Luca Pinello, Matthew C. Canver, Megan D. Hoban, Stuart H. Orkin, Donald B. Kohn, Daniel E. Bauer*, Guo-Cheng Yuan*

CRISPResso: sequencing analysis toolbox for CRISPR-Cas9 genome editing

Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute and Harvard School of Public Health, Boston, Massachusetts 02115, USA
Luca Pinello, Guo-Cheng Yuan

Division of Hematology/Oncology, Boston Children's Hospital, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA
Matthew C. Canver, Stuart H. Orkin, Daniel E. Bauer

Department of Microbiology, Immunology, and Molecular Genetics, Eli & Edythe Broad Center of Regenerative Medicine & Stem Cell Research, University of California, Los Angeles, Los Angeles, California, 90095, USA
Megan D. Hoban, Donald B. Kohn

Howard Hughes Medical Institute, Boston Massachusetts 02115, USA
Stuart H. Orkin

*Correspondence: daniel.bauer@childrens.harvard.edu, gcyuan@jimmy.harvard.edu

To the Editor:

Recent progress in genome editing technologies, in particular the CRISPR-Cas9 nuclease system, has provided new opportunities to investigate the biological functions of genomic sequences by targeted mutagenesis [1-4]. Briefly, Cas9 may be directed by a chimeric single guide RNA (sgRNA) to a target genomic sequence upstream of a protospacer adjacent motif (PAM) for cleavage. Double strand breaks (DSBs) resulting from site-specific Cas9 cleavage can be resolved by endogenous DNA repair pathways such as non-homologous end joining (NHEJ) or homology-directed repair (HDR). These repair mechanisms result in a spectrum of diverse outcomes including insertions, deletions, nucleotide substitutions, and, in the case of HDR, recombination of extrachromosomal donor sequences [1-3, 5,6]. Deep sequencing of amplified genomic regions or whole genome sequencing (WGS) allows for quantitative and sensitive detection of targeted mutations. However, to date no standard analytic tool has been developed to systematically enumerate and visualize these events, resulting in inconsistencies between different experiments and across laboratories. Challenging issues for the interpretation of CRISPR-Cas9 edited sequences include amplification or sequencing errors, experimental variation in sequence quality, ambiguous alignment of variable length indels, deconvoluting mixed HDR/NHEJ outcomes, analyzing large datasets from WGS experiments and analyzing pooled experiments where many different target sites are present in a single sequencing library. To solve these issues with an aim to standardize data analysis, we developed CRISPResso as a robust and easy-to-use computational pipeline (**Supplementary Note 1** and **Supplementary Fig. 1**). CRISPResso enables accurate quantification and visualization of CRISPR-Cas9 outcomes, as well as comprehensive evaluation of effects on coding sequences, noncoding elements and selected off-target sites.

CRISPResso is a suite of computational tools that provides an integrated, user-friendly interface that can be operated by biologists and bioinformaticians alike (**Supplementary Fig. 2**). Compared to existing tools [7], CRISPResso offers several novel features, including: batch sample analysis via command line interface, integration with other pipelines, tunable parameters of sequence quality and alignment fidelity, discrete measurement of insertions, deletions, and nucleotide substitutions (ignored by other methods), tunable windows around the cleavage site to minimize false positive classification, quantification of frameshift versus in-frame coding mutations, and distinction between NHEJ, HDR, and mixed mutation events. In addition, CRISPResso accommodates single or pooled amplicon deep sequencing and WGS datasets. CRISPResso automates the following steps: 1. filtering low quality reads, 2. trimming adapters, 3. aligning the reads to a reference amplicon, 4. quantifying the proportion of HDR and NHEJ outcomes, and 5. determining the proportion of frameshift and in-frame mutations as well as detecting potential splice site mutations. A graphical report is generated to visualize mutagenesis profiles (**Fig. 1, Supplementary Figs. 3-5**), and plain text output files are also produced for further integrative analyses (**Supplementary Note 2**). This pipeline can be used for assessment of on-target editing efficacy as well as of off-target editing at selected loci [8, 9].

