., K.C., A.J.B., M.H., M.B., H.S., M.L., J.D.S. and S.U.-R. designed and developed software; S.O.M.D. developed ethics and policy research; M.F. and M.C. designed and developed the Beacon Network;

P.F. provided security review; M.F. and L.D. wrote the manuscript with contributions from all other authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41587-019-0046-x>.



Open Access This article is licensed under a Creative Commons Attribution 4.0

International License, which permits

use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

CRISPResso2 provides accurate and rapid genome editing sequence analysis

To the Editor — The field of genome editing is advancing rapidly¹, most recently exemplified by the advent of base editors that enable changing single nucleotides in a predictable manner^{2–4}. For the validation and characterization of genome editing experiments, targeted amplicon sequencing has become the gold standard⁵. Here we present a substantially updated version of our CRISPResso tool^{6–9} to facilitate the analysis of data that would be difficult to handle with existing tools^{6–9}.

CRISPResso2 introduces five key innovations: first, comprehensive analysis of sequencing data from base editors; second, a batch mode for analyzing and comparing multiple editing experiments; third, allele-specific quantification of heterozygous or polymorphic references; fourth, a biologically informed alignment algorithm; and fifth, ultrafast processing time. We discuss each of these in turn below.

Our updated software allows users to readily quantify and visualize amplicon sequencing data from base-editing experiments. It takes as input raw FASTQ sequencing files and outputs reports describing frequencies and efficiencies of base editing activity, plots showing base substitutions across the entire amplicon region (Fig. 1a), and nucleotide substitution frequencies for a region specified by the user (Fig. 1b). Users can also specify the nucleotide

substitution (for example, C→T or A→G) that is relevant for the base editor used, and the software produces publication-quality plots for nucleotides of interest with heat maps showing conversion efficiency.

We also improved processing time and memory usage of CRISPResso2 to enable users to analyze, visualize and compare results from hundreds of genome editing experiments using batch functionality. This is particularly useful when many input FASTQ files must be aligned to the same amplicon or have the same guides, and the genome editing efficiencies and outcomes can be visualized together. In addition, CRISPResso2 generates intuitive plots to show the nucleotide frequencies and indel rates at each position in each sample. This allows users to easily visualize the results and extent of editing in their experiments for different enzymes (Fig. 1c).

In cases where the genome editing target contains more than one allele (for example, when heterozygous single nucleotide polymorphisms (SNPs) are present), genome editing on each allele must be quantified separately, even though reads from both alleles are amplified and mixed in the same input FASTQ file. Current strategies are not capable of analyzing multiple reference alleles and may lead to incorrect quantification. CRISPResso2 enables allele-specific quantification by

aligning individual reads to each allelic variant and assigning each read to the most closely aligned allele. Downstream processing is performed separately for each allele so that insertions, deletions or substitutions that distinguish each allele are not confounded with genome editing. To demonstrate the utility of our approach, we reanalyzed amplicon sequencing data from a mouse with a heterozygous SNP at the *Rho* gene in which an engineered SaCas9-KKH nuclease was directed to the P23H mutant allele¹⁰. CRISPResso2 deconvoluted reads, quantified insertions and deletions from each allele, and produced intuitive visualizations of experimental outcomes (Fig. 1d).

Existing amplicon sequencing analysis toolkits ignore the biological understanding of genome editing and instead optimize the alignment on the basis of sequence identity only. However, this can lead to incorrect quantification of indel events, especially in sequences with short repetitive subsequences where the location of indels may be ambiguous as a result of multiple alignments with the same best score. In such cases, it is reasonable to assume that indels should overlap with the predicted nuclease cleavage site. Our improved alignment algorithm extends the Needleman–Wunsch algorithm with a mechanism to incentivize the assignment of insertions or deletions to

