CRISPR-Detector: Fast and Accurate Detection, Visualization, and Annotation of Genome Wide Mutations Induced by Gene Editing Events

Authors

Lei Huang[1†], Dan Wang[2†*], Haodong Chen[3], Jinnan Hu[3], Xuechen Dai[2], Chuan Liu[2], Anduo Li[2], Xuechun Shen[5], Chen Qi[1], Haixi Sun[2], Dengwei Zhang[2], Tong Chen[4*] Yuan Jiang[2*]

Affiliations

1 College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China.

2 BGI-Shenzhen, Shenzhen, China.

3 Sentieon Inc, San Jose, CA, USA.

4 National Resource Center for Chinese Materia Medica, China Academy of Chinese Medical Sciences, Beijing, China.

5 School of Biology and Biological Engineering, South China University of Technology, Guangzhou, China.

*****To whom correspondence should be addressed. Fax: 408.648.2161; Email: yjiang@completegenomics.com.

Correspondence may also be addressed to Email: chent@nrc.ac.cn.

Correspondence may also be addressed to Email: wangdan6@genomics.cn.

† These authors contributed equally to this work.

Abstract

The leading edge of genome editing, driven by CRISPR/Cas technology, is revolutionizing biotechnologies. In pace with the burst of novel genome editing methods, bioinformatic tools to monitor the on/off-target events need to be more

evolved as well. Existing tools suffer from limitations in speed and scalability, especially struggle in the whole genome sequencing data analysis which is believed to replace amplicon sequencing and become mainstream data type due to the rapidly decreasing sequencing cost. Here, we introduce our CRISPR-detector tool, which is a web-hosted or locally deployable pipeline with five key innovations: 1) optimized scalability allowing for whole genome sequencing data analysis beyond BED file-defined regions; 2) improved accuracy benefited from haplotype based variant calling to handle sequencing errors; 3) treated and control sample co-analysis to remove background variants existing prior to genome editing; 4) integrated structural variation (SV) calling; 5) functional and clinical annotation of editing-induced mutations.

Introduction

Since the first report of CRISPR/Cas9 editing technology in 2012,[1] various novel CRISPR-based genome editing tools have been introduced by different groups.[2-6] From the initial methods that could only induce gene knockout to base editing,[1] which enables the direct conversion of one nucleotide to another without producing DSBs in the target sequence and the recent prime editing which can make targeted indels and all 12 types of point mutations,[7-10] genome editing methods have progressively become more flexible and precise. This opens up new exciting opportunities for a plethora of life science applications including therapeutic treatment, drug discovery, biofuel producing, agriculture and etc. Early clinical trials demonstrated CRISPR's general safety and utility and many more trials are under way.[11-13] The increased pace in CRISPR based treatment sets precision analysis, aka detecting on and off target events, a top priority for regulators including the FDA. In fact, FDA published guidance related to genome editing products in March 2022 (https://www.fda.gov/regulatory-information/search-fda-guidance-documents/human-gene-therapy-products-incorporating-human-genome-editing), where it raised

concerns for detection of gene editing events type and frequency multiple times. Therefore, the bioinformatic tools to monitor the on/off-target events need to be more accurate and capable of detecting true editing-induced mutations.

Several tools have been developed to analyze targeted amplicon next-generation sequencing (NGS) data. However, there are four main drawbacks for existing tools. The first limitation involves speed and scalability; most tools are designed to consider single or only a few amplicons for analysis. However, with the increasing need to comprehensively cover all potential off-target sites, large panel or even whole genome sequencing analysis should be conducted. Secondly, tools including CRISPR-GA,[14] AGEseq,[15] ampliconDIVider,[16] Cas-Analyzer and CRISPResso1/2 do not support the co-analysis of paired treatment/control sequencing data,[17-19] and thus previously existing genetic background SNPs (Single Nucleotide Polymorphisms) and indels could be interpreted as CRISPR/Cas-induced mutations and lead to the incorrect quantification of mutation events. Third, beyond SNPs and indels, structural variation is currently ignored by analysis. The fourth drawback is that, to our knowledge, no existing pipeline estimates the functional or clinal consequences resulting from editing-induced mutations. This annotation is quite important for researchers whose genome editing projects are for therapeutic purposes.

To deal with these limitations, we herein propose a one-stop CRISPR-detector platform. CRISPR-detector accepts raw FASTQ sequencing files from respective CRISPR/Cas-treatment and control experiments. Each input FASTQ file includes sequencing data from one or multiple amplicons, e.g., one on-target and its several off-target amplicons. The output of the web-hosted version of CRISPR-detector includes comprehensive analysis, intuitive visualization, and locus comparison. Specifically, CRISPR-detector primarily outputs: i) a summary webpage comparing and visualizing editing outcomes of the input amplicons; ii) detailed information on editing outcomes including frequency of SNPs and indels and a histogram of the indel sizes across the reference amplicon in the quantification window; iii) annotation of

editing-induced mutations showing whether encoded protein sequences are changed and filter-based annotation of mutations of clinical significance sourced from the ClinVar database,[20] including pathogenic, drug-response, histocompatibility, and others.

In addition to the web-hosted version, a locally deployable CRISPR-detector pipeline focusing on whole genome sequencing analysis and SV calling was developed. The pipeline not only calls regular SVs caused by the editing process but also virus vector insertion from viral-mediated gene editing. The whole genome analysis capability available with the locally deployable version is unique and not available in other gene editing pipelines.

Methods

In the analysis module, reads are mapped using Sentieon Minimap2, and short variants are called using Sentieon TNscope module.[21] Sentieon Minimap2 is an accelerated version of the corresponding opensource Minimap2,[22] returning identical results two times faster. Sentieon TNscope is developed on top of GATK Mutect2 v3.8 and is specifically optimized for low allele frequency variant calling and integrated structural variation calling functions. Later in the annotation stage, variants are annotated using ANNOVAR (version 2019 Oct24) with the ClinVar (version 20210501) database.[23]

Structural variations are called using TNscope, which is based on evidence from soft clips of mapped reads and the discordance of read pairs. Specifically, the calling occurs in three stages. In the first stage, the caller identifies all abnormally aligned reads from treatment and control samples (if available). Concurrently, it builds a local assembly around the split reads and filters out all haplotypes that exist in the control samples and noisy haplotypes with little read support. In the second stage, it analyzes the candidate haplotypes assembled around split reads and identifies consistent haplotype pairs. These reads represent potential break-ends of the structural

variants. In the third stage, the caller looks for evidence from the discordant read pairs.