CRISPResso qualitatively and quantitatively evaluates the outcomes of genome editing experiments in which target loci are subject to deep sequencing. We initially assessed the performance and limitations of CRISPResso, performing simulations with various genome editing outcomes with and without sequencing errors included (**Supplementary Note 3**, **Supplementary Figs. 6-8**). We found that CRISPResso, even in the presence of sequencing errors, robustly and accurately recovers editing events with a negligible false positive rate (<0.2%). Then we applied CRISPResso to actual paired end deep sequencing data from cells expressing Cas9 and sgRNA-1 targeted to a coding sequence with the intent to create gene knockout by frameshift mutations (experiment 1) or cells expressing Cas9, an extrachromosomal homologous donor template, and sgRNA-2 (experiment 2) or sgRNA-1 (experiment 3) with the intent of targeted introduction of four nucleotide substitutions (**Supplementary Note 4** and **Supplementary Figs. 9, 3-5**). For experiment 1, CRISPResso provides a quantification of the proportion of NHEJ occurrences, mutated allele size distribution and precise mutation localization with respect to the reference amplicon (**Fig. 1a-c**). When coding sequences are provided as an optional input, the software quantifies frameshift and in-frame mutations as well as predicts splice site mutations (**Fig. 1d, Supplementary Fig. 10**). When an expected HDR amplicon sequence is provided (experiments 2 and 3), CRISPResso is able to deconvolve and characterize unmodified, NHEJ-modified, and HDR-modified alleles as distinct outcomes (**Fig. 1e, f, Supplementary Note 1**, **and Supplementary Figs. 3-5**). In addition, it identifies mixed alleles that may result from sequential cleavages initially resulting in HDR and later NHEJ repair (**Fig. 1f**). In a case when the donor sequence disrupts the guide RNA seed sequence or PAM, the relative fraction of mixed events appears substantially reduced, consistent with the effect of these HDR alleles on resisting subsequent cleavage (**Supplementary Fig. 11,12**). By specifying the sequence identity required to classify an event as HDR, the user can control the specificity of HDR and sensitivity of mixed HDR-NHEJ allele detection (**Supplementary Fig. 8,9**). CRISPResso can be run either as a stand-alone command line utility (**http://github.com/lucapinello/CRISPResso**) or web application (**www.crispresso.rocks**, **Supplementary Note 2**).

The CRISPResso suite provides two additional utilities: (1) CRISPREssoPooled, a tool for the analysis of pooled amplicon experiments, that first preprocesses the input data to highlight and remove PCR amplification or trimming artifacts. In this mode, sequences may be first aligned to a reference genome and then those mapping to unexpected genomic locations removed, which may help resolve alignment artifacts (such as amplification of pseudogenes). Alternatively, all the regions discovered in the alignment to the reference genome can be considered. Then a separate CRISPResso report is created for each target region, with detailed mapping statistics (**Supplementary Note 5**, **Supplementary Fig 13**. and **Supplementary Table 1-2**). (2) CRISPRessoWGS, a tool for the analysis of WGS data, that provides detailed CRISPResso reports for any set of sites throughout the genome (for example, potential off-target sites) and separate .bam files (for discrete visualization in a genome browser) (**Supplementary Note 6** and **Supplementary Fig. 14**). The CRISPResso suite offers flexible and powerful tools to evaluate and quantitate genome editing outcomes from sequencing experiments, and for standardizing and streamlining analyses that currently require development of custom in-house algorithms.

References

1. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F: Multiplex genome engineering using CRISPR/Cas systems. Science 2013, 339:819–23.

2. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, Norville JE, Church GM: RNA-guided human genome engineering via Cas9. Science 2013, 339:823–6.

3. Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F: Genome engineering using the CRISPR-Cas9 system. Nat Protoc 2013, 8:2281–308.

4. Sander JD, Joung JK: CRISPR-Cas systems for editing, regulating and targeting genomes. Nat Biotechnol 2014.

5. Canver MC, Bauer DE, Dass A, Yien YY, Chung J, Masuda T, Maeda T, Paw BH, Orkin SH: Characterization of Genomic Deletion Efficiency Mediated by Clusted Regularly Interspaced Palindromic Repeats (CRISPR)/Cas9 Nuclease System in Mammalian Cells. J Biol Chem 2014, 289:21312–21324.

6. Ran FA, Hsu PD, Lin CY, Gootenberg JS, Konermann S, Trevino AE, Scott DA, Inoue A, Matoba S, Zhang Y, Zhang F: Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. Cell 2013, 154:1380–9.

7. Güell M, Yang L, Church GM: Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). Bioinformatics 2014, 30:1–3.

8. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, Cradick TJ, Marraffini LA, Bao G, Zhang F: DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol 2013, 31:827–32.

9. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD: High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol 2013, 31:822–826.

