CHOPOFF: symbolic alignments enable fast and sensitive CRISPR off-target detection

AUTHORS
Kornel Labun[1 *], Oline Rio[2], Håkon Tjeldnes[3], Michał Swirski[4], Anna Zofia Komisarczuk[2], Emma Haapaniemi[2], Eivind Valen[1, 5, 6, *]

1. Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway
2. Centre for Molecular Medicine Norway, University of Oslo, Oslo, Norway
3. School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland
4. Institute of Genetics and Biotechnology, University of Warsaw, Warsaw, Poland
5. Michael Sars Centre, University of Bergen, Bergen, Norway
6. Department of Biosciences, University of Oslo, Oslo, Norway

* Corresponding authors: kornel.labun@gmail.com eivind.valen@ibv.uio.no

Abstract

CRISPR/Cas is a revolutionary technology for genome editing. Although hailed as a potential cure for a wide range of genetic disorders, CRISPR/Cas translation faces severe challenges due to unintended off-target editing. Predicting these off-targets are difficult and necessitates trade-offs between speed and sensitivity. Here, we develop the original concept of symbolic alignments to efficiently identify off-targets without sacrificing sensitivity. We also introduce data structures that allow near-instant alignment-free probabilistic ranking of guides based on their off-target counts. Implemented in the tool CHOPOFF, these innovations support mismatches, bulges and genomic sequence variation for personalized genomes while outperforming state-of-the-art methods in both speed and accuracy.

Availability: The CHOPOFF is available at https://github.com/JokingHero/CHOPOFF.jl

Main

CRISPR/Cas is a transformative genome editing technology that has dramatically improved our ability to introduce precise changes into genes and genomes. Due to its versatility and relative ease of use, CRISPR/Cas is widely employed in basic research, most notably functional genomics, disease modeling, synthetic biology, and many others. Additionally, given its capacity to edit genetic errors, CRISPR/Cas is also being evaluated for a broad range of medical applications, recently resulting in the regulatory approval of the first CRISPR-based therapeutics [1].

However, despite its revolutionary status, CRISPR/Cas technology does have notable limitations and liabilities. The most important limitation of CRISPR/Cas is the potential for off-target editing, whereby CRISPR/Cas cuts DNA at unintended locations. Such off-target (OT) edits can distort the interpretation of functional experiments and introduce noise and variability, thus decreasing reliability of experimental results and functional conclusions. Importantly, OT activity is especially dangerous in therapeutic applications of CRISPR where even a very low frequency OT editing may have profoundly disastrous outcomes [2,3]. To address this challenge, many efforts in the field have focused on improving guideRNA (gRNA) design to ensure target specificity [4] and engineering Cas variants with improved fidelity [5]. In parallel, methods to measure OT effects, such as GUIDE-seq [6], CIRCLE-seq [7] and SITE-seq [8], have also been instrumental in increasing our capacity to quantify and rationalize OT edits. Additionally, the ability to predict OTs has been of increasing importance for the field, leading to the development of a wide range of computational methods for predicting OT sites.

CRISPR/Cas OT prediction algorithms work by analyzing gRNA sequence and conducting a genome-wide search to identify sites with high complementarity with the gRNA and the presence of protospacer adjacent motif (PAM) sequence, which is essential for Cas binding. Given that gRNA has been shown to edit sites with up to 6 mismatches and even small bulges (insertions or deletions), predicting CRISPR OT effects represents a significant challenge [6,9]. Identifying potential OT sites is therefore equivalent to finding all short sequences in a genome that are likely to engage with a given gRNA despite the presence of mismatches, insertions, and deletions. Currently, commonly used sgRNA design tools such as CHOPCHOP [10] and CRISPOR [11] restrict OT searches to 3-4 mismatches and do not support bulges. Several tools that can handle mismatches and bulges, such as Cas-OFFinder [12], CRISPRitz [13] and SWOffinder [14], have been developed but they are computationally intensive and their running time makes these approaches impractical in larger experiments or web tools. Additionally, support for natural genomic variation (SNPs, insertions, and deletions) in existing software is limited [10–12,14], which restricts the utility of CRISPR/Cas technology for precision medicine applications.

