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

Authors

Lei Huang[1†](huanglei@genomics.cn), Dan Wang[2†*](wangdan6@genomics.cn), Haodong Chen[3](haodong.chen@sentieon.com), Jinnan Hu[3](frank.hu@sentieon.com), Xuechen Dai[2](a7068a@126.com), Chuan Liu[2](liuchuan2@genomics.cn), Anduo Li[2](lianduo@genomics.cn), Xuechun Shen[5](shenxuechun@genomics.cn), Chen Qi[1](qichen@genomics.cn), Haixi Sun[2](sunhaixi@genomics.cn), Dengwei Zhang[2](zhangdw@connect.hku.hk), Tong Chen[4*](chent@nrc.ac.cn), Yuan Jiang[2*](yjiang@completegenomics.com)

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 (WGS) data analysis which has the potential to detect off-target mutations at genome-wide level. Here, we introduce 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 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] Genome editing methods have progressively become more flexible and precise, such as varied base editors which enable the direct conversion of one nucleotide to another without producing DSBs in the target sequence and the prime editing which can make targeted Indels and all 12 types of point mutations. [7-10] 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, bioinformatic tools to monitor the on/off-target events need to be more accurate and capable of detecting true editing-induced mutations which refers to mutations introduced during gene editing event, but not germline or previously existing somatic variants.

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 (WGS) 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. While CRISPResso2 offers batch mode to output and visualize variants on the same location from different samples, yet the ability to directly filter germline and other existing variants from control samples is still missing. Third, beyond SNPs and Indels, structural variations (SVs) are currently ignored by analysis. The fourth drawback is that, most of existing pipelines do not estimate 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 samples. 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 mutation 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

Pipeline Overall 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 mandatory and sgRNA sequence(s) is optional to input. After variants called, web hosted CRISPR-detector supports mutation clinical annotation of human sample, and functional annotation for other 9 species. Local deployable pipeline supports user customized database 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 mandatory as input. The BED (Browser Extensible Data) file is optional for calling variants at regions of interest such as on/off target sites. CRISPR-detector also calls SVs in addition to SNPs and Indels, and outputs information in a VCF (Variant Call Format) file. If virus vector (e.g. rAAV) sequence(s) is provided, CRISPR-detector is capable of reporting the location of vector insertion events as well.

Input Webpage Module

On the home page (https://db.cngb.org/crispr-detector/) of the web hosted version, users can upload FASTQ files, 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 either sequencing data from a single amplicon or pooled amplicons.

Data Analysis Module

CRISPR-detector analyzes the input data in two stages: “Preprocessing” stage and “Variants Identification” stage. In the preprocessing stage, reads from edited samples

or control samples are firstly mapped to either amplicon sequences or human reference genome using Sentieon Minimap2, which is an accelerated version of the corresponding opensource Minimap2,[21] faster and returning identical result. Next, all mapped reads are sorted by Sentieon "sort" utility when reads rank are adjusted by their genomic locations. Then a quality control step is conducted on the sorted BAM files by Sentieon "Quality Yield" utility, to calculate quality score distribution of the aligned reads. A warning will be given to user if Q30 value (percentage of mapped bases with quality score larger than 30) is lower than 0.75 by default. The claimed Q30 of MiSeq is around 0.85 so a lower than 0.75 Q30 value may indicate abnormal sequencing quality issues. The last step of this stage is deduplication by Sentieon "Dedup and LocusCollector" utility, which marks reads with same starting and ending position to avoid the duplicated counting. This dedup step is necessary for WGS or probe hybridization based captured sequencing but should be skipped for amplicon sequencing.

The second stage "Variant Identification" transforms prepared BAM files into variants. Firstly, Sentieon TNscope with default parameters is used to generate SNP and Indel candidate variant.[22] 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.[23] Then structural variations are called by TNscope using default parameters if pipeline is run as a local version. Genetic background filtering is the next step, when TNscope generates candidate variant from control samples as well. Each of candidate variants from edited sample and control sample is tested in a fisher's exact test to calculate a p-value and determine if their supporting features reach significant difference. Variants from edited samples are filtered out if the p-value is higher than the user determined cutoff (default at 0.05). The next step is to filter out low quality reads. Specifically, reads marked by TNscope other than "PASS", "triallelic_site", or "alt_allele_in_normal" are filtered out, and

variants with lower than input AF cutoff (default at 0.005) or with less than input supporting read depth (default at 500) for amplicon data are filtered out. The last step is to filter out variants that are not likely to be introduced by editing event. When uploading amplicon datasets, users could provide guide RNA sequence(s) and input "cleavage_offset" parameter to define center of quantification window with respect to the 3' end of the provided sgRNA sequence excluding the PAM (Protospacer Adjacent Motif). And "window_size" parameter to define the size (in bp) of the quantification window extending from the center of quantification window defined by "cleavage_offset" parameter. Default value of "window_size" parameter is 0, meaning analysis the whole amplicons. If the quantification window is defined, then any variants that are not overlap with the window are filtered out.