Figure Legends

Figure 1: **Quantification and visualization of NHEJ and HDR mutagenesis profiles. a-d**, An example of NHEJ-mediated disruption of a coding sequence by CRISPR-Cas9 (experiment 1). **a**, Quantification of editing frequency as determined by the percentage and number of sequence reads showing unmodified and modified alleles. When no donor sequence is provided, CRISPResso classifies any mutation as an NHEJ event**. b,** Frequency distribution of alleles with indels (shown in blue) and without indels (in red). Length-conserving substitutions are not classified as indels in this plot. In this example, the indels are dominated by small deletions, consistent with the anticipated CRISPR-Cas9 effect. **c,** NHEJ reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon. For insertions, the positions immediately adjacent to the insertion are indicated. In this example, the mutations cluster around the predicted cleavage position, consistent with the anticipated CRISPR-Cas9 effect. A low level of substitutions apparent throughout the amplicon suggests low-level technical error. **d,** Frameshift analysis of coding sequence reads affected by modifications. Frameshift and in-frame mutations include any mutations that partially or fully overlap coding sequences as input by the user, with any non-overlapping mutations classified as noncoding (see also Supplementary Fig. 6). **e-f**, An example of HDR-mediated recombination of an extrachromosomal donor sequence resulting in four substitutions relative to the reference amplicon (experiment 2)**. e,** When an expected HDR amplicon is provided, CRISPResso classifies sequence reads as HDR if they preferentially align to the expected HDR amplicon sequence and NHEJ (or unmodified) if they preferentially align to the reference amplicon. An alignment threshold may be provided to distinguish HDR alleles from those showing evidence of mixed HDR-NHEJ repair. **f**, Mapping of mutation position to reference amplicon of reads classified as NHEJ (left), HDR (center), and mixed HDR-NHEJ (right). In this example with the alignment threshold set to 100% sequence identity, the HDR alleles show only the four expected substitutions (see Supplementary Fig. 3) while the mixed HDR-NHEJ alleles show additional indels at the predicted cleavage position, consistent with sequential cleavages initially repaired by HDR and subsequently by NHEJ.

Figure 1

a
Unmodified
(308468 reads)
29.6%
70.4%
NHEJ
(734542 reads)
b
Indel size distribution
No indel
Indel
Sequences (no.)
400000
350000
300000
250000
200000
150000
100000
50000
−60
−40
−20
0
20
40
60
80
Indel size (bp)
c
Mutation position distribution
Insertions
Deletions
Substitutions
Predicted cleavage position
Sequences (no.)
514996
411996
308997
205998
102999
50%
40%
30%
20%
10%
Sequences (% total)
0
25
50
75
100
125
150
Reference amplicon position (bp)
d
Frameshift mutation
(569959)
77.3%
2.7%
Noncoding mutation
(19676 reads)
20.0%
In-frame mutation
(147160 reads)
Amplicon sequence
Coding sequence/s
Predicted Cas9 cleavage site/s
e
Unmodified
(813570 reads)
77.1%
2.1%
HDR
(22515 reads)
17.2%
NHEJ
(181738 reads)
3.5%
Mixed HDR-NHEJ
(37201 reads)
f
Mutation position distribution of NHEJ
Sequences (no.)
47842
38273
28705
19136
9568
4.5%
3.6%
2.7%
1.8%
0.9%
Sequences (% total)
0
25
50
75
100
125
150
Reference amplicon position (bp)
Mutation position distribution of HDR
Sequences (no.)
22515
18012
13509
9006
4503
2.1%
1.7%
1.3%
0.9%
0.4%
Sequences (% total)
0
25
50
75
100
125
150
Reference amplicon position (bp)
Mutation position distribution of mixed HDR-NHEJ
Sequences (no.)
37162
29729
22297
14864
7432
3.5%
2.8%
2.1%
1.4%
0.7%
Sequences (% total)
0
25
50
75
100
125
150
Reference amplicon position (bp)

Supplementary Material

Figure Legends

Supplementary Figure 1: Overview of the CRISPResso pipeline.

Supplementary Figure 2: The front page of the CRISPResso web application user interface. Required input files include a FASTQ file (or two FASTQ files for paired end reads) containing the raw reads and the reference amplicon sequence. Additional information may be provided as optional input to customize the analysis (details are described in Supplementary Note 3 and on the online manual available at: www.crispresso.rocks/help).

Supplementary Figure 3: Screenshot of the graphic report obtained from the web version of CRISPResso for experiment 1. Here the intent is to create gene knockout by frameshift mutations. The targeted locus was amplified and the amplicon subjected to paired end deep sequencing. CRISPResso was run with default parameters, except for the average bp quality filter parameter that was set to 30 on the Phred33 scale.

Supplementary Figure 4: Screenshot of the graphic report obtained from the web version of CRISPResso for experiment 2. Here the intent is to induce four substitutions to the target locus by HDR, introducing an extrachromosomal homologous donor sequence to the cells. The targeted locus was amplified and the amplicon subjected to paired end deep sequencing. CRISPResso was run with default parameters, except for the average bp quality filter parameter that was set to 30 on the Phred33 scale. Note that the frameshift analysis includes all modified alleles so the rate of in-frame alleles will appear high in experiments such as this one in which the rate of HDR is high and the HDR allele remains in-frame.

Supplementary Figure 5: Screenshot of the graphic report obtained from the web version of CRISPResso for experiment 3. The targeted locus was amplified and the amplicon subjected to paired end deep sequencing. CRISPResso was run with default parameters, except for the average bp quality filter parameter that was set to 30 on the Phred33 scale. In this case, mixed HDR-NHEJ alleles appear to be reduced relative to HDR alleles consistent with the HDR donor being relatively resistant to recognition by sgRNA-1.

