Molecular Therapy
Methods & Clinical Development
Original Article

CRISPAltRations: a validated cloud-based approach for interrogation of double-strand break repair mediated by CRISPR genome editing

Gavin Kurgan,[1] Rolf Turk,[1] Heng Li,[2] Nathan Roberts,[1] Garrett R. Rettig,[1] Ashley M. Jacobi,[1] Lauren Tso,[1] Morgan Sturgeon,[1] Massimo Mertens,[3] Roel Noten,[3] Kurt Florus,[3] Mark A. Behlke,[1] Yu Wang,[1] and Matthew S. McNeill[1]

[1]Integrated DNA Technologies, Coralville, IA 52241, USA; [2]Department of Data Sciences, Dana-Farber Cancer Institute, Boston, MA 02215; [3]Illumina, Inc., San Diego, CA 92122

CRISPR systems enable targeted genome editing in a wide variety of organisms by introducing single- or double-strand DNA breaks, which are repaired using endogenous molecular pathways. Characterization of on- and off-target editing events from CRISPR proteins can be evaluated using targeted genome resequencing. We characterized DNA repair fingerprints that result from non-homologous end joining (NHEJ) after double-stranded breaks (DSBs) were introduced by Cas9 or Cas12a for >500 paired treatment/control experiments. We found that building biological understanding of the repair into a novel analysis tool (CRISPAltRations) improved the quality of the results. We validated our software using simulated, targeted amplicon sequencing data (11 guide RNAs [gRNAs] and 603 on- and off-target locations) and demonstrated that CRISPAltRations outperforms other publicly available software tools in accurately annotating CRISPR-associated indels and homology-directed repair (HDR) events. We enable non-bioinformaticians to use CRISPAltRations by developing a web-accessible, cloud-hosted deployment, which allows rapid batch processing of samples in a graphical user interface (GUI) and complies with HIPAA security standards. By ensuring that our software is thoroughly tested, version controlled, and supported with a user interface (UI), we enable resequencing analysis of CRISPR genome editing experiments to researchers no matter their skill in bioinformatics.

INTRODUCTION

The use of programmable, targeted endonucleases has revolutionized the field of therapeutic genetic engineering.[1] CRISPR enzymes form a ribonucleoprotein (RNP) when hybridized with either a 2-part CRISPR RNA (crRNA) + *trans*-activating CRISPR RNA (tracrRNA) or a single guide RNA (sgRNA), enabling flexible targeting to genomic loci. With either approach, a short, ~20-nucleotide spacer sequence, which is part of the guide RNA (gRNA), targets DNA with complementarity to the gRNA sequence and introduces a double-strand break (DSB), which can be repaired by non-homologous end joining (NHEJ) or homology-directed repair (HDR).[2] The NHEJ pathway ligates broken DNA ends and may modify broken ends to find a biochemically favorable ligation product, generating insertions, deletions, and substitutions.[3] The accurate detection and quantification of these editing events at both on- and off-target locations is paramount to ensuring safety for therapeutic applications of CRISPR.

Producing safety information for genome editing therapeutics first involves nomination and interrogation of a set of putative affected off-target genomic loci utilizing *in vivo*,[4,5] *in vitro*,[6,7] and/or *in silico*[8] methods. After off-target nomination has been performed, alterations in gRNA structure, delivery mechanism, and endonuclease properties can decrease off-target editing effects.[9] Importantly, the use of high-activity and -specificity nucleases[10–13] in combination with delivery mechanisms that limit nuclease exposure time (e.g., RNP delivery) can reduce off-target editing down to levels that are below the standard Illumina next-generation sequencing (NGS) noise rates.[13] During therapeutic optimization, simultaneous quantification of editing at on- and off-target loci can then be used to expediently determine when sufficient efficacy and specificity have been achieved.

A number of methods have been developed to quantify the population of alleles after editing, including heteroduplex cleavage assays,[14–16] capillary electrophoresis,[17] Sanger deconvolution (TIDE/ICE),[18,19] and NGS.[20–23] Limitations have been described for non-NGS-based detection methods, including limited effective editing range,[24] low sensitivity,[25,26] indel size and type limitations,[14,18] low allelic frequency resolution,[26] and reliance on high-quality Sanger traces.[19,26] Thus, NGS has become the gold standard for high-throughput accurate genome editing detection,[27] and it is the only

Received 20 November 2020; accepted 29 March 2021;
https://doi.org/10.1016/j.omtm.2021.03.024.

Correspondence: Matthew S. McNeill, Integrated DNA Technologies, Coralville, IA 52241, USA.
E-mail: mmcneill@idtdna.com

method capable of simultaneously quantifying editing at both on- and off-target locations in highly multiplexed samples.

Specialized software tools have been developed to characterize and quantify allelic diversity after a CRISPR experiment from NGS data, but these tools have not yet been comprehensively validated using a genomic scale ground truth.[20–22] These tools generally align NGS reads to a reference sequence by scoring matches, mismatches, and missing (gap) aligned nucleotides, selecting the highest scoring of the possible alignments, and annotating allelic variants within a certain distance from the predicted enzyme cut site.[20–23] These tools are challenged by the occurrence of repetitive components in the reference or edited sequences, requiring the algorithm to arbitrarily choose between multiple equally scored alignment options (i.e., secondary alignments), which affect the accuracy of the results.[21] Recently developed tools partially overcome this challenge by prioritizing selection of indel events at the predicted cut site,[21,22] but this approach has not yet been comprehensively validated by examining alleles resulting from Cas9 (blunt cut 3 bp from 3′ gRNA end) or Cas12a (two variable nick positions, staggered 4–5 bp from 3′ gRNA end) DSB repair events.[28–31]

In this work, we developed a software tool, CRISPAltRations, for the analysis of NGS data generated from amplicon resequencing of CRISPR edited DNA. We characterized the editing profiles of 516 unique on-target guides for two CRISPR-Cas systems: Cas9 and Cas12a. We demonstrated a novel CRISPR-Cas enzyme-specific aligner and optimized application parameters to characterize indel profiles, which together improve the quality of the results. We validated this software tool by benchmarking it against other popular NGS analysis software tools using synthetic NGS data generated to represent 11 gRNAs with a total of 603 GUIDE-Seq[4] nominated on- and off-target pairs that span a wide variety of genomic sequence features with experimentally modeled indels. Finally, we developed a web-accessible graphical user interface (GUI) to run CRISPAltRations with cloud resources to empower scientists to securely analyze data and visualize results.

RESULTS

Iterative characterization and refinement of Cas9/Cas12a editing profiles

Software tool iteration 1

We hypothesized that implementation of an enzyme-specific alignment program will improve characterization of CRISPR enzyme activity. To begin, we created a pipeline with no preferential indel realignment, prior to characterization of the positional prevalence and type of edits (i.e., population alleles resulting from DSB repair) induced by Alt-R S.p. Cas9 V3 (Cas9) or Alt-R A.s. Cas12a Ultra V3 (Cas12a) in Jurkat cells (Figures S2 and S3). For Cas9 (n = 273; average read depth = 17,518), indel mutations generally intersected the cut site (median 66% of insertions and 80% of deletions). For Cas12a (n = 243; average read depth = 7,416), insertions were generally identified (median insertion frequency > 2%) within a −10 to +2 bp window from the (protospacer adjacent motif) PAM-distal nick site (median 3%–9% per position) (Figure S3). A median of 85% of deletions overlapped with either the PAM-proximal or PAM-distal nick site for Cas12a (Figure S3).

Software tool iteration 2

Upon observing that reads containing indels often had equally scored secondary alignments, we performed a round of iterative optimization using our novel position-specific Needleman-Wunsch (psnw) alignment algorithm (Figure 1). We used psnw to re-align the NGS reads described above to the reference sequence using a modified position-specific gap-open/extension vector (scoring vector), which positively scores alignments at or overlapping the cut site or PAM-distal nick site (for Cas12a), similar to previous work[21] (Figure S4). For Cas9, this increased the prevalence of insertions intersecting the cut site (median 95%), but indels remained identified at other positions (Figure S4). For example, a median of 1.8% of total insertion events were identified −2 bp of the Cas9 cut position. For Cas12a, this increased the prevalence of insertions intersecting the PAM-distal nick site from a median of 7% to 24%. Indels were identified at positions other than cut sites for both Cas9 and Cas12a, and variability of insertion start positions was higher for Cas12a compared to Cas9 (Figures S4 and S5). Cas12a indels were identified between the two nick sites and as far as −5 bp of the PAM-proximal cut to +4 bp of the PAM-distal cut. Deletion position profiles for the two enzymes mostly remained the same after this iteration (Figures S4 and S5).

Software tool iteration 3

With these characterized indel profiles, we further improved the position-based gap-open/extension scoring vector with bonuses that spanned the entire variant detection window (±20 nucleotides around the cut sites) to ensure secondary alignments are selected within the variant detection window. Additionally, we provided larger bonuses to insertion positions enriched in experimental data to move indels closer to the cut/nick site(s) (Figures 2C and 2F; Figures S4 and S5). This increased indels that were identified at the −2 bp position of Cas9 cleavage to a median of 2.6% of events (Figure 2A). For Cas9, the majority of insertion events remained at the cut site (median 95%) and −2 bp position (median 2.6%), with rare events of insertions at the −3 bp (median 0.7%) or +1 bp (median 0.4%) positions (Figures 2A and 2B). For Cas12a, a median of 24% of insertion events occurred at the putative PAM-proximal nick sites. We observed that a median of 52% of insertion events did not occur at either the PAM-distal or PAM-proximal nick site (Figure 2D). Overall, this round of optimization brought indels closer to the cut site(s).

