Innovation, Sciences et Développement économique Canada
Office de la Propriété Intellectuelle du Canada

Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

CA 3144106 A1 2021/01/07

(21) 3 144 106

(12) DEMANDE DE BREVET CANADIEN
CANADIAN PATENT APPLICATION

(13) A1

(86) Date de dépôt PCT/PCT Filing Date: 2020/07/02
(87) Date publication PCT/PCT Publication Date: 2021/01/07
(85) Entrée phase nationale/National Entry: 2021/12/16
(86) N° demande PCT/PCT Application No.: US 2020/040621
(87) N° publication PCT/PCT Publication No.: 2021/003343
(30) Priorités/Priorities: 2019/07/03 (US62/870,426); 2019/07/03 (US62/870,471); 2019/12/23 (US62/952,603); 2019/12/23 (US62/952,598)

(51) Cl.Int./Int.Cl. *C12N 9/22* (2006.01), *C12N 15/10* (2006.01), *G16B 20/20* (2019.01), *G16B 30/10* (2019.01)

(71) Demandeur/Applicant:
INTEGRATED DNA TECHNOLOGIES, INC., US

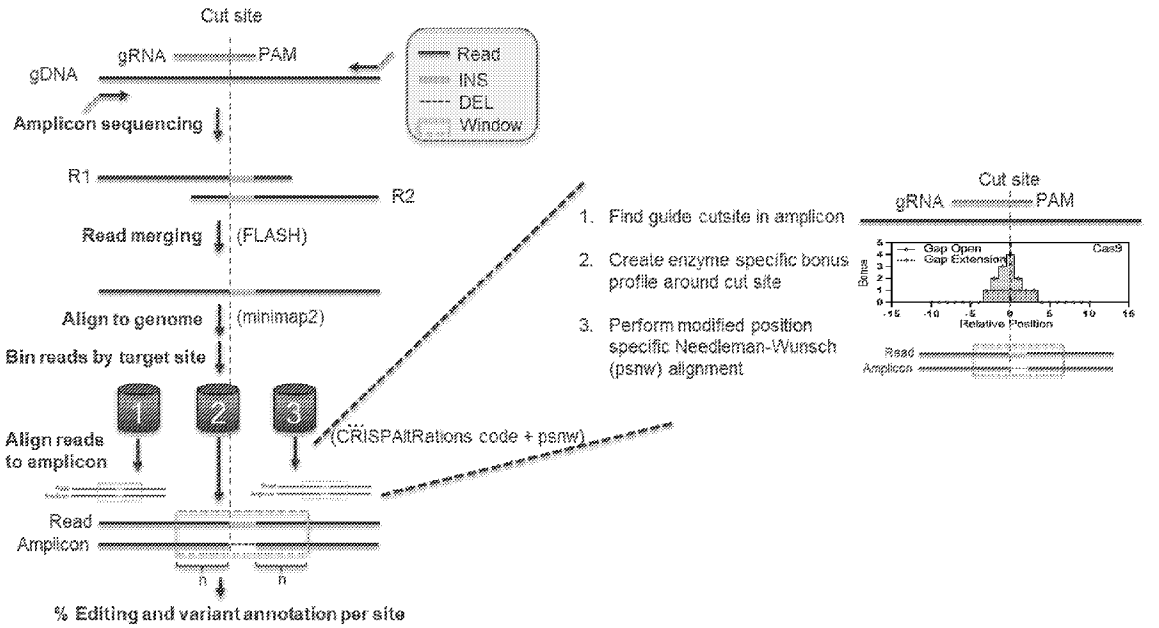
(72) Inventeurs/Inventors:
LI, HENG, US;
KURGAN, GAVIN, US;
MCNEILL, MATTHEW, US;
WANG, YU, US

(74) Agent: ROBIC

(54) Titre : IDENTIFICATION, CARACTERISATION ET QUANTIFICATION DE REPARATIONS DE CASSURE D'ADN DOUBLE BRIN INTRODUITES PAR CRISPR

(54) Title: IDENTIFICATION, CHARACTERIZATION, AND QUANTITATION OF CRISPR-INTRODUCED DOUBLE-STRANDED DNA BREAK REPAIRS

FIG. 1



(57) Abrégé/Abstract:
Described herein is a system and process for identifying and characterizing double-stranded DNA break repair sites that are based on biological information and have improved accuracy. Also described is a sequence alignment process that uses biological data to inform the alignment matrix for position specific alignment scoring, resulting in the identification of noncanonical target sites.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
07 January 2021 (07.01.2021)

WIPO | PCT

(10) International Publication Number
WO 2021/003343 A1

(51) International Patent Classification:
C12N 9/22 (2006.01) *G16B 20/20* (2019.01)
C12N 15/10 (2006.01) *G16B 30/10* (2019.01)

(21) International Application Number:
PCT/US2020/040621

(22) International Filing Date:
02 July 2020 (02.07.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

62/870,426	03 July 2019 (03.07.2019)	US
62/870,471	03 July 2019 (03.07.2019)	US
62/952,603	23 December 2019 (23.12.2019)	US
62/952,598	23 December 2019 (23.12.2019)	US

(71) Applicant: **INTEGRATED DNA TECHNOLOGIES, INC.** [US/US]; 1710 Commercial Park, Coralville, Iowa 52241 (US).

(72) Inventors: **LI, Heng**; 11 Oak St., Unit 43, Wellesley, Massachusetts 02482 (US). **KURGAN, Gavin**; 3053 Hastings Ave., Iowa City, Iowa 52245 (US). **MCNEILL, Matthew**; 135 N. Westminster St., Iowa City, Iowa 52245 (US). **WANG, Yu**; 8 Bridle Ridge Drive, North Grafton, Massachusetts 01536 (US).

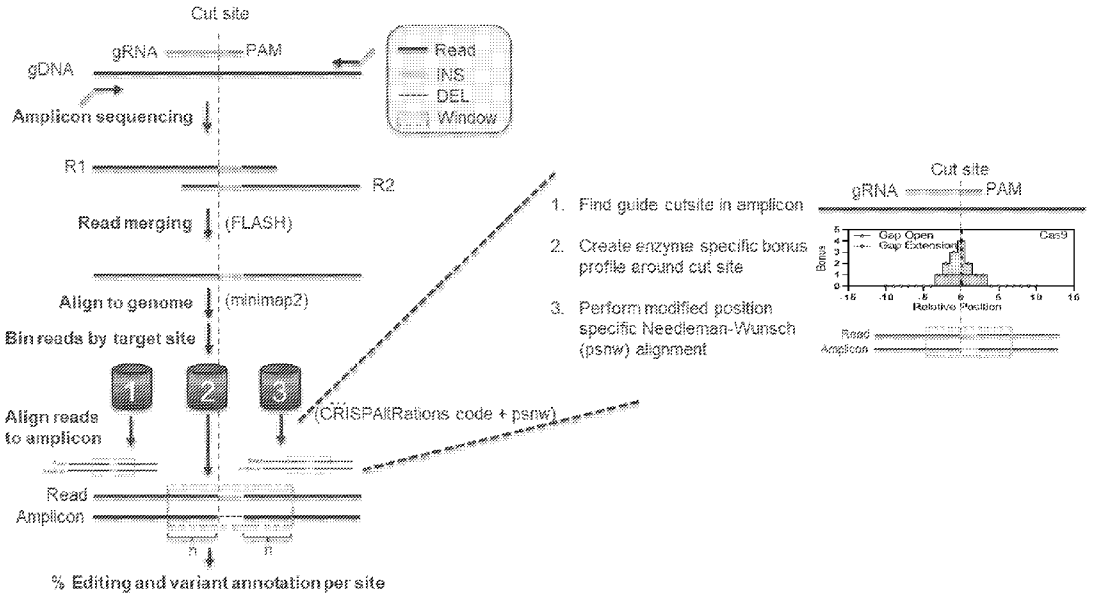
(74) Agent: **BROWN II, Bernard A.** et al.; 790 N Water Street, Suite 2500, Milwaukee, Wisconsin 53202 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,

(54) Title: IDENTIFICATION, CHARACTERIZATION, AND QUANTITATION OF CRISPR-INTRODUCED DOUBLE-STRANDED DNA BREAK REPAIRS

FIG. 1



(57) Abstract: Described herein is a system and process for identifying and characterizing double-stranded DNA break repair sites that are based on biological information and have improved accuracy. Also described is a sequence alignment process that uses biological data to inform the alignment matrix for position specific alignment scoring, resulting in the identification of noncanonical target sites.

[Continued on next page]

WO 2021/003343 A1

TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

IDENTIFICATION, CHARACTERIZATION, AND QUANTITATION OF CRISPR-INTRODUCED DOUBLE-STRANDED DNA BREAK REPAIRS

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims priority to U.S. Provisional Patent Application Nos. 62/870,426 and 67/870,471, both filed on July 3, 2019 and 62/952,603 and 62/952,598, both filed on December 23, 2019, the contents of which are incorporated by reference herein in their entirety.

TECHNICAL FIELD

Described herein is a system and process for identifying and characterizing double-stranded DNA break repair sites that are based on biological information and have improved accuracy. Also described is a sequence alignment process that uses biological data to inform the alignment matrix for position specific alignment scoring, resulting in the accurate identification of noncanonical target sites.

BACKGROUND

The use of targeted nucleases, such as CRISPR proteins, has transformed genome editing. CRISPR enzymes form ribonucleoproteins (RNPs) when hybridized with either 2-part crRNA and tracrRNA or single guide RNAs (sgRNAs). With either approach, a short protospacer sequence (a guide RNA or "gRNA") targets a specific sequence in a complementary molecule. Upon finding a match, these enzymes introduce a break in one or both DNA (or RNA) strands. CRISPR enzymes targeting DNA (e.g., Cas9, Cas12a/Cpf1) introduce double stranded breaks (DSBs) at predictable genomic positions relative to the gRNA's hybridization target. DNA DSBs are repaired by intracellular machinery, but the repair process often results in insertions and deletions (indels), substitutions, and other suboptimal allelic variants.

Because each cell in an affected population must repair itself independent of adjacent cells, and the specific outcome can contain different resulting alleles, the population of cells is likely to contain a plurality of alleles at the targeted location. Additionally, the targeting capability of these nucleases is often somewhat non-specific, which results in undesired mutations at other, off-target, genomic locations.

It is highly desirable to characterize and quantify the plurality of alleles at both on-target and off-target locations. Researchers often use DNA sequencing (such as Illumina next generation sequencing; NGS) to observe the resulting allelic diversity. Multiplexed polymerase chain reaction (PCR) can be performed to amplify and enrich all targeted locations. The resulting

amplicons can be sequenced. The plurality of alleles can be characterized and counted using specialized software.