Results

To inform development of an algorithm for OT prediction that overcomes these challenges, we asked how different can a sequence be to have a reasonable chance of being an OT?. As has been previously demonstrated, OTs with deletions and insertions are frequently cut [9,15], a fact we readily verified through a simple cell-free screen (**Supplementary Figure 1**). To investigate this at a larger scale, we reprocessed OT data from GUIDE-seq [6], dual-target [16], and SURRO-seq [17]. From this we calculated the presence of OT cuts and estimated cutting efficiency by normalizing to on-target efficiency (**Supplementary Figure 2**). This data revealed that although short bulges are less tolerated than mismatches, they are still cut at a significant level (**Supplementary Figure 3**). This argues that OT similarity should account for insertions and deletions and therefore use the Levenshtein metric[18] (edit distance) rather than the more commonly used Hamming metric (mismatches).

The Levenshtein metric expresses similarity between gRNA sequence and potential OT sequence by calculating the minimum number of single sequence changes (substitutions, deletions or insertions) needed to transform the gRNA target sequence into the potential OT sequence. While the number of potential OT targets grows exponentially with Levenshtein distance, we observed that the probability of each site being cut decreases dramatically with increased distance (**Figure 1A**, **Table 1**). This leads to the futile situation where the vast majority of predictions based on the use of Levenshtein distance that exceeds a certain threshold are false positives, i.e. sites that an algorithm would predict as “cuttable” that are in practice not going to be cut. Indeed, targets at an edit distance of 4 are extremely unlikely to be cut (0.36%) and if they are cut it is typically at a negligible efficiency (0.78%, **Figure 1A**, **Table 1**). Taken together, this analysis suggests that an OT prediction algorithm must include the capacity to account for short bulges, in addition to mismatches, but at the same time minimize false positives by not searching for OTs that are too dissimilar.