Supplementary Figure 6: Evaluation of CRISPResso pipeline on synthetic datasets. **a,** Quantification of editing efficiency for unmodified and modified reads without error. **b,** Same experiment with sequencing errors simulated from the Illumina Miseq platform. **c,** Same experiment with sequencing errors simulated from the Illumina Miseq platform with a filtering window of 10 bp around the cleavage site (5 bp on each side).

Supplementary Figure 7: Evaluation of CRISPResso pipeline on synthetic datasets. Same experiments as in Supplementary Figure 3. **a-c,** Location and frequency of insertions, deletions and substitutions on the reference amplicon.

Supplementary Figure 8: Evaluation of CRISPResso pipeline on synthetic datasets. Same experiments as in Supplementary Figure 3. **a-c,** Position and average length of insertions and deletions.

Supplementary Figure 9: Schematic of *HBB* locus (for which modifications are depicted in Fig. 1), with positions of sgRNAs, genomic cleavage, primers, exons, and substitutions in HDR donor sequence indicated. **a**, Reference amplicon sequence (positions 1-158) with sites of sequence complementarity (lines) and anticipated cleavage sites (arrowheads) for sgRNAs indicated. sgRNA-2 (red line) results in cleavage in *HBB* 5'UTR whereas sgRNA-1 (green line) results in cleavage of exon 1 coding sequence. Primers for amplification in blue. **b**, Zoom showing the genomic sequence of the cells used in this experiment (corresponding to hg19 chr11:5248206-5248248, minus strand) with the cleavage site of sgRNA-1 indicated by arrowhead and sites of substitution within the HDR donor sequence indicated by larger font. **c**, Schematic of expected HDR amplicon sequence. **d**, Zoom showing the four substitutions (larger font in red) relative to the genomic sequence. Positions listed are relative to the reference amplicon input to CRISPResso.

Supplementary Figure 10: Frameshift analysis for experiment 1. **a**, Reads with frameshift and in-frame mutations are classified based on any mutation that partially or fully overlaps coding sequence, with mutations that do not overlap coding sequences classified as noncoding. Mutations affecting coding sequence are classified as in-frame or frameshift according to how they alter the final coding sequence length. **b**, Frameshift and in-frame mutagenesis profiles indicating position affected by modification. **c**, Reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position exclusively in noncoding region/s (that is, without mutations affecting coding sequences). The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted. In an experiment in which Cas9 cleavage is directed to coding sequence, noncoding mutations would be expected to be infrequent and these may reflect amplification or sequencing errors rather than true mutations. This plot directly illustrates the position of these mutations which could help identify any recurrent mutations. One approach to minimize the impact of sequencing errors would be to adjust the CRISPResso parameters for read quality or to utilize a window to search for mutations surrounding the cleavage site. Another, highly recommended, approach would be to empirically estimate the rate of sequence errors by preparing a library with a non-edited control sample sequenced in parallel.

Supplementary Figure 11: Effect of variation of the HDR sequence identity parameter on experiment 2. A read that better aligns to the expected HDR amplicon as compared to the reference amplicon is classified as either an HDR or mixed HDR-NHEJ allele. If the read has identity to the expected HDR amplicon sequence exceeding the HDR identity parameter, it is classified as an HDR allele; if the identity is below the HDR identity parameter, it is classified as a mixed HDR-NHEJ allele. Increasing the sequence identity value reduces tolerance of mismatches, with increased numbers of reads quantified as mixed HDR-NHEJ and a reduction in the number of reads quantified as HDR. Tuning the sequence identity parameter has no effect on NHEJ quantification. The default value for this parameter is 98%. Setting the HDR identity parameter to 98% results in a substantial number of reads classified as HDR showing small indels precisely at the cleavage site, consistent with sequential cleavages with HDR followed by NHEJ repair. Increasing the HDR identity parameter to 100% increases the specificity of HDR allele classification and sensitivity of mixed HDR-NHEJ allele classification. The tool allows the user to easily visualize results as the parameter is adjusted.

Supplementary Figure 12: Effect of variation of the HDR sequence identity parameter on experiment 3. Setting the identity parameter to 98% results in classification of most alleles as HDR with only a small fraction of mixed HDR-NHEJ alleles. By increasing the identity parameter to 100%, the rate of false-positive mixed HDR-NHEJ alleles and false negative HDR alleles might increase due to technical errors of sequencing. The tool allows the user to easily visualize results as the parameter is adjusted.

Supplementary Figure 13: Results from CRISPRessoPooled using data from ref [8], using the Mixed mode (Amplicons + Genome). Quantification and localization of insertion, deletion and substitution on a set of 11 unique amplicons from a pooled amplicon experiment.