Many specialized software tools have been developed to characterize the allelic variants resulting from DSBs. Prior tools include *CRISPResso* [1], *crispRvariants* [2], and *Amplican* [3]. These tools generally work by aligning each sequence read against expected amplicon targets using the Needleman-Wunsch, bwa, or custom alignment algorithms. The algorithm generates a list of possible read:target alignments. Each alignment is scored based on the number of matching, mismatching, and missing nucleotides (gaps). The best scoring alignment is used for downstream data processing.

Alignment algorithms sometimes generate equally valued query:target alignments, which are most likely to occur when the query contains an insertion or deletion. From the equally valued options, alignment methods will return all or select one option. When selecting, some methods make the selection at random. Without a good predictive model or set of heuristic rules to make the selection, the alignment choice is variable, which can lead to incorrect indel annotations and lower accuracy results.

What is needed is an algorithm and process for identifying and characterizing double-stranded DNA break repair sites that are based on biological information and have improved accuracy.

SUMMARY

One embodiment described herein is a computer implemented process for identifying and characterizing double-stranded DNA break repair sites with improved accuracy, the process comprising executing on a processor the steps of: receiving sample sequence data comprising a plurality of sequences; analyzing and merging of the sample sequence data and outputting merged sequences; developing target-site sequences containing predicted outcomes of repair events when a single-stranded or a double-stranded DNA oligonucleotide donor is provided and outputting the target predicted outcomes; binning the merged sequences with the target-site sequences or the optional target predicted outcomes using a mapper and outputting target-read alignments; re-aligning the binned target-read alignments to the target-site using an enzyme specific position-specific scoring matrix derived from biological data that is applied based on the position of a guide sequence and a canonical enzyme-specific cut site and producing a final alignment; analyzing the final alignment and identifying and quantifying mutations within a pre-defined sequence distance window from the canonical enzyme-specific cut sites; outputting the final alignment, analysis, and quantification results data as tables or graphics. In one aspect, the

sequence data comprises sequences from a population of cells or subjects. In another aspect, the enzyme-specific cut site comprises one or more of Cas9, Cas12a, or other Cas enzymes. In another aspect, the pre-defined sequence distance window is enzyme specific and comprises between 1 nt to about 15 nt. In another aspect, the results show the percent editing, percent insertion, percent deletion, or a combination thereof. In another aspect, the accuracy of identifying variant target sites is improved by about 15 to about 20% as compared to comparable processes.

Another embodiment described herein is a computer implemented process for aligning biological sequences, the process comprising executing on a processor the steps of: receiving sample sequence data comprising a plurality of sequences; aligning the sequence data to a predicted target sequence using a matrix based on an enzyme specific position-specific scoring of a specific nuclease target site; outputting the alignment results as tables or graphics. In one aspect, the sequence data comprises sequences from a population of cells or subjects. In another aspect, the specific nuclease target sequence comprises a target site for one or more of Cas9, Cas12a, or other Cas enzymes. In another aspect, the matrix uses position-specific gap open and extension penalties.

Another embodiment described herein is a method for identifying and characterizing double-stranded DNA break repair sites with improved accuracy, the process comprising: extracting genomic DNA from a population of cells or tissue from a subject; amplifying the genomic DNA using multiplex PCR to produce amplicons enriched for target-site sequences; sequencing the amplicons and obtaining sample sequence data; subsequently executing on a processor, the steps of: receiving sample sequence data comprising a plurality of sequences; analyzing and merging of the sample sequence data and outputting merged sequences; developing target-site sequences containing predicted outcomes of repair events when a single-stranded or a double-stranded DNA oligonucleotide donor is provided and outputting the target predicted outcomes; binning the merged sequences with the target-site sequences or the optional target predicted outcomes using a mapper and outputting target-read alignments; re-aligning the binned target-read alignments to the target-site using an enzyme specific position-specific scoring matrix derived from biological data that is applied based on the position of a guide sequence and a canonical enzyme-specific cut site and producing a final alignment; analyzing the final alignment and identifying and quantifying mutations within a pre-defined sequence distance window from the canonical enzyme-specific cut sites; outputting the final alignment, analysis, and quantification results data as tables or graphics. In one aspect, the enzyme-specific cut site comprises one or more of Cas9, Cas12a, or other Cas enzymes. In another aspect, the pre-defined sequence distance window is enzyme specific and comprises between 1 nt to about 15 nt. In another

aspect, the results show the percent editing, percent insertion, percent deletion, or a combination thereof. In another aspect, the accuracy of identifying variant target sites is improved by about 15 to about 20% as compared to comparable processes.

DESCRIPTION OF THE DRAWINGS

The patent or application contains at least one drawing executed in color. Copies of this patent application publication or patent with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

FIG. 1. General workflow for *CRISPAltRations*. Edited genomic DNA is extracted and amplified using targeted multiplex PCR to enrich for the on- and predicted off-target loci. Amplicons are sequenced on an Illumina MiSeq. Read pairs are merged into a single fragment (FLASH), mapped to the genome (minimap2) and binned by their alignment to expected amplicon positions. Reads in each bin are re-aligned to the expected amplicon sequence after finding the cut site and creating a position specific gap open/extension bonus matrix to preferentially align indels closer to the cut site/expected indel profiles for each enzyme (*CRISPAltRations* code + psnw). Indels that intersected with a window upstream or downstream of the cut site were annotated. Percent editing is the sum of reads containing indels / total observed.

FIG. 2. Directed acyclic grapihic of the *CRISPRAltRations* pipeline. Dashed boxes represent steps within the pipeline, and each step may contain one or more software tools. Lines and arrows show the flow of information through the pipeline. Two important steps are minimap2_orig_reads (orange) uses *minimap2* [4] to align sequence reads against a reference genome, which is an optional step. Later, the *minimap2* is used to align sequence reads against their expected target regions. The *CRISPy* Python tool (developed internally), re-aligns sequence reads against their target regions by calling a custom-modified version of *psnw*, which performs a cut-site re-alignment against the target region. *CRISPy* is also responsible for characterizing the detected indels in aligned reads.

FIG. 3. MinION sequencing read data aligned to a reference target demonstrates a blunt insertion. The drop in expected coverage (grey highlight) indicates a large insertion. The mismatches on the internal edges of the grey highlight indicate mismatches between the observed read data and the expected reference.

FIG. 4. Cartoon examples representing the kinds of structural variants that could occur when using a DNA oligo as a template for homology directed repair (HDR). These examples are not comprehensive. In the cartoon, blue represents the reference sequence, green represents the homology arms, and orange represents the desired inserted sequence. (A) and (B) Example

double stranded and single stranded template oligos, respectively. (1) [Prefect repair] The region containing the DSB is repaired without any structural variants being introduced even when single or double stranded template oligos are present. (2) [HDR-mediated repair] The template oligo directs HDR, resulting in the desired insertion. Only the desired inserted sequence is observed in the repaired DNA. (3) [Non-homologous end joining (NHEJ) repair] the template oligo is inserted bluntly after the DSB. (4) [NHEJ repair with duplicate insertion] the template oligo is inserted bluntly multiple times after the DSB. Examples 3 and 4 are most likely to occur with a double stranded oligo used as a repair template; the homology arms present in the donor sequence are also inserted into genome.

FIG. 5. Position of occurrence of editing events in Cas9 data. Position of indel occurrence from the nuclease cut site using a set of 274 Cas9 guides editing unique genomic targets in a Jurkat cell line. Both (A) deletion and (B) insertion position events were normalized by quantifying them as % total of the respective indel event within the sample. Only sites with >50 reads and >5% indels were used for analysis to limit low-confidence signal and remove noise. Indels were quantified within a 20 bp window from the cut site. Outliers were largely found to be sites that had a non-reference indel present in the Jurkat cell line that did not appear to be caused by DSB activity.

FIG. 6. Position of occurrence of editing events in Cas12a data. Position of indel occurrence from the nuclease cut site using a set of 199 Cas12a guides editing unique genomic targets in a Jurkat cell line. Both (A) deletion and (B) insertion position events were normalized by quantifying them as % total of the respective indel event within the sample. Only sites with >50 reads and >5% indels were used for analysis to limit low-confidence signal noise. Indels were quantified within a 20-bp window from the cut site. Outliers were largely found to be sites that had a non-reference indel present in the Jurkat cell line that did not appear to be caused by DSB activity.

FIG. 7. Gap open/extension penalty vectors near the nuclease introduced DSB. To positively weight indels near the cut site in sequence alignments, we use a position specific matrix of values representing variable gap open or extension penalties are used, where the length of the array is equal to the integer positions of each nucleic acid in the target (thick blue lines). The vector values (red circles and blue diamonds) are varied based on the proximity to the nuclease cut site (vertical black dashed line). Thus, indels that are closest to the cut site have the least gap open or extension penalty.

FIG. 8. Choosing an optimal window to observe true CRISPR-Cas9 editing events. (A) The editing window is a nucleotide distance around the CRISPR cut site used to identify indels.

True and false editing events were calculated as % indels from targeted sequencing of cells treated or untreated for (B) Cas9 (n = 263 pairs of treatment vs. control) or (C) Cas12a (n = 384 pairs of treatment vs. control), respectively. Most true editing events are collected with a 4 nt window for Cas9 or 7 nt for Cas12a; while expanding the window only further collects additional false editing.

FIG. 9. Screen shot of deduplicated read alignments by observed frequency. Total region coverage is indicated by the height of the vertical gray bars (top), and reads are horizontal colored bars. Brighter colored reads indicate more frequently observed indels. Horizontal thin lines indicate deletions; vertical purple "I" symbols show insertions, and colored bars within reads indicate mismatched bases.

FIG. 10. *CRISPAltRations* reliably finds the correct indel. Each bar reports the percentage of correctly returned indels. Using synthetically generated data representing 12,060 unique indel events, evenly distributed between each size bin (x-axis) and 603 unique amplicons. Error bars represent the 95% confidence interval across targets.

FIG. 11. *CRISPRAltRations* more accurately reports the total % editing at the 603 synthetic targets with simulated (A) Cas9 or (B) Cas12a editing. Dots represent the percent editing observed at each target site, and the horizontal line represents the percent indels introduced at each target in the generated synthetic data. *CRISPResso2*Pooled and *CRISPResso1*Pooled were run using default pipeline parameters for each panel with the required minimal read depth for quantification removed (to prevent incorrect drop-out of amplicons).