Optimization of the variant detection window limits noise

The variant detection window is a common configurable parameter for CRISPR genome editing analysis software. It limits variant calling to a set distance from a predicted DSB, which reduces the number of collected false-positive events. False-positive events in this context can result from errors introduced in sample preparation or sequencing and are indicated by the presence of indels in

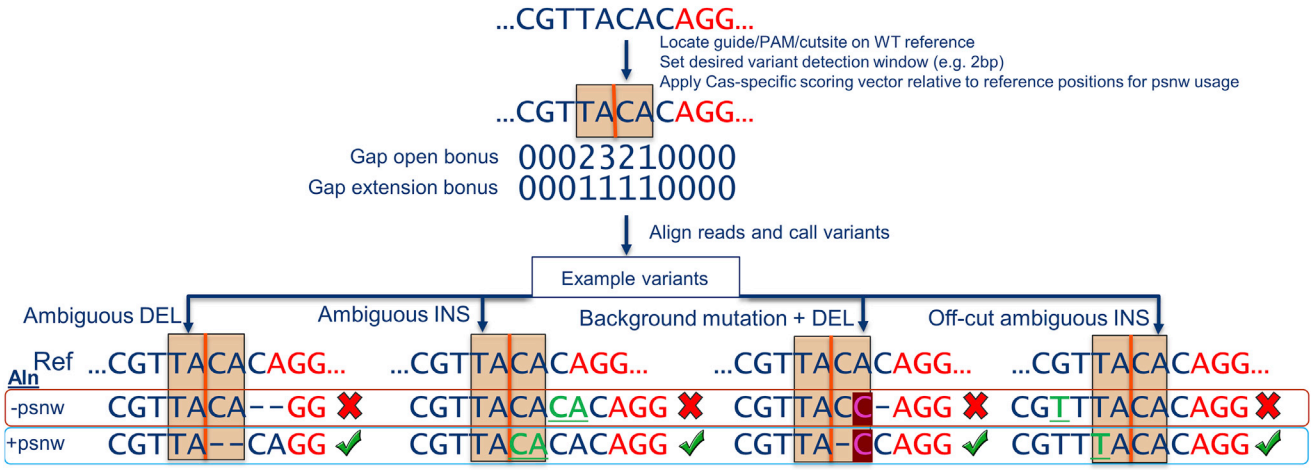

Figure 1. Resolving secondary alignments using psnw alignment with a Cas-specific gap-open/extension scoring vector

To quantify variants, CRISPAltRations locates the guide, PAM (red letters), cut site (orange vertical line), and positions for variant quantification (orange box) in the wild-type reference sequence of interest. Reads are then realigned to the reference sequence using psnw with an applied gap-open and gap-extension bonus scoring vector (indicated as a number per position) to resolve secondary alignments prior to calling variants. This shifts alignments to favor positioning indel variants near predicted cleavage sites, enabling quantification of deletions and insertions in repetitive regions. The scoring vector is applied based on positions in the reference sequence enabling appropriate functionality independent of unexpected mutations in the genetic background. Examples of successful (green check) and failed (red X) quantification of variants are displayed with (+psnw) or without (−psnw) the gap-open/extension matrix applied in software iteration #3 for reference. Ref, reference sequence; Aln, alignment; INS, insertion; DEL, deletion.

unedited control samples. To provide a recommended window for quantifying CRISPR editing events in CRISPAltRations, we compared observed indels in Jurkat cell samples treated with Cas9 or Cas12a against controls collected within a ±20-bp window around the cut site (or PAM-distal cut site for Cas12a). We determined the optimal window size to be the size at which the median difference of calculated indel editing between treatment and control samples was less than 0.1%. Using this rationale, we find that an optimal window can be defined as ±8 bp for Cas9 (Figure 3A) and ±12 bp for Cas12a (Figure S6). However, we found that if the center of the Cas12a window is shifted −3 bp from the PAM distal cut site, the optimal variant window can be decreased to ±9 bp (Figure 3B). Application of this optimal window results in a median decrease in total false-positive indel signal from control samples by 60% as compared to a window size of ±20 for both Cas9 and Cas12a while retaining >98% total indel results from treated cells (Figures 3C and 3D). We set these window sizes as the recommended defaults for variant detection in CRISPAltRations.

Benchmarking of pipeline on- and off-target specificity performance using synthetic datasets

We created a multiplex, synthetic specificity dataset, containing 603 targets, representing performance of 11 gRNAs with indels modeled on observed Cas9 or Cas12a repair events and Illumina MiSeq v3 noise (Figure S7). We created 4,000 synthetic reads per target (50% edited), and we modeled 100 insertion (1–15 bp) and 100 deletion (1–25 bp) events for a total 120,600 unique indel events (Figure S7). We then validated the performance of CRISPAltRations, and we compared performance with Amplican, CRISPResso, and CRISPResso2.

CRISPAltRations calculated the indel percentage within 0.1% of the expected editing level for 99.5% (600/603) of synthetic Cas9 and Cas12a targets (Figure 4). The three erroneous targets were the result of poor paired-end read merging, a critical early step in CRISPAltRations, in regions containing long stretches of homopolymers or repetitive sequence. Observed editing at affected targets deviated from the expected indel percentage by <2% using CRISPAltRations. Mean precision for characterizing indel-containing reads from both Cas9 and Cas12a sites was 0.999 using CRISPAltRations.

We examined the same targets using comparable software tools. The percentage of targets that exceed 2% deviation from the expected Cas9/Cas12a indel percentage for alternative software tools were 72.4%/73.5% (Amplican), 94.5%/99.2% (CRISPResso), 22.4%/100.0% (CRISPResso2), and 1.7%/1.7% (CRISPResso2 with the optimized window parameter derived from Figures 3 and S6) (Figure 4). Mean precision for characterizing indel-containing reads from Cas9/Cas12a sites using these tools was 0.741/0.730 (Amplican), 0.845/0.528 (CRISPResso), 0.960/0.592 (CRISPResso2), and 0.994/0.994 (CRISPResso2 with the optimized window parameter derived from Figure 3 and Figure S6).

Benchmarking of pipeline on-target HDR accuracy

We created a second synthetic Cas9 on-target dataset (a subset of 91 targets from the previous dataset with equivalent performance between tools) to simulate the performance of the two best-performing pipelines, CRISPResso2 and CRISPAltRations, at quantifying HDR rates with a ground truth. We excluded CRISPResso and Amplican from this

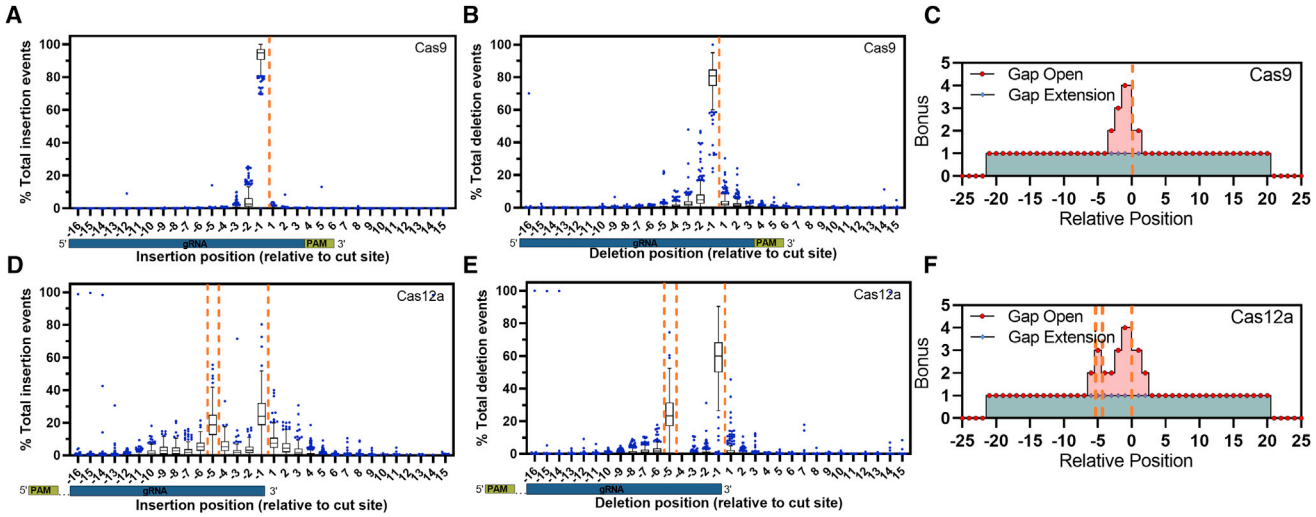

Figure 2. Characterization of Cas9- and Cas12a-specific indel profiles for aligner creation (software iteration #3)

(A–F) Tukey box-and-whisker plot of (A and D) insertion position and (B and E) deletion position relative to the cut/nick site(s) (orange dashed line) derived using (C and F) an integrated scoring vector to apply a position-specific bonus to gap-open and gap-extension events to preferentially select secondary alignments representing the most likely event to occur biologically for Alt-R S.p. Cas9 V3 (n = 273 guides) and Alt-R A.s. Cas12a Ultra V3 (n = 243 guides) editing events delivered via ribonucleoprotein electroporation into Jurkat cells analyzed using software iteration #3. (A,B,D,E) boxes represent the interquartile range (IQR); horizontal line is the median, and whiskers are 1.5x the IQR. Additional points are outside of the IQR.

analysis based on the prior observation that several selected sites were already incorrectly characterized using the synthetic specificity dataset (Figure 4). This dataset contained each target with a heterogeneous set of events including non-edited events (15%), NHEJ indel events (25%), non-HDR donor integration (15%), imperfect HDR events (15%), and a perfect HDR event (30%). HDR donors were designed to either generate deletions (3, 10, 20, 40 bp) or insertions (3, 25, 50, 100 bp) within 8 bp of the cut site (Materials and methods). The CRISPResso2 software tool was not able to complete data processing on 4 target sites (4.3% total sites) when providing an HDR event to the program due to an unhandled exception in the "CRISPResso" analysis mode that was not previously present when using the "CRISPRessoPooled" analysis mode on the same sites in the synthetic specificity dataset (data not shown). These data points were excluded from represented analysis results for CRISPResso2 (Figure 5).

CRISPAltRations correctly characterized the percent perfect HDR repair at 100% of sites with <2% deviation from truth. CRISPResso2 overestimates the percent perfect HDR repair events by >2% at 43% of sites (Figure 5A). CRISPResso2 does not account for any unexpected (single nucleotide polymorphism) SNPs in or near the HDR event in its annotation of percent perfect HDR, which means that any sequencing or polymerase error, naturally occurring mutations, or incomplete HDR events (e.g., 3 out of 4 SNPs successfully incorporated) are not accounted for in its quantification (leading to overestimation). In contrast, synthetic HDR-mediated insertions of 50 and 100 bp cause the percent perfect HDR of CRISPAltRations to deviate 1%–2% below expectation due to the increased probability of SNPs from sequencing errors to occur in these regions (Figure 5B). Both software tools correctly characterize the proportion of CRISPR-edited cells at 100% of targets, demonstrating that these differences are not previously identified issues in annotating editing efficiency (Figure 5C). CRISPAltRations also outperforms CRISPResso2 in its ability to characterize an event as derived from the HDR (imperfect) versus NHEJ pathway at 27 targets (30% of sites) (Figure 5). Overall, CRISPAltRations better characterized HDR editing events in the dataset.