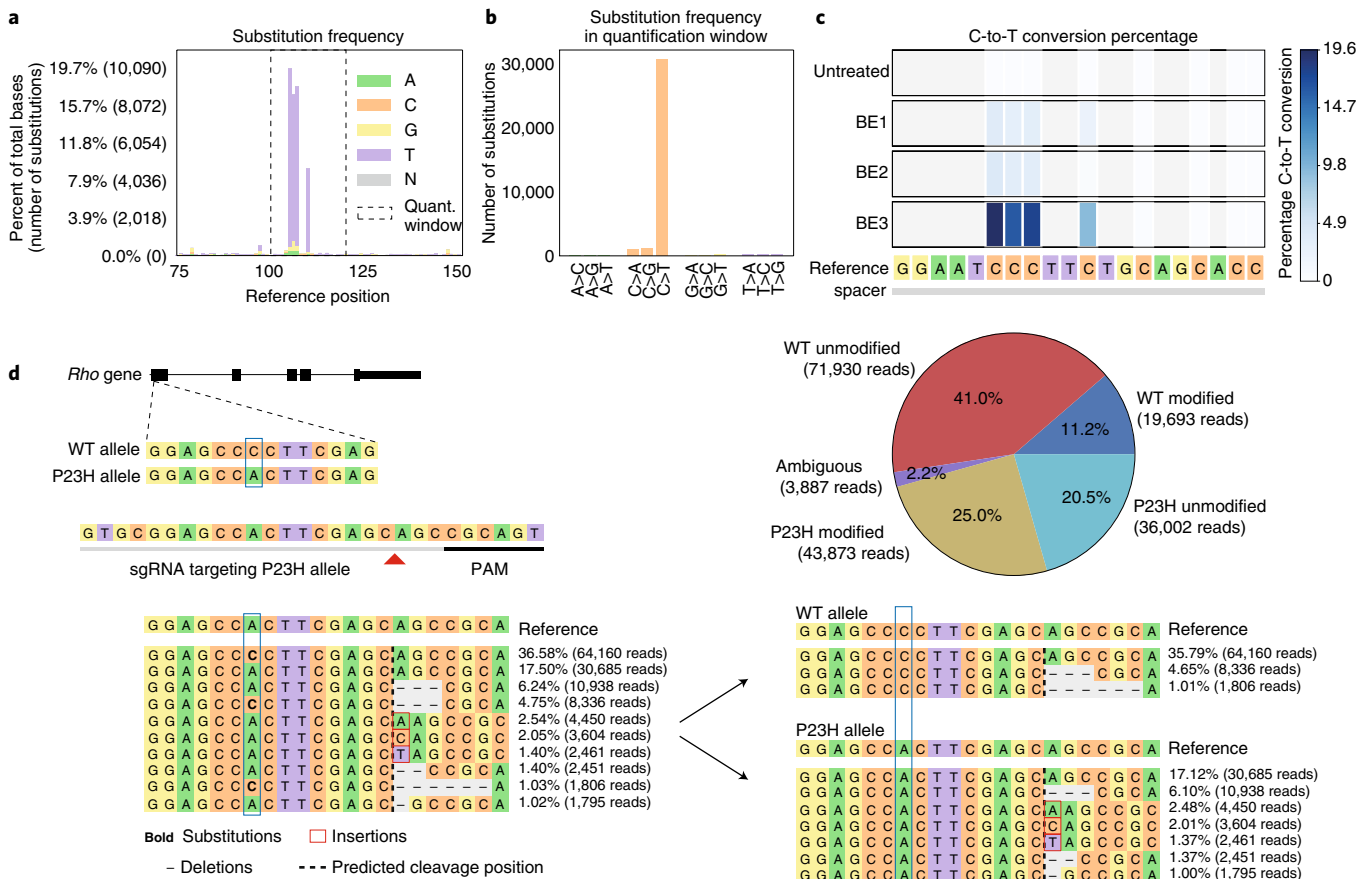


Fig. 1 | Novel features of CRISPResso2. **a-c**, CRISPResso2 analysis of base editing data. **a**, Locations of substitutions across the *FANCF* reference sequence for the BE3 base editor². At each position, the number of substitutions from the reference base to each non-reference base are shown. The quantification (quant.) window is outlined by the dashed box. **b**, Bar plot showing the frequency of substitution from a reference base to a non-reference base including only bases in the quantification window from **a**. **c**, Batch output mode comparing the editing efficiencies of three base editors and an untreated control at the *FANCF* locus². C>T conversion rates are shown at each cytosine overlapping the guide. **d**, Allele-specific editing outcomes of SaCas9-KKH editing of the *Rho* gene in P23H heterozygous mice¹⁰. Reads (left) can be assigned to each allele using CRISPResso2 (right) to achieve accurate quantification of genome editing at genomic loci with multiple alleles. The pie chart shows the assignment of each read to the wild-type (red and dark blue) allele or to the P23H allele (yellow and turquoise). Ambiguous alignments that could not be attributed uniquely to one of the alleles (e.g., owing to a deletion at the SNP location) are shown in slate blue.

specific indices in the reference amplicon sequence. These indices are chosen on the basis of guide sequence, predicted cleavage site and nuclease properties (Supplementary Note 1). This approach increases the accuracy of indel calling and produces alignments that reflect our current understanding of the editing mechanism. We compared our improved alignment algorithm to those used in other amplicon-based genome editing analysis software and found that our algorithm avoids the incorrect alignment to regions distal from the predicted cut site observed for other software tools (Supplementary Note 2).

To study putative off-targets, it is often necessary to analyze large-scale, pooled sequencing datasets that profile hundreds of sites to assess the potential safety of genome editing interventions⁵. These and other

large datasets have created a need for faster, more accurate and efficient analysis tools. To accelerate performance and decrease processing time, we designed an efficient implementation of our biologically informed alignment algorithm. Further optimization of other components of the processing pipeline has reduced processing time tenfold for large datasets, so that an experiment analyzed using modern high-throughput sequencing technologies can be processed in under a minute (Supplementary Fig. 1). We tested the accuracy of our improved alignment algorithm and other optimizations using an extensive set of simulations with various mutational profiles and in the presence of sequencing errors and found that CRISPResso2 accurately recovered editing events with a negligible false-positive rate (<0.01) limited only

by current sequencing technologies (Supplementary Note 3).