DETAILED DESCRIPTION

One embodiment described herein is an analytical pipeline called *CRISPAltRations*. *See* FIGs. 1–2. Briefly, this pipeline takes in FASTQ files, and builds a merged R_1/R_2 consensus using *FLASH*. Simultaneously, a target site reference is built, which describes the sequences for all expected on-target locations. Optionally, a target is built that contains an expected outcome of a homology directed repair (HDR) event. Next, the merged sequence reads are aligned to the target reference sequences using *minimap2*, (which was originally developed for rapid alignment of long reads (e.g., those generated by the Oxford Nanopore Technologies MinION). Reads aligning to each target are then re-aligned using a modified form of the Smith-Waterman aligner. The modified aligner allows us to improve detection of insertions and deletions resulting from DSB repair. All observed variants within a pre-defined distance of the DSB location are characterized and quantified. Finally, the results are summarized in tables and graphs. The various described programs, tools, and file types (as well as those mentioned below) are familiar to and readily

accessible to those having ordinary skill in the art. It should be understood that these programs, tools, and file types are exemplary and are not intended to be limiting. Other tools and file types could be used to practice the described processing and analysis.

In this analytical pipeline, the following improvements over prior methods are described. First, the use of *minimap2* [4] enables alignment of reads generated from both short and long read sequencers. Second, by constructing the expected outcome of the homology directed repair event, the ability to characterize perfect (i.e., correctly occurring) HDR events is improved. Third, use of the modified Needleman-Wunsch aligner that can accept a Cas-specific bonus matrix enables significantly improved indel characterization and percent (%) editing quantification over prior methods. Fourth, graphical visualization of the introduced allelic variants is improved. Fifth, a predicted repair event, as described in a prior tool [5], is compared against the observed repair, and the molecular pathways involved in the repair can be described.

In one embodiment, the processes described herein have the following advantageous uses:

- Accurate characterization of indel profiles resulting from DSBs.
- The fraction of reads containing an indel after a DSB is repaired is used to calculate the percentage of editing. This metric (% editing) is used to determine the effectiveness of a gRNA for use in CRISPR-Cas gene editing.
- Accurate characterization of the resulting indel similarly improves the ability to identify the percentage of cellular chromosomes in a population of cells containing a frame-shifting mutation. Frame-shifting mutations modify proteins encoded by affected genes.
- Accurate characterization of inserted sequences.
- Accurate characterization of multiple mutations resulting from multiple gRNA/Cas9 (i.e., ribonucleoprotein complex) deliveries or dual-guide region modifications.
- Analysis of indels sequenced on a long-read platform, such as MinION. Additionally, it allows phased characterization of both ends of large (>400 nt) insertion of deletion events, which occur after DSB repair.
- Improved result visualization.

One embodiment described herein is a computer implemented process for identifying and characterizing double-stranded DNA break repair sites with improved accuracy, the process comprising executing on a processor the steps of: receiving sample sequence data comprising a plurality of sequences; analyzing and merging of the sample sequence data and outputting merged sequences; developing target-site sequences containing predicted outcomes of repair

events when a single-stranded or a double-stranded DNA oligonucleotide donor is provided and outputting the target predicted outcomes; binning the merged sequences with the target-site sequences or the optional target predicted outcomes using a mapper and outputting target-read alignments; re-aligning the binned target-read alignments to the target-site using an enzyme specific position-specific scoring matrix derived from biological data that is applied based on the position of a guide sequence and a canonical enzyme-specific cut site and producing a final alignment; analyzing the final alignment and identifying and quantifying mutations within a pre-defined sequence distance window from the canonical enzyme-specific cut sites; outputting the final alignment, analysis, and quantification results data as tables or graphics.

In one embodiment, edited genomic DNA is extracted and amplified using targeted multiplex PCR to enrich for the on- and predicted off-target loci. Amplicons are sequenced on an Illumina MiSeq. Read pairs are merged into a single fragment (FLASH), mapped to the genome (minimap2) and binned by their alignment to expected amplicon positions. Reads in each bin are re-aligned to the expected amplicon sequence after finding the cut site and creating a position specific gap open/extension bonus matrix to preferentially align indels closer to the cut site/expected indel profiles for each enzyme (CRISPAltRations code + psnw). Indels that intersected with a window upstream or downstream of the cut site were annotated. Percent editing is the sum of reads containing indels / total observed.

In some embodiments, the process described herein uses *minimap2* [4], which enables alignment of reads generated from both short and long read sequencers. Prior tools typically only accept short read sequencing data, such as those that are generated by Illumina sequencers. Others have used long read sequencing data to examine large insertions or deletions [6–8], but no stand-alone publicly available tools are believed to exist. Long read data handling is partially enabled by use of the *minimap2* aligner. For example, the alignment results can be visualized, which shows identification of a blunt molecular insertion in DNA after a DSB repair (FIG. 3). Another embodiment sorts between real and noise-derived indels using a method similar to a previously published tool [7] in which small indels are ignored.

In another embodiment, by constructing the expected outcome of the HDR event, the ability to characterize perfect HDR events is improved. A reference file, in FASTA format, contains each expected sequence target and modified sequence targets as well. The first step toward constructing this file involves creating a reference sequence index that enables reads to be aligned to each expected structural variant. For example, if one interrogates a region targeted for a DSB and double stranded DNA donor oligo to enable HDR, there are multiple different likely biological repair outcomes:perfect repair (FIG. 4-1), HDR-mediated repair (FIG. 4-2), NHEJ repair

(FIG. 4-3), and NHEJ repair with duplicate insertion (FIG. 4-4). Other outcomes, such as template fragment or triple template insertions, are also possible (not shown). A similar reference file construction approach has been used by other tools, such as UDiTaS™ [9].

In another embodiment, a modified version of the Needleman-Wunsch algorithm is used to re-align reads against their expected target. The method described herein increases accuracy of alignments containing an indel (as annotated in alignment's CIGAR string). It significantly improves indel characterization and % editing quantification over prior methods. DNA sequence aligners such as *minimap2* and Needleman-Wunsch approaches weigh indel alignments using fixed penalties for opening and extending gaps. This method is improved upon by re-aligning reads to their targets using position-specific gap open and extension penalties (enabled in a tool called "*psnw*") such that alignments with indels favor positioning them overlapping or near the predicted DSB. This position specific matrix is set to reflect the actual characterized indel profile of the specific Cas enzyme being used for editing (FIGs. 5–6). Thus, indel base alignments are most highly favored at or near the predicted target cut site (variable scoring strategy; FIG. 7). This method enables accurate realignment of indels, particularly those that occur in repetitive regions in the reference sequence. This approach improves the ability to identify the most biologically likely result.

A recently developed tool (*CRISPResso2* [11]), uses a cut-site aware alignment strategy. However, the processes described herein implements the cut-site aware alignment method using a full gap open/extension matrix during alignment that is tuned by actual editing data at Cas9/Cas12a sites and implemented in C++. In contrast, *CRISPResso2* uses a method that only enables a single bonus at the cut site and is implemented in Python.

In another embodiment, the processes described herein collect indels nearby the nuclease cut site and tag indels that intersect the cut site, or within a fixed distance. Some published accounts suggest a 1–2 nt fixed distance, but the data supporting those choices has been limited. In developing the embodiments described herein, the optimal distance (i.e., window size) around the cut site was studied using a set of Cas9-RNP treated and paired untreated control samples. It was observed that a 4-nt window for Cas9 or a 7-nt window for Cas12a provided the greatest sensitivity and provided an acceptable specificity (FIG. 4). The larger window requirement for Cas12a is likely due to the mechanism of action; Cas12a implements a double strand break by producing two single strand breaks 5-bp away (leaving "sticky" ends) [12]. Thus, the process described herein can be expanded to other nucleases (e.g., CasX) [13] having biological data to inform the target window size and enzymatic mechanism of action.

In another embodiment, graphical visualization of the allelic variation is improved. Downstream of the alignment step, several other analyses are performed that are unique to the described method. To generate an improved visualization, reads are deduplicated based on the identity of identified indel sequences within the CRISPR editing window post-alignment. Deduplicated reads are written back to a BAM file, and the frequency of each deduplicated read within the original population of reads is written to an associated BAM tag. After the file is indexed, indels in deduplicated reads and their associated frequencies can be visualized using the commonly available *IGV* tool [10] (FIG. 9).

In another embodiment, a predicted repair event is compared, as described in a prior tool [5], against the observed repair, and can be used to determine the molecular pathways involved in the repair. The system described herein also adds the ability to compare the observed indel profile against expected indel profiles, which enables rapid discernment about whether an experimental treatment modified the intracellular mechanisms of DNA repair.

The utility of the system and methods described herein is demonstrated by generating a synthetic set of 603 gRNA:amplicon pairs. At each target, 4000 read pairs (2 × 150 bp) are synthetically generated with a simulated Illumina MiSeq v3 platform error profile. In half of the reads, random indels are introduced based on a model generated off the observed editing profile for Cas9 and Cas12a (FIGs. 4–5). The synthetic data is analyzed using the *CRISPRAltRations* system described herein and the previously published *CRISPResso1* and *CRISPResso2* tools [11]. By implementing the method described herein, the ability to correctly characterize indels is improved by ~15–20% (FIG. 10). The algorithm described herein has increased accuracy because it produces a biologically informed selection of the best alignment in targets where multiple equally scored alignments are possible. Additionally, the method described herein more accurately calculates the percentage of modified DNA molecules (FIG. 11). The process and strategy described herein is an important enhancement toward characterizing and quantifying indels introduced after DSB repair.

Another embodiment described herein is a computer implemented process for aligning biological sequences, the process comprising executing on a processor the steps of: receiving sample sequence data comprising a plurality of sequences; aligning the sequence data to a predicted target sequence using a matrix based on an enzyme specific position-specific scoring of a specific nuclease target site; outputting the alignment results as tables or graphics. In one aspect, the sequence data comprises sequences from a population of cells or subjects. In another aspect, the specific nuclease target sequence comprises a target site for one or more of Cas9,

Cas12a, or other Cas enzymes. In another aspect, the matrix uses position-specific gap open and extension penalties.

Another embodiment described herein is a method for identifying and characterizing double-stranded DNA break repair sites with improved accuracy, the process comprising: extracting genomic DNA from a population of cells or tissue from a subject; amplifying the genomic DNA using multiplex PCR to produce amplicons enriched for target-site sequences; sequencing the amplicons and obtaining sample sequence data; subsequently executing on a processor, the steps of: receiving sample sequence data comprising a plurality of sequences; analyzing and merging of the sample sequence data and outputting merged sequences; developing target-site sequences containing predicted outcomes of repair events when a single-stranded or a double-stranded DNA oligonucleotide donor is provided and outputting the target predicted outcomes; binning the merged sequences with the target-site sequences or the optional target predicted outcomes using a mapper and outputting target-read alignments; re-aligning the binned target-read alignments to the target-site using an enzyme specific position-specific scoring matrix derived from biological data that is applied based on the position of a guide sequence and a canonical enzyme-specific cut site and producing a final alignment; analyzing the final alignment and identifying and quantifying mutations within a pre-defined sequence distance window from the canonical enzyme-specific cut sites; outputting the final alignment, analysis, and quantification results data as tables or graphics.