Using CRISPAltRations to describe mutation profiles of Cas9/Cas12a

We characterized enzyme-dependent (Jurkat/Cas9 versus Jurkat/Cas12a) and cell-line-dependent (Jurkat/Cas9 versus HAP1/Cas9) effects on mutation profiles (i.e., indel sizes/types and putative repair pathway) resulting from gene editing using the improved mutation dissemination present in CRISPAltRations.

Across the 273 targets, Cas9 indel profiles were cell-line dependent. Editing efficiency was >50% in >92% of Cas9 targets for HAP1 and Jurkat cell lines (Figure S8). The most prevalent mutations in Jurkat cells edited with Cas9 were insertions (median 81%), and a 2 bp insertion (median 16%) was the most prominent indel event overall (Figure 6). In contrast, deletions were most prevalent in HAP1 cells (median 75%), and a 1 bp insertion (median 18%) was the most prominent indel event overall (Figure 6). Templated insertions (duplication of 1+ nucleotides adjacent to the DSB site) are thought to be a primary mechanism by which insertions are introduced into the genome from repair of DSB events.[32] Insertions in HAP1 cells are predominantly introduced by templated repair events (median

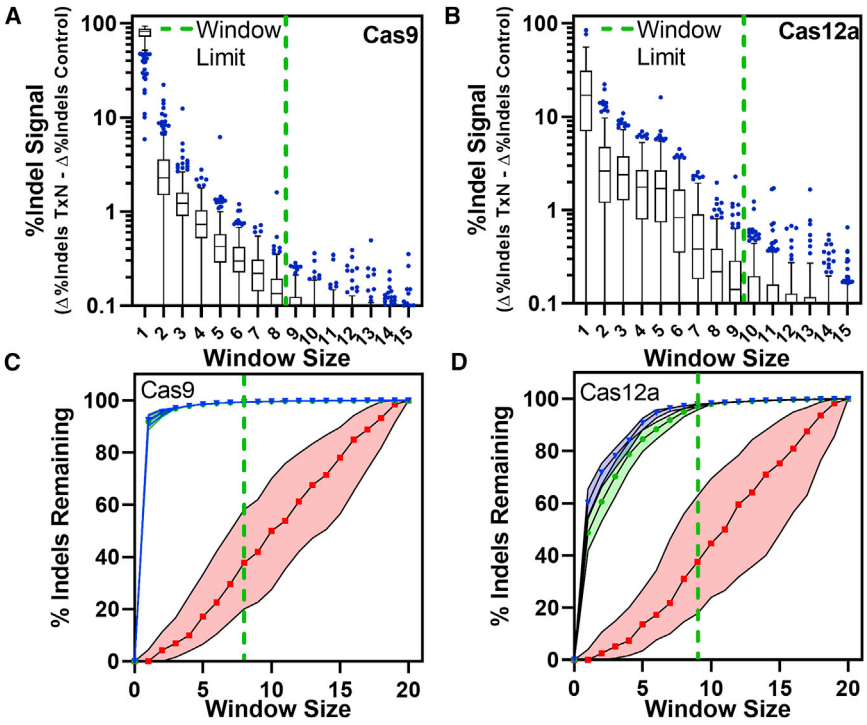

Figure 3. Selection of an optimal variant detection window size

(A and B) An optimal limit for the variant detection window size (green dashed line) for annotating variants was selected for (A) Alt-R S.p. Cas9 V3 (n = 273), and (B) Alt-R A.s. Cas12a Ultra V3 (n = 243; Cas12a window center shifted −3 bp 5′ from PAM-distal nick site) at which median indel signal differences between treatment and control samples was <0.1%. (C and D) The effects of window size on total indels annotated (relative to a window size of 20) was calculated (median ± IQR) for unedited (red), edited samples with software iteration #2 (green), and edited samples with software iteration #3 (blue). (A andB) boxes represent the interquartile range (IQR); horizontal line is the median, and whiskers are 1.5x the IQR. Additional points are outside of the IQR.

74%). In contrast, insertions in Jurkat cells are introduced by templated repair less frequently (median 8%; Figure 6A). A fraction of insertion events (median 16%) were derived from a non-templated insertion of a repeat of guanine and cytosine nucleotides (GC insertions) of >1 bp, an event that did not appear as often in HAP1 cells (median <1%; Figure 6A). Both cell types derive a fraction of the total deletions from microhomology-mediated end joining (MMEJ) events (deletions with >1 bp of exact microhomology; Materials and methods). Deletions mediated by MMEJ were higher in HAP1 (median 43%) compared to Jurkat cells (median 21%; Figure 6A).

Comparison of Cas9 targets to the 243 Cas12a targets demonstrated that indel profiles in Jurkat are enzyme dependent (Figure 6; Figure S9). The most prevalent mutations in Jurkat cells edited with Cas12a were deletions (median 90%), and a 1 bp deletion was the most prevalent event (median 8%; Figure 6). Insertions mediated by Cas12a editing in Jurkat cells had low frequencies of templated insertions (median 12%). GC insertions were also observed to occur (median 18%) with Cas12a editing (Figure 6A). The normalized abundance of GC insertions was not significantly different ($p > 0.05$) in Jurkat cells whether Cas9 or Cas12a was used for editing (Figure 6A). DSB repairs mediated by MMEJ were higher with Cas12a (median 31%) compared to Cas9 (median 21%; Figure 6A). Deletion mutations resulting from Cas12a editing were also 6-fold larger than that of Cas9 in Jurkat cells (Figure 6B).

To better understand if mutation profiles could be predicted *a priori*, we compared the spectrum of indels observed to predictions made by *in silico* repair profile prediction tools, inDelphi[33] and FORECasT,[34] for all previous targets in Jurkat and HAP1 cells. Both tools perform best when compared to DSB repair events in HAP1 cells with Cas9. In general, FORECasT more accurately predicted the most prevalent mutation, while inDelphi more accurately predicted the spectrum of which indels were observed (Figure S10). For HAP1 cells, FORECasT and inDelphi correctly predict the top mutation event 47% and 41% of the time, respectively (Figure S10B). Both FORECasT and inDelphi predict the outcomes of Jurkat cells edited with Cas9 less accurately and only predicted the most prevalent mutation type 14% and 10% of the time, respectively (Figure S10B). We used the symmetric Kullback-Leibler (KL) divergence to calculate similarity of predicted and observed allelic frequencies, as has been done previously.[34] Both tools predict the repair profiles for Jurkat cells treated with Cas12a (median KL = 0.9) better than Cas9 (median KL = 2.0–2.5; Figure S10A). All predictions made at the canonical cut site of these enzymes are better than those made away from the cut site (−3 bp 5′ of cut site) in the same sequence (Figure S10). Predicted frameshift frequencies of both tools correlate with observed results ($R^2 > 0.6$), although FORECasT outperforms inDelphi for all cell line/enzyme combinations (Figure S10).

Recommendations for experimental read depth requirements and tool limits

We analyzed and subsampled CRISPR NGS data from a series of on- and off-target amplicon sequencing panels (2 panels; 91 and 50 targets) with a wide range of editing frequencies to determine the relationship between read depth and precision. There is an inverse relationship between editing efficiency and number of reads needed to accurately quantify editing (Figure 7). The absolute deviation of % indels is dependent on concentration, but the relative change in % indels plateaus for frequencies investigated as read depth surpasses 1,000 paired-end reads (Figure 7B). We find that with target coverage >1,000 paired reads per site, 0.5% indels can be calculated with deviation of approximately ± 0.2% indels (Figure 7B).

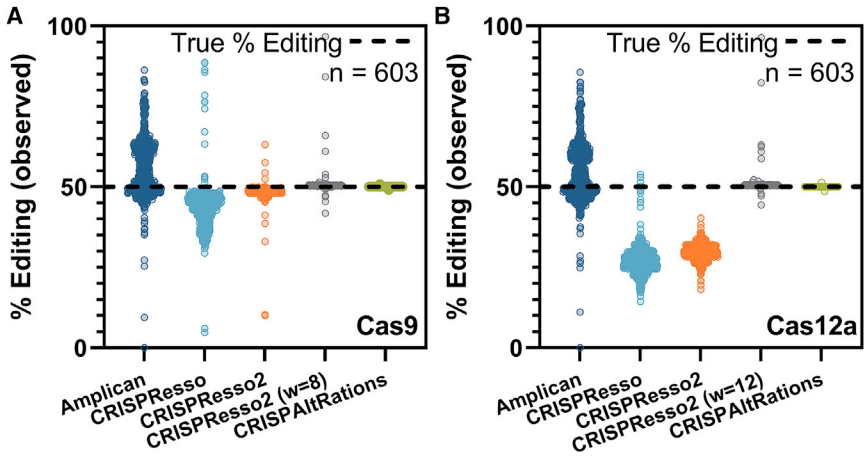

Figure 4. Benchmarking current pipelines supporting multiplex on/off-target analysis

(A and B) Publicly available tools that easily support multiplex analysis were compared to CRISPAltRations using synthetic data (Figure S7; n = 603 sites) generated for (A) Cas9 and (B) Cas12a for the ability to accurately determine % editing at each site (open circles) with a ground truth of 50% editing (black dashed line). w, window size.

We evaluated detection limitations using serially diluted DNA standards with multiplex amplification enrichment performed using rhAmpSeq library preparation (Integrated DNA Technologies [IDT], Coralville, IA, USA) and sequenced using Illumina paired-end sequencing. Without any type of background subtraction, the fraction of indels deviated by ~0.2% from the expected standard concentration as indel editing efficiencies approach <1% (Figure S11). After accounting for the indel error rate in a wild-type template using background subtraction, indel editing correlates with expectation (<0.1% deviation) down to 0.1% indel editing (Figure S11).

To better understand the background indel frequencies at diverse genomic loci, we evaluated the indel percentages in unedited control samples at all 273 unique gRNA sites for Cas9 in both HAP1 and Jurkat cell lines. Background indel mutation rates ranged between 0.0%–1.0%, depending on genomic locus. Indel frequencies in control samples were found to exceed 0.1% indels ~45% of the time. However, 98% of control samples had indel frequencies ranging from 0.0%–0.4% indels (Figure S12). This demonstrates that background indel frequencies can exceed 0.5%, which is above the reported noise rate of Illumina MiSeq instruments[35] (Figure S12).