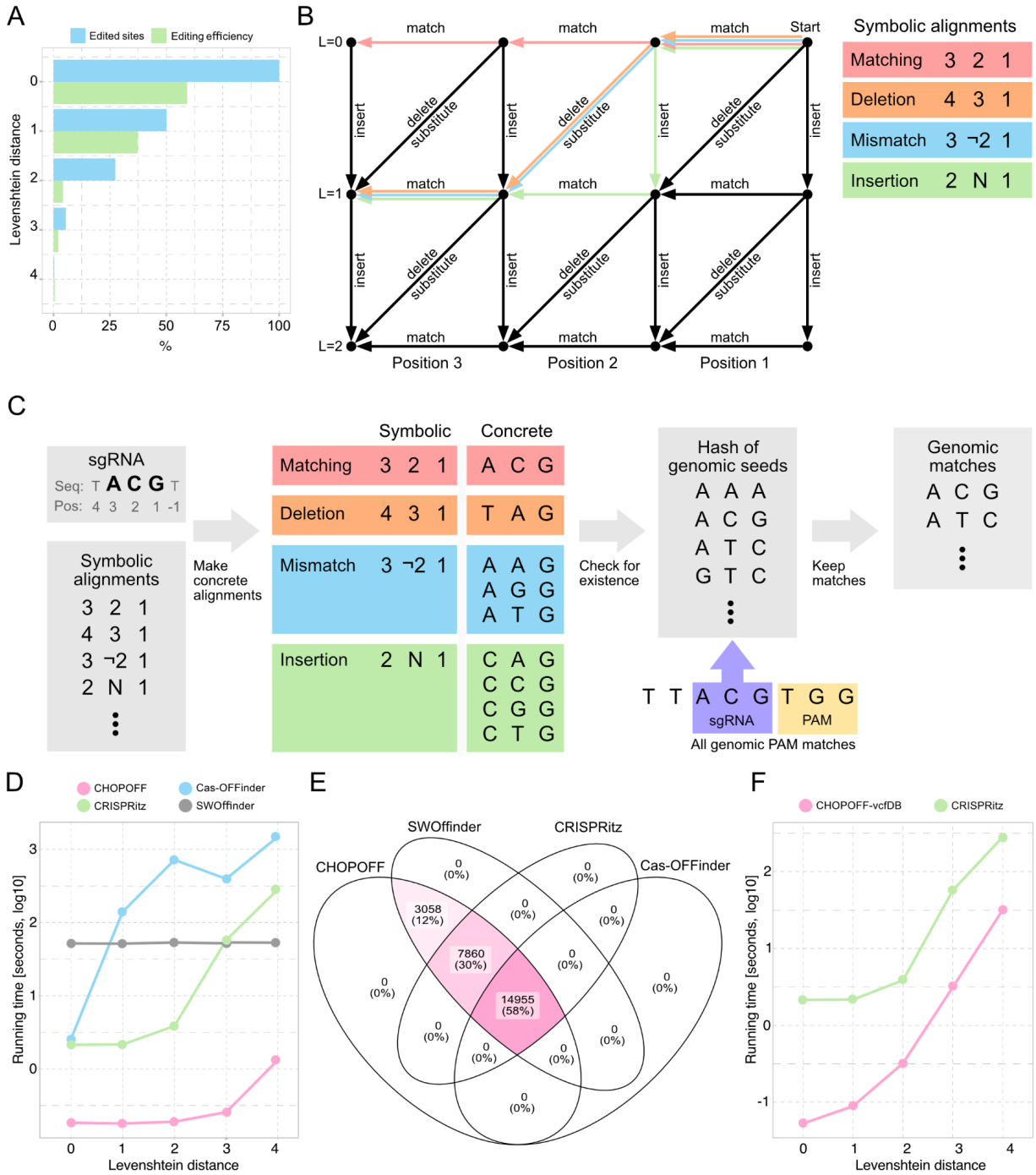

Figure 1. **Precomputed symbolic alignments combined with hashing are fast and precise at finding all off-targets. A)** GUIDE-seq data for 51 sgRNA shows the percentage of sites edited (blue) and the mean editing efficiency of those sites (green) with increasing Levenshtein distance (y-axis). **B)** All possible alignments for a sgRNA sequence can be pre-computed through enumeration of all paths in a graph (left). Red path highlights the only zero distance (L=0) alignment, while orange, blue and green paths show example alignments at distance one. Right panel shows examples of precomputed symbolic alignments corresponding to the paths highlighted in the graph. **C)** All symbolic alignment paths are pre-computed. When presented with a concrete sgRNA sequence the symbolic alignments are

substituted with the nucleotides of the sgRNA giving a list of all possible off-target sequences for that sgRNA. These are checked against a genomic index and if a match is found a Needleman-Wunch alignment is made between the sgRNA and the loci. **D)** Running time for methods supporting bulges with increasing Levenshtein distance. **E)** Venn diagram showcasing sensitivity of OT search by each tool. **F)** Running time for methods supporting genomic variation.

Aligning sgRNA sequences with a genome at arbitrary Levenshtein distance is a time-consuming computational operation. However, since the length of the sgRNA and PAM positioning are known prior to the search, and given our conclusions regarding the optimal Levenshtein distance, we reasoned that precomputing all possible OT sequences in advance would allow us to identify OTs without the need to align all sgRNAs with the entire genome. We implemented this precomputing framework into CHOPOFF, a standalone software and framework for identifying potential OTs based on the idea of preprocessing both the sgRNA and the genome. CHOPOFF is an order of magnitude faster than comparable methods while supporting arbitrary Levenshtein distance, genomic variation and ambiguous bases through IUPAC codes.

Symbolic alignments enable fast and sensitive CRISPR off-target detection

The genome is preprocessed by identifying all potential target sites defined by being adjacent to a valid PAM sequence (~230 * 10^6 for SpCas9 in humans). These sites are stored using two bit encoding in a lookup table of k-mer prefixes starting from the PAM site. Since bulges can result in sgRNAs of different lengths and to limit space usage for the index we restrict the prefix length to 16nt (0.5 GB for human genome). Since 16nt is shorter than a typical sgRNAs we store all suffixes of each prefix that is present in the genome. This genomic index enables instant lookup of k-mers adjacent to the PAM followed by an efficient local alignment step to the complete target site (prefix and suffix). If this alignment is within the predefined distance L a potential OT is identified.

This genomic index allows instant lookup of k-mers, but can not identify OTs at Levenshtein distance greater than 0. To accomplish this we instead preprocess the sgRNA to generate all possible k-mers within a given Levenshtein distance L. Generating these corresponds to following all paths through an alignment graph (**Figure 1B,** showing L=2) where the number of paths grows exponentially with L (**Supplementary Fig 4**). Generating these k-mers is therefore a costly operation. However, a key insight is that these paths are identical for all sgRNAs and can therefore be abstracted and reused for all sgRNAs. Exploiting this property we created a symbolic sgRNA defined not in terms of nucleotides, but by the index of the positions relative to the PAM. We then followed all paths through the alignment graph performing a k-mer expansion on this symbolic sgRNA to obtain a generalized “symbolic” k-mer outcome termed “symbolic alignments” (**Figure 1C**). Using this precomputed list of symbolic alignments allows us to skip the expansion step for any concrete sgRNA and instead substitute the sgRNA nucleotides at their respective positions in the symbolic alignments, thereby obtaining a specific “concrete” k-mer expansion for that sgRNA. Such substitution is efficient and takes a fraction of the time it takes to perform graph traversal for a given sgRNA. When complete information of all OTs is not

required, early stopping can be used to further increase speed and abort searches after a maximum number of OTs has been reached.