Many different arrangements of the various components and processes described herein as well as components or processes not shown, are possible without departing from the spirit and scope of the present disclosure. It should be understood that embodiments or aspects may include and otherwise be implemented by a combination of various hardware, software, or electronic components. For example, various microprocessors and application specific integrated circuits ("ASICs") can be utilized, as can software of a variety of languages Also, servers and various computing devices can be used and can include one or more processing units, one or more computer-readable mediums, one or more input/output interfaces, and various connections (e.g., a system bus) connecting the components.

It will be apparent to one of ordinary skill in the relevant art that suitable modifications and adaptations to the compositions, formulations, methods, processes, and applications described herein can be made without departing from the scope of any embodiments or aspects thereof. The compositions and methods provided are exemplary and are not intended to limit the scope of any of the specified embodiments. All the various embodiments, aspects, and options disclosed herein can be combined in any variations or iterations. The scope of the methods and

processes described herein include all actual or potential combinations of embodiments, aspects, options, examples, and preferences herein described. The methods described herein may omit any component or step, substitute any component or step disclosed herein, or include any component or step disclosed elsewhere herein. Should the meaning of any terms in any of the patents or publications incorporated by reference conflict with the meaning of the terms used in this disclosure, the meanings of the terms or phrases in this disclosure are controlling. Furthermore, the specification discloses and describes merely exemplary embodiments. All patents and publications cited herein are incorporated by reference herein for the specific teachings thereof.

REFERENCES

1. Pinello, L. et al., "Analyzing CRISPR genome-editing experiments with *CRISPResso*." *Nat Biotechnol*. 34(7): 695–697 (2016).
2. Lindsay, H. et al., "CrispRVariants: precisely charting the mutation spectrum in genome engineering experiments," *Nat. Biotechnol.* 34(7): 701–703 (2015).
3. Labun, K. et al., "Accurate analysis of genuine CRISPR editing events with ampliCan Kornel," *bioRxiv* 249474 (2018); now published in *Genome Research* 29: 843-847 (2019)
4. Li, H., "Minimap2: Pairwise alignment for nucleotide sequences," *Bioinformatics* 34(18): 3094–3100 (2018).
5. Shen, M. W. et al., "Predictable and precise template-free CRISPR editing of pathogenic variants," *Nature* 563 (7733): 646–651 (2018).
6. Hendel, A. et al., "Quantifying genome-editing outcomes at endogenous loci with SMRT sequencing." *Cell Rep*. 7(1): 293–305 (2014).
7. Iyer, S. et al., "Precise therapeutic gene correction by a simple nuclease-induced double-stranded break," *Nature* 568 (7753): 561–565 (2019).
8. Vu, G. T. H. et al., "Endogenous sequence patterns predispose the repair modes of CRISPR/Cas9-induced DNA double-stranded breaks in *Arabidopsis thaliana*," *Plant J*. 92(1): 57–67 (2017).
9. Giannoukos, G. et al., "UDiTaS™, a genome editing detection method for indels and genome rearrangements," *BMC Genomics* 19: 212 (2018).
10. Robinson, J., "Integrated genomics viewer," *Nat. Biotechnol.* 29(1), 24–26 (2012).
11. Clement, K. et al., "Analysis and comparison of genome editing using *CRISPResso2*," *bioRxiv* 1–20 (2018). Now published in *Nat. Biotechnol*. 37(3): 224-226 (2019)

12. Zetsche, B. et al., "Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System," *Cell* 163(3): 759–771 (2015).
13. Liu, J. J. et al., "CasX enzymes comprise a distinct family of RNA-guided genome editors," *Nature* 566(7743): 218–223 (2019).

COMPUTER CODE

Exemplary code used to generate a 1D scoring matrix for gap open bonus during alignment using *psnw*.

```
def get_gap_open_bonus(self, cut_up, cut_down, rel_cut_pos, ref_amplicon, nuclease,
direction):
    """
    This function takes in an minimum and maximum distance to look for CRISPR variants
(window size)
    and creates a string to introduce weights for gap open bonuses into the psnw
alignment algorithm.
    If there are multiple cuts, any position between the two cuts is weighted equally

    e.g. one cut Cas9
    :param cut_up: 2
    :param cut_down: 6
    :param rel_cut_pos: 4
    :param ref_amplicon: ACTGATCA
    :return: 002342100

    e.g. two cut Cas9
    :param cut_up: 2
    :param cut_down: 6
    :param rel_cut_pos: 4
    :param ref_amplicon: ACTGATCA
    :return: 001111100
    """
    gap_open_string = []
    cpf1_enz = list(["CAS12A", "CPF1", "MAD7"])
    try:
        # First the program looks to see if there is a second cut site
        if not None in rel_cut_pos:
            for i in range(0,len(ref_amplicon)):
                # If there is a second cut site, everything between the two cuts gets
a +1 bonus
                if cut_down <= i <= cut_up:
                    gap_open_string.append("1")
                # If the position is not between the two cuts, it gets no bonus
                else:
                    gap_open_string.append("0")

        # Execute these functions if there is only one cut site
        else:
            #If the enzyme is Cas9 nuclease, perform the following functions
            if "CAS9" in nuclease.upper():
                for i in range(0, len(ref_amplicon)):
                    # The cut site gets a +4 gap open bonus automatically
                    if i == int(rel_cut_pos[0]):
                        gap_open_string.append("4")
                    # Sites adjacent by 1bp to the cut site and distal from PAM get a
+3 gap open bonus
                    elif ((direction[0] == "forward") and (i == (int(rel_cut_pos[0])-
1))) or \
                            ((direction[0] == "reverse") and (i ==
(int(rel_cut_pos[0]) + 1))):
                        gap_open_string.append("3")
                    # Sites adjacent by 2bp to the cut site distal from PAM get a +2
gap open bonus
                    elif ((direction[0] == "forward") and (i == (int(rel_cut_pos[0]) -
2))) or \
```

```
                            ((direction[0] == "reverse") and (i ==
(int(rel_cut_pos[0]) + 2))):
                        gap_open_string.append("2")
                    # Sites adjacent by 1bp to the cut site proximal from PAM get a +2
gap open bonus
                    elif ((direction[0] == "forward") and (i == (int(rel_cut_pos[0]) +
1))) or \
                            ((direction[0] == "reverse") and (i ==
(int(rel_cut_pos[0]) - 1))):
                        gap_open_string.append("2")
                    # Sites between the upstream and downstream window get a +1 gap
open bonus
                    elif cut_down <= i <= cut_up:
                        gap_open_string.append("1")
                    # All other positions get no bonus
                    else:
                        gap_open_string.append("0")

            elif any(enzyme in nuclease.upper() for enzyme in cpf1_enz):
                for i in range(0, len(ref_amplicon)):

                    if direction[0] == "forward":
                        #Sites at the PAM distal cutsite get a +4 bonus
                        if i == (int(rel_cut_pos[0])):
                            gap_open_string.append("4")
                        #Site at the PAM proximal cutsite get a +3 bonus
                        elif i == (int(rel_cut_pos[0]) - 4):
                            gap_open_string.append("3")
                        # Sites adjacent by 1bp to either cut site get a +2 gap open
bonus
                        elif (i == (int(rel_cut_pos[0])-5)) \
                                or (i == (int(rel_cut_pos[0]-3))) \
                                or (i == (int(rel_cut_pos[0]) - 1)) \
                                or (i == (int(rel_cut_pos[0]) + 1)):
                            gap_open_string.append("2")
                        # Sites between the upstream and downstream window get a +1
gap open bonus
                        elif cut_down <= i <= cut_up:
                            gap_open_string.append("1")
                        # All other positions get no bonus
                        else:
                            gap_open_string.append("0")

                    elif direction[0] == "reverse":
                        # Sites at the PAM distal cutsite get a +4 bonus
                        if i == (int(rel_cut_pos[0])):
                            gap_open_string.append("4")
                        # Site at the PAM proximal cutsite get a +3 bonus
                        elif i == (int(rel_cut_pos[0]) + 4):
                            gap_open_string.append("3")
                        # Sites adjacent by 1bp to either cut site get a +2 gap open
bonus
                        elif (i == (int(rel_cut_pos[0]) + 5)) \
                                or (i == (int(rel_cut_pos[0]) + 1)) \
                                or (i == (int(rel_cut_pos[0]) - 1)) \
                                or (i == (int(rel_cut_pos[0]) + 3)):
                            gap_open_string.append("2")
                        # Sites between the upstream and downstream window get a +1
gap open bonus
                        elif cut_down <= i <= cut_up:
                            gap_open_string.append("1")
                        # All other positions get no bonus
                        else:
```

```
                        gap_open_string.append("0")

        else:
            print("Nuclease is not accepted: {}").format(nuclease)
            raise ValueError
except ValueError:
    print("This value is not valid")
return "".join(gap_open_string)
```

Exemplary code used to generate 1D scoring matrix for gap extension bonus during alignment using *psnw*.

```
def get_gap_ext_bonus(self, cut_up, cut_down, rel_cut_pos, ref_amplicon):
    """
    This function takes in an minimum and maximum distance to look for CRISPR variants
(window size)
    and creates a string to introduce weights for gap extension bonuses into the psnw
alignment algorithm.
    If there are multiple cuts, any position between the two cuts is weighted equally

    e.g. one cut Cas9
    :param cut_up: 2
    :param cut_down: 6
    :param rel_cut_pos: 4
    :param ref_amplicon: ACTGATCA
    :return: 001111100

    e.g. two cut Cas9
    :param cut_up: 2
    :param cut_down: 6
    :param rel_cut_pos: 4
    :param ref_amplicon: ACTGATCA
    :return: 001111100
    """
    gap_ext_string = []
    try:
        # First the program looks to see if there is a second cut site
        if not None in rel_cut_pos:
            for i in range(0,len(ref_amplicon)):
                # If there is a second cut site, everything between the two cuts gets
a +1 bonus
                if cut_down <= i <= cut_up:
                    gap_ext_string.append("1")
                # If the position is not between the two cuts, it gets no bonus
                else:
                    gap_ext_string.append("0")

        # Execute these functions if there is only one cut site
        else:
            for i in range(0, len(ref_amplicon)):
                # If there is a second cut site, everything between the two cuts gets
a +1 bonus
                if i == int(rel_cut_pos[0]):
                    gap_ext_string.append("1")
                elif (i == (int(rel_cut_pos[0])-1)) or (i == (int(rel_cut_pos[0])+1)):
                    gap_ext_string.append("1")
                elif cut_down <= i <= cut_up:
                    gap_ext_string.append("1")
                else:
                    gap_ext_string.append("0")
    except ValueException:
        print("This value is not valid")
    return "".join(gap_ext_string)
```

CLAIMS

What is claimed

1. A computer implemented process for identifying and characterizing double-stranded DNA break repair sites with improved accuracy, the process comprising executing on a processor the steps of:
receiving sample sequence data comprising a plurality of sequences;
analyzing and merging of the sample sequence data and outputting merged sequences;
developing target-site sequences containing predicted outcomes of repair events when a single-stranded or a double-stranded DNA oligonucleotide donor is provided and outputting the target predicted outcomes;
binning the merged sequences with the target-site sequences or the optional target predicted outcomes using a mapper and outputting target-read alignments;
re-aligning the binned target-read alignments to the target-site using an enzyme specific position-specific scoring matrix derived from biological data that is applied based on the position of a guide sequence and a canonical enzyme-specific cut site and producing a final alignment;
analyzing the final alignment and identifying and quantifying mutations within a pre-defined sequence distance window from the canonical enzyme-specific cut sites;
outputting the final alignment, analysis, and quantification results data as tables or graphics.