Supplementary Figure 14: Results from CRISPRessoWGS using data from ref [11]. Analysis of a region of interest from WGS data. **a,** Mutation position distribution (left panel) and average length of insertion and deletion (center and right panel) for a control experiment in which Cas9/sgRNA were not delivered to the cells. In this experiment the cells are known to have a pre-existing homozygous 1 bp deletion relative to the reference genome. The vertical dotted line represents the predicted cleavage site. **b,** Mutation position distribution (left panel) and average length of insertion and deletion (center and right panel) for a companion experiment in which Cas9/sgRNA were utilized. 50% of the alleles show one 1 bp insertion and 50% of the alleles show a different 1 bp insertion consistent with an edited clone that has compound heterozygous NHEJ repaired alleles. The vertical dotted line represents the predicted cleavage site. **c.** IGV genome browser screenshot of the reads extracted by CRISPRessoWGS for the analysis and aligned to the mm9 reference genome.

Supplementary note 1 - Details of the CRISPResso pipeline

Quality filtering

In order to reliably quantify mutation events and avoid false positives due to sequencing errors, it may be important to take into account read quality. A common standard for quality is the Phred 33 scale that represents scores proportional to the base-calling error probabilities [1]. CRISPResso allows filtering the reads based on the average quality score, defined as the average score of each single bp quality score. Additionally, it is possible to filter reads excluding any read for which a single bp is below a predefined threshold. Suggested values for these parameters, on the Phred33 scale, are: 30 for the average read quality and 20 for the single bp quality. If not explicitly specified by the user, the reads are not filtered.

Trimming

Although some sequencing facilities provide sequence reads already trimmed of adapter sequences, this is not always the case. Since the trimming step is a prerequisite to reliably quantify modifications, CRISPResso provides the option to trim sequences if the user provides the adapter sequences used (common ones for Illumina sequencing machines are Nextera and TrueSeq 2/3). Internally, the reads are trimmed using *Trimmomatic* [2]. A custom adapter for particular experimental designs can also be provided.

Merging paired end reads

When paired end reads are provided, it is possible to “merge” the reads since usually the amplicon sequence provided is shorter than twice the read length. This step is performed in order to increase the confidence of the region sequenced in the overlapping region. Internally CRISPResso uses the *FLASH* method [3].

Alignment

To align the filtered reads to the reference amplicon CRISPResso uses *needle* from the *EMBOSS* suite [4], an optimal global sequence aligner that can easily account for gaps. The main advantage of this aligner is that it is a strict implementation of the Needleman-Wunsch algorithm instead of simple heuristics, providing predictable and accurate results. In addition, its rich output can be easily parsed and integrated in other pipelines.

Quantification of mutation

After alignment to quantify mutation occurrences, the sequence identity scores of both the reference amplicon and the expected HDR amplicon (in case a donor sequence was used in the experiment) are considered. If the read aligns better to the HDR expected amplicon, the read is classified either as HDR or mixed HDR-NHEJ. If the sequence identity is 100% (or above a user-defined threshold, default 98%) the read is classified as HDR; if sub-threshold alignment, the read is classified as mixed HDR-NHEJ. The reads that align better to the reference amplicon than the expected HDR amplicon are considered unmodified if they have 100% sequence identity with it, otherwise the reads are classified as NHEJ. Mutation spectra of each (total, HDR, mixed HDR-NHEJ, and NHEJ) are then plotted along with expected cleavage position to help the user interpret allelic outcomes. To make the process efficient, some pre-compiled regular expressions are built on the special characters obtained from the alignment step (for insertion, deletion or nucleotide substitution) with respect to the reference amplicon. Subsequently, for a fast localization of each event in the reference amplicon, a lookup table for the positions is used. This allows an efficient quantification of the occurrences with a low memory footprint. To build the average profiles, occurrences for all the reads are summed and then divided by the total number of reads mapped. In the case of the combined profile, for each position any of the possible mutation types (insertion, deletion or nucleotide substitution) are counted.

Frameshift quantification

To quantify frameshift and in-frame occurrences, CRISPResso determines if reads with mutations inside the coding sequences alter the total coding sequence length. Alterations that conserve the length or change the length by a multiple of three are considered as in-frame while the rest are considered frameshift. Mutations that are only outside the coding sequence but leave coding sequence intact are reported as noncoding mutations. If an sgRNA is designed to direct Cas9 cleavage within coding sequences, the proportion of noncoding sequence mutations would be expected to be very low. CRISPResso generates a plot specifically to show the position and distribution of noncoding mutations as well as the indel size distribution of frameshift and in-frame mutations. An example is shown in Supplementary Fig. 10.

Potential splice site quantification

To quantify potential splice sites, CRISPResso determines for each read the presence of mutations at either of the two intronic positions adjacent to the exon boundary. This analysis is performed on all the reads, and is independent of the frameshift quantification step.