Functional and clinically important variants in the treatment group data are annotated using the published ANNOVAR (version 2019 Oct24) tool and ClinVar (version 20210501) databases.[24] 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 "Example" button on the webpage, example results for real dataset are shown. 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 frequency of the identified SNPs and Indels in each position (Fig. 2A) and Indel size distribution (Fig. 2B). Co-analysis is conducted if control sample are available. Furthermore, detailed information on mutations including 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 has also been developed and provided to the community (available at https://github.com/hlcas/CRISPR-detector).

There are three differences between the web version and this local 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, as the pipeline only outputs VCF format files.

In this local version, 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

Performance Benchmarks using Simulated Data

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 reads representing actual sequencing data from on/off-target site amplicons and compared the reported variant frequencies with three other popular pipelines including “Cas-Analyzer”, “AGEseq”, and “CRISPResso2”. Specifically, 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 using MiSeq profile provided by ART simulation tool.[25] Average running time is calculated from 3 repeated runs. The speed of Cas-Analyzer and AGEseq are not shown as these two were only tested on their web version but not our local standard benchmark hardware environment.

Compared to CRISPResso2, CRISPR-detector showed 6 times faster processing speed, and the reported mutation rate was closer to the expected value at 0% for control dataset and 50% for all other datasets (Table 2). It should be noted that TNscope, as the variant calling module of CRISPR-detector, utilized haplotype assembling method to filter out sequencing error, which more easily eliminates sequencing errors compared with simple pileup-based read statistics utilized by CRISPResso2.

Then we performed more detailed variant detection testing, to evaluate CRISPR-detector’s performance on Indels at different frequency and length. CRISPR-detector can only detect editing events with > 0.1% frequency, because it is known that sequencing errors from modern sequencers are at also at 0.1% level. 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 reported the exact expected frequency value, while CRISPResso2 incorrectly reported sequencing errors into mutated reads frequency (Table 3).

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. Frequencies reported by CRISPR-detector across 1-50 bp insertions, and 1-71 bp deletions almost aligned with expected value. As comparison, CRISPResso2's reported value was higher than expected, and failed to detect deletions larger than 50 bp (Fig. 3).

One unique feature of CRISPR-detector is co-analysis with treatment and control samples 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 the simulated co-analysis testing dataset we generated, different type of variants and sequencing errors were inserted in both treatment and control reads to benchmark the pipeline's ability to detect a "True positive" mutation (1 bp insertion at position 105 bp) . 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). This failed result shows a limitation of CRISPR-detector in correctly mapping reads around multiple long insertions.

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 poses a requirement for analysis tools to consider speed and scalability during development. To deal with this need, CRISPResso2 provides two dedicated pipelines, CRISPRessoPooled and CRISPRessoWGS, for pooled amplicons and WGS. In comparison, CRISPR-detector utilizes high-performance modules and can handle small and large datasets from one single pipeline, to provide uncompromised accuracy and scalability.

For scalability 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 all 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 for larger datasets especially WGS, alignment and variant calling are two critical steps affecting pipeline speed, and CRISPR-detector removes these speed bottlenecks.

Whole Genome Sequencing Performance

The ability to process WGS data 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. In our test, 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 region size to analysis (3 GB vs < 1 KB), and sequencing depth (< 100x vs > 10,000x). The larger covered size of WGS 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, due to the sequencing depth limitation, only approximately 990 variants with above 1% AFs are inserted into 100x WGS dataset, and approximately 967 variants with above 3.3% AFs are inserted into 30x WGS. To ensure at least 3 reads to support a reported variant, we only reported of variants with AFs above 5% for 100x WGS, and with AFs above 15% for 30x WGS. 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, especially specificities (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 is significantly lower than that from amplicons, this is partially because 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 value and may be filtered out by AF cutoff.