Integration of CRISPAltRations into a cloud platform with a versatile web UI

Running computational pipelines can be time consuming on personal machines and non-intuitive for those unfamiliar with programming interfaces. Thus, we created a website utilizing cloud-hosted computational resources to run the CRISPAltRations software tool. The website enables either single or batch file upload of demultiplexed sequencing data files (FASTQ) directly into a cloud-based storage system from a drag-and-drop interface or streamed directly from a sequencer, hard drive, or cloud backup location into the website. In addition, batch sample analysis is enabled by providing a configuration file (i.e, comma-separated values), and results are summarized in a single report. The website enables interactive visualization of run metrics, including percent editing/frameshift/repair pathway information, percent SNPs for base editing experiments, and a heatmap pileup of all allelic frequencies aligned to the reference sequence for visualizing the variant population (Figure 8). In addition to gene-editing event summarization, we provide information regarding the performance of the sequencing library and library preparation technique used including percentage reads passing quality control filters, primer-dimers, uniformity (for multiplex amplification panels), and troubleshooting documents to enable end-users to identify and troubleshoot problematic samples or sequencing runs.

We compared runtime performance metrics between CRISPAltRations and publicly available tools processing two synthetic multiplex samples from our on- and off-target benchmarking dataset at various read depths (Figure S13). On common, local hardware, our software runtime is comparable to CRISPResso2 (<40% difference) or outperforms CRISPResso1/Amplican by ~200%–750%. Amplican failed time benchmarking on highly multiplexed samples due to a potential unhandled parallelization error (Figure S13). Using the CRISPAltRations website implementation, runtime is slower (17 min to completion) than the local instance on a run with 14 targets (12,000 reads/target), but it remains ~10-fold faster than the CRISPResso2 website implementation (~4 h to completion) (Figure S13). The CRISPResso2 web solution also failed to complete analysis on highly multiplexed (196 targets) or large datasets (>100 MB file size), representing an additional limitation (Figure S13). In addition, our website implementation enables batch runs of thousands of samples simultaneously, while the current CRISPResso2 website implementation has a maximum of only 4 samples in "batch mode."

DISCUSSION

In this work, we develop a software tool, CRISPAltRations, for the analysis of NGS data generated from CRISPR editing experiments. We incorporated knowledge of characterized indel profiles of Cas9/Cas12a into the algorithm, which enhances CRISPR indel detection accuracy. We furthermore show that optimization of the variant detection window reduces false-positive rates and increases true-positive variant calling in Cas9 and Cas12a editing experiments. We benchmark this pipeline against other publicly available, NGS-compatible software solutions using a large synthetic dataset modeled after real Cas9 and Cas12a editing profiles. We demonstrate that our software tool outperforms other available tools. We further

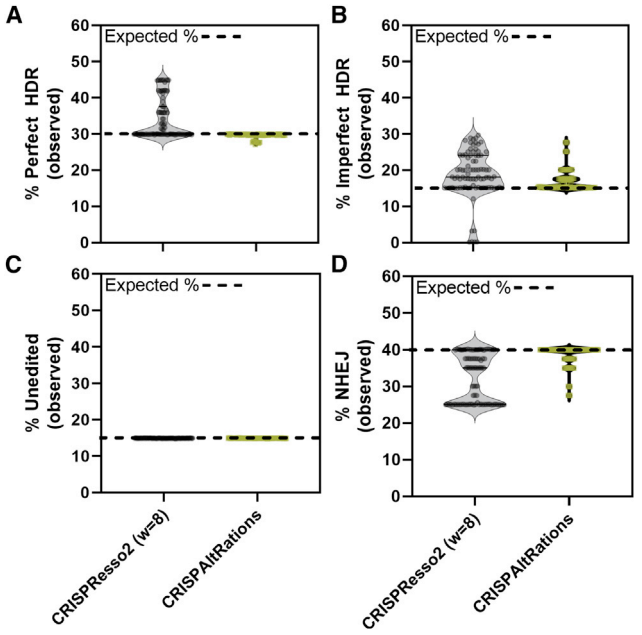

Figure 5. Benchmarking on-target HDR annotation accuracy

(A–D) CRISPResso2 and CRISPAltRations were compared using a synthetic dataset (n = 91 sites) for the ability to accurately determine the percentage of events derived from (A) perfect HDR, (B) imperfect HDR (HDR event with any unintended mutations), (C) wild type, and (D) NHEJ at all edited sites. w, window size.

demonstrate the utility in the ability of CRISPAltRations to characterize repair profile information, by showing that DSB repair profiles are both enzyme and cell-line specific. Lastly, we provide general experimental recommendations grounded in data for performing CRISPR NGS experiments and access to our tool via a distributed cloud-based web solution with an easy-to-use website.

Insertions through the NHEJ pathway are primarily introduced at a DSB site. These insertions can be derived from a number of molecular mechanisms, including misalignment of microhomologies in cleaved DNA products, staggered overhangs from the cleavage event followed by gap filling, and/or template-independent polymerase extension.[36] Our quantification of positional insertion prevalence provides unambiguous evidence that insertion events are observed at non-canonical cut site positions. It was recently found that Cas9 endonucleolytic cleavage of the non-targeted DNA by the RuvC domain can vary in position relative to the HNH domain cut site to generate a staggered DSB.[37] Combinations of variable endonucleolytic cleavage, 5′ to 3′ end processing, and extension by template-independent polymerases may explain the positional occurrence of insertions during repair of DSBs introduced by Cas9. For Cas12a, we observe a diverse spectrum of positions between +3 bp of the PAM distal cut site and −5 bp of the PAM proximal cut site where insertions occur, suggesting a wide range of locations involved in endonucleolytic cleavage and repair. This provides an increased level of resolution on previous work, which has shown that Cas12a cleavage products are diverse and both enzyme and sequence specific.[28–30] This leads us to the conclusion that Cas9 and Cas12a genome editing lead to DSB repair events that cannot be found if only narrow windows (i.e., 1–2 bp) around cut sites are interrogated for variants, a challenge that CRISPAltRations solves with optimized parameter defaults. To the best of our knowledge, this is the first report of the positional prevalence of repair products of Cas9/Cas12a across a wide variety of target sites.

We also demonstrate that indel repair profiles vary with cell and enzyme type. Previous work performing in-depth characterization of Cas9 repair profiles in K562 and HEK293 cell lines showed indel size profiles similar to our observations for HAP1.[33,34,38] Our results support other findings that repair of DSBs caused by Cas12a is prone to larger deletions on average when compared to Cas9.[39] Larger deletions have also been shown to be indicative of MMEJ-related repair events.[40] In agreement with this, we find that putative MMEJ events are more predominant in Cas12a deletions compared to Cas9, within the same cell line, suggesting that the DSB mechanism contributes to repair pathway preference. Additionally, deletions derived from Cas9 editing in HAP1 cells appear to be more prone to MMEJ than Jurkat cells, suggesting MMEJ prevalence is cell-line dependent due to differences in repair pathway expression/activity. Other mutations such as templated insertions have been reported after Cas9 editing, and they are thought to be the main mechanism by which insertions are introduced during DSB repair.[32] Here we provide evidence that templated insertion prevalence after DSB repair is largely dependent on cell type, too. The Jurkat cell line has a relatively low frequency of templated insertions, but Jurkat cells had a higher frequency of >1 bp insertions containing primarily GC motifs. Future work should address if this type of mutation pattern is widespread in clinically relevant cell types and identify if it is a result of a nucleotide bias in a template-independent polymerase. These and other less-characterized repair events are poorly predicted in the current generation of *in silico* indel prediction tools as well, leading to poor performance on Jurkat cells where template-independent mutations are most prevalent. This is likely due to limited repair profile diversity in cell types used for training these models. In the future, these or new tools could be improved by identifying biomarkers predictive of differential repair outcomes to ensure sufficiently diverse modeling data are generated.

Validation and stability of software has traditionally been an overlooked aspect in bioinformatics program development.[41] Two of the publicly available software tools we evaluated generated uncaught exceptions or run failures at the command line and web interface on runs that would be reasonably generated for an individual experiment. Additionally, all evaluated software tools were found to inaccurately annotate variants in our benchmarking datasets. Issues resulting in software tool inaccuracies include, but are not limited to, (1) improper target:read assignment, (2) suboptimal read merging strategies, (3) suboptimal alignment strategies, (4) problematic filters/defaults, and (5) general programming errors. Amplican's performance on this dataset was particularly surprising, and it is primarily caused by the chosen read:target assignment strategy using a string

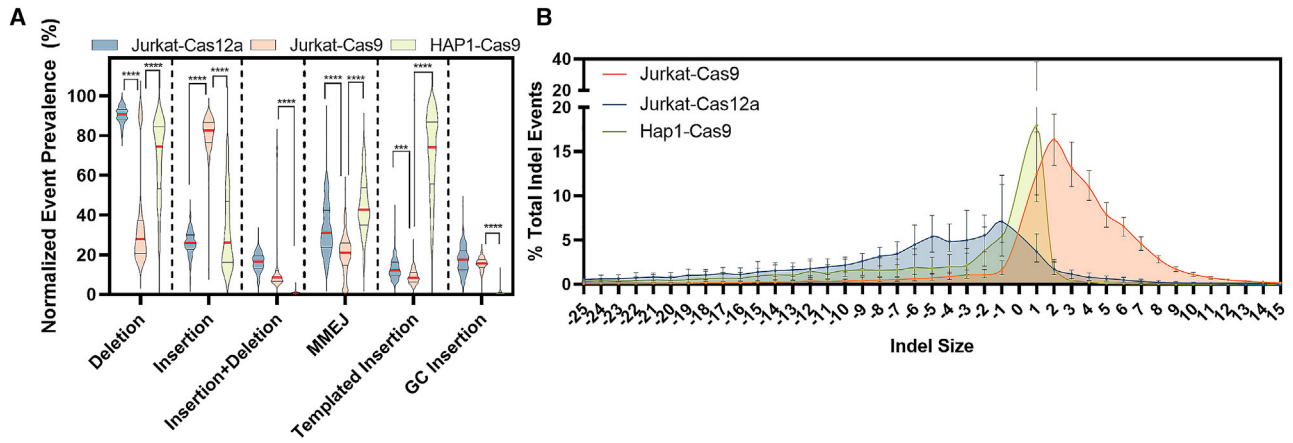

Figure 6. Characterization of cell-line/enzyme-specific repair pathways

(A and B) Normalized occurrence of different characterized indel repair events (A) and median indel size ± interquartile range (IQR) B) for Alt-R S.p. Cas9 V3 or Alt-R A.s. Cas12a Ultra V3 delivered to Jurkat or HAP1 cells. MMEJ, microhomology-mediated end joining. (A) The horizontal red line is the median, and black horizontal lines represent the first and third quartiles. Significance was evaluated using a 2-way ANOVA with a post hoc Tukey multiple comparisons test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$) for indel profile differences between Cas9 (Jurkat), Cas9 (HAP1), and Cas12a (Jurkat) treatments.

match of the primer binding site based on exact read content. Although we enabled an extra 1 bp of ambiguous content (primer_mismatch = 1) in an attempt to account for modeled sequencing errors, enough reads were still lost due to inaccurate annotation. Enabling higher amounts of ambiguity in matches leads to increases in memory requirements, which can cause the program to crash (data not shown). CRISPResso1 and CRISPResso2 without an optimized window parameter are mainly affected by the prevalence of CRISPR-associated indel events occurring outside of the default annotation window. Once the annotation window is extended, suboptimal read merging, alignment, and program annotation of variants seem to be primary causes of misannotation.