Web application

The web application was built using several open source projects. The server component is written in Flask (http://flask.pocoo.org/link), a Python framework library for the rapid prototyping of web applications and served using the Apache webserver (http://httpd.apache.org). In order to make the machine scalable for many users, requests and data from users are asynchronously processed by a messaging queue system (RabbitMQ, https://www.rabbitmq.com/link)) that calls the CRISPResso command line utility. The front end component was built using HTML5, Javascript and the Twitter Bootstrap framework library (http://getbootstrap.com/ink) for the user interface elements.

Code availability

CRISPResso source code is freely available at: https://github.com/lucapinello/CRISPResso/ and is released under the GNU Affero General Public License v3.

Supplementary Note 2 - Installation and usage of CRISPResso

Installation

To use the web version of CRISPResso, no installation is required. Please open the website: http://crispresso.rocks/. The documentation of the website is available online here: http://crispresso.rocks/help.

To install the command line version of CRISPResso, some dependencies need to be installed before running the setup:

1. Python 2.7 Anaconda: http://continuum.io/downloads
2. Java: http://java.com/download

3. C compiler / make. For Mac with OSX 10.7 or greater, open the terminal app and type and execute the command `make`, which will trigger the installation of OSX developer tools. Windows systems are not officially supported, although CRISPResso may work with Cygwin (https://www.cygwin.com/).

After checking that the all required software is installed, install CRISPResso from the official Python repository following these steps:

1. Open a terminal window
2. Type the command:

```
pip install CRISPResso --verbose
```

3. Close the terminal window and open a new one (this is important in order to setup correctly the `PATH` variable in the system).

Alternatively, to install the package without the Python pip utility:

1. Download the setup file: https://github.com/lucapinello/CRISPResso/archive/master.zip and decompress it
2. Open a terminal window and go to the folder where the zip file has been decompressed
3. Type the command: `python setup.py install`
4. Close the terminal window and open a new one (this is important in order to setup correctly the `PATH` variable in the system).

The setup will try to install the following software:

1. Trimmomatic (tested with v0.33): http://www.usadellab.org/cms/?page=trimmomatic
2. Flash (tested with v1.2.11): http://ccb.jhu.edu/software/FLASH/
3. Needle from the EMBOSS suite (tested with 6.6.0): ftp://emboss.open-bio.org/pub/EMBOSS/

If this setup fails, may need to install the above software manually following the instructions for your specific machine and include these utilities/binary files in the path following the instructions provided in their documentation.

To check that the installation worked, open a terminal window and execute:

```
 CRISPResso –help
```

A help page should appear.

NOTE: The setup will create automatically a folder in the home folder called `CRISPresso_dependencies`. Do not delete this folder or CRISPResso will stop working. To put the folder in a different location, set the environment variable:

```
CRISPRESSO_DEPENDENCIES_FOLDER
```

For example to put the folder in `/home/lpinello/other_stuff` write in the terminal **BEFORE** the installation:

```
export CRISPRESSO_DEPENDENCIES_FOLDER=/home/lpinello/other_stuff
```

Usage

CRISPResso requires two inputs: (1) paired-end reads (two files) or single-end reads (single file) in .fastq format (fastq.gz files are also accepted) from a deep sequencing experiment and (2) a reference amplicon sequence to assess and quantify the efficiency of the targeted mutagenesis. The amplicon sequence expected after HDR can be provided as an optional input to assess HDR frequency. One or more sgRNA sequences (without PAM sequences) can be provided to compare the predicted cleavage position/s to the position of the observed mutations. Coding sequence/s may be provided to quantify frameshift and potential splice site mutations.

NHEJ quantification

The required inputs are:

- Two files for paired-end reads or a single file for single-end reads in fastq format (fastq.gz files are also accepted). The reads are assumed to be already trimmed for adapters. If reads are not trimmed, please use the --*trim_sequences* option and the --*trimmomatic_options_string* if using an adapter different than Nextera.
- The reference amplicon sequence must also be provided.

Example:

```
CRISPResso -r1 reads1.fastq.gz -r2 reads2.fastq.gz -a
GCTTACACTTGCTTCTGACACAACTGTGTTCACGAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGAATGCCGTCACC
ACCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGA
```

HDR quantification

The required inputs are:

- Two files for paired-end reads or a single file for single-end reads in fastq format (fastq.gz files are also accepted). The reads are assumed to be already trimmed for adapters.
- The amplicon sequence expected after HDR must also be provided.