To further validate WGS performance using real edited dataset, a pair of 30x whole genome sequencing samples (“Mutant” and “Rescued”) from a previous study was used.[29] “Mutant” refers to control cells having a pre-existing homozygous 1 bp disease causing deletion in the coding sequence of the gene *Crygc*, “Rescued” refers to the edited clone rescued by CRISPR-Cas9 inducing 1 bp insertions 6-7 bp aside of 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, 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 aside of the deletion (chr1 65118232; chr1 65118233).

Assessing Real World Clinical Data

rhAMP-seq Dataset

In addition to above simulated datasets, we also utilized 4 published clinical datasets to evaluate accuracy of CRISPR-detector, comparing with either original result published with datasets, or CRISPResso2 as alternative. The 4 datasets included are 1) rhAMP-seq from IDT; 2) ABE8e-NRCH Dataset;[30] 3) SPACE Dataset;[31] 4) GUIDE-seq Dataset.[32-33]

The first 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),[34] CRISPR-detector reached 100% accuracy with no false positive or false negative report (Supplementary Table 1).

ABE8e-NRCH Dataset

CRISPR-detector is also suitable for assessing base editing outcomes. We analyzed 480 off-target sites from the third primer pool identified by CIRCLE-seq and Cas-OFFinder,[35] and amplified from genomic DNA from CD34+ SCD cells edited by ABE8e-NRCH or unedited control SCD donor cells using rhAMP-Seq (IDT).

Following the original study's methods, we calculated the statistical significance of ABE8e-NRCH mRNA off-target editing to control samples, by applying a Chi-square test. The 2x2 contingency table was constructed based on the number of edited reads and unedited reads, in treated and control groups. FDR was calculated using the

Benjamini/Hochberg method. We applied the same cutoff thresholds as the original study used: 1) FDR < 0.05 and 2) Difference in editing frequency between treated and control > 0.5%.

As a result, among the 479 sites, 14 sites in replicate 1 sample, and 29 sites in replicate 2 sample were called containing A•T-to-G•C mutations of significance. In replicate 1 sample, CRISPR-detector and CRISPResso2 agreed on all 479 sites. And in replicate 2 sample, the two pipelines showed discordances on only 3 sites (Fig. 5, Table 6).

Specifically, site "91469BDFFDAD44DZ0Z (Off-target site 54)" was called positive only by CRISPResso2. The difference in editing frequency between treated and control was 0.8%. There are two sites that are called positive only by CRISPR-detector. The difference in editing frequency between treated and control of site were just around the 0.5% cutoff value. While the truth is not available to determine which calls are correct, current info indicates these calls are on edges between positive and negative (Supplementary Table 2).

SPACE Dataset

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. CRISPR-detector's performance was compared with CRISPResso2 on 28 on target sites edited by SPACE as treatment and nCas9 as control.

Each potential A•T-to-G•C and C•G-to-T•A position evaluated were categorized into 5 levels based on reported frequency: < 1%, 1%-10%, 10%-30%, 30%-60%, > 60%, and a highest frequency within the 20 bp protospacer was chosen to represent overall

editing level for each on target site. While all 28 nCas9 treated sites showed no targeted substitution, each SPACE treated site returned multiple substitutions within the window.

CRISPR-detector and CRISPResso2 reported concordance editing levels in all 28 SPACE treated sites (Fig. 6), despite 10 of 560 (1.8%) evaluated positions showed discordant levels (Supplementary Table 3). At those discordant positions CRISPR-detector typically reported slightly lower frequencies than result of CRISPResso2, which was the same trend of our observations from simulated dataset, where CRISPResso2 incorrectly reported sequencing errors as editing introduced mutations.

GUIDE-seq Dataset

GUIDE-seq event identification is different from other gene editing events because the inserted dsODN sequences is known, and some of the editing happened below 0.1% AF level. To handle this challenge, we designed a separate module to analysis GUIDE-seq datasets aside of the main pipeline, and benchmarked pipeline using datasets from a published study. Firstly, reads were mapped to the reference amplicons using Sentieon Minimap2. Mapped reads of each amplicon were then extracted and grouped into different read types. Next, all read type sequences were compared against dsODN sequence using BLAST.[36] Number of read types containing dsODN sequence and number of total read types were used to calculate targeted tag integration frequencies.