Previously developed CRISPR NGS software tools have relied on limited synthetic data or focused on experimentally derived datasets with limited resolution on "truth," leading to large discrepancies in accuracy of different software solutions. More established applications of variant calling software tools have experienced similar shortcomings, such as for somatic variant calling in cancer genomics,[42] and consortiums/researchers have developed a series of best practices, nomenclature standardizations, and gold-standard datasets for benchmarking software tools.[43–45] With this work we provide a more comprehensive simulated CRISPR NGS benchmarking dataset to identify limitations in analysis software tools and provide evidence that similar best practices and standards should be established for the genome editing community. In addition, the sensitivity of many of these CRISPR NGS tools has been stated in previous work ranging from 0.01%–0.1% editing.[20,21] Although we show that annotation of ~0.1% indel editing events is possible under ideal scenarios, this is a misleading sensitivity measurement because it does not account for processes that may introduce variable levels of false-positive editing signals, which may impact reliability in calling variants. This includes variability in methods used for DNA extraction, library preparation, sequencing/technical artifacts, and sequence context. In other fields, such as cancer genomics, detecting variants even below 5% allelic frequency with high precision/recall is considered challenging.[46] Sophisticated methods incorporating unique molecular identifiers (UMIs), paired treatment/control background subtraction, and more have all been applied within the cancer genomics field to enable high-specificity detection of variants at sub 1% allelic frequencies.[47,48] We highlight here for CRISPR NGS analysis that even background editing signal can vary dramatically, further emphasizing the need for advanced methods for confident detection of low editing levels. Future work will need to incorporate error-correction sequencing strategies (e.g., UMIs) and more sophisticated background subtraction methods to increase accuracy of editing annotation.

As genome editing therapies enter clinical trials, it becomes a necessity that software and sequencing methods are thoroughly vetted to prevent incorrect conclusions or exclusion of variant information. This has become clear with accumulating evidence that dsDNA donor (e.g., plasmids, adeno-associated virus vectors) integrations,[49,50] translocations,[51] and large indels/rearrangements[52] all take place from DSB-mediated genome editing. We show that for small dsDNA donors, CRISPAltRations more accurately discriminates and quantifies NHEJ, imperfect HDR, and perfect HDR than existing pipelines using simulated data. However, detection of many larger events requires advances in the use of long read sequencing and targeted hybridization/capture-based sequencing, enrichment protocols, and analysis tools. Additionally, novel genome engineering tools such as base editors[53] and prime editors[54] currently lack optimized/validated computational detection strategies, which should be a focus for improvement. By testing, versioning, and deploying CRISPRAltRations within a cloud-hosted UI with reproducible code production environments and security certifications, we aim to provide a plug-and-play hardware-

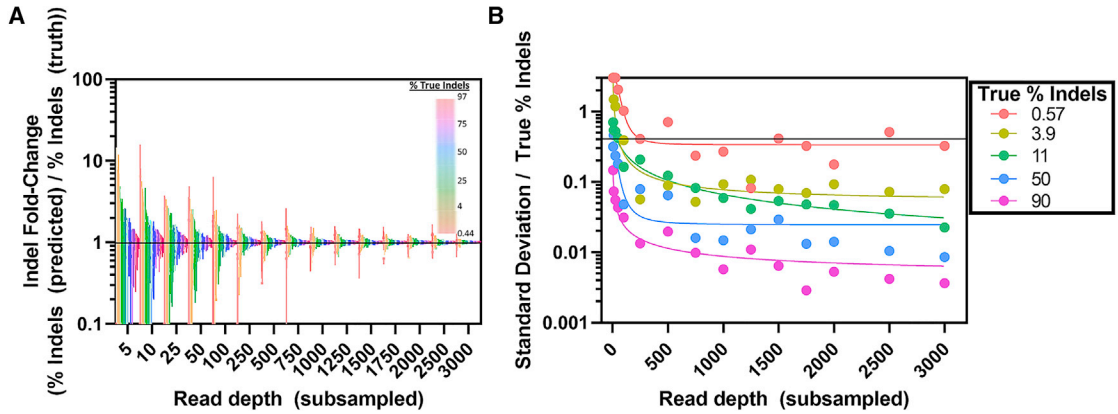

Figure 7. Read depth requirements for variable levels of precision

(A and B) Subsampling of 284 CRISPR editing experiments with varying editing efficiencies (>0.5% editing) to variable read depths in triplicate with comparison of (A) subsampled % indels and (B) standard deviation to unsubsampled (i.e., full depth) results. (A) bars show the full range of each measured value; (B) points represent measured values; colored lines show trendlines drawn through data points of the corresponding color.

independent solution to generate high-quality genome editing specificity data.

MATERIALS AND METHODS

RNP complex formation

Cas9 gRNAs were prepared by mixing equimolar amounts of Alt-R crRNA and Alt-R tracrRNA (Integrated DNA Technologies, Coralville, IA, USA) in IDT Duplex Buffer (30 mM HEPES, pH 7.5, 100 mM potassium acetate; Integrated DNA Technologies), heating to 95°C and slowly cooling to room temperature or using Alt-R sgRNA (Integrated DNA Technologies) hydrated in IDTE (pH 7.5) (10 mM Tris, pH 7.5, 0.1 mM EDTA; Integrated DNA Technologies). Cas12a gRNAs consisted of Alt-R Cas12a crRNAs (Integrated DNA Technologies) hydrated in IDTE (pH 7.5). RNP complexes were assembled by combining the CRISPR-Cas nuclease (Alt-R S.p. Cas9 Nuclease V3 or Alt-R A.s. Cas12a Ultra V3; Integrated DNA Technologies) and the Alt-R gRNA at a 1.2:1 molar ratio of gRNA:protein and incubating at room temperature for 10 min. The target-specific sequences of the gRNAs used in this study are listed in Tables S1 for Cas9 and S2 for Cas12a. The guides chosen were either within the same general genetic context (same amplicon sequencing space; enzyme dependent) or identical between the two cell lines (cell-line dependent) used in this study.

Cell culture

HAP1 cells were purchased from Horizon Discovery (Cambridge, UK). Jurkat E6-1 cells were purchased from ATCC (Manassas, VA, USA). Cells were maintained in RPMI-1640 (Jurkat) or IMDM (HAP1) (ATCC), each supplemented with 10% fetal bovine serum and 1% penicillin-streptomycin (Thermo Fisher Scientific, Carlsbad, CA, USA). Cells were incubated in a 37°C incubator with 5% CO_2. HAP1 cells were used for transfection at 50%–70% confluency. Jurkat cells were used for transfection at 5–8 × 10^5 cells/mL density. After transfection, cells were allowed to grow for 48–72 h in total, after which genomic DNA was isolated using QuickExtract DNA Extraction Solution (Epicenter, Madison, WI, USA). We chose HAP1 and Jurkat since they are derived from human chronic myelogenous leukemia and T lymphocyte cell lines, which are derived from cell types that are similar to those that have been best studied in the context of predicting Cas9 repair profiles.[33,34,55]

Delivery of genome-editing reagents by nucleofection

Electroporation was performed using the Lonza Nucleofector 96-well Shuttle System (Lonza, Basel, Switzerland). For each nucleofection, cells were washed with 1× PBS and resuspended in 20 µL of solution SF or SE (Lonza). HAP1 experiments used ~350,000 cells per nucleofection, and Jurkat experiments used ~500,000 cells per nucleofection. Then, cell suspensions were combined with an RNP complex. For Cas9, the RNP concentration was 4 µM with 4 µM Alt-R Cas9 Electroporation Enhancer. For Cas12a, the RNP concentration was a suboptimal dose of 0.2 µM with 3 µM Alt-R Cas12 Electroporation Enhancer (Integrated DNA Technologies) to provide a more diverse range of editing frequencies. This mixture was transferred into one well of a Nucleocuvette Plate (Lonza) and electroporated using manufacturer's recommended protocols. After nucleofection, 75 µL pre-warmed culture media was added to the cell mixture in the cuvette, mixed by pipetting, and 25 µL was transferred to a 96-well culture plate with 175 µL pre-warmed culture media. Transfection plates were incubated at 37°C and 5% CO_2.

Quantification of editing by NGS

On-target editing efficiency for Cas9/Cas12a nucleofected cells was measured by NGS. Amplicon sequencing libraries were prepared using a previously described rhAmpSeq amplification-based method.[56] Briefly, the first round of PCR was performed using target-specific primers. A second round of PCR was used to incorporate P5 and P7 Illumina adapters to the ends of the amplicons for universal amplification. Libraries were purified using Agencourt AMPure XP system (Beckman Coulter, Brea, CA, USA) and quantified with qPCR before loading onto the Illumina MiSeq platform (Illumina, San Diego, CA,

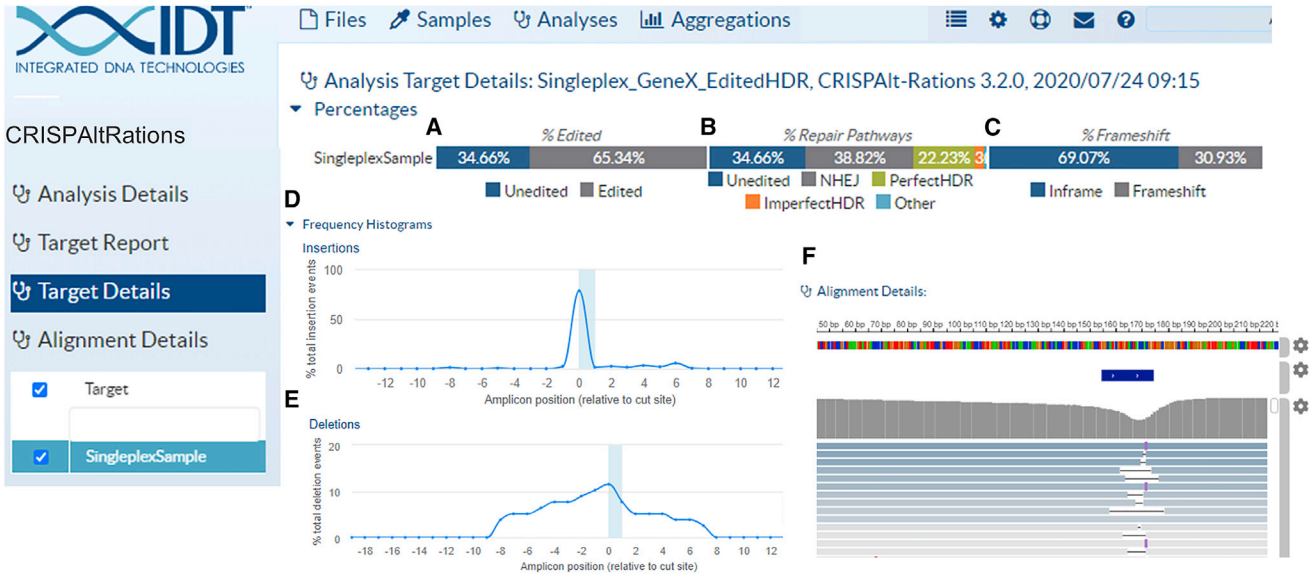

Figure 8. Example of cloud-hosted UI with interactive graphics. As an example, a single on-target HDR experiment is displayed.