2. The process of claim 1, wherein the sequence data comprises sequences from a population of cells or subjects.

3. The process of claim 1, wherein the enzyme-specific cut site comprises one or more of Cas9, Cas12a, or other Cas enzymes.

4. The process of claim 1, wherein the pre-defined sequence distance window is enzyme specific and comprises between 1 nt to about 15 nt.

5. The process of claim 1, wherein the results show the percent editing, percent insertion, percent deletion, or a combination thereof.

6. The process of claim 1, wherein the accuracy of identifying variant target sites is improved by about 15 to about 20% as compared to comparable processes.

7. A computer implemented process for aligning biological sequences, the process comprising executing on a processor the steps of:
receiving sample sequence data comprising a plurality of sequences;
aligning the sequence data to a predicted target sequence using a matrix based on an enzyme specific position-specific scoring of a specific nuclease target site;
outputting the alignment results as tables or graphics.

8. The process of claim 7, wherein the sequence data comprises sequences from a population of cells or subjects.

9. The process of claim 7, wherein the specific nuclease target sequence comprises a target site for one or more of Cas9, Cas12a, or other Cas enzymes.

10. The process of claim 7, wherein the matrix uses position-specific gap open and extension penalties.

11. A method for identifying and characterizing double-stranded DNA break repair sites with improved accuracy, the process comprising:
extracting genomic DNA from a population of cells or tissue from a subject;
amplifying the genomic DNA using multiplex PCR to produce amplicons enriched for target-site sequences;
sequencing the amplicons and obtaining sample sequence data;
subsequently executing on a processor, the steps of:
receiving sample sequence data comprising a plurality of sequences;
analyzing and merging of the sample sequence data and outputting merged sequences;
developing target-site sequences containing predicted outcomes of repair events when a single-stranded or a double-stranded DNA oligonucleotide donor is provided and outputting the target predicted outcomes;
binning the merged sequences with the target-site sequences or the optional target predicted outcomes using a mapper and outputting target-read alignments;

re-aligning the binned target-read alignments to the target-site using an enzyme specific position-specific scoring matrix derived from biological data that is applied based on the position of a guide sequence and a canonical enzyme-specific cut site and producing a final alignment;

analyzing the final alignment and identifying and quantifying mutations within a pre-defined sequence distance window from the canonical enzyme-specific cut sites;

outputting the final alignment, analysis, and quantification results data as tables or graphics.

12. The process of claim 1, wherein the enzyme-specific cut site comprises one or more of Cas9, Cas12a, or other Cas enzymes.

13. The process of claim 1, wherein the pre-defined sequence distance window is enzyme specific and comprises between 1 nt to about 15 nt.

14. The process of claim 1, wherein the results show the percent editing, percent insertion, percent deletion, or a combination thereof.

15. The process of claim 1, wherein the accuracy of identifying variant target sites is improved by about 15 to about 20% as compared to comparable processes.

FIG. 1

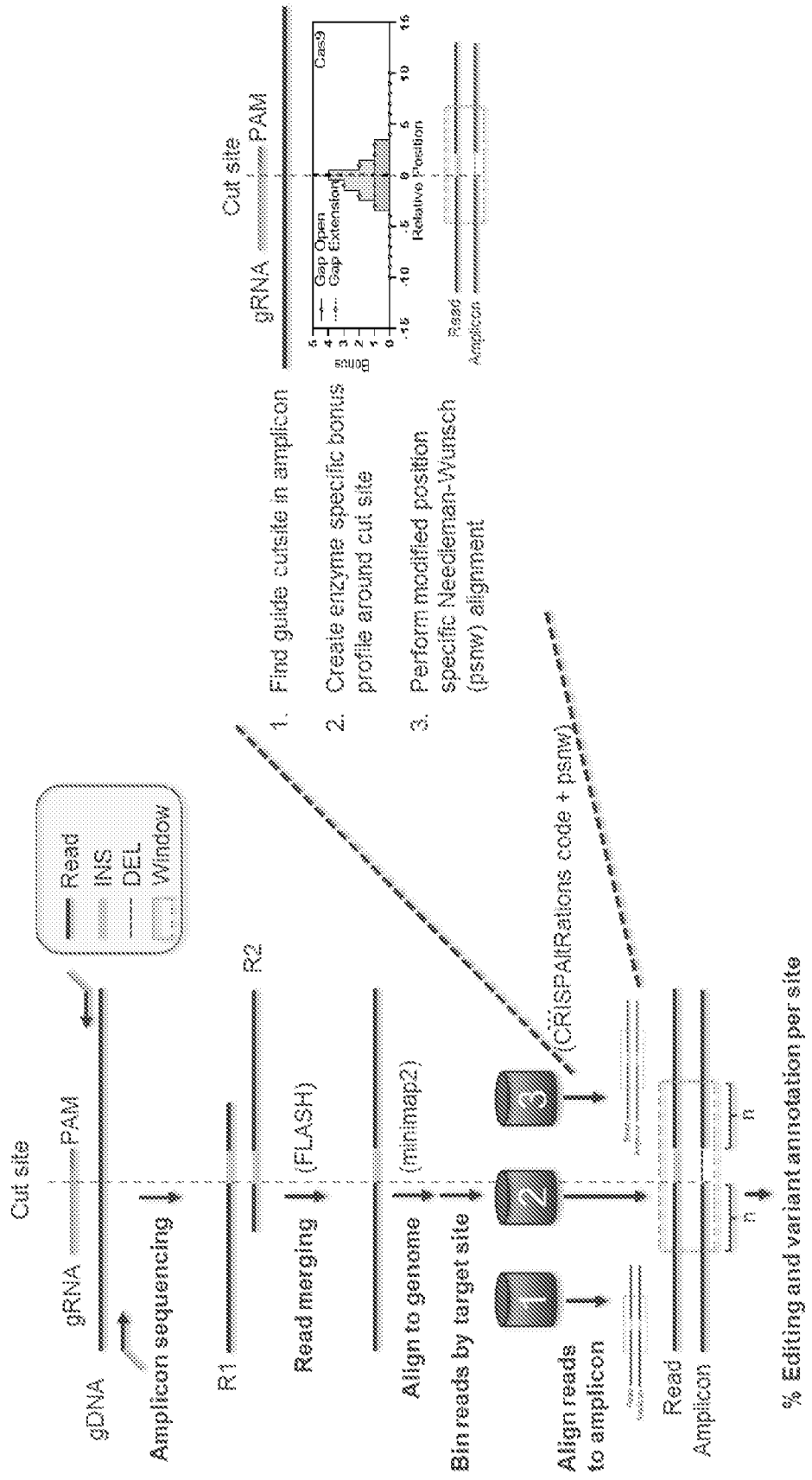

Cut site
gRNA
PAM
gDNA
Amplicon sequencing
R1
R2
Read merging
(FLASH)
Align to genome
(minimap2)
Bin reads by target site
Align reads to amplicon
(CRISPAIRations code + psnw)
Read
Amplicon
% Editing and variant annotation per site
Read
INS
DEL
Window
1. Find guide cutsite in amplicon
2. Create enzyme specific bonus profile around cut site
3. Perform modified position specific Needleman-Wunsch (psnw) alignment
Gap Open
Gap Extension
Cas9
Bonus
Relative Position