The performance of CRISPR-detector and CRISPResso2 on three replicate datasets were benchmarked. Out of 156 tested sites (1 on-target and 51 off-target sites in three replicates), CRISPR-detector and CRISPRess2 returned 6 discordant identifications with CRISPResso2 at 6 sites (3.8%), where only CRISPR-detector reported positive result (Fig. 7, Table 7). Four of these discordant sites were reported positive in the

original study, which suggested potential false positive result of CRISPResso2. Regarding the other 2 sites (off-target #3 in replicate 1, and off-target #20 in replicate 2) where only CRISPR-detector returned positive result, by manually checking read types we identified multiple inserted dsODN sequences with no mismatches (Supplementary Table 4), and thus confirmed they were both true positive identifications. CRISPResso2's lower sensitivity was partly due to it was not designed for such ultra-low AF detection. In manuscript of CRISPResso1/2, only events with 1% and above AF were showed, and no surprised to see its accuracy drops when the targeted GUIDE-seq AFs are down to 0.1%.

In conclusion, CRISPR-detector showed identical accuracy as the IDT original analysis pipeline on rhAMP-seq dataset, same level accuracy of CRISPResso2 on ABE8e-NRCH dataset and SPACE dataset, and higher sensitivity on GUIDE-seq dataset.

Structural Variation Identification

Gene editing not only introduces short variants but also structural variations (SVs) that should not be excluded from analysis. A major purpose for conducting whole genome sequencing 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 events occurrence and functional consequences. To benchmark SV 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. [37] 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 understand the limitations of short read-based SV calling compared to other technologies such as co-barcoded reads and long reads, we also compared the result with other pipelines from a recently published study that used the same testing dataset.[38] 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 8).

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.[39] 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, 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 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 could correctly 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 simulate five partial insertion scenarios, including insertion, reverse insertion, non-AAV sequence insertion, partial swap, and insertion with foreign sequences. In all 5 scenarios CRISPR-detector returned correct insertion location and always captured at least one of the breakpoints (Fig. 8), guiding users 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 dataset, and found 1 variant returning annotation from ClinVar database. A stopgain C•G-to-T•A mutation on *FANCF* site 1 was annotated to be pathogenic by ClinVar database, and associated with a rare disease "Fanconi anemi". CRISPR-detector is convenient to predict the pathogenic variants caused by genome editing tools at on/off target sites.

Our pathogenicity annotations completely depend on the ClinVar database and are not intended for direct diagnostic use or medical decision-making purpose. ClinVar database is one of the largest NCBI maintained public databases for human genomic variants and clinical interpretations. Our tool just integrates function of ClinVar annotation and provides users with a more convenient way to look into clinical interpretation of identified variants. We also added a clear statement in our user agreement saying "The pathogenicity annotations are based on ClinVar database."

Discussion

Comparison with Other Analysis Tools

In pace with the development of genome engineering, analysis toolkit to assess 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 improve the accuracy and directly showed the possible function and clinical consequences. Several widely-used tools for analyzing genome editing NGS data were compared (Table 9).

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 very rare bases below Q30 and thus by default the read filtering step is not included. However, a quality check step is included in pipeline, to warn users if less than 75% aligned bases returned Q30 or lower. We also suggest users to conduct additional QC before input data 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. analysed 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 and GUIDE-seq datasets available as the GEO reference series SRP045395 and SRP103697. 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. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018; 34(18):3094-3100. DOI:10.1093/bioinformatics/bty191.

22. 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.

23. David B, Takuto S, Kristian C, et al. Calling somatic SNVs and Indels with Mutect2. 2019; https://www.biorxiv.org/content/10.1101/861054v1.

24. 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.

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. 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.

31. 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.

32. Tsai SQ, Nguyen NT, Malagon-Lopez J, et al. 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.

33. Tsai SQ, Zheng Z, Nguyen NT, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol*. 2015;33(2):187-197. DOI:10.1038/nbt.3117.

34. Poplin R, Ruano-Rubio V, DePristo MA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. 2018; https://www.biorxiv.org/content/10.1101/201178v3

35. 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.

36. Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-10. DOI:10.1016/S0022-2836(05)80360-2.

37. 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.

38. 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.