(A–C) After completing data processing in the cloud, graphics are automatically created to display high-level metrics like (A) editing frequency, (B) repair pathway utilization, and (C) frameshift frequency. (D–F) Additionally, graphics are generated to display positional occurrence of (D) insertions, (E) deletions, and (F) an IGV visualization of the collapsed variants and their allelic frequencies, and more. Some graphics are artificially condensed to fit in this figure.

USA). Paired-end, 150 bp reads were sequenced using V2 chemistry. Data were demultiplexed using Picard tools v2.9 (https://github.com/broadinstitute/picard).

CRISPAltRations algorithm

We developed the CRISPRAltRations software tool in python, and it plus other software tools are together managed by a snakemake (https://github.com/snakemake/snakemake) or CWL (https://github.com/common-workflow-language/common-workflow-language) workflow manager (Figures S1 and S14).[57,58] The software is hosted with a front-end GUI at https://idtdna.com/pages/tools/rhampseq-crispr-analysis-tool. The UI enables the end-user to specify run information, which is used to partition computational resources hosted in the cloud to perform all data processing using the CRISPAltRations software tool. Results can be visualized and downloaded from the UI. Sequencing data stored in the cloud (AWS, BaseSpace, Google) or on local data stores can be automatically synced with the platform using or uploaded through a drag-and-drop mechanism within the UI. Data are processed in region-specific data centers, duplicated, and protected in a manner that is GDPR (General Data Protection Regulation), HIPAA (Health Insurance Portability and Acountability Act), DSPT (Data Security and Protection Toolkit), PHIPA (Personal Health Information Protection Act), and PIPEDA (Personal Information Protection and Electronic Documents Act) compliant.

The CRISPAltRations software tool workflow starts from demultiplexed FASTQ files as input along with guide and amplicon information in the form of strings or six-column BED-formatted genomic coordinates. The pipeline assumes that the end-user has generated Illumina sequencing data (single or paired end) in FASTQ format and that the reads completely span the cut site in both directions after merging of R1/R2 pairs (if applicable). If genomic coordinates are provided in BED file format, amplicon and guide sequences are extracted from the selected genome and paired using bedtools.[59] Next, low-quality reads and Illumina sequencing adapters are removed using FASTP[60] (–adapter_sequence=AGATCGGAAGAGCACACGTCTGAACTCCAGTCA; –adapter_sequence_r2=AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT; -L; -n=10; -q=15; -u=30). If paired-end data were used, read pairs are merged into a single fragment using FLASH[61] (-O flag used). Putative primer dimers are identified based on a size limit (<60 bp reads), annotated by homology to known amplicon sequences, and removed from downstream analysis. The remaining reads are then mapped to all potential amplicon targets using minimap2[62] (default parameters). The mapped reads are separated into amplicon target-specific BAM files using bamtools[63] to enable parallel processing of all targets. If an HDR donor was supplied, the theoretically perfect HDR event is recreated by iterating through a Needleman-Wunsch alignment with a high gap-open penalty implemented in biopython[64,65] (match=2; mismatch=1; gap open=-30; gap extension=0) at all potential amplicons, choosing the optimal query:target assignment, reconstructing the hypothetical sequence based on the alignment, and adding the hypothetical sequence to the mappable amplicons reference file. By adding this hypothetical perfect HDR sequence to the mappable amplicons, a mapping algorithm can better classify variant alleles as being derived from the NHEJ or HDR pathway based on sequence similarity. Reads are collapsed based on exact sequence identities

and re-mapped to the mappable amplicons reference file using minimap2[62] (-w=1; -k=11; -A=2; -B=4; -O=8; -E = 5; –secondary=no; –no-end-flt;–max-chain-iter=100000) to bin reads appropriately between events derived from HDR versus NHEJ repair pathways. Mapped reads containing indels are re-aligned using a modified Needleman-Wunsch algorithm we call psnw (https://github.com/lh3/psnw) that attributes an alignment score bonus to placement of gap open or extension in specific locations in the alignment. Psnw extends the features of Needleman-Wunsch to include an elevated match/mismatch/gap-open/gap-extension scoring matrix (multiplied by a scalar) and a customizable position specific gap-open/extension vector giving a configurable bonus to alignments that place these features in specific positions. The scoring matrix enables the algorithm to select alignments that have gap open/extensions at desired positions. Our Cas-specific scoring matrices were selected to give maximal gap-open bonuses at the positions with the greatest insertion prevalence in experimental data (Figures S2 and S3). All reads with a mutation that begins within a set distance from the predicted cut site(s) are annotated and summarized in the results, with a number of other visualizations and reports.

Variant annotation

Annotation of variants is performed in a stepwise process with custom python code. First, variants are collapsed based on their annotated nucleotide changes within range of the cut site window. Then, if an HDR donor is supplied, a variant is determined to be derived from the HDR versus NHEJ repair pathway based on the reference amplicon that the read mapped to (wild type versus theoretical HDR event). Next, a variant is annotated as an imperfect HDR event if any SNP or indel is found within the pre-defined window from the cut site or from the location of the first mutation incorporated from the HDR event to the last, whichever is larger. Next, insertions, deletions, and insertion + deletion frequencies are quantified relative to the reference sequence.

Insertions are further characterized by inspecting the sequence of the insertion and surrounding genomic context. If the sequence of an insertion is found to be an exact repeat of DNA adjacent to its insertion, it is described as a templated insertion.[66] If the sequence of an insertion is not found to be a templated insertion, and it is found to be composed of >1 nucleotide and contain only guanine/cytosine nucleotides, it is described as a GC insertion. These events are represented as percentages of the total number of insertions to enable easy comparison between targets.

Deletions are further characterized by inspecting surrounding genomic context of the deletion. If a deletion is >1 nucleotide in length and found to contain >1 nucleotide of exact microhomology from the start of the deletion to the 3′ end of the remaining genomic sequence or from the end of a deletion to the 5′ end of the remaining genomic sequence (accounting for secondary alignments), it is annotated as a MMEJ event. MMEJ events are represented as a percentage of the total number of deletions to enable easy comparison between targets. Any events with both insertions + deletions are excluded from this analysis.

Indel mutations that are not multiples of 3 bp are annotated as frameshifting events, independent of whether they intersect known coding sequences. For identification of the position of mutations, an insertion position is described as the 5′ reference base position adjacent to the insertion. For deletions, the position is considered to be the position closest to the cut site at which a reference base is missing. Additionally, a deletion was considered to intersect the cut site if the base directly 5′ of the cut/nick site was missing in the variant. Since the cut site(s) of A.s. Cas12a Ultra V3 with a 21 bp spacer have not been explicitly defined, we annotated the PAM-proximal and PAM-distal nick sites to be the position between the sites where the most insertion events were observed prior to algorithm optimization. Additionally, for Cas12a two PAM-proximal nick sites that could result in a 5 bp and 4 bp overhang after cleavage were considered in quantification.

Synthetic read generation for on- and off-target editing validation

To create a synthetic benchmarking dataset reminiscent of CRIPSR editing, we used VarSim[67] for generating the defined variants in a paired-end amplicon sequencing read format with an Illumina MiSeq v3 error profile and ART[68] to generate unmodified reads with MiSeq v3 error profiles to enable addition of “wildtype” reads with desired error profiles. We used this to generate a synthetic dataset using sequence space from 11 real rhAmpSeq panels (Table S3) representing GUIDE-seq nominated Cas9 on- and off-target sites (n = 603 on- and off-target sites) with indels modeled based on our real Cas9/Cas12a editing events in Jurkat cells. To do this, median mutation size, position, and frequency of event types across these two datasets were used to create a series of mutation probability vectors that describe the probability of observing different editing events relative to the canonical cut site in a random guide. To create indels, mutation probability vectors were sampled to create 100 unique insertion and deletion events for each guide, each unique event with a read depth of 10 (4,000 reads per target; 50% indels; 2 × 150 reads). It should be noted that the Cas12a sites are not true experimentally determined Cas12a off-targets or binding sites but were merely created at the same genomic positions as the Cas9 dataset to recapitulate the challenge to bin reads between on- and off-target sites with similar genomic context.

Synthetic read generation for on-target HDR quantification validation

To create a synthetic benchmarking dataset representing the ability to perform on-target HDR quantification, we took all of on- and off-targets from the RAG1 Cas9 GUIDE-Seq panel and separated these out as single targets (91 total).[25] The RAG1 panel was chosen because (1) no target processing problems were found when using CRISPResso2, and (2) the genomic sequence around the targets included homopolymers and other events that represent challenging genomic regions to annotate. We then created dsDNA donors *in silico* (as sequence strings) with 40 bp homology arms using the same

synthetic generator previously described. Donors were designed to synthetically introduce a mutation at each of these sites as a deletion (3, 10, 20, 40 bp) or insertion (3, 25, 50, 100 bp) within 8 bp from the expected cut site. We modeled the dataset with simulated dsDNA donors, since this introduces an additional potential complication of the actual donor sequence being directly ligated into the cleavage site, which is an important event to describe.[4,49] We made all sites have a heterogeneous set of events including non-edited events (15%), 10 unique NHEJ indel events (25%), 5 unique non-HDR donor integrations (15%), 5 unique imperfect HDR events (15%), and 1 perfect HDR event (30%). NHEJ indel events were modeled using the mutation probability from Jurkat with Cas9 (see above). Integration of the donor (non-HDR donor integration) was modeled with one perfect integration of the complete dsDNA donor at the cut site and 4 imperfect integrations. Imperfect integration events were modeled with random sizes of truncations of the integration event (not to exceed 40% the full dsDNA donor size) or SNPs within the integrated donor. Imperfect HDR events were similarly modeled with either truncated events (deletion or insertion HDR events) or SNPs (insertion HDR events) within the portion of DNA that was intended to be altered by the HDR donor. Reads were simulated with MiSeq v3 noise profiles (4,000 reads per target, 2 × 250 reads).