The combination of these two preprocessing steps (the genomic index and the symbolic alignments) generates a highly efficient sgRNA target search algorithm. When compared to the leading tools SWOffinder, CRISPRitz and Cas-OFFinder, CHOPOFF is an order of magnitude faster (**Figure 1D**) without losing sensitivity (**Figure 1E**). In fact, most tools do not correctly retrieve the full set of OTs at a given distance and for our test cases only CHOPOFF and SWOffinder are capable of correctly retrieving all OTs.

We expanded this algorithm further by allowing ambiguous bases, mismatches and indels to support genomic variation and methods for precision medicine. CHOPOFF-*vcfDB* accomplishes this by extending the genomic index to keep track of all possible ambiguous base combinations allowing for multiple overlapping variants of any type (insertion/deletion/mismatch). This comes at the cost of a modest decrease in speed, although the process is still significantly faster than the fastest other tool, CRISPRitz (**Figure 1F**).

Binary Fuse Filters allow for near-instant estimation of off-target events

When a very large number of guides is evaluated, heuristics based on alignment-free pre-filtering of low-quality sgRNAs can produce near-instant results. The recently developed Binary Fuse Filters (BFF) provide fast approximate set membership while using very little memory [19]. This probabilistic filtering offers a guarantee that it will never miss a sgRNA that is in the set (i.e. no false negatives), but may produce a false positive with low probability. To encode all unique sgRNA and their counts we created a stack of BFFs where each BFF encodes all sgRNAs with a specific number of OTs (**Figure 2A**). We define sgRNA with 10 or more OTs to be ineligible for editing purposes and store these sgRNAs in one BFF as a final tenth layer. This data structure can be stored using only 503MB of data for the human genome (hg34, UInt16 encoding), as compared to a complete dictionary of all k-mers with 3.1GB. The vast majority of all sgRNA are unique (95.61%), and their counts follow a long-tailed distribution enabling a layered database design (**Supplementary Figure 5**).