39. 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	Expected value (%)	CRISPR-detector		CRISPResso2		Cas-Analyzer	AGEseq
		Mutated reads (%)	Average running time (s)	Mutated reads (%)	Average running time (s)	Mutated reads (%)	Mutated reads (%)
unmodified	0.0	0.0	2	8.3	14	18.3	14.4
1 substitution	50.0	49.9	2	54.5	13	58.8	57.3
2 substitutions	50.0	50.0	2	53.6	14	58.5	57.1
3 substitutions	50.0	50.0	2	54.9	13	59.6	57.0

5 bp insertion	50.0	49.9	2	54.1	14	59.9	57.5
10 bp insertion	50.0	49.9	2	54.9	13	58.8	56.3
50 bp insertion	50.0	48.9	2	54.9	14	56.7	56.9
5 bp deletion	50.0	49.9	2	54.9	14	58.9	57.5
10 bp deletion	50.0	49.9	2	55.5	13	58.9	57.5
50 bp deletion	50.0	48.9	2	53.9	13	59.3	57.2

Table 2. Accuracy and processing time of a single amplicon. Compared to CRISPResso2, CRISPR-detector showed 7 times faster processing speed, and reported mutated reads % closer to expected value.

Indel frequencies (%) (Expected value)	CRISPR-detector		CRISPResso2	
	Mutated Reads* (%)	Indel** (%)	Mutated Reads (%)	Indel (%)
0.10	0.10	0.10	9.01	0.12
1.00	1.00	1.00	10.16	1.01
10.00	10.01	10.01	18.46	10.03
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, as the reported mutated reads frequencies are more close to the expected value.

*** Mutated Reads (%) is the percentage of mutated reads among all the reads;**

**** Indel (%) is the percentage of reads with Indels among all the reads**

Simulated Reads Type		Variant Detected				
Control Reads Type	Inserted Reads Type	Expected Value (%)	POS	REF	ALT	Mutated reads (%)
1 substitution	1 substitution & 1 bp insertion	50.0	105	-	C	49.6
2 substitutions	2 substitutions & 1 bp insertion	50.0	105	-	C	47.9
3 substitutions	3 substitutions & 1 bp insertion	50.0	105	-	C	47.9
5 bp deletion	5 bp deletion & 1 bp insertion	50.0	105	-	C	49.8
10 bp deletion	10 bp deletion & 1 bp insertion	50.0	105	-	C	65.8
50 bp deletion	50 bp deletion & 1 bp insertion	50.0	105	-	C	64.0
5 bp insertion	5 bp insertion & 1 bp insertion	50.0	105	-	C	48.5
10 bp insertion	10 bp insertion & 1 bp insertion	50.0	105	-	C	48.2
50 bp insertion	50 bp insertion & 1 bp insertion	50.0	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%	SNPs		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%	**SNPs**		**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. SNPs and Indels are tested. Co-analysis showed better accuracy than only analyzing edited sample, especially specificities. Due to a much larger analysis region and lower depth, accuracy of WGS is lower than that from amplicons.

Site	CRISPR-detector					CRISPResso2				
	mRNA: edited reads (%)	Control: edited reads (%)	mRNA & Control difference (%)	FDR	Editing Observed?	mRNA: edited reads (%)	Control: edited reads (%)	mRNA & Control difference (%)	FDR	Editing Observed?
91469BDFFDAD44DZ0Z	0.25	0.26	-0.01	0.74	No	24.67	23.88	0.79	0.00	Yes
69C4F321559C433Z0Z	0.59	0.00	0.59	0.00	Yes	0.80	0.43	0.37	0.30	No
C3FD98908B0A460Z0Z	0.51	0.00	0.51	0.00	Yes	1.10	0.61	0.48	0.00	No

Table 6. The discrepant sites between CRISPR-detector and CRISPResso2 on ABE8e-NRCH datasets. Frequencies of mRNA and control samples, and FDRs are shown to support the positive/negative call.

Site	Sample	GUIDE-seq	CRISPR-detector	CRISPResso2

EMX1_off-target#3	Replicate #1	0.00	0.006	0.00
EMX1_off-target#5	Replicate #3	0.02	0.563	0.00
EMX1_off-target#19	Replicate #3	0.01	0.004	0.00
EMX1_off-target#20	Replicate #2	0.00	0.003	0.00
EMX1_off-target#20	Replicate #3	0.01	0.007	0.00
EMX1_off-target#48	Replicate #3	0.02	0.006	0.00

Table 7. Frequency of the 6 sites where CRISPR-detector and CRISPResso2 returned discordant result on GUIDE-seq datasets. CRISPR-detector displayed three decimal places where GUIDE-seq and CRISPResso2 only reported two decimal places in original publication. GUIDE-seq result agreed with CRISPR-detector's positive calles on 4 sites.