Determination of required read depth levels

To provide recommendations for target sequencing read depth requirements, we re-analyzed previously published CRISPR NGS data from a series of rhAmpSeq panels designed for on/off-target sites of guides targeting the RAG1/RAG2 loci with a wide range of editing frequencies, obtainable at the Sequence Read Archive (SRA) under accession number SRA: PRJNA628100.[25] Reads from these samples were subsampled, without replacement, in triplicate with random seeds to a range between 5–3,000 reads pairs per site and quantified using CRISPAltRations with optimized parameters. Indel frequencies and standard deviation among all three read-depth replicates were then compared to the frequency obtained using all reads for the corresponding on- and off-target site to determine deviation from expectation.

DNA standard titration for evaluating rhAmpSeq accuracy

Synthetic dsDNA templates were generated as gBlocks (Integrated DNA Technologies, Coralville, IA, USA) using simulated events at an HPRT1 Cas9 genomic locus (Table S4). Templates were quantified using qPCR before being pooled at equimolar concentrations. These synthetic events consisted of 10 deletions, 10 insertions, and 3 SNPs spiked in to create a known mixture (43.5:43.5:13). Serial dilution was performed with varying levels of wild-type sequence ranging from 0%–100% (Table S4) and subjected to the previously stated library preparation procedure followed by NGS.

Statistical and data analysis

Data collected from experiments were analyzed and statistics generated using GraphPad Prism 8. Editing data for Cas9/Cas12a experiments were only used if a sample had >100 merged reads obtained, and the treated sample had >5% editing. Significance was evaluated using a 2-way ANOVA with a post hoc Tukey multiple comparisons test (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, and ****$p < 0.0001$) for indel profile differences between Cas9 (Jurkat), Cas9 (HAP1), and Cas12a (Jurkat) treatments.

Software versions and parameters utilized

For benchmarking analyses, the following software and versions were used: CRISPResso (1.0.13), CRISPResso2 (2.0.40), and Amplican (1.6.2). When using Amplican, the following non-default parameters were used: average_quality = 15, min_quality = 1, primer_mismatch = 1, min_freq = 0.000001. These parameters were used to prevent Amplican from artificially scoring lower on benchmarking datasets due to stringent default read quality requirements or low allelic frequency filters. It should be noted that the tolerance for primer mismatches could not easily be extended above one due to computational memory requirements. Similarly, for CRISPResso2, the window for detection ("w") was modified to improve results quality in detecting indel events based on our findings (Figure 3; Figure S6). For comparison of *in silico* repair profile prediction tools, the following software versions were used: inDelphi (GitHub commit tag: 9ab67ca53ebb91e49aeb4530ec1e999ee9827ca1) and FORECasT (GitHub commit tag: 019a2f52ba8437528298523c79c224c205146f00). For both models, the K562 model was used for comparing performance.

Availability

The CRISPAltRations pipeline is available via a cloud-hosted web UI at https://idtcrispr.bluebee.com/idtcrispr/ using an account created at https://idtdna.com/pages/tools/rhampseq-crispr-analysis-tool?c=US. Deployed versions of CRISPAltRations can be found in the interface, and desired versions can be configured using the Analysis Mode button prior to run initiation. Access to the CRISPAltRations software is provided through the purchase of rhAmpSeq CRISPR Library Preparation kits from IDT. For more information, please contact crispr@idtdna.com. The psnw aligner is available at https://github.com/lh3/psnw. All Cas-specific gap-open/extension scoring vectors (for psnw) and parameters for publicly available tools are disclosed in Materials and methods for reproducibility. Sequencing data associated with this work have been deposited in the SRA under accession number SRA: PRJNA675792. Example outputs of the tool can be found in Data S1.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtm.2021.03.024.

ACKNOWLEDGMENTS

We would like to thank the Molecular Genetics group at IDT for many helpful discussions. We would also like to thank all of the individuals who participated in testing phases of the CRISPAltRations web platform and their helpful feedback that led to an ultimately better user interface.

AUTHOR CONTRIBUTIONS

G.K. and M.S.M. performed all back-end bioinformatics pipeline creation. H.L. created the psnw alignment algorithm. M.S.M. designed all Cas9 and Cas12a guides. R.T. and G.R.R. planned and performed all cell culture, nucleofection, and library preparation. G.K. performed optimization of different algorithm components and performed data analysis. N.R. performed gBlock template dilution experiments. L.T. and G.K. developed code testing framework. G.K. generated synthetic *in silico* validation datasets. A.M.J., M.S., and G.K. led efforts to test UI framework and generate feedback for improvements. M.M., R.N., and K.F. created front-end web UI. Y.W., M.A.B., M.S.M., and G.R.R. coordinated research group efforts and oversaw development. G.K. wrote an initial draft.

DECLARATION OF INTERESTS

G.K., A.M.J., M.S.M., R.T., G.R.R., N.R., H.L., L.T., M.S., Y.W., and M.A.B. are employees or paid contractors/consultants of Integrated DNA Technologies (IDT), which sells reagents used or similar to those used in this manuscript. M.M., K.F., and R.N. are employees of Illumina Inc., which provides a productized cloud-computing platform for doing NGS analysis.

REFERENCES

1. Porteus, M.H. (2019). A new class of medicines through DNA editing. N. Engl. J. Med. *380*, 947–959.
2. Carroll, D. (2014). Genome engineering with targetable nucleases. Annu. Rev. Biochem. *83*, 409–439.
3. Stinson, B.M., Moreno, A.T., Walter, J.C., and Loparo, J.J. (2020). A Mechanism to Minimize Errors during Non-homologous End Joining. Mol. Cell *77*, 1080–1091.e8.
4. Tsai, S.Q., Zheng, Z., Nguyen, N.T., Liebers, M., Topkar, V.V., Thapar, V., Wyvekens, N., Khayter, C., Iafrate, A.J., Le, L.P., et al. (2015). GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat. Biotechnol. *33*, 187–197.
5. Wienert, B., Wyman, S.K., Richardson, C.D., Yeh, C.D., Akcakaya, P., Porritt, M.J., Morlock, M., Vu, J.T., Kazane, K.R., Watry, H.L., et al. (2019). Unbiased detection of CRISPR off-targets in vivo using DISCOVER-Seq. Science *364*, 286–289.
6. Tsai, S.Q., Nguyen, N.T., Malagon-Lopez, J., Topkar, V.V., Aryee, M.J., and Joung, J.K. (2017). CIRCLE-seq: a highly sensitive in vitro screen for genome-wide CRISPR-Cas9 nuclease off-targets. Nat. Methods *14*, 607–614.
7. Lazzarotto, C.R., Malinin, N.L., Li, Y., Zhang, R., Yang, Y., Lee, G., Cowley, E., He, Y., Lan, X., Jividen, K., et al. (2020). CHANGE-seq reveals genetic and epigenetic effects on CRISPR-Cas9 genome-wide activity. Nat. Biotechnol. *38*, 1317–1327.
8. Cradick, T.J., Qiu, P., Lee, C.M., Fine, E.J., and Bao, G. (2014). COSMID: A web-based tool for identifying and validating CRISPR/Cas off-target sites. Mol. Ther. Nucleic Acids *3*, e214.
9. Vakulskas, C.A., and Behlke, M.A. (2019). Evaluation and reduction of crispr off-target cleavage events. Nucleic Acid Ther. *29*, 167–174.
10. Kim, D., Kim, J., Hur, J.K., Been, K.W., Yoon, S.H., and Kim, J.S. (2016). Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. Nat. Biotechnol. *34*, 863–868.
11. Schmid-Burgk, J.L., Gao, L., Li, D., Gardner, Z., Strecker, J., Lash, B., and Zhang, F. (2020). Highly Parallel Profiling of Cas9 Variant Specificity. Mol. Cell *78*, 794–800.e8.
12. Kleinstiver, B.P., Sousa, A.A., Walton, R.T., Tak, Y.E., Hsu, J.Y., Clement, K., Welch, M.M., Horng, J.E., Malagon-Lopez, J., Scarfò, I., et al. (2019). Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. Nat. Biotechnol. *37*, 276–282.
13. Vakulskas, C.A., Dever, D.P., Rettig, G.R., Turk, R., Jacobi, A.M., Collingwood, M.A., Bode, N.M., McNeill, M.S., Yan, S., Camarena, J., et al. (2018). A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells. Nat. Med. *24*, 1216–1224.
14. Mashal, R.D., Koontz, J., and Sklar, J. (1995). Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. Nat. Genet. *9*, 177–183.
15. Ota, S., Hisano, Y., Muraki, M., Hoshijima, K., Dahlem, T.J., Grunwald, D.J., Okada, Y., and Kawahara, A. (2013). Efficient identification of TALEN-mediated genome modifications using heteroduplex mobility assays. Genes Cells *18*, 450–458.
16. Bhattacharya, D., and Van Meir, E.G. (2019). A simple genotyping method to detect small CRISPR-Cas9 induced indels by agarose gel electrophoresis. Sci. Rep. *9*, 4437.
17. Ramlee, M.K., Yan, T., Cheung, A.M.S., Chuah, C.T.H., and Li, S. (2015). High-throughput genotyping of CRISPR/Cas9-mediated mutants using fluorescent PCR-capillary gel electrophoresis. Sci. Rep. *5*, 15587.
18. Hsiau, T., Maures, T., Waite, K., Yang, J., Kelso, R., Holden, K., and Stoner, R. (2018). Inference of CRISPR Edits from Sanger Trace Data. bioRxiv. https://doi.org/10.1101/251082.
19. Brinkman, E.K., and van Steensel, B. (2019). Rapid Quantitative Evaluation of CRISPR Genome Editing by TIDE and TIDER. Methods Mol. Biol. *1961*, 29–44.
20. Pinello, L., Canver, M.C., Hoban, M.D., Orkin, S.H., Kohn, D.B., Bauer, D.E., and Yuan, G.C. (2016). Analyzing CRISPR genome-editing experiments with CRISPResso. Nat. Biotechnol. *34*, 695–697.
21. Clement, K., Rees, H., Canver, M.C., Gehrke, J.M., Farouni, R., Hsu, J.Y., Cole, M.A., Liu, D.R., Joung, J.K., Bauer, D.E., and Pinello, L. (2019). CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat. Biotechnol. *37*, 224–226.
22. Labun, K., Guo, X., Chavez, A., Church, G., Gagnon, J.A., and Valen, E. (2019). Accurate analysis of genuine CRISPR editing events with ampliCan. Genome Res. *29*, 843–847.
23. Lindsay, H., Burger, A., Biyong, B., Felker, A., Hess, C., Zaugg, J., Chiavacci, E., Anders, C., Jinek, M., Mosimann, C., and Robinson, M.D. (2016). CrispRVariants charts the mutation spectrum of genome engineering experiments. Nat. Biotechnol. *34*, 701–702.
24. Vouillot, L., Thélie, A., and Pollet, N. (2015). Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. G3, *5* (Bethesda), pp. 407–415.
25. Shapiro, J., Iancu, O., Jacobi, A.M., McNeill, M.S., Turk, R., Rettig, G.R., Amit, I., Tovin-Recht, A., Yakhini, Z., Behlke, M.A., and Hendel, A. (2020). Increasing CRISPR Efficiency and Measuring Its Specificity in HSPCs Using a Clinically Relevant System. Mol. Ther. Methods Clin. Dev. *17*, 1097–1107.
26. Sentmanat, M.F., Peters, S.T., Florian, C.P., Connelly, J.P., and Pruett-Miller, S.M. (2018). A Survey of Validation Strategies for CRISPR-Cas9 Editing. Sci. Rep. *8*, 888.
27. Miller, J.C., Patil, D.P., Xia, D.F., Paine, C.B., Fauser, F., Richards, H.W., Shivak, D.A., Bendaña, Y.R., Hinkley, S.J., Scarlott, N.A., et al. (2019). Enhancing gene editing specificity by attenuating DNA cleavage kinetics. Nat. Biotechnol. *37*, 945–952.
28. Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., Van Der Oost, J., Regev, A., et al. (2015). Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell *163*, 759–771.
29. Li, S.Y., Zhao, G.P., and Wang, J. (2016). C-Brick: A New Standard for Assembly of Biological Parts Using Cpf1. ACS Synth. Biol. *5*, 1383–1388.
30. Strohkendl, I., Saifuddin, F.A., Rybarski, J.R., Finkelstein, I.J., and Russell, R. (2018). Kinetic Basis for DNA Target Specificity of CRISPR-Cas12a. Mol. Cell *71*, 816–824.e3.
31. Lei, C., Li, S.Y., Liu, J.K., Zheng, X., Zhao, G.P., and Jin, W. (2017). The CCTL (Cpf1-assisted Cutting and Taq DNA ligase-assisted Ligation) method for efficient editing of large DNA constructs in vitro. Nucleic Acids Res. *45*, e74.
32. Lemos, B.R., Kaplan, A.C., Bae, J.E., Ferrazzoli, A.E., Kuo, J., Anand, R.P., Waterman, D.P., and Haber, J.E. (2018). CRISPR/Cas9 cleavages in budding yeast reveal templated insertions and strand-specific insertion/deletion profiles. Proc. Natl. Acad. Sci. U S A *115*, E2040–E2047.
33. Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., and Sherwood, R.I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. Nature *563*, 646–651.