FIG. 2

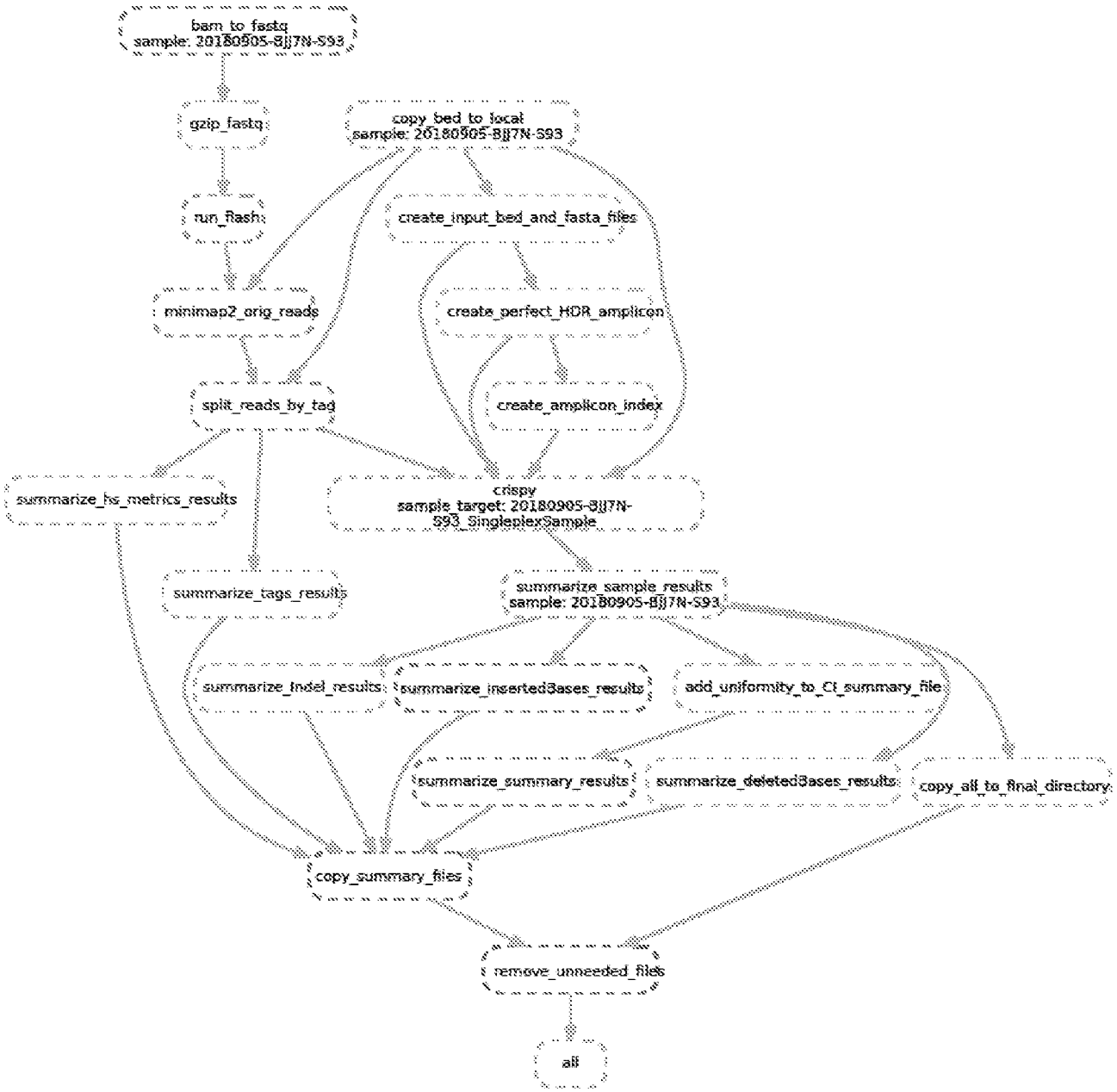

bam_to_fastq
sample: 20180905-BJ7N-S93
gzip_fastq
copy_bed_to_local
sample: 20180905-BJ7N-S93
run_flash
create_input_bed_and_fasta_files
minimap2_orig_reads
create_perfect_HDR_amplicon
split_reads_by_tag
create_amplicon_index
summarize_hs_metrics_results
crispy
sample_target: 20180905-BJ7N-S93_SingleplexSample
summarize_tags_results
summarize_sample_results
sample: 20180905-BJ7N-S93
summarize_indel_results
summarize_insertedBases_results
add_uniformity_to_CI_summary_file
summarize_summary_results
summarize_deletedBases_results
copy_all_to_final_directory
copy_summary_files
remove_unneeded_files
all

FIG. 3

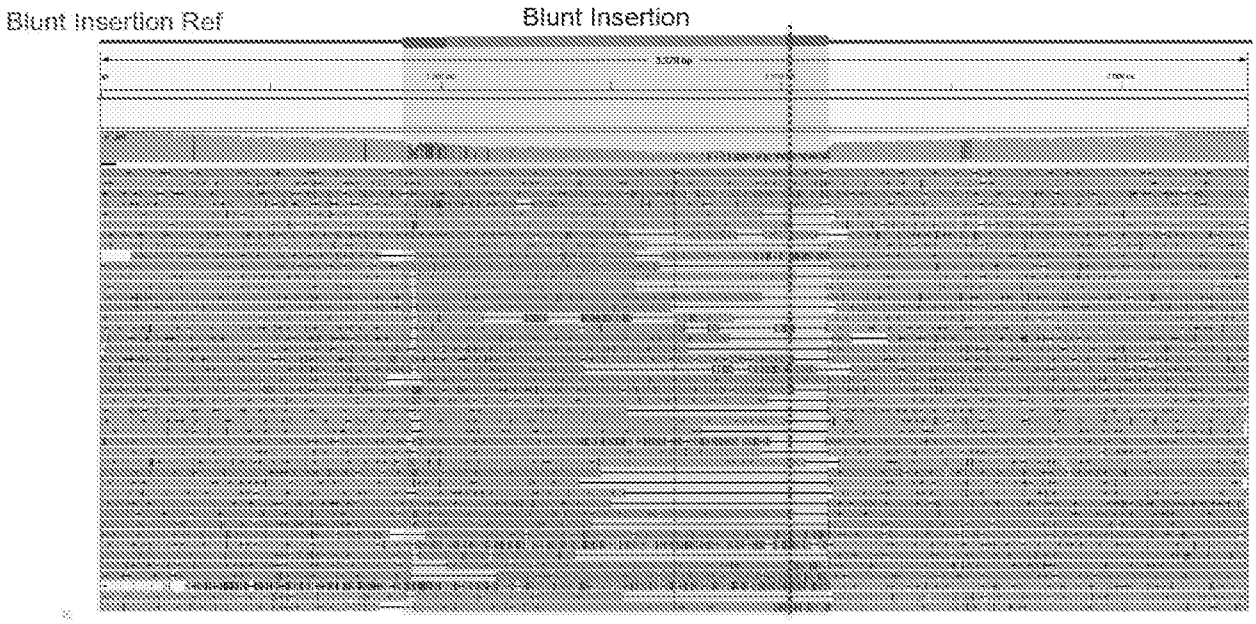


FIG. 4

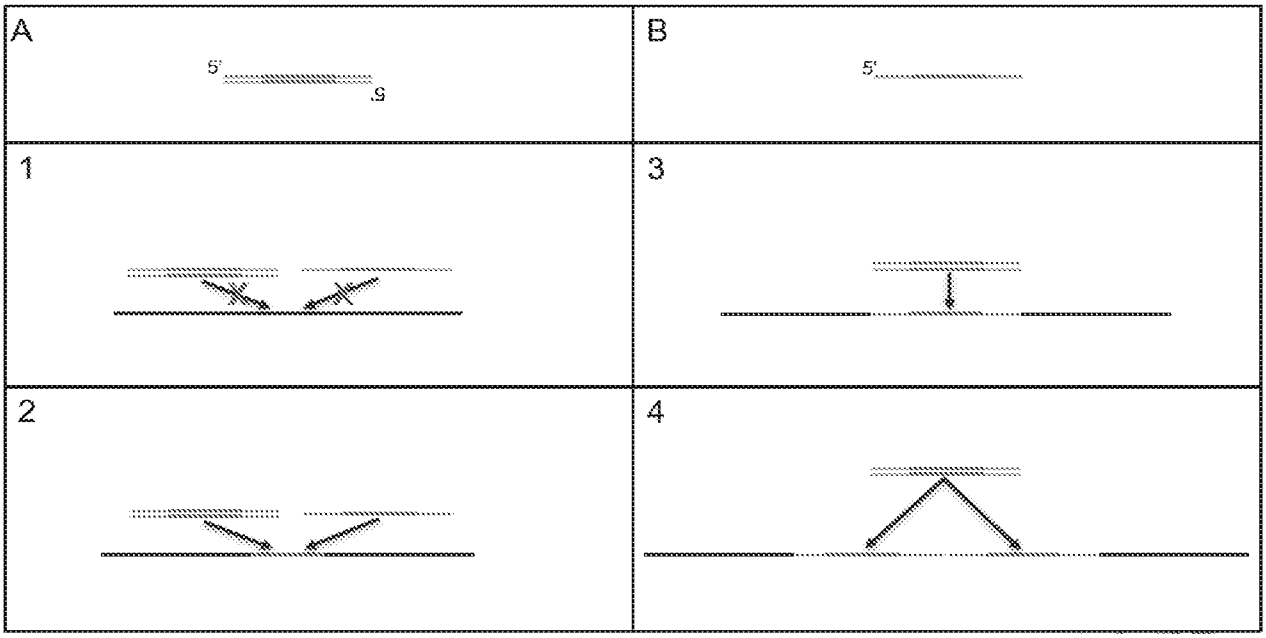


FIG. 5

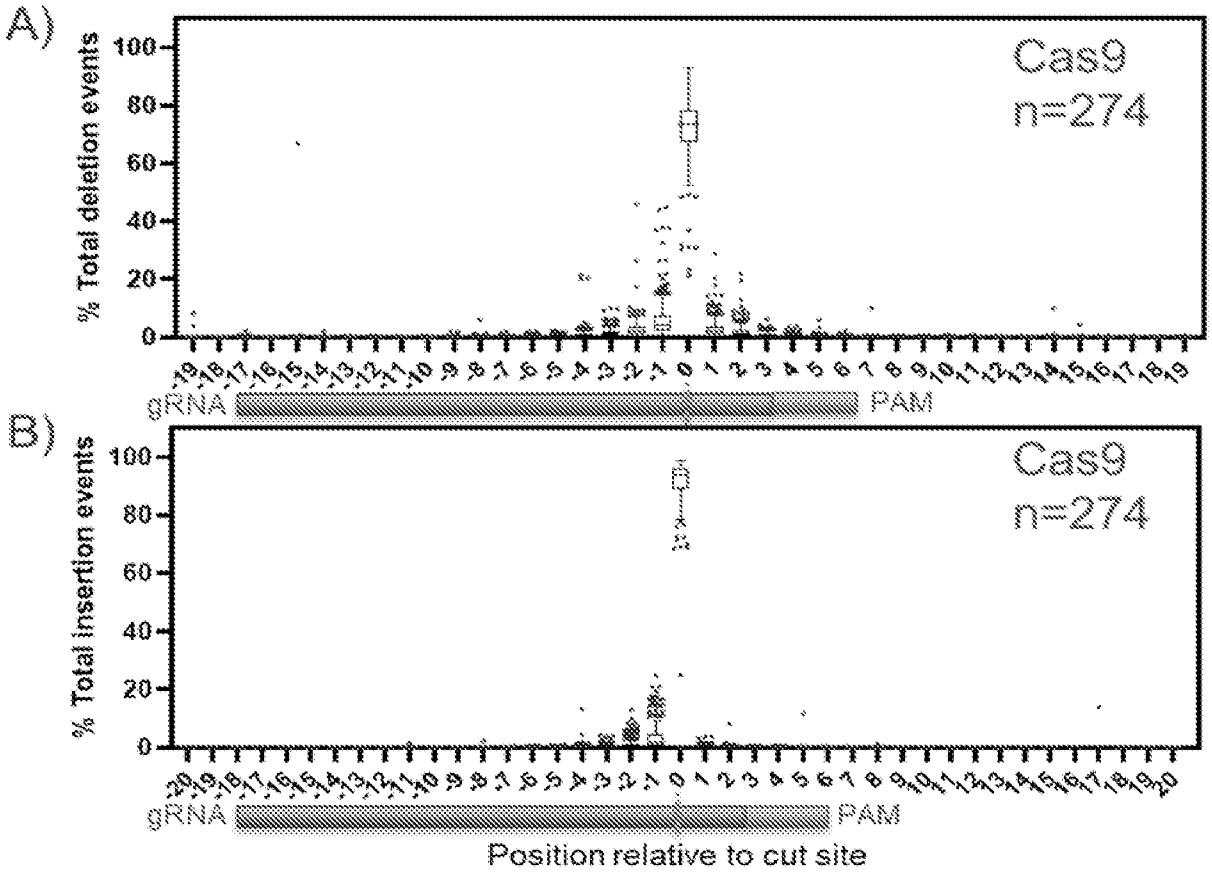

A)
% Total deletion events
Cas9
n=274
gRNA
PAM
B)
% Total insertion events
Cas9
n=274
gRNA
PAM
Position relative to cut site