Results

Pipeline Description

As shown in CRISPR-detector flowchart (Fig. 1), CRISPR-detector accepts single/pooled amplicon deep-sequencing as well as whole genome sequencing data as input. Input reads are mapped to amplicon sequence(s) or reference genome to call variants. If control group sample is provided, CRISPR-detector will utilize control group sample to remove background variants existing prior to genome editing for treatment group sample.

For targeted amplicon sequencing data, basic information including amplicon sequence(s) is compulsory and sgRNA sequence(s) is optional to input. After variants called, CRISPR-detector web server supports clinical annotation of mutations of human cell experiments sample, and functional annotation for other 9 species. Local deployable pipeline supports specific database built by users for mutations annotations. CRISPR-detector outputs graphs and tables showing quantifications and visualizations of the position and type of mutation(s) within the amplicon sequence(s).

For whole genome sequencing data analysis, reference genome is compulsory to input. The BED (Browser Extensible Data) format text file is optional and used for calling variants at regions of interest such as on/off target sites. CRISPR-detector is capable of calling variants involving SNPs and SVs. CRISPR-detector outputs VCF (Variant Call Format) format text file contains informations about variants called. As efficient tools for in vivo gene transfer virus vectors like AAV vectors (rAAV) may nonspecifically integrate into the host cell genome thus may cause insertional mutagenesis events. If virus vector sequence(s) is provided, CRISPR-detector will be capable of reporting the location of vector insertion events.

Input Webpage Module

On the home page, users can upload FASTQ files, provide basic information, and set parameters. FASTQ sequences from the treatment group are mandatory, whereas sequences from a control group are optional, but recommended. Each of the input FASTQ files can be sequencing data from a single amplicon or pooled amplicons.

Data Analysis Module

CRISPR-detector analyzes the input data in three main steps. In the first step, reads from both treatment and control (if available) samples are mapped to a reference sequence uploaded by the user, and then a quality control step is conducted when the quality scores of aligned reads are calculated and users will be warned if reads quality is lower than expected. Next, variants within quantification windows are called. In the treatment/control co-analysis mode, only editing-induced variants are called. In the third step, functional and clinically important variants in the treatment group data are annotated using the published ANNOVAR tool and ClinVar databases. Functional variants may potentially result in frameshift, nonsynonymous, start-loss, or stop-gain mutations, and clinical variants are those marked pathogenic or likely pathogenic in the ClinVar databases, which provides interpretation of variant correlation with disease.

Result Visualization Module

By clicking the "Demo" button on the webpage, example results for real dataset sequencing one on-target locus of the human *KRAS* gene and its corresponding 14 off-target loci is shown. This example result is from a published study,[24] and CRISPR-detector was used in analysis. The results webpage includes three parts:

1) A downloadable table summarizing the number of processed reads and the frequency of different types of mutations on all the input loci (Table 1).

2) A batch output mode comparing the editing outcomes from all input loci, producing figures comparing the general SNPs and indel ratios. It should be noted that ratios CRISPR-detector reports are considered as the maximum mutation ratio across the amplicon or users defined quantification window.
3) Results for individual input locus. The example figure shows the edited number position of the identified SNPs and indels (Fig. 2A) and indel size distribution (Fig. 2B). The reported positions are from co-analysis if control sample data are available. Furthermore, detailed information on mutations including the annotation on this locus are summarized in a downloadable table (Fig. 2C).

Locally Deployable Version

Due to data upload speed limitations, our web-hosted pipeline is inconvenient for processing whole genome sequencing data. To provide an alternative solution, a locally deployable version that accepts NGS data larger than 100 MB, including WGS sequencing data, has also been developed and is provided to the community (available at https://github.com/hlcas/CRISPR-detector).

There are three differences between the web version and this locally deployable version: 1) the local pipelines accept WGS and other large datasets; 2) structural variants can be called; and 3) visualization modules and annotation functions are not included, because the pipeline outputs VCF format files.

Performance Benchmarks

SNP/Indel Identification from Single Amplicon

The accurate detection of SNPs and indels is one of the key features for a gene editing analysis pipeline. To benchmark the accuracy of CRISPR-detector, we generated simulated testing reads representing actual sequencing data from on/off-target site amplicons and compared the accuracy with three other popular pipelines including “Cas-Analyzer”, “AGEseq”, and “CRISPResso2”. We introduced SNPs (1-3 per read)

and indels (5-50 bp length) into the simulated 150 bp paired-end reads at 50% frequency together with errors introduced from MiSeq default profile provided by ART simulation tool.[25] Accuracy is calculated by comparing the pipeline reported mutation percentage to the introduced frequency (50%). Average running time is calculated from 3 repeated runs. The speed of Cas-Analyzer and AGEseq are not shown as they were tested on different computation platforms.

Compared to CRISPResso2, CRISPR-detector showed 6 times faster processing speed, and the nearly 100% accuracy also outperformed CRISPResso2 (Table 2). It should be noted that CRISPR-detector's variant calling pipeline is based on haplotype assembly, which more easily eliminates sequencing errors compared with simple pileup-based read statistics.

Then we performed more detailed variant detection testing, to evaluate CRISPR-detector's performance on indels at different frequency and length. It should be noted that TNscope, as the variant calling module of CRISPR-detector, utilized haplotype assembling method to filter out sequencing error, and this strategy also makes it difficult to identify variant below 0.1% considering sequencing errors at same frequency level. So CRISPR-detector is capable of detecting editing events with >0.1% frequency. We designed 5 bp indel in a single simulated amplicon dataset, at 4 different frequencies starting from 0.1%. The testing dataset was processed by CRISPR-detector and CRISPResso2. CRISPR-detector almost reached 100% accuracy while 0.02% sequencing error indels reduced CRISPResso2's accuracy by 20% on the 0.1% indel frequency datapoint (Table 3).

In terms of indels at different length, we simulated additional datasets to test CRISPR-detector's limitation. Due to limitation brought by short reads length, CRISPR-detector's indel detection length is set as 50 bp for insertion, and 71 bp for deletion. To demonstrate this performance, we simulated 50 amplicon datasets with insertions ranging from 1-50 bp, and 72 amplicons with deletions ranging from 1-72 bp, all at 50% frequencies. CRISPR-detector maintains almost 100% accuracy across

1-50 bp insertions, and 1-71 bp deletions. As comparison, CRISPResso2's accuracy is around 95%, and failed to detect deletions larger than 50 bp (Fig. 3).