Example:

```
CRISPResso -r1 reads1.fastq.gz -r2 reads2.fastq.gz -a
GCTTACACTTGCTTCTGACACAACTGTGTTCACGAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGAGGAGAAGAATGCCGTCACC
ACCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGA -e
GCTTACACTTGCTTCTGACACAACTGTGTTCACGAGCAACCTCAAACAGACACCATGGTGCATCTGACTCCTGTGGAAAAAAACGCCGTCACG
ACGTTATGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGTTGGTATCAAGGTTACAAGA
```

IMPORTANT: You must input the entire reference amplicon sequence (’Expected HDR Amplicon sequence’ is the reference for the sequenced amplicon, not simply the donor sequence). If only the donor sequence is provided, an error will result.

Troubleshooting

- It is important to check if reads are trimmed or not. CRISPResso assumes that the reads are already trimmed. If reads are not trimmed, use the option `--trim_sequences` (or mark the correct option in the webpage). The default adapter file used is the Nextera. To specify a custom adapter use the option `--trimmomatic_options_string`.
- It is possible to use CRISPResso with single end reads. In this case, omit the option -r2 for specifying the second fastq file.
- The command line CRISPResso tool requires OS 10.7 or greater on Mac computers. It also requires that command line tools are installed on the machine. After the installation of Anaconda, open the Terminal app and type `make`, this should prompt to install command line tools (requires internet connection). Once installed, simply typing CRISPResso into any new terminal should load CRISPResso (the CRISPResso cup image will appear)
- Paired end sequencing files requires overlapping sequence from the paired sequencing data (at least 1 bp).
- The CRISPResso default setting is to output analysis files into the directory, otherwise the --output parameter may be used.

Output format

Examples of the website reports are shown in Supplementary Figures 3-5. Full reports, with the complete output, are provided as supplementary files (Supplementary Files 1-3).

Inside the report it is possible to find .pdf and .png files for all the figures. In addition, the execution log and all the processed raw data used to generate the figures are produced.

- *CRISPResso_RUNNING_LOG.txt*: the log of the external utilities called;
- Mapping_statistics.txt: this file contains number of: reads in input, reads after preprocessing (merging or quality filtering) and reads properly aligned.
- *Quantification_of_editing_frequency.txt*: quantification of editing frequency (number of reads aligned, reads with NHEJ, reads with HDR, and reads with mixed HDR-NHEJ);
- *Frameshift_analysis.txt*: number of modified reads with frameshift, in-frame and noncoding mutations;
- *Splice_sites_analysis.txt*: number of reads corresponding to potential affected splicing sites;
- *effect_vector_combined.txt*: location of mutations (including deletions, insertions, and substitutions) with respect to the reference amplicon;
- *effect_vector_deletion.txt* : location of deletions;
- *effect_vector_insertion.txt*: location of insertions;
- effect_vector_substitution.txt: location of substitutions.
- position_dependent_vector_avg_insertion_size.txt: average length of the insertions for each position.
- position_dependent_vector_avg_deletion_size.txt: average length of the deletions for each position.

Understanding the parameters of CRISPResso

Required parameters

To run CRISPResso, only 2 parameters are required for single end reads, or 3 for paired end reads:

-r1 or --fastq_r1

This parameter allows for the specification of the first fastq file.

-r2 or --fastq_r2 FASTQ_R2

This parameter allows for the specification of the second fastq file for paired end reads.

-a or --amplicon_seq

This parameter allows the user to enter the amplicon sequence used for the experiment.

Optional parameters

In addition to the required parameters explained in the previous section, several optional parameters can be adjusted to tweak your analysis, and to ensure CRISPResso analyzes your data in the best possible way.

-g or --guide_seq

This parameter allows for the specification of the sgRNA sequence. If more than one sequence are included, please separate by comma/s. If the guide RNA sequence is entered, then the position of the guide RNA and the cleavage site will be indicated on the output analysis plots. Note that the sgRNA needs to be input as the guide RNA sequence (usually 20 nt) immediately 5' of the PAM sequence (usually NGG for SpCas9). If the PAM is found on the opposite strand with respect to the Amplicon Sequence, ensure the sgRNA sequence is also found on the opposite strand. The CRISPResso convention is to depict the expected cleavage position using the value of the parameter cleavage_offset nt 3' from the end of the guide. In addition, the use of alternate nucleases to SpCas9 is supported. For example, if using the Cpf1 system, enter the sequence (usually 20 nt) immediately 3' of the PAM sequence and explicitly set the cleavage_offset parameter to 1, since the default setting of -3 is suitable only for SpCas9. (default:None)

-e or --expected_hdr_amplicon_seq

This parameter allows for the specification of the amplicon sequence expected after HDR. If the data to be analyzed were derived from an experiment using a donor repair template for homology-directed repair (HDR for short), then you have the option to input the sequence of the expected HDR amplicon. This sequence is necessary for CRISPResso to be able to identify successful HDR events within the sequencing data.