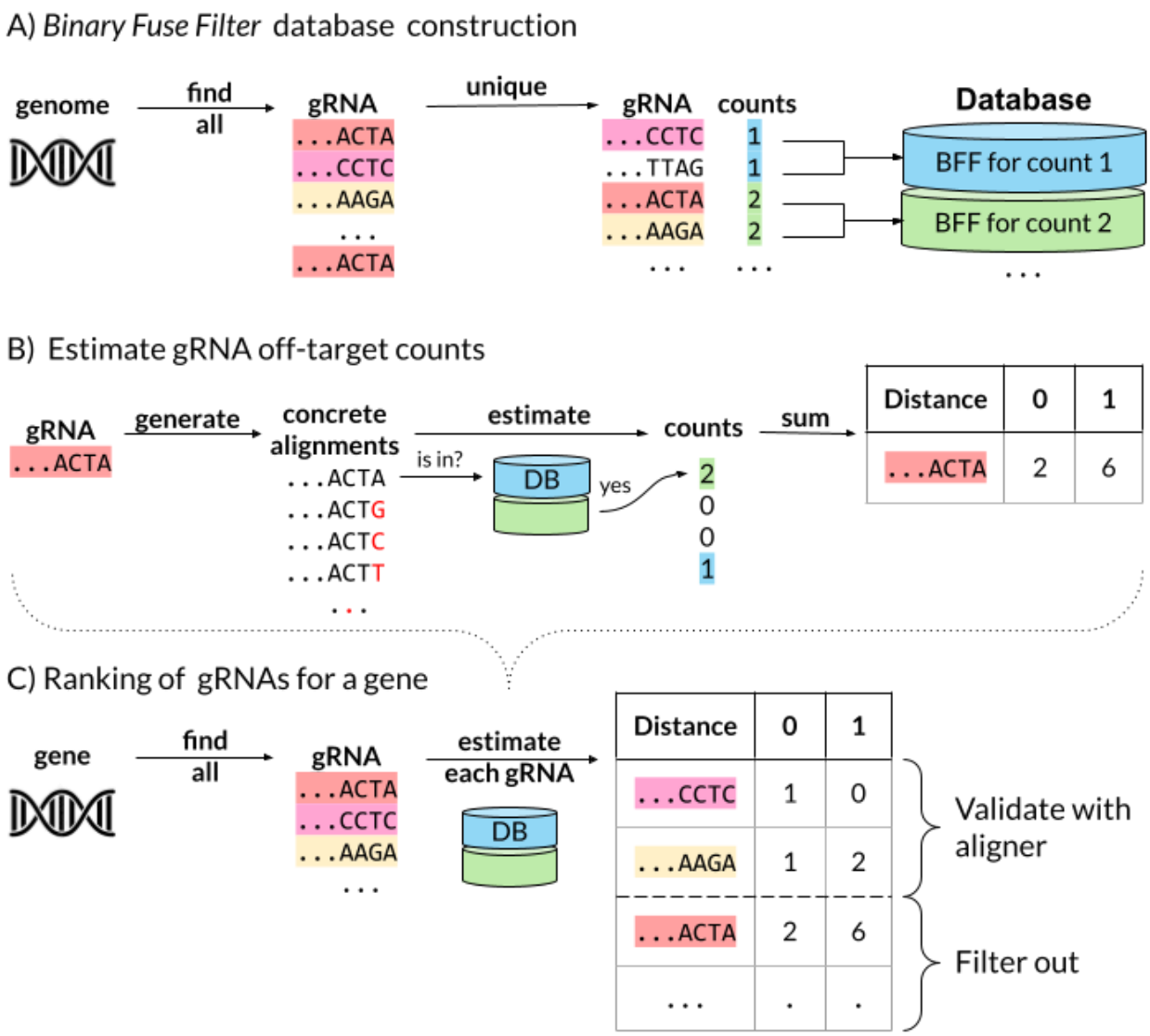

Figure 2. **Near-instant alignment-free estimation of off-target counts with CHOPOFF. A**) Grouping guide sequences by the number of occurrences in the genomer allows them to be stored in probabilistic and space efficient data structures - binary fuse filters (BFF). Each filter stores a guide with a specific number of occurrences in the genome from 1 to 9 and a final BFF that stores 10 and above. **B)** When a sequence is queried it is first expanded using symbolic alignments to its concrete alignments. These are sequentially checked for presence in each of the BFFs. The first hit gives an estimate of the number of occurrences for this concrete alignment. The sum of occurrences for all concrete alignments gives an estimate of off-targets. BFFs only make false positive errors at low probability, and there is no risk of false negatives. **C**) For a gene of interest, all sgRNAs can rapidly be filtered for low-quality sgRNAs with high off-target estimates. sgRNAs that pass the filter can be verified with our previously described genomic index approach or with an aligner supporting large distances.

When the OT count of a sgRNA needs to be determined, the guide is first expanded to all possible concrete alignments and each of these are evaluated in the layered BFFs from low counts to high counts (**Figure 2B**). Due to the probabilistic nature of BFFs this may result in

underestimating the off-target counts for some sgRNAs, but will likely not reject any sgRNAs that have low off-target counts. Importantly, when sgRNA is reported as off-target free it is guaranteed to be true. Estimating OTs for a sgRNA is achieved through enumerating the occurrence counts for all concrete alignments for a given sgRNA (**Figure 2B,C**). For larger search operations such as all sgRNA candidates in a gene this allows for near-instant filtering of all sgRNAs that have a high OT count and restricts exhaustive search to only those sgRNAs that have a potential to be highly specific. For larger servers where space and memory is not an issue CHOPOFF also supports use of a traditional dictionary data structure replacing the BFF, which decreases estimation errors.

Discussion

CHOPOFF introduces methods for OT prediction that are significantly faster and more accurate than the current state-of-the-art methods. Thus, CHOPOFF allows for more sensitive searches and higher quality, particularly relevant for large-scale experiments. It should be noted that CHOPOFF does not provide a general sequence alignment technique; rather it is a strategy optimized for sgRNAs, as efficiency comes at the expense of requiring separate indices for each PAM. It is also restricted to a limited prefix length that is intractable for mapping longer sequences. However, symbolic alignments introduce a paradigm which can be generalized to speed up other seed-based aligners.

CHOPOFF also have implementations of other proof-of-concepts algorithms (see **Supplementary Material** or CHOPOFF documentation) such as Vantage-point tree [20], pigeonhole principle [21], 01*0 seed [22], which were implemented as part of our research into the most performant solution. The methods are open source (**Table 2**), implemented in Julia and compatible with crisprVerse [23] through the crisprCHOPOFF R package. All CHOPOFF methods have above 95% test coverage ensuring robustness and high quality of the software.