CRISPR-detector can perform co-analysis with control sample data as input to address existing germline mutations and possible somatic or hematopoiesis-related mutations that were not introduced by a gene editing event.[26] In this simulated co-analysis testing dataset, some of the variants were inserted in both treatment and control reads to benchmark the pipeline's ability to detect "True positive" mutations (1 bp insertion at position 105 bp) in a noisy background with 7% sequencing error.

For all read types except the 50 bp insertion, CRISPR-detector successfully identified true editing events without false positive from variants existing in control samples (Table 4). We tried to run the failed 50 bp insertion sample without a control group and it ran through. This result shows a limitation of CRISPR-detector in correctly mapping reads around such long and multiple insertions.

CRISPResso2 offers batch mode to output and visualize variants on the same location from different samples. While this brings convenience for comparing more than 2 samples together, the ability to directly filter germline and other existing variants from control samples, which significantly increases users' efficiency in recognizing and confirming target variants, is missing.

Scalability on pooled amplicons

In a few years' time, the scope of NGS-based on/off-target editing detection has expanded from only a few amplicons, to hundreds of amplicons, to the coverage of all potential off-target sites, and even to whole genome sequencing to ensure that every portion of the genome is screened. From a recent review, current mainstream in silico off-target prediction algorithms reach 95% sensitivity if the top 200 potential off-target sites are considered.[27] This finding indicates a requirement for analysis tools to consider speed and scalability during development. To deal with this need, CRISPResso2 provides two specific pipelines, CRISPRessoPooled and

CRISPRessoWGS, for these two input types. In comparison, CRISPR-detector's optimized high-performance modules are able to handle small and large datasets from one single pipeline to provide uncompromised accuracy and scalability.

For benchmarking, we generated 200 amplicons in the format of 150 bp paired-end reads, to represent a large panel in which SNPs, indels, and sequencing errors were introduced, similar to our single amplicon benchmark.

Accuracy and speed performance are no different than the single amplicon test, and CRISPR-detector consistently outperforms CRISPResso2 (Fig. 4). It should be noted that processing speed is not a bottleneck for small data because input/output limits the whole pipeline speed. However, for larger datasets, especially whole genome sequencing, alignment and variant calling are two critical steps affecting pipeline speed, and CRISPR-detector utilizes and benefits from high-performance modules.

Whole Genome Sequencing Performance

The ability to process whole genome sequencing is a unique feature of CRISPR-detector; CRISPRessoWGS requires a BED file to run WGS samples, and we believe the goal of WGS is to cover all regions and identify potential off-target events as opposed to considering only predicted high-probability regions. Actually CRISPRessoWGS was not run through even with BED file input. Therefore, we did not compare CRISPRessoWGS in the current benchmark but will include it in future benchmarking when the new version can be run through, and the BED file requirement can be removed.

Main difference of WGS and amplicon dataset are the regions to analysis (~3GB vs <1KB), and sequencing depth (<100x vs >10,000x). The whole genome screening brings higher false positive result, and less than sufficient sequencing depth makes false native unavoidable. To evaluate the accuracy of CRISPR-detector on a WGS analysis, we inserted 1000 SNPs and indels into one chromosome sequence using

BAMSurgeon.[28] Frequencies of inserted variants are ranging from 0.1% to 100%. However, considering much lower sequencing depth of WGS comparing to amplicon, we limited AF (Allele Frequency) for 100x WGS at 5% and above, and AF for 30x WGS at 15% and above. This is to ensure at least 3 reads to support a reported variant. Inserted reads and original reads are processed in co-analysis mode. Sensitivity is calculated by TP/(TP+FN), and specificity is calculated by TP/(TP+FP).

We found that co-analysis result is always better than only analyzing edited sample (Table 5). Accuracy of SNPs are lower than that from amplicons, but still acceptable, especially co-analysis on SNPs with AF >15%. Accuracy of indels are significantly lower than that from Amplicons, this is partially due to the fact that reads with indels are more difficult to be mapped to the correct location genome-wide, thus the reported indel frequency is lower than the correct number and may be filtered out by AF cutoff.

To validate this performance, a pair of 30x whole genome sequencing samples ("Mutant" and "Rescued") from a previous study was used.[29] "Mutant" means control cells having a pre-existing homozygous 1 bp disease causing deletion relative to the reference genome in the coding sequence of the mutant gene *Crygc*, "Rescued" means the edited clone rescued by CRISPR-Cas9 inducing 1 bp insertions 6-7 bp from the deletion.

CRISPR-detector completed the "Mutant" group WGS dataset in 3 repeated runs using a consistent and reasonable 4h19m running time with 15 threads allocated (in total 65 core-hours), and it successfully detected the homozygous 1 bp deletion (chr1 65118239) as expected. Similarly, the pipeline completed the "Rescued" group sample in 5h36m and detected 2 insertions 6-7 bp from the deletion (chr1 65118232; chr1 65118233).

Assessing Real World Clinical Data

CRISPR-detector is simultaneously suitable for assessing base editing outcomes. Published study reported a dual-deaminase base editor called SPACE (Synchronous Programmable Adenine and Cytosine Editor) which can concurrently introduce A•T-to-G•C and C•G-to-T•A substitutions.[30] CRISPR-detector's performance was compared with CRISPResso2 on 28 on target sites by SPACE as treatment and nCas9 as control. Among the 28 on target sites, CRISPR-detector and CRISPResso2 showed concordance in 23 sites. The other 5 sites (ABE site 17, ABE site 18, ABE site 19, HEK site 2 and HEK site 3) were suggested as a form of MNPs (Multiple Nucleotide Polymorphisms), which were shown as a combination of one deletion and one insertion instead of complex substitutions, due to haplotype assembly process (Supplementary Table 1).

Another dataset benchmarked was rhAMP-seq sample data provided by IDT website. For the true editing positions provided by IDT (detected by GATK4 with a tuned parameter), CRISPR-detector reached 100% accuracy with no false positive or false negative report (Supplementary Table 2).