Reads Type		**30x Short Reads**	**100x Co-barcoded Reads**		**100x Long Reads**
Deletion Length	**Pipeline**	**CRISPR-detector**	**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 8. Structural variants calling accuracy benchmark. In the 50 bp - 1 kb group, CRISPR-detector reaches 26.87% recall and 80.76% precision and slightly outperforms Smoove. For longer deletions (> 1 kb), CRISPR-detector's overall accuracy 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 9. Comparing functions of tools that assess genome editing results. Several widely-used tools for analyzing genome editing NGS data.

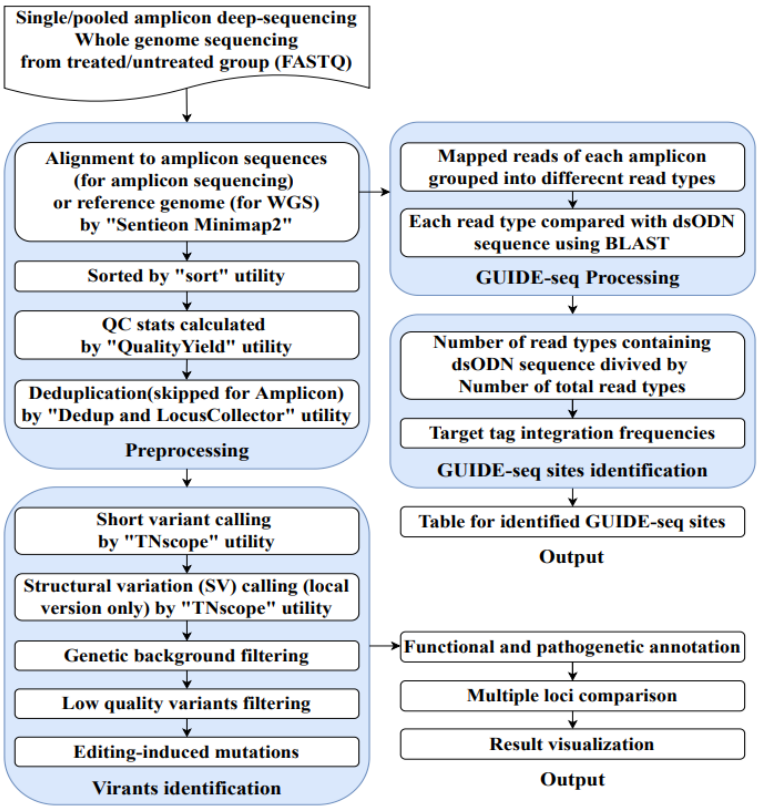


Fig. 1. Flowchart of CRISPR-detector. CRISPR-detector accepts single/pooled amplicon deep-sequencing as well as whole genome sequencing data as input. Control group sample is utilized for removing background variants existing prior to genome editing for treatment group sample. CRISPR-detector outputs visualizations results and functional annotations of editing-induced mutations for targeted amplicon sequencing data, and VCF format text file for whole genome sequencing data. GUIDE-seq analyzed by a dedicated pipeline.

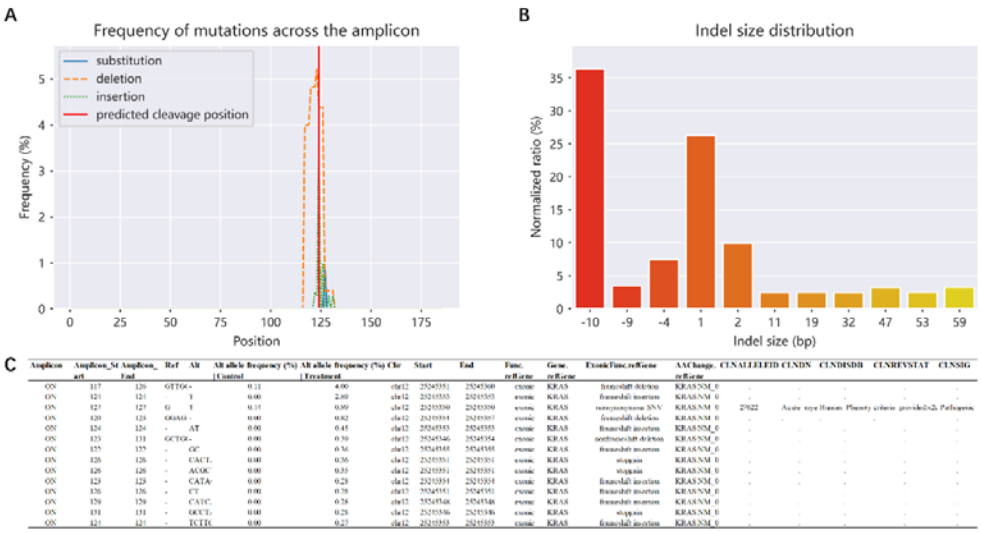


Fig. 2. Example results for single locus. (A) Frequency of identified SNPs and Indels across the amplicon; **(B)** Indel size distribution; **(C)** Detailed information on mutations including the annotation on this locus.