Taken together, CHOPOFF is a fast and sensitive method that addresses the main challenges associated with predicting potential off-target sites for CRISP/Cas technology. Importantly, CHOPOFF supports prediction of OT in the context of genomic variation, which has important implications for the use of CRISPR/Cas in personalized medicine.

Online Methods

Confirming off-target potential for small bulges

To make a simple experiment to demonstrate off-target (OT) potential for bulges we performed a simple cell-free screening for 27 variable guides on the same on-target loci - where alignment between sgRNA and target will create variable sgRNA-OT conditions such as mismatches, bulges and deletions (**Supplementary Figure 1A**). The outcome was measured by gel electrophoresis. As can be seen (**Supplementary Figure 1B**) some deletions and insertions between sgRNA-OT are tolerated and create detectable cut rates. This is merely mechanistic Cas9 cutting potential and does not account for cell like environment and repair mechanisms.

Reanalysis of existing datasets

At a larger scale we reanalyzed OT datasets from three different studies: GUIDE-seq [24], dual-target [16], SURRO-seq [17]. We used the data already processed by those studies and available in the supplementary files. Our analyses and datasets are publicly available in the gitlab repository (https://git.app.uib.no/valenlab/offtarget_reanalysis). Dual-target and SURRO-seq datasets consist of artificial synthetic oligo sets of sgRNA-OT pairs, while GUIDE-seq data is analysis of actual genomic OTs.

We wanted to understand the relation between editing efficiency at the OT site and Levenshtein distance to the on-target sgRNA. Current studies (including the data sources) typically simplify the relationship between sgRNA and OT sites by only considering Hamming distance, which results in the inaccurate classification of some OT loci. We therefore realigned the sgRNA and OT sites using CHOPOFF with Levenshtein distance as a metric. During our analysis we made some notable observations listed below.

GUIDE-seq can be used to estimate off-target editing efficiency

GUIDE-seq data is highly correlated with PCR based methods of OT validation, which was shown in the original paper [6] and later confirmed [24]. As can be seen on **Supplementary Figure 2A** we replicated this relationship in our analysis, however GUIDE-seq reads can't be used directly to evaluate editing efficiency. We noticed that simply normalizing the GUIDE-seq data by the on-target editing efficiency measured independently (with PCR for example) approximates true OT editing efficiencies. This relationship is described in **Equation 1** below and validated with actual OT editing efficiencies in **Supplementary Figure 2B**. This simple normalization that uses on-target editing efficiency will allow us to repurpose GUIDE-seq reads from many studies and give us a more comprehensive dataset of real OT editing efficiencies. Applying similar normalization for CHANGE-seq does not yield predictive results.

$$frequency_i = \frac{readcount_i \times frequency_o}{readcount_o}$$

Where:

- $frequency_i$ - Editing efficiency at off-target i
- $readcount_i$ - Read count at off-target i, estimated by GUIDE-seq
- $frequency_o$ - Editing efficiency at on-target site, estimated by rhAmpSeq/PCR
- $readcount_o$ - Read count at on-target site, as estimated by GUIDE-seq

Equation 1. Inferring indel frequency on the off-target sites using GUIDE-seq reads.

Modeling relationship of Levenshtein distance and indel frequency

In **Supplementary Figure 3** we realigned sgRNA and OT pairs using CHOPOFF to get Levenshtein and Hamming distances for three different biological datasets (GUIDE-seq [6,24],

dual-target [16], SURRO-seq [17]. Each of these datasets carries its own independent biases, but they all show similar conclusions. For the majority of OT loci both Hamming and Levenshtein metrics are identical (blue line on the plot), these loci therefore represent a ground truth, regardless of the metric. GUIDE-seq has 1033 ground truth loci, dual-target has 3314, while SURRO-seq has 2654.

Contrasting to these, OT loci where metrics differ can be used to highlight which metric is better at capturing the relationship to editing efficiency. The datasets have 229 of these loci for GUIDE-seq, 155 for dual-target and 126 for SURRO-seq. The metric that more closely resembles the ground truth is therefore better suited to better capturing editing efficiency. The Levenshtein metric is always equal to or lower in distance compared to Hamming. This property arises since Levenshtein distance is an extension of and can be reduced to Hamming distance, if there are only mismatches. In contrast, there can be loci where Hamming of distance can be very high (e.g. >10) that can be explained with a single DNA bulge, resulting in a Levenshtein distance of one. As can be seen from **Supplementary Figure 3** the Levenshtein distance more closely resembles the ground truth and we therefore believe this to be a better metric for capturing real OT sites.