34. Allen, F., Crepaldi, L., Alsinet, C., Strong, A.J., Kleshchevnikov, V., De Angeli, P., Páleníková, P., Khodak, A., Kiselev, V., Kosicki, M., et al. (2018). Predicting the mutations generated by repair of Cas9-induced double-strand breaks. Nat. Biotechnol. *37*, 64–72, 2019.

35. Glenn, T.C. (2011). Field guide to next-generation DNA sequencers. Mol. Ecol. Resour. *11*, 759–769.

36. Lieber, M.R. (2010). The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. Annu. Rev. Biochem. *79*, 181–211.

37. Stephenson, A.A., Raper, A.T., and Suo, Z. (2018). Bidirectional Degradation of DNA Cleavage Products Catalyzed by CRISPR/Cas9. J. Am. Chem. Soc. *140*, 3743–3750.

38. van Overbeek, M., Capurso, D., Carter, M.M., Thompson, M.S., Frias, E., Russ, C., Reece-Hoyes, J.S., Nye, C., Gradia, S., Vidal, B., et al. (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. Mol. Cell. *63*, 633–646.

39. Gao, Z., Fan, M., Das, A.T., Herrera-Carrillo, E., and Berkhout, B. (2020). Extinction of all infectious HIV in cell culture by the CRISPR-Cas12a system with only a single crRNA. Nucleic Acids Res. *48*, 5527–5539.

40. Owens, D.D.G., Caulder, A., Frontera, V., Harman, J.R., Allan, A.J., Bucakci, A., Greder, L., Codner, G.F., Hublitz, P., McHugh, P.J., et al. (2019). Microhomologies are prevalent at Cas9-induced larger deletions. Nucleic Acids Res. *47*, 7402–7417.

41. Mangul, S., Mosqueiro, T., Abdill, R.J., Duong, D., Mitchell, K., Sarwal, V., Hill, B., Brito, J., Littman, R.J., Statz, B., et al. (2019). Challenges and recommendations to improve the installability and archival stability of omics computational tools. PLoS Biol. *17*, e3000333.

42. Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., and Dry, J.R. (2016). VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. *44*, e108.

43. Li, M.M., Datto, M., Duncavage, E.J., Kulkarni, S., Lindeman, N.I., Roy, S., Tsimberidou, A.M., Vnencak-Jones, C.L., Wolff, D.J., Younes, A., et al. (2017). Standards and Guidelines for the Interpretation and Reporting of Sequence Variants in Cancer. J. Mol. Diagn. *19*, 4–23.

44. Krusche, P., Trigg, L., Boutros, P.C., Mason, C.E., De La Vega, F.M., Moore, B.L., Gonzalez-Porta, M., Eberle, M.A., Tezak, Z., Lababidi, S., et al. (2019). Best practices for benchmarking germline small-variant calls in human genomes. Nat. Biotechnol. *37*, 555–560.

45. Craig, D.W., Nasser, S., Corbett, R., Chan, S.K., Murray, L., Legendre, C., Tembe, W., Adkins, J., Kim, N., Wong, S., et al. (2016). A somatic reference standard for cancer genome sequencing. Sci. Rep. *6*, 24607.

46. Marx, V. (2016). Cancer: hunting rare somatic mutations. Nat. Methods *13*, 295–299.

47. Wang, T.T., Abelson, S., Zou, J., Li, T., Zhao, Z., Dick, J.E., Shlush, L.I., Pugh, T.J., and Bratman, S.V. (2019). High efficiency error suppression for accurate detection of low-frequency variants. Nucleic Acids Res. *47*, e87.

48. Kamps-Hughes, N., McUsic, A., Kurihara, L., Harkins, T.T., Pal, P., Ray, C., and Ionescu-Zanetti, C. (2018). ERASE-Seq: Leveraging replicate measurements to enhance ultralow frequency variant detection in NGS data. PLoS ONE *13*, e0195272.

49. Hanlon, K.S., Kleinstiver, B.P., Garcia, S.P., Zaborowski, M.P., Volak, A., Spirig, S.E., Muller, A., Sousa, A.A., Tsai, S.Q., Bengtsson, N.E., et al. (2019). High levels of AAV vector integration into CRISPR-induced DNA breaks. Nat. Commun. *10*, 4439.

50. Norris, A.L., Lee, S.S., Greenlees, K.J., Tadesse, D.A., Miller, M.F., and Lombardi, H.A. (2020). Template plasmid integration in germline genome-edited cattle. Nat. Biotechnol. *38*, 163–164.

51. Stadtmauer, E.A., Fraietta, J.A., Davis, M.M., Cohen, A.D., Weber, K.L., Lancaster, E., Mangan, P.A., Kulikovskaya, I., Gupta, M., Chen, F., et al. (2020). CRISPR-engineered T cells in patients with refractory cancer. Science *367*, 80.

52. Kosicki, M., Tomberg, K., and Bradley, A. (2018). Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements. Nat. Biotechnol. *36*, 765–771.

53. Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature *533*, 420–424.

54. Anzalone, A.V., Randolph, P.B., Davis, J.R., Sousa, A.A., Koblan, L.W., Levy, J.M., Chen, P.J., Wilson, C., Newby, G.A., Raguram, A., and Liu, D.R. (2019). Search-and-replace genome editing without double-strand breaks or donor DNA. Nature *576*, 149–157.

55. Leenay, R.T., Aghazadeh, A., Hiatt, J., Tse, D., Roth, T.L., Apathy, R., Shifrut, E., Hultquist, J.F., Krogan, N., Wu, Z., et al. (2019). Large dataset enables prediction of repair after CRISPR-Cas9 editing in primary T cells. Nat. Biotechnol. *37*, 1034–1037.

56. Jacobi, A.M., Rettig, G.R., Turk, R., Collingwood, M.A., Zeiner, S.A., Quadros, R.M., Harms, D.W., Bonthuis, P.J., Gregg, C., Ohtsuka, M., et al. (2017). Simplified CRISPR tools for efficient genome editing and streamlined protocols for their delivery into mammalian cells and mouse zygotes. Methods *121–122*, 16–28.

57. Köster, J., and Rahmann, S. (2018). Snakemake-a scalable bioinformatics workflow engine. Bioinformatics *34*, 3600.

58. Amstutz, P., Crusoe, M.R., Tijanić, N., Chapman, B., Chilton, J., Heuer, M., Kartashov, A., Leehr, D., Ménager, H., Nedeljkovich, M., et al. (2016). Common Workflow Language Specifications, v1.0, https://figshare.com/articles/dataset/Common_Workflow_Language_draft_3/3115156.

59. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842.

60. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890.

61. Magoč, T., and Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics *27*, 2957–2963.

62. Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100.

63. Barnett, D.W., Garrison, E.K., Quinlan, A.R., Strömberg, M.P., and Marth, G.T. (2011). Bamtools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics *27*, 1691–1692.

64. Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

65. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and de Hoon, M.J. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

66. Schimmel, J., van Schendel, R., den Dunnen, J.T., and Tijsterman, M. (2019). Templated Insertions: A Smoking Gun for Polymerase Theta-Mediated End Joining. Trends Genet. *35*, 632–644.

67. Mu, J.C., Mohiyuddin, M., Li, J., Bani Asadi, N., Gerstein, M.B., Abyzov, A., Wong, W.H., and Lam, H.Y.K. (2015). VarSim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. Bioinformatics *31*, 1469–1471.

68. Huang, W., Li, L., Myers, J.R., and Marth, G.T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics *28*, 593–594.