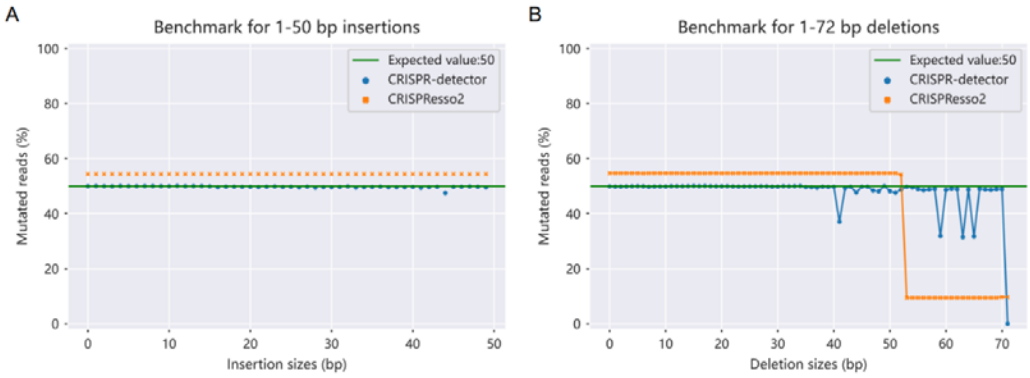


Fig. 3. Benchmark for 1-50 bp insertions and 1-72 bp deletions. (A) Both CRISPR-detector and CRISPResso2 correctly report 1-50 bp insertions, while CRISPResso2's reported value was higher than expected value. **(B)** CRISPR-detector is capable of reporting 1-71 bp deletions while CRISPResso2 only reports 1-52 bp deletions.

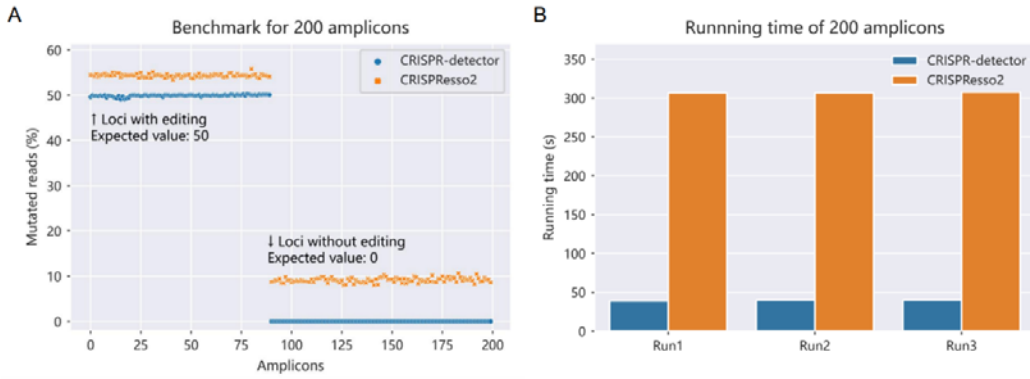


Fig. 4. Benchmark for 200 amplicons testing. (A) CRISPR-detector outperforms CRISPResso2 on all 200 amplicons tested; **(B)** CRISPR-detector runs 8 times faster than CRISPResso2.