From **Supplementary Figure 3** one can estimate the tradeoff between sensitivity and specificity for detection of the OT loci. All datasets agree in a substantial loss in editing efficiency with increasing distance. On average a distance of 3 is on the edge of detection by all methods, there is however large variance and some outliers. For instance, in these datasets there exists outliers with a distance of 4 with relative editing rates above 0.5, but this is no longer true for distances above 4. Larger and more comprehensive datasets can, in the future, give a more comprehensive overview and show whether these observations hold in general.

Computational predictions at greater distances produce many false positives

Based on the GUIDE-seq dataset from Lazzarotto et al. 2020 and CHOPOFF OT search we summarized how many sites are actually edited in relation to all potential sites for 51 sgRNAs **Table 1**. To combine results from many sgRNAs we normalized GUIDE-seq counts to the on-target site as described above. From the computational OT detection perspective we recommend not searching with a Levenshtein distance greater than 3.

Distance	Mean editing efficiency [%]	Mean relative editing efficiency [%]	Number of edited OTs sites	Number of OTs sites within distance	Number of edited sites as fraction [%]
0	61.27	100.00	51	51	100
1	38.57	68.55	4	8	50
2	7.92	17.87	77	281	27.40
3	2.57	5.73	390	7117	5.48
4	0.78	1.76	468	129297	0.36

Table 1. Summary of GUIDE-seq for 51 sgRNAs and their off-targets.

Benchmarks

For the benchmark all tools used the same reference for OT search, the human genome (GENCODE hg38 v34, Dec 2013), which each tool preprocessed to its own format. Database construction is not part of the benchmark. CHOPOFF prefixHashDB method was run with early stopping conditions set to a maximum of 10^6 OTs for each sgRNA for each distance, which ensured it found all OTs and did not use its early stopping condition. We searched with the same 61 sgRNAs for each tool and reported the mean running time. For the purpose of the analysis we allowed a Levenshtein distance up to 4, which ensured inclusion of bulges and mismatches. Cas-Offinder was installed from github (v 3.0.0, supporting bulges) and run with GPU (TigerLake-H GT1, faster than 15 threads on CPUs), while CRISPRitz (v 2.6.6) was installed through conda and run with 15 threads, CHOPOFF (v1.0.0) was also run with 15 threads. Benchmarks were run on a regular laptop machine with 32GB RAM and 16 processors 11th Gen Intel(R) Core(TM) i7-11850H @ 2.50GHz. Code link is available at **Table 2** below.

Output comparison of off-target searches

We restricted our comparison of the different tools to results at Levenshtein distance <=3 because CRISPRitz produces very redundant output showing many alignments for the same OT sites and for distance 4 its output of 160GB is too large to parse. Each program outputs OTs positions differently, and we gave the benefit of the doubt allowing for 5bp imprecision in reporting the OT position. All of CHOPOFF' algorithms return exactly the same result, in the same format. SWOffinder reports identical OT sites to CHOPOFF. Both CRISPRitz and Cas-OFFinder do not find any OT sites that are not also found by CHOPOFF and SWOffinder. However, CRISPRitz does not find 3058 sites and Cas-OFFinder does not find an additional

7860 OTs. We verified these OTs through independent string searching showing that the targets found by CHOPOFF and SWOffinder are correct. In the repository (chopoff-benchmark) we provide file `compare_software_results.html` which contains examples of these OT sites that are not detected. We notice that all OTs not reported by CRISPRitz and Cas-OFFinder are for distances 2 and 3, they come from a variety of sgRNAs and from different sites. While we have not further explored the details of their omission, we note that CRISPRitz missed sites are always in the form of a combination of bulges between reference and sgRNA.

Data availability

Repository	Description
https://github.com/JokingHero/CHOPOFF.jl DOI: 10.5281/zenodo.14039564	CHOPOFF software, Julia package
https://jokinghero.github.io/CHOPOFF.jl/	CHOPOFF documentation
https://github.com/JokingHero/crisprCHOPOFF DOI: 10.5281/zenodo.14039583	R package with CHOPOFF interface
https://git.app.uib.no/valenlab/chopoff-benchmark	Benchmark code and results
https://git.app.uib.no/valenlab/offtarget_reanalysis	Re-analysis of available datasets

Table 2. Code is freely available at github and gitlab.