Additionally, we analyzed 480 off-target sites from the third primer pool identified by CIRCLE-seq and CasOFFinder,[31-32] and amplified from genomic DNA from CD34+ SCD cells edited by ABE8e-NRCH or unedited control SCD donor cells using rhAMP-Seq (IDT).[33] Among the 480 sites, 27 sites were observed containing A•T-to-G•C mutations. CRISPR-detector showed discordance on "Off-target site 54", which we believe the signal was too weak to be considered as a real editing event. Therefore, we did not report this site (Supplementary Table 3).

Structural Variation Identification

Gene editing not only introduces short variants but additional structural variations that should not be excluded from analysis. A major purpose for conducting whole genome sequencing on edited samples is to better identify off-target structural variants. In

addition, in viral-mediated gene transfer events, it is important to know the exact genome insertion location of the virus vectors to evaluate off-target event occurrence and functional consequences. To benchmark the accuracy, we designed two benchmark datasets for both general SV calling and vector insertion events.

In the first part, we utilized the well-established PrecisionFDA dataset from HG002, in which a total 12,745 isolated indel calls are available, to serve as a truth set.[34] SVs were called from the CRISPR-detector pipeline and compared with the NIST truth set using Truvari (https://github.com/spiralgenetics/truvari). To better evaluate the performance of CRISPR-detector and to better understand the limitations of short read-based SV calling compared to co-barcoded reads and long reads, we also compared the result with some pipelines from a recently published study that used the same testing dataset.[35] Because insertions are more challenging and Truvari can only recognize the exact match to the truth set, which significantly underestimates true positives, we only showed the performance on deletions (Table 6).

Deletions were grouped into 2 categories based on their length. In the 50 bp - 1 kb group where the majority of deletions exist, CRISPR-detector reaches 26.87% recall and 80.76% precision and slightly outperforms Smoove (https://brentp.github.io/post/smoove/) on 100x co-barcoded reads. Long Ranger (https://github.com/10XGenomics/longranger) returned similar accuracy to CRISPR-detector on 100x co-barcoded reads, and Sniffles returns better recall than any other callers.[36] For longer deletions (>1k), CRISPR-detector's accuracy score is slightly lower than co-barcoded or long-read callers. Overall, it is quite clear that long read sequencing and co-barcoded read sequencing technologies are more capable of identifying large deletions, especially for decent recall, but the CRISPR-detector pipeline on short reads data still generates useful results. Nevertheless, we recommend conducting long read whole genome sequencing for SV detection, and CRISPR-detector will aim to adapt to these long read types in a future release.

Next, we simulated 150 bp paired-end reads with mimic ~4.7k AAV vector insertion in a 30x WGS dataset to see if CRISPR-detector is able to report the insertion location. Because the inserted sequences are known, we modified the analysis pipeline to overcome the incompetence of insertion detection from the general SV calling pipeline. The AAV vector sequence was added into the human reference genome file as a dummy chromosome. Reads thus can be mapped to both human and vector sequences, and a translocation event can be reported to show the insertion location. We also introduced AAV partial sequences to better mimic partial insertion and simulated five scenarios, including insertion, reverse insertion, non-AAV sequence insertion, partial swap, and insertion with foreign sequences.

In all 5 scenarios CRISPR-detector returned the correct insertion location and always captured at least one of the breakpoints (Fig. 5), giving users the ability to take a closer look at the reported region and confirm the exact insertion format.

Functional annotation of editing induced mutations

We examined annotation result of the processed SPACE and ABE8e-NRCH datasets, and found 1 variant returning annotation from ClinVar database. Annotating mutations from SPACE datasets, a stopgain C•G-to-T•A mutation on *FANCF* site 1 were annotated to be pathogenic by ClinVar database, and associated with Fanconi anemi. CRISPR-detector is convenient for users to predict the pathogenic variants may be caused by genome editing tools both for on target and off target sites.

Our pathogenicity annotations totally depend on the ClinVar database and may be different from the real pathogenicity. Our tool just assembles the function of ClinVar annotation and provides a more convenient way for users who are interested in whether the off-target happens in the ClinVar database which is one of the largest public databases on human genomic variants and interpretations of their relationships to diseases and other conditions maintained by NCBI. We also have added a clear statement in our user agreement that “The pathogenicity annotations are based on

ClinVar database." The annotations are not intended for direct diagnostic use or medical decision-making purpose.

Discussion

Comparison with other analysis tools

In pace with the development of genome engineering, analysis toolkit to assess the quality of genome editing experiment have been developed and made available to community. Early published tools like CRISPR-GA (2014), ampliconDIVider (2015) and Cas-Analyzer (2016) only support single amplicon sequencing data, while AGEseq (2015) and CRISPResso2 (2019) support pooled amplicon sequencing data but do not support unbiased analysis of WGS data as well as SV calling. Besides, CRISPR-detector is more comprehensive with functions including genetic background removal and mutations annotations that other listed toolkits are not capable of, which significantly improved the accuracy and directly showed the possible function and clinical consequences. Several widely-used tools for analyzing genome editing NGS data were compared (Table 7).

Other potentially useful computation modules

We considered whether read trimming and filtering steps should be included in the pipeline and decided they were unnecessary. Sequencing quality has improved greatly recently, and most modern sequencers are capable of generating raw reads with only rare bases below Q30, and thus by default the read filtering step is not included. We suggest that users conduct QC before input into CRISPR-detector if they expect sequencing quality issues. The other step is adapter and primer trimming, and we also decided to exclude these steps from the pipeline because primers are typically not included in the reference target regions in this application. Adapters may exist in raw data, but they do not affect the mapping result to a noticeable extent, and thus removing them may not worth the computational cost.

Benefits of variant clinical annotation

Genomic variants may lead to clinically meaningful changes to protein function and phenotypes, and it is important for off-target detection to evaluate the potential clinical outcome of these introduced variants. The key here is to separate pathogenic variants from others. Instead of requiring users to take variant output and conduct annotation on other platforms, we integrated the functional and clinical annotation within the CRISPR-detector platform so that users can directly access clinical results.

Acknowledgments: We would like to thank Dr. Gao Qianqian for advising on our web server and providing sequencing data and colleagues at EHBIO Gene Technology (Beijing) for the guidance and help in constructing the web server. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding: This work was supported by the Fundamental Research Funds for the Central Public Welfare Research Institutes (ZZ13-YQ-095 and ZZXT201708).