In summary, CRISPResso2 is a software tool for the comprehensive analysis, visualization and comparison of sequencing data from genome editing experiments. In addition to accurate indel analysis from nucleases, such as Cas9, CRISPResso2 offers analysis tools for recent base editors, support for multiple alleles, increased computational speed, an improved alignment algorithm, and a batch functionality for analyzing and comparing genome editing experiments (Supplementary Note 4).

Code availability

CRISPResso2 is available online at <http://crispresso2.pinellolab.org>, where users can run up to four samples simultaneously. The command line version without any

limitations and with additional tools is available as a Docker image at <https://hub.docker.com/r/pinellolab/crispresso2/> (Supplementary Note 4). The source code is available as Supplementary Software and online at <https://github.com/pinellolab/CRISPResso2>.

Data availability

Fig. 1a–c shows data available as SRR3305546 (untreated), SRR3305543 (BE1), SRR3305544 (BE2) and SRR3305545 (BE3)². Figure 1d shows data obtained from the authors of ref. ¹⁰. □

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

Kendell Clement^{1,2,3}, Holly Rees^{4,5,6}, Matthew C. Canver^{1,2,3}, Jason M. Gehrke^{2,3}, Rick Farouni^{1,2,3}, Jonathan Y. Hsu^{1,2,3}, Mitchel A. Cole⁷, David R. Liu^{4,5,6}, J. Keith Joung^{1,2,3} , Daniel E. Bauer^{1,7,8,9*} and Luca Pinello^{1,2,3*} 

¹Broad Institute of MIT and Harvard, Cambridge, MA, USA. ²Molecular Pathology Unit, Center for Cancer Research and Center for Computational and Integrative Biology, Massachusetts General Hospital, Charlestown, MA, USA. ³Department of Pathology, Harvard Medical School, Boston, MA, USA.

⁴Merkin Institute of Transformative Technologies in Healthcare, Broad Institute of MIT and Harvard,

Cambridge, MA, USA. ⁵Howard Hughes Medical Institute, Harvard University, Cambridge, MA, USA.

⁶Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA, USA.

⁷Division of Hematology/Oncology, Boston Children's Hospital; Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA.

⁸Department of Pediatrics, Harvard Medical School, Boston, MA, USA. ⁹Harvard Stem Cell Institute, Cambridge, MA, USA.

*e-mail: daniel.bauer@childrens.harvard.edu; lpinello@mgh.harvard.edu

Published online: 26 February 2019
<https://doi.org/10.1038/s41587-019-0032-3>

References

1. Tsai, S. Q. & Joung, J. K. *Nat. Rev. Genet.* **17**, 300–312 (2016).
2. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. *Nature* **533**, 420–424 (2016).
3. Komor, A. C. et al. *Sci. Adv.* **3**, eaao4774 (2017).
4. Kim, Y. B. et al. *Nat. Biotechnol.* **35**, 371–376 (2017).
5. Akcakaya, P. et al. *Nature* **561**, 416–419 (2018).
6. Pinello, L. et al. *Nat. Biotechnol.* **34**, 695–697 (2016).
7. Wang, X. et al. *Bioinformatics* **33**, 3811–3812 (2017).
8. Park, J., Lim, K., Kim, J.-S. & Bae, S. *Bioinformatics* **33**, 286–288 (2017).
9. Lindsay, H. et al. *Nat. Biotechnol.* **34**, 701–702 (2016).
10. Li, P. et al. *CRISPR J.* **1**, 55–64 (2018).

Acknowledgements

D.R.L. is supported by DARPA HR0011-17-2-0049; US NIH RM1 HG009490, R01 EB022376 and R35 GM118062; and the HHMI. J.K.J. is supported by DARPA HR0011-17-

2-0042, NIH R35 GM118158 and NIH RM1 HG009490. D.E.B. is supported by the NIDDK (R03DK109232), NHLBI (DP2OD022716, P01HL32262), Burroughs Wellcome Fund, Doris Duke Charitable Foundation and St. Jude Children's Research Hospital Collaborative Research Consortium. L.P. is supported by NHGRI (R01HG008399), DARPA HR0011-17-2-0042, and the Centers for Excellence in Genomic Science of the National Institutes of Health under award number RM1HG009490 through a New Collaborator Grant subaward.

Author contributions

K.C. and L.P. conceived the project, led the study and wrote the software. K.C. analyzed experimental data. All authors contributed input on measurement and visualization of genome editing outcomes and provided input on the manuscript.

Competing interests

At the time of manuscript preparation, J.M.G. was a consultant for Beam Therapeutics, and now is employed by Beam Therapeutics. J.K.J. has financial interests in Beam Therapeutics, Editas Medicine, Endcadia, EpiLogic Therapeutics, Pairwise Plants, Poseida Therapeutics and Transposagen Biopharmaceuticals. J.K.J.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies. J.K.J. is a member of the Board of Directors of the American Society of Gene and Cell Therapy. J.M.G. and J.K.J. are co-inventors on patents and patent applications that describe gene editing technologies. D.R.L. is a consultant and cofounder of Editas Medicine, Pairwise Plants and Beam Therapeutics, companies that use genome editing.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41587-019-0032-3>.