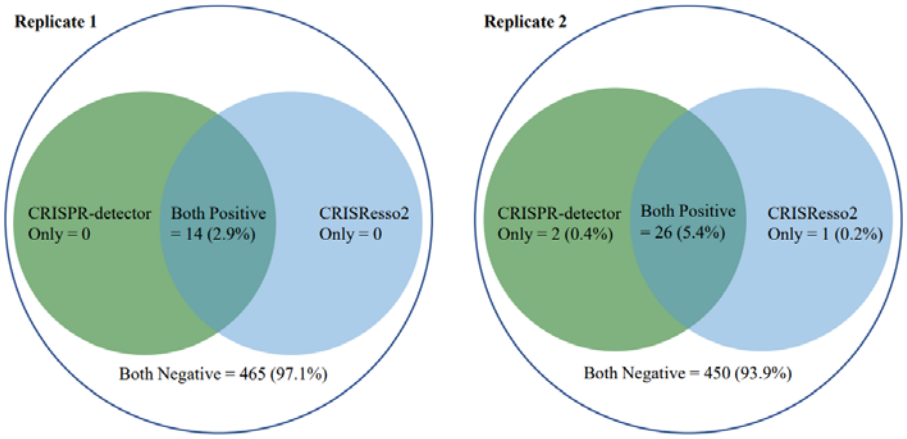


Fig. 5. Venn diagram of CRISPR-detector and CRISPResso2 calling on ABE8e-NRCH datasets. In replicate 1 sample, CRISPR-detector and CRISPResso2 agreed on all 479 sites. And in replicate 2 sample, the two pipelines showed disconcordances on only 3 sites.

SPACE HEK293T on-target sites		
Sites	**CRISPResso2**	**CRISPR-detector**
ABE site 2	52.1	52.1
ABE site 3	66.2	65.5
ABE site 4	68.1	66.5
ABE site 7	47.0	46.9
ABE site 8	35.8	35.5
ABE site 9	28.0	27.0
ABE site 10	44.3	38.0
ABE site 11	55.7	55.7
ABE site 14	61.7	61.3
ABE site 15	61.7	61.6
ABE site 17	59.0	53.6
ABE site 18	58.6	45.0
ABE site 19	53.3	45.5
***EMX* site 1**	16.7	16.3
***EMX* site 2**	27.9	27.9
***FANCF* site 1**	33.3	33.2
HEK site 2	79.2	74.2
HEK site 3	71.6	67.7
HEK site 4	48.6	48.2
***PDCD1* site 1**	38.0	37.7
***PPP1R12C* site 2**	55.7	54.1
***PPP1R12C* site 3**	55.3	55.2
***PPP1R12C* site 4**	42.5	42.4
***PPP1R12C* site 5**	16.0	15.9
***PPP1R12C* site 6**	32.3	31.8
***PPP1R12C* site 7**	44.6	44.6
***RNF2* site 1**	66.5	66.2
***VEGFA* site 4**	26.2	25.9

Frequency	**A•T-to-G•C**	**C•G-to-T•A**
> 60%		
30% - 60%		
10% - 30%		
1 - 10%		
< 1%		

Fig. 6. Maximum editing frequency across the protospacers. CRISPR-detector and CRISPResso2 reported concordance editing levels in all 28 SPACE treated sites.

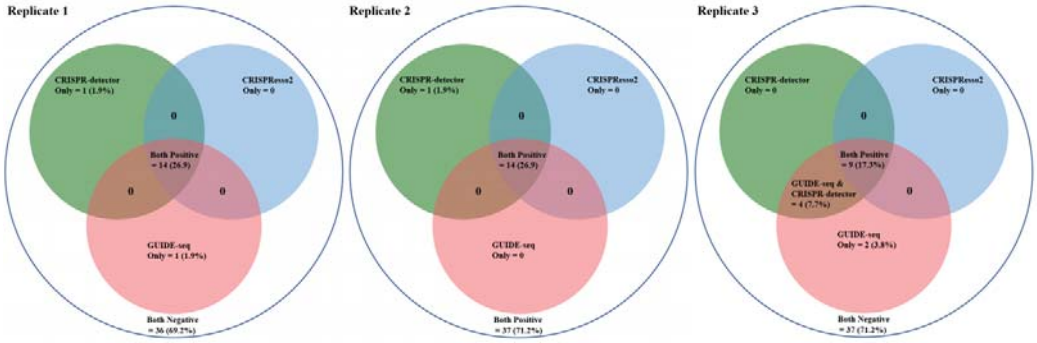
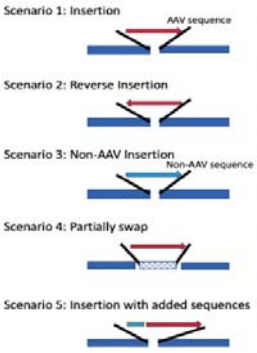


Fig. 7. Venn diagram of CRISPR-detector and CRISPResso2 calling on GUIDE-seq datasets. Out of 156 tested sites, CRISPR-detector and CRISPRess2 returned 6 discordant identifications at 6 sites (3.8%), where only CRISPR-detector reported positive results.



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. 8. 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.