Author contributions: Y.J., T.C. and D.W. designed the study. L.H., D.W., H.C. and D.Z. wrote the bioinformatic pipelines. L.H., D.W. and H.S. designed the web server. L.H., H.C., X.D., C.L., A.L., X.S. and C.Q. acquired the data. L.H., J.H. and H.C. analyzed and interpreted the data. D.W., J.H. and L.H. drafted the manuscript. Y.J. and D.W. supervised the project and edited the manuscript. Y.J. provided the funding. J.H. and D.W. revised the manuscript critically for important intellectual content. All authors read and approved the final manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Source code and scripts to simulate sequencing data are available at GitHub (https://github.com/hlcas/CRISPR-detector). WGS datasets available as as the GEO reference series SRP045395. SPACE and ABE8e-NRCH datasets can be accessed at the NCBI Sequence Read Archive (PRJNA609075; PRJNA725249).

References

1. Jinek M, Chylinski K, Fonfara I, et al. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* 2012;337(6096):816-821. DOI:10.1126/science.1225829.

2. Cong L, Ran FA, Cox D, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* 2013;339(6121):819-823. DOI:10.1126/science.1231143.

3. Zetsche B, Gootenberg J S, Abudayyeh O O, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 2015;163(3):759-771. DOI:10.1016/j.cell.2015.09.038.

4. Abudayyeh O O, Gootenberg J S, Konermann S, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 2016;353(6299):aaf5573. DOI:10.1126/science.aaf5573.

5. Liu J J, Orlova N, Oakes B L, et al. CasX enzymes comprise a distinct family of RNA-guided genome editors. *Nature* 2019;566(7743):E8-E10. DOI:10.1038/s41586-019-0908-x.

6. Pausch P, Al-Shayeb B, Bisom-Rapp E, et al. CRISPR-CasΦ from huge phages is a hypercompact genome editor. *Science* 2020;369:333-337. DOI:10.1126/science.abb1400.

7. Komor AC, Kim YB, Packer MS, et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 2016;533(7603):420-424. DOI:10.1038/nature17946.

8. Kim Y B, Komor A C, Levy J M, et al. Increasing the genome-targeting scope and precision of base editing with engineered Cas9-cytidine deaminase fusions. *Nat Biotechnol* 2017;35(4):371-376. DOI:10.1126/science.abb1400.

9. Gaudelli NM, Komor AC, Rees HA, et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 2017;551(7681):464-471. DOI:10.1038/nature24644.

10. Anzalone AV, Randolph PB, Davis JR, et al. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* 2019;576(7785):149-157. DOI:10.1038/s41586-019-1711-4.

11. Lu Y, Xue J, Deng T, et al. Safety and feasibility of CRISPR-edited T cells in patients with refractory non-small-cell lung cancer. *Nat Med* 2020;26:732–740. DOI:10.1038/s41591-020-0840-5.

12. Frangoul H, Altshuler D, Cappellini MD, et al. CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *N Engl J Med* 2021;384(3):252-260. DOI:10.1056/NEJMoa2031054.

13. Gillmore JD, Gane E, Taubel J, et al. CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *N Engl J Med* 2021;385(6):493-502. DOI:10.1056/NEJMoa2107454.

14. Güell M, Yang L, Church G M. Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). *Bioinformatics* 2014;30(20): 2968-2970: DOI:10.1093/bioinformatics/btu427.

15. Xue L J, Tsai C J. AGEseq: analysis of genome editing by sequencing. *Mol Plant* 2015;8(9): 1428-1430. DOI:10.1016/j.molp.2015.06.001.

16. Varshney GK, Pei W, LaFave MC, et al. High-Throughput Gene Targeting and Phenotyping in Zebrafish Using CRISPR/Cas9. *Genome Res* 2015;25(7):1030-42. DOI:10.1101/gr.186379.114.

17. Park J, Lim K, Kim J S, et al. Cas-analyzer: an online tool for assessing genome editing results using NGS data. *Bioinformatics* 2017;33(2): 286-288. DOI:10.1093/bioinformatics/btw561.

18. Pinello L, Canver MC, Hoban MD, et al. Analyzing CRISPR genome-editing experiments with CRISPResso. *Nat Biotechnol* 2016;34(7):695-697. DOI:10.1038/nbt.3583.

19. Clement K, Rees H, Canver M C, et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat Biotechnol* 2019;37(3):224-226. DOI:10.1038/s41587-019-0032-3.

20. Landrum M J, Lee J M, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Re*s 2018;46(D1): D1062-D1067. DOI:10.1093/nar/gkx1153.

21. Freed D, Pan R, Aldana R. TNscope: Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering. 2018; https://biorxiv.org/content/10.1101/250647v1.

22. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018; 34(18):3094-3100. DOI:10.1093/bioinformatics/bty191.

23. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38(16):e164-e164. DOI:10.1093/nar/gkq603.

24. Gao Q, Ouyang W, Kang B, et al. Selective targeting of the oncogenic KRAS G12S mutant allele by CRISPR/Cas9 induces efficient tumor regression. *Theranostics* 2020;10(11):5137. DOI:10.7150/thno.42325.

25. Huang W, Li L, Jason R. Myers, et al. ART: a next-generation sequencing read simulator, *Bioinformatics* 2012;28(4);593-594. DOI:10.1093/bioinformatics/btr708.

26. García-Nieto, P.E., Morrison, A.J. & Fraser, H.B. The somatic mutation landscape of the human body. *Genome Biol* 2019;20:298. DOI:10.1186/s13059-019-1919-5.

27. Bao X R, Pan Y, Lee C M, et al. Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nat Protoc* 2021;16(1):10-26. DOI:10.1038/s41596-020-00431-y.

28. Ewing, A., Houlahan, K., Hu, Y. et al. Combining tumor genome simulation with crowdsourcing to benchmark somatic single-nucleotide-variant detection. *Nat Methods* 2015;12:623-630. DOI:10.1038/nmeth.3407.

29. Wu, Y., Zhou, H., Fan, X. et al. Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells. *Cell Res* 2015;25:67-79. DOI:10.1038/cr.2014.160.

30. Grünewald J, Zhou R, Lareau C A, et al. A dual-deaminase CRISPR base editor enables concurrent adenine and cytosine editing. *Nat Biotechnol* 2020;38(7): 861-864. DOI:10.1038/s41587-020-0535-y.

31. Tsai SQ, Nguyen NT, Malagon-Lopez J, Topkar VV, Aryee MJ, Joung JK. CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat Methods* 2017;14(6):607-614. DOI:10.1038/nmeth.4278.