FIG. 6

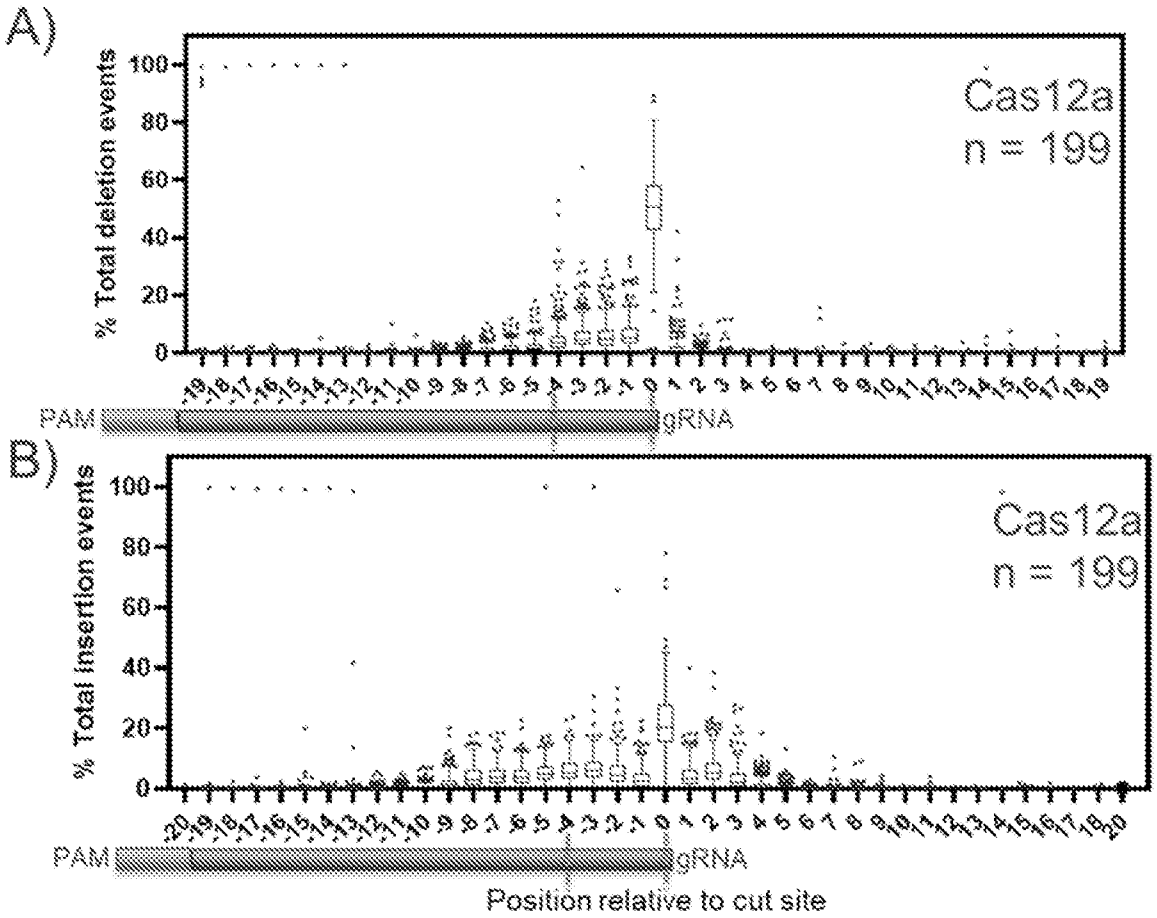

A)
% Total deletion events
Cas12a
n = 199
PAM
gRNA
B)
% Total insertion events
Cas12a
n = 199
PAM
gRNA
Position relative to cut site

FIG. 7

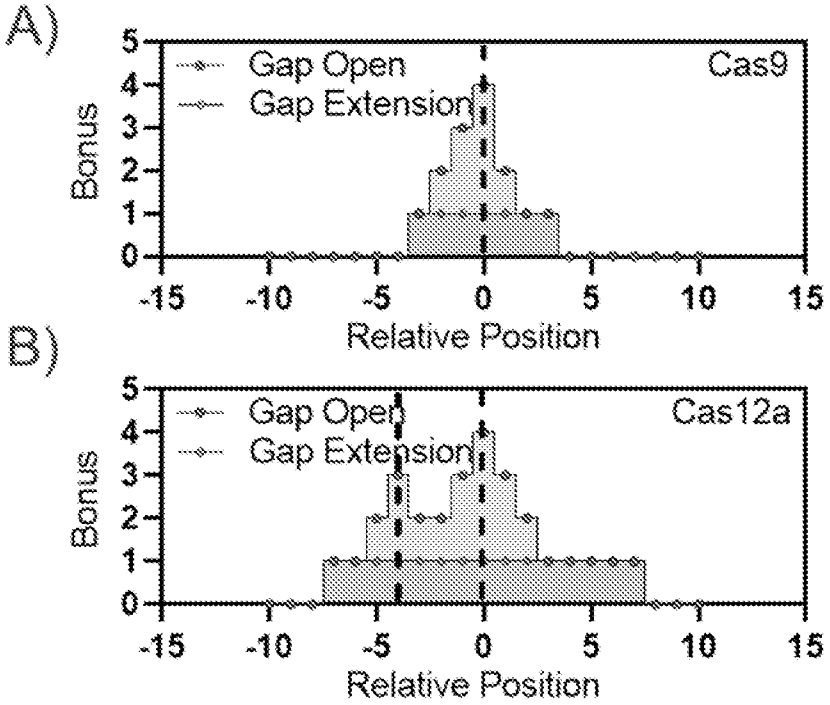

A)
Gap Open
Gap Extension
Cas9
Bonus
Relative Position
B)
Gap Open
Gap Extension
Cas12a
Bonus
Relative Position

FIG. 8

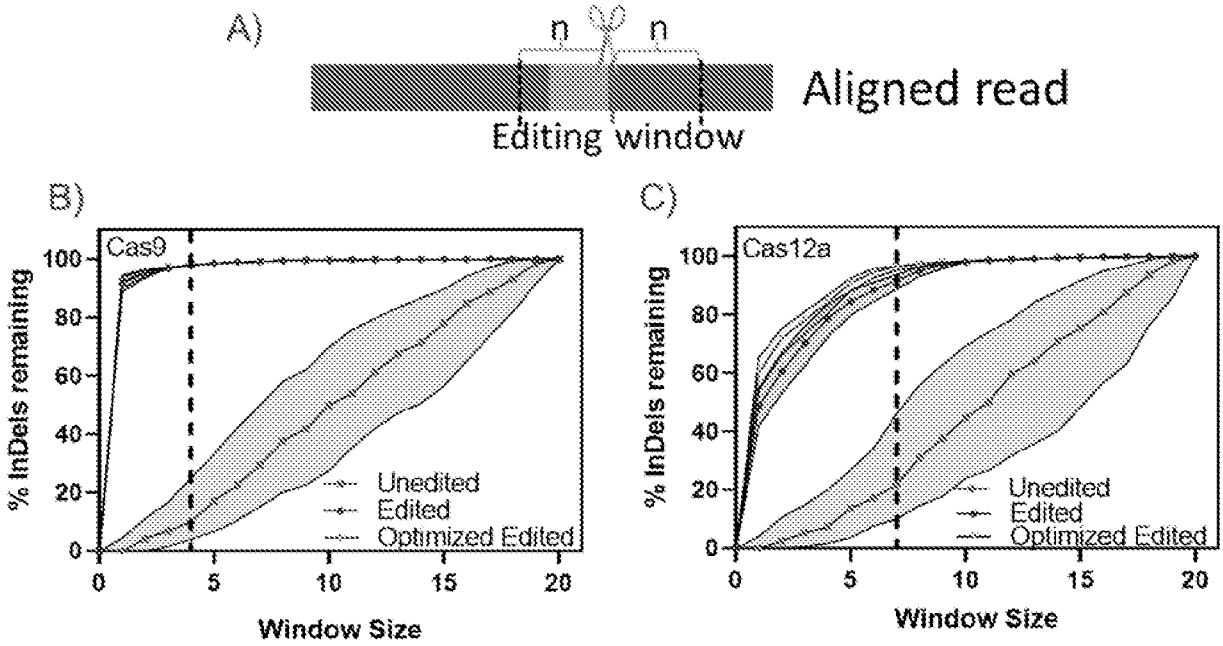

A)
n
n
Aligned read
Editing window
B)
Cas9
% InDels remaining
Unedited
Edited
Optimized Edited
Window Size
C)
Cas12a
% InDels remaining
Unedited
Edited
Optimized Edited
Window Size