Acknowledgments
We wish to thank all the members of the Valen and Haapaniemi labs and our funding sources, the Research Council of Norway (Project #331912) and The Norwegian Cancer Society (Project #190290).

Author contributions
Kornel Labun: Conceptualization, Implementation, Validation, Writing. Oline Rio and Anna Zofia Komisarczuk: Validation in the lab, Writing. Håkon Tjeldnes and Michał Swirski: R package and crisprVerse integration, Writing. Emma Haapaniemi and Eivind Valen: Conceptualization and Writing.

Competing interests
None declared.

REFERENCES

1. Sheridan, C. The world's first CRISPR therapy is approved: who will receive it? *Nat. Biotechnol.* **42**, 3–4 (2024).
2. Ikeda, M. *et al.* Unexpected Mutations by CRISPR-Cas9 CTG Repeat Excision in Myotonic Dystrophy and Use of CRISPR Interference as an Alternative Approach. *Mol Ther Methods Clin Dev* **18**, 131–144 (2020).
3. Cancellieri, S. *et al.* Human genetic diversity alters off-target outcomes of therapeutic gene editing. *Nat. Genet.* **55**, 34–43 (2023).
4. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
5. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nature Biotechnology* Preprint at https://doi.org/10.1038/s41587-020-0561-9 (2020).
6. Tsai, S. Q. *et al.* GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
7. Tsai, S. Q. *et al.* CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR--Cas9 nuclease off-targets. *Nat. Methods* **14**, 607 (2017).
8. Cameron, P. *et al.* Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods* **14**, 600–606 (2017).
9. Lin, Y. *et al.* CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
10. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research* Preprint at https://doi.org/10.1093/nar/gkz365 (2019).
11. Haeussler, M. *et al.* Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).

12. Bae, S., Park, J. & Kim, J.-S. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* **30**, 1473–1475 (2014).
13. Cancellieri, S., Canver, M. C., Bombieri, N., Giugno, R. & Pinello, L. CRISPRitz: rapid, high-throughput and variant-aware in silico off-target site identification for CRISPR genome editing. *Bioinformatics* **36**, 2001–2008 (2020).
14. Yaish, O., Malle, A., Cohen, E. & Orenstein, Y. SWOffinder: Efficient and versatile search of CRISPR off-targets with bulges by Smith-Waterman alignment. *iScience* **27**, 108557 (2024).
15. Yaish, O. & Orenstein, Y. Generating, modeling and evaluating a large-scale set of CRISPR/Cas9 off-target sites with bulges. *Nucleic Acids Res.* **52**, 6777–6790 (2024).
16. Fu, R. *et al.* Systematic decomposition of sequence determinants governing CRISPR/Cas9 specificity. *Nat. Commun.* **13**, 474 (2022).
17. Pan, X. *et al.* Massively targeted evaluation of therapeutic CRISPR off-targets in cells. *Nat. Commun.* **13**, 4049 (2022).
18. Berger, B., Waterman, M. S. & Yu, Y. W. Levenshtein distance, sequence comparison and biological database search. *IEEE Trans. Inf. Theory* **67**, 3287–3294 (2021).
19. Graf, T. M. & Lemire, D. Binary Fuse Filters: Fast and Smaller Than Xor Filters. *ACM J. Exp. Algorithmics* **27**, 1–15 (2022).
20. Yianilos, P. Data structures and algorithms for nearest neighbor search in general metric spaces. *ACM-SIAM Symposium on Discrete Algorithms* (1993).
21. Herstein, I. N. *Topics In Algebra*. (Blaisdell Publishing Company, 1964).
22. Vroland, C., Salson, M., Bini, S. & Touzet, H. Approximate search of short patterns with high error rates using the 01_*0 lossless seeds. *J. Discrete Algorithms* **37**, 3–16 (2016).
23. Hoberecht, L., Perampalam, P., Lun, A. & Fortin, J.-P. A comprehensive Bioconductor ecosystem for the design of CRISPR guide RNAs across nucleases and technologies. *Nat.*

Commun. **13**, 6568 (2022).

24. Lazzarotto, C. R. *et al.* CHANGE-seq reveals genetic and epigenetic effects on CRISPR-Cas9 genome-wide activity. *Nat. Biotechnol.* **38**, 1317–1327 (2020).