32. Bae S, Park J, Kim JS. Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics* 2014;30(10):1473-1475. DOI:10.1093/bioinformatics/btu048.

33. Newby, G.A., Yen, J.S., Woodard, K.J. et al. Base editing of haematopoietic stem cells rescues sickle cell disease in mice. *Nature* 2021;595:295-302. DOI:10.1038/s41586-021-03609-w.

34. Zook, J.M., Hansen, N.F., Olson, N.D. et al. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol* 2020;38:1347-1355. DOI:10.1038/s41587-020-0538-8.

35. Guo J, Shi C, Chen X, et al. stLFRsv: A Germline Structural Variant Analysis Pipeline Using Co-barcoded Reads. *Front. Genet* 2021;12:636239. DOI:10.3389/fgene.2021.636239.

36. Sedlazeck F J, Rescheneder P, Smolka M, et al. Accurate detection of complex structural variations using single-molecule sequencing. *Nat Methods* 2018;15(6):461-468. DOI:10.1038/s41592-018-0001-7.

Figures and Tables

amplicon	n_reads|treatment	n_reads|control	substitution%	insertion%	deletion%	indel%	modified%
ON	51326	28370	0.89	3.62	5.21	8.00	8.00
OT1	31944	16035	0.00	0.00	0.00	0.00	0.00
OT2	14963	12116	0.00	0.00	0.00	0.00	0.00
OT3	30303	16208	0.00	0.00	0.00	0.00	0.00
OT4	18801	20543	9.63	0.00	0.00	0.00	9.63
OT5	27888	35844	0.00	0.00	0.00	0.00	0.00
OT7	42631	42635	0.00	0.00	0.00	0.00	0.00
OT8	14040	29633	0.00	0.00	0.00	0.00	0.00
OT9	12976	25703	0.00	0.00	0.00	0.00	0.00
OT10	27942	19998	0.00	0.00	0.00	0.00	0.00
OT11	37896	45899	0.00	0.00	0.00	0.00	0.00
OT12	32091	65750	0.00	0.00	0.00	0.00	0.00
OT13	58203	25622	0.00	0.00	0.00	0.00	0.00
OT14	26453	35578	0.00	0.00	0.00	0.00	0.00

Table 1. Example result for all loci. A downloadable table summarizing the number of processed reads and the frequency of different types of mutations on all the input loci.

Reads Type	CRISPR-detector			CRISPResso2			Cas-Analyzer		AGEseq	
	Mutated reads (%)	Accuracy (%)	Average Running time (s)	Mutated reads (%)	Accuracy (%)	Average Running time (s)	Mutated reads (%)	Accuracy (%)	Mutated reads (%)	Accuracy (%)
unmodified	0.0	100.0	2	8.3	91.7	14	18.3	81.7	14.4	85.6
1 substitution	49.9	99.9	2	54.5	95.5	13	58.8	91.3	57.3	92.7
2 substitutions	50.0	100.0	2	53.6	96.4	14	58.5	91.5	57.1	93.0
3 substitutions	50.0	100.0	2	54.9	95.1	13	59.6	90.5	57.0	93.1
5 bp insertion	49.9	99.9	2	54.1	95.9	14	59.9	90.1	57.5	92.6
10 bp insertion	49.9	99.9	2	54.9	95.1	13	58.8	91.3	56.3	93.7
50 bp insertion	48.9	98.9	2	54.9	95.1	14	56.7	93.3	56.9	93.1
5 bp deletion	49.9	99.9	2	54.9	95.1	14	58.9	91.2	57.5	92.5
10 bp deletion	49.9	99.9	2	55.5	94.5	13	58.9	91.1	57.5	92.5
50 bp deletion	48.9	98.9	2	53.9	96.1	13	59.3	90.8	57.2	92.8

Table 2. Accuracy and processing time of a single amplicon. Compared to CRISPResso2, CRISPR-detector showed 7 times faster processing speed, and the nearly 100% accuracy also outperformed CRISPResso2.

Indel frequencies (%)	CRISPR-detector			CRISPResso2		
	Mutated Reads (%)	Indel (%)	Accuracy (%)	Mutated Reads (%)	Indel (%)	Accuracy (%)
0.1	0.10	0.10	100.00	9.01	0.12	80.00
1	1.00	1.00	100.00	10.16	1.01	99.00
10	10.01	10.01	99.90	18.46	10.03	99.70
100	100.00	100.00	100.00	100.00	100.00	100.00

Table 3. Performance on 0.1%-100% indel frequencies. CRISPR-detector outperformed CRISPResso2 at low frequency indels detection accuracy.

Simulated Reads Type		Variant Detected				
Control Reads Type	**Inserted Reads Type**	**POS**	**REF**	**ALT**	**Mutated reads (%)**	**Accuracy (%)**
1 substitution	1 substitution & 1 bp insertion	105	-	C	49.6	99.6
2 substitutions	2 substitutions & 1 bp insertion	105	-	C	47.9	97.9
3 substitutions	3 substitutions & 1 bp insertion	105	-	C	47.9	97.9
5 bp deletion	5 bp deletion & 1 bp insertion	105	-	C	49.8	99.8
10 bp deletion	10 bp deletion & 1 bp insertion	105	-	C	65.8	84.2
50 bp deletion	50 bp deletion & 1 bp insertion	105	-	C	64.0	86.0
5 bp insertion	5 bp insertion & 1 bp insertion	105	-	C	48.5	98.5
10 bp insertion	10 bp insertion & 1 bp insertion	105	-	C	48.2	98.2
50 bp insertion	50 bp insertion & 1 bp insertion				Failed	

Table 4. Identification of insertion at position 105 using co-analysis dataset. For all read types except the 50 bp insertion, CRISPR-detector successfully identified true editing events without false positive from variants existing in control samples.

AF > 5%	**SNV**		**Insertions (1-50 bp)**		**Deletions (1-72 bp)**	
	Sensitivity%	**Specificity%**	**Sensitivity%**	**Specificity%**	**Sensitivity%**	**Specificity%**
100x co-analysis	87.7	82.4	58.3	39.3	56.5	37.0
100x edited only	96.4	41.4	55.6	13.9	71.7	18.3
AF > 15%	**SNV**		**Insertions (1-50 bp)**		**Deletions (1-72 bp)**	
	Sensitivity%	**Specificity%**	**Sensitivity%**	**Specificity%**	**Sensitivity%**	**Specificity%**
30x co-analysis	98.8	99.3	13.3	50.0	23.7	66.7
30x edited only	97.3	70.9	16.7	18.8	23.7	27.3

Table 5: Performance on synthetic WGS datasets. SNVs and indels are tested. Co-analysis showed better accuracy than only analyzing edited sample, especially specificities. Due to a much larger analysis region, accuracy of WGS is lower than that from amplicons.