FIG. 9

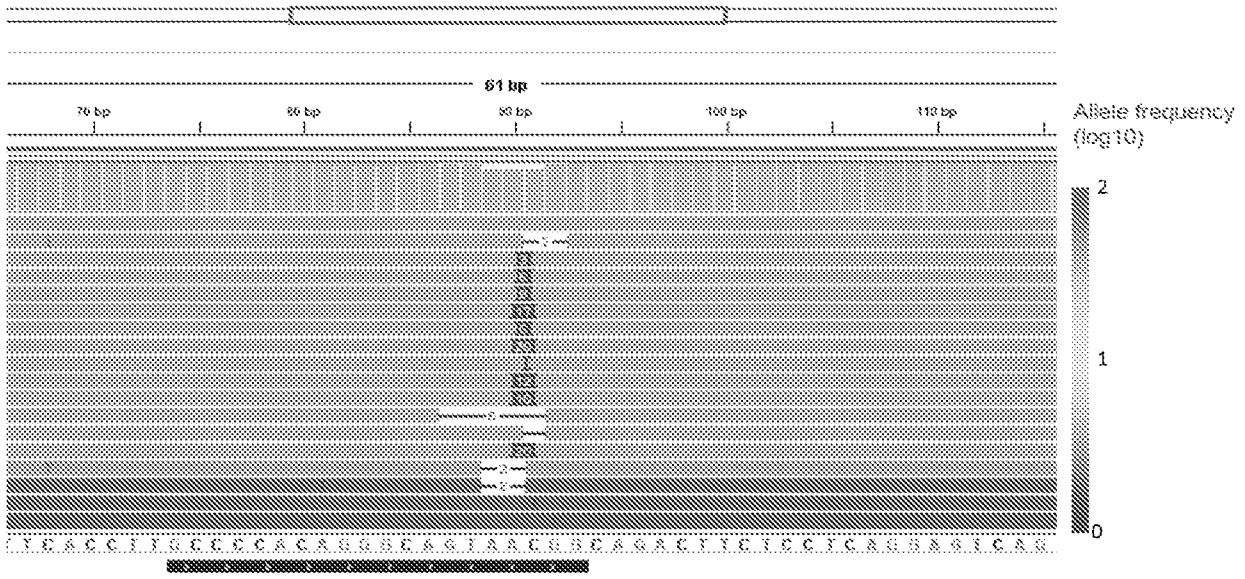

61 bp
70 bp
80 bp
90 bp
100 bp
110 bp
Allele frequency
(log10)
2
1
0

FIG. 10

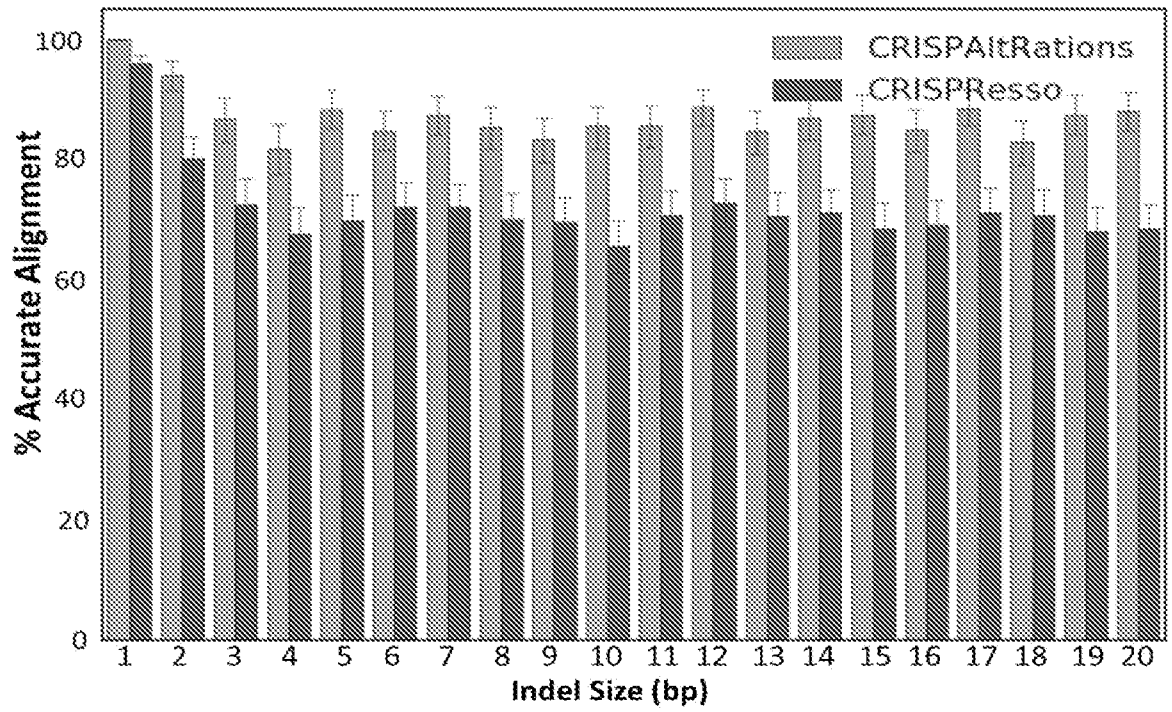

CRISPAltRations
CRISPResso
% Accurate Alignment
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Indel Size (bp)

FIG. 11

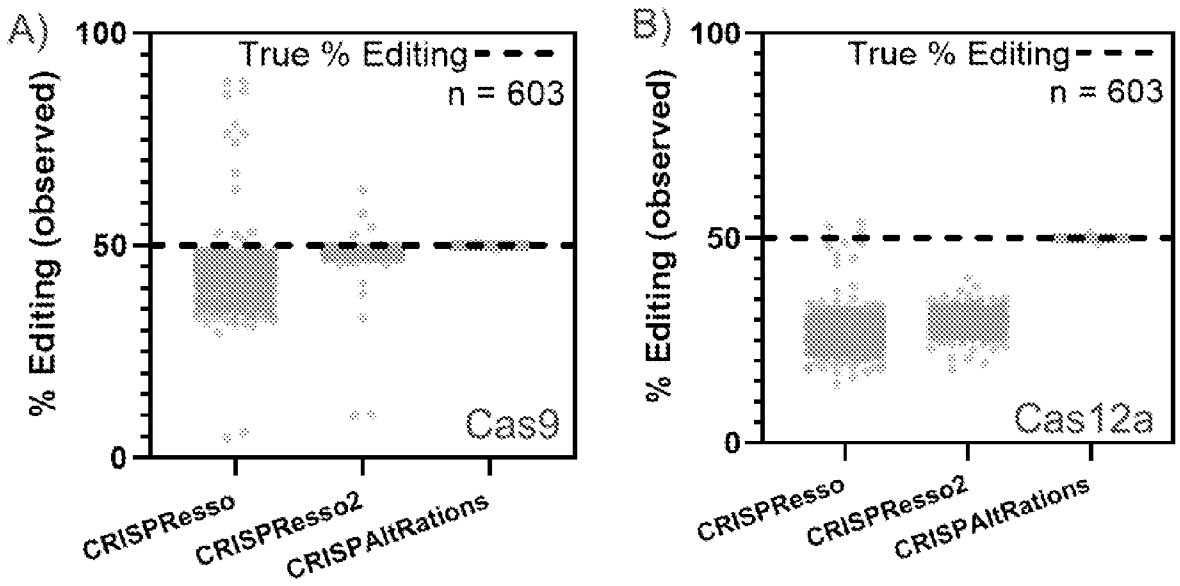
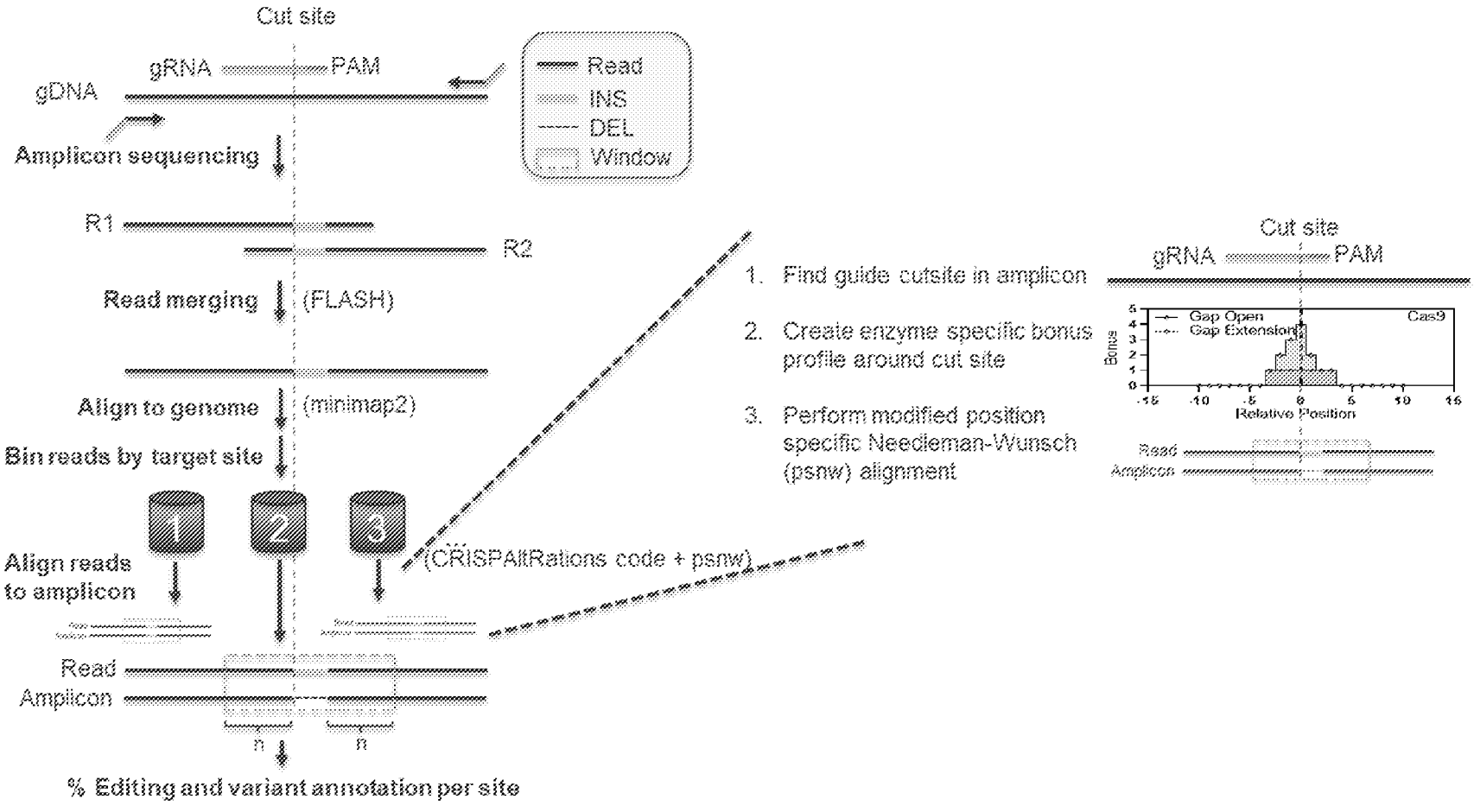

A)
100
50
0
% Editing (observed)
True % Editing
n = 603
Cas9
CRISPResso
CRISPResso2
CRISPAltRations
B)
100
50
0
% Editing (observed)
True % Editing
n = 603
Cas12a
CRISPResso
CRISPResso2
CRISPAltRations

FIG. 1


Cut site
gRNA
PAM
gDNA
Amplicon sequencing
Read
INS
DEL
Window
R1
R2
Read merging
(FLASH)
Align to genome
(minimap2)
Bin reads by target site
1
2
3
Align reads to amplicon
(CRISPAltRations code + psnw)
Read
Amplicon
n
n
% Editing and variant annotation per site
1. Find guide cutsite in amplicon
2. Create enzyme specific bonus profile around cut site
3. Perform modified position specific Needleman-Wunsch (psnw) alignment
Cut site
gRNA
PAM
Gap Open
Gap Extension
Cas9
Bonus
Relative Position
Read
Amplicon