Reads Type		30x Short Reads	100x Co-barcoded Reads		100x Long Reads
SV Length	**Pipeline**	**TNscope**	**Smoove**	**Long Ranger**	**Sniffles**
50-1000 bp	TP	1268	724	2304	4168
	FP	302	248	1279	5285
	FN	3451	3995	2415	551
	Recall	26.87%	15.34%	48.82%	88.32%
	Precision	80.76%	74.49%	64.30%	44.09%
> 1000 bp	TP	445	463	558	570
	FP	212	162	385	447
	FN	172	154	59	47
	Recall	72.12%	75.04%	90.44%	92.38%
	Precision	67.73%	74.08%	59.17%	56.05%

Table 6. SV accuracy benchmark. In the 50 bp - 1 kb group, CRISPR-detector reaches 26.87% recall and 80.76% precision and slightly outperforms Smoove. Long Ranger returned similar accuracy to CRISPR-detector, and Sniffles returns better recall than any other callers. For longer deletions (>1k), CRISPR-detector's accuracy score is slightly lower than co-barcoded or long-read callers.

	CRISPR-detector	CRISPResso1/2	Cas-Analyzer	AGEseq	CRISPR-GA	ampliconDIVider
Web-based	√	√	√	√	√	
Editing tools	Any	Any	Cas9/Cpf1/ZFN/TALEN/dCas9-Fok1	Any	Any	Cas9
Pooled amplicon sequencing data	√	√		√		
WGS data	Whole genome anaylsis	Regions of interest				
Mutation types	Substitution; Indels; SV	Substitution; Indel	Indel	Substitution; Indel	Indel	Indel
Vector detection	√					
Inter-sample comparison	√	√				
Genetic background removal	√					
Functional annotations	Frameshift; Non-synonymous substitution; Startloss/Stopgain	Frameshift; Splice site				
Pathogenic annotation	√					

Table 7. Comparing functions of tools that assess genome editing results. Several widely-used tools for analyzing genome editing NGS data.

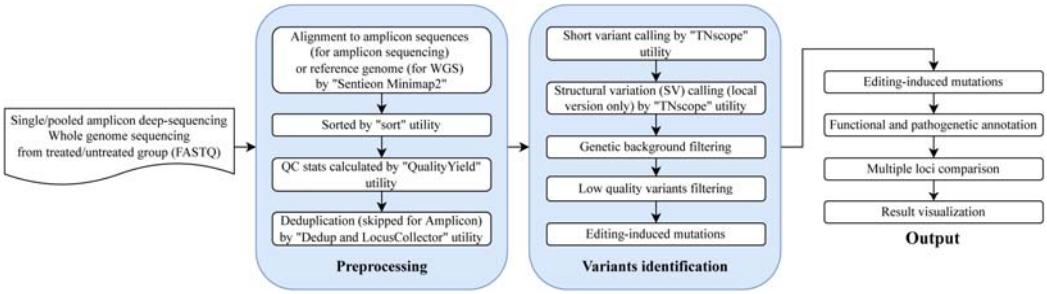
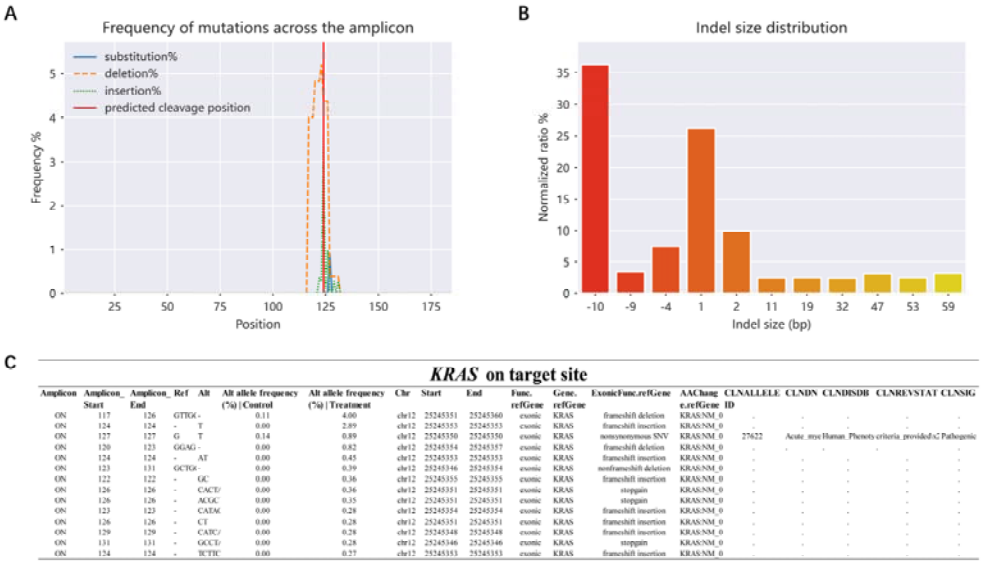


Fig. 1. Flowchart of CRISPR-detector. CRISPR-detector accepts single/pooled amplicon deep-sequencing as well as whole genome sequencing data as input. Control group sample is utilized for removing background variants existing prior to genome editing for treatment group sample. CRISPR-detector outputs visualizations results

and functional annotations of editing-induced mutations for targeted amplicon sequencing data, and VCF format text file for whole genome sequencing data.



Amplicon	Amplicon_Start	Amplicon_End	Ref	Alt	Alt allele frequency (%) | Control	Alt allele frequency (%) | Treatment	Chr	Start	End	Func.refGene	Gene.refGene	ExonicFunc.refGene	AAChange.refGene	CLNALLELEID	CLNDN	CLNDISDB	CLNREVSTAT	CLNSIG
ON	117	126	GTTG	-	0.11	4.00	chr12	25245351	25245360	exonic	KRAS	frameshift deletion	KRAS:NM_0	-	-	-	-	-	-
ON	124	124	-	T	0.00	2.89	chr12	25245353	25245353	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-
ON	127	127	G	T	0.14	0.89	chr12	25245350	25245350	exonic	KRAS	nonsynonymous SNV	KRAS:NM_0	27622	Acute_myel	Human_Phenoty	criteria_provided	2	Pathogenic
ON	120	123	GGAG	-	0.00	0.82	chr12	25245354	25245357	exonic	KRAS	frameshift deletion	KRAS:NM_0	-	-	-	-	-	-
ON	124	124	-	AT	0.00	0.45	chr12	25245353	25245353	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-
ON	123	131	GCTG	-	0.00	0.39	chr12	25245346	25245354	exonic	KRAS	nonframeshift deletion	KRAS:NM_0	-	-	-	-	-	-
ON	122	122	-	GC	0.00	0.36	chr12	25245355	25245355	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-
ON	126	126	-	CACT	0.00	0.36	chr12	25245351	25245351	exonic	KRAS	stopgain	KRAS:NM_0	-	-	-	-	-	-
ON	126	126	-	ACGC	0.00	0.35	chr12	25245351	25245351	exonic	KRAS	stopgain	KRAS:NM_0	-	-	-	-	-	-
ON	123	123	-	CATAC	0.00	0.28	chr12	25245354	25245354	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-
ON	126	126	-	CT	0.00	0.28	chr12	25245351	25245351	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-
ON	129	129	-	CATC	0.00	0.28	chr12	25245348	25245348	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-
ON	131	131	-	GCCT	0.00	0.28	chr12	25245346	25245346	exonic	KRAS	stopgain	KRAS:NM_0	-	-	-	-	-	-
ON	124	124	-	TCTTC	0.00	0.27	chr12	25245353	25245353	exonic	KRAS	frameshift insertion	KRAS:NM_0	-	-	-	-	-	-

Fig. 2. Example result for single locus. (A) Frequency of identified substitutions and indels across the amplicon; **(B)** Indel size distribution; **(C)** Detailed information on mutations including the annotation on this locus.

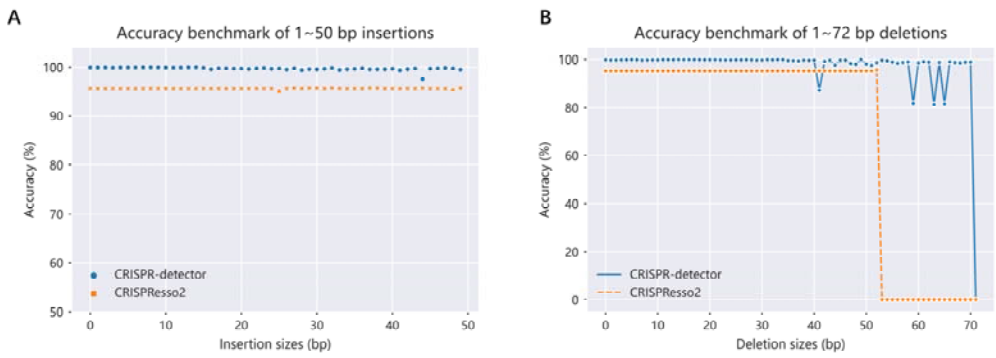


Fig. 3. Accuracy benchmark on 1-50 bp insertions and 1-72 bp deletions. (A) Both CRISPR-detector and CRISPResso2 correctly report 1-50 bp insertions, while CRISPR-detector shows slightly higher accuracy. **(B)** CRISPR-detector is capable of reporting 1-71 bp deletions while CRISPResso2 only reports 1-52 bp deletions.

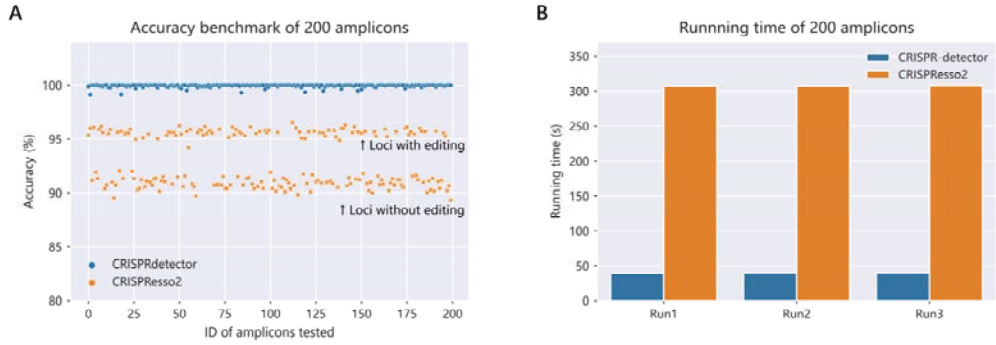
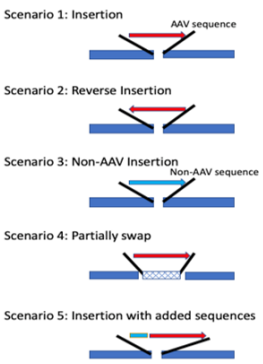


Fig. 4. Accuracy and speed for 200 amplicon testing. (A) CRISPR-detector outperforms CRISPResso2 on accuracy; **(B)** CRISPR-detector runs 8 times faster than CRISPResso2.



Break Point	Inserted Chromosome	Inserted Position	AAV position	Supported Reads Number
bnd_0	chr22	29807035	pAAV-MCS:101	11
bnd_2	chr22	29807038	pAAV-MCS:2102	14

Break Point	Inserted Chromosome	Inserted Position	AAV position	Supported Reads Number
bnd_4	chr22	30910099	pAAV-MCS:3500	10

Break Point	Inserted Chromosome	Inserted Position	AAV position	Supported Reads Number
None				

Break Point	Inserted Chromosome	Inserted Position	AAV position	Supported Reads Number
bnd_8	chr22	45434256	pAAV-MCS:3501	9
bnd_10	chr22	45434753	pAAV-MCS:4200	10

Break Point	Inserted Chromosome	Inserted Position	AAV position	Supported Reads Number
bnd_14	chr22	48273825	pAAV-MCS:1246	22
bnd_12	chr22	48273826	pAAV-MCS:2308	15

Fig. 5. AAV sequence insertion detection. Five simulated scenarios tested, including insertion, reverse insertion, non-AAV sequence insertion, partial swap, and insertion with foreign sequences. CRISPR-detector returned the correct insertion location.