--hdr_perfect_alignment_threshold

Sequence homology percentage for an HDR occurrence (default: 98.0). This parameter allows for the user to set a threshold for sequence homology for CRISPResso to count instances of successful HDR. This is useful to improve the analysis allowing some tolerance for technical artifacts present in the sequencing data such as sequencing errors or single nucleotide polymorphisms (SNPs) in the cells used in the experiment. Therefore, if you have a read that exhibits successful HDR but has a SNP or sequencing error within the amplicon, you can lower the sequence homology in order allow CRISPResso to count the read as a successful HDR event. If the data are completely free of sequencing errors or polymorphisms, then consider to set parameter to 100.

-d or -donor_seq

This parameter allows the user to highlight the critical subsequence of the expected HDR amplicon in plots. This parameter does not have any effect on the quantification of HDR events.

-c, --coding_seq

This parameter allows for the specification of the subsequence/s of the amplicon sequence covering one or more coding sequences for the frameshift analysis. If more than one (for example, split by intron/s), please separate by comma. (default: None)

-q, or --min_average_read_quality

This parameter allows for the specification of the minimum average quality score (phred33) to include a read for the analysis.(default: 0, minimum: 0, maximum: 40). This parameter is helpful to filter out low quality reads. If filtering based on average base quality is desired, a reasonable value for this parameter is greater than 30.

-s or --min_single_bp_quality

This parameter allows for the specification of the minimum single bp score (phred33) to include a read for the analysis (default: 0, minimum: 0, maximum: 40). This parameter is helpful to filter out low quality reads. This filtering is more aggressive, since any read with a single bp below the threshold will be discarded. If you want to filter your reads based on single base quality to have very high quality reads, a reasonable value for this parameter is greater than 20.

--min_identity_score

This parameter allows for the specification of the min identity score for the alignment (default: 60.0). In order for a read to be considered properly aligned, it should pass this threshold. We suggest to lower this threshold only if really large insertions or deletions are expected in the experiment (>40% of the amplicon length).

-n or --name

This parameter allows for the specification of the output name of the report (default: the names is obtained from the filename of the fastq file/s used in input).

-o or --output_folder

This parameter allows for the specification of the output folder to use for the analysis (default: current folder).

--trim_sequences

This parameter enables the trimming of Illumina adapters with Trimmomatic (default: False)

--trimmomatic_options_string

This parameter allows the user the ability to override options for Trimmomatic (default: ILLUMINACLIP:/Users/luca/anaconda/lib/python2.7/site-packages/CRISPResso-0.8.0-py2.7.egg/CRISPResso/data/NexteraPE-PE.fa:0:90:10:0:true). This parameter is useful to specify different adaptor sequences used in the experiment if you need to trim them.

--min_paired_end_reads_overlap

This parameter allows for the specification of the minimum required overlap length between two reads to provide a confident overlap during the merging step. (default: 4, minimum: 1, max: read length)

-w ,--window_around_sgrna

This parameter allows for the specification of a window(s) in bp around each sgRNA to quantify the indels. The window is centered on the predicted cleavage site specified by each sgRNA. Any indels not overlapping this window are excluded. A value of -1 will disable this filter. (default: -1). This parameter is important since sequencing artifacts and/or SNPs can lead to false positives or false negatives in the quantification of indels and HDR occurrences. Therefore, the user can choose to create a window around the predicted double strand break site of the nuclease used in the experiment. This can help limit sequencing or amplication errors or non-editing polymorphisms from being inappropriately quantified in CRISPResso analysis. Note: any indels that fully or partially overlap the window will be quantified.

--cleavage_offset

This parameter allows for the specification of the cleavage offset to use with respect to the provided sgRNA sequence. Remember that the sgRNA sequence must be entered without the PAM. The default is -3 and is suitable for the SpCas9 system. For alternate nucleases, other cleavage offsets may be appropriate, for example, if using Cpf1 set this parameter to 1. (default: -3, minimum:1, max: reference amplicon length).

--exclude_bp_from_left

Exclude bp from the left side of the amplicon sequence for the quantification of the indels (default: 5). This parameter is helpful to avoid artifacts due to imperfect trimming of the reads.

--exclude_bp_from_right

Exclude bp from the right side of the amplicon sequence for the quantification of the indels (default: 5). This parameter is helpful to avoid artifacts due to imperfect trimming of the reads.

--needle_options_string

This parameter allows the user to override options for the Needle aligner (default: -gapopen=10 -gapextend=0.5 -awidth3=5000). More information on the meaning of these parameters can be found in the needle documentation (http://embossgui.sourceforge.net/demo/manual/needle.html). We suggest that only experienced users modify these values.

--keep_intermediate

This parameter allows the user to keep all the intermediate files (default: False). We suggest keeping this parameter disabled for most applications, since the intermediate files (processed reads and alignments) can be really large.

--dump

This parameter allows to dump numpy arrays and pandas dataframes to file for debugging purposes (default: False).

--save_also_png

This parameter allows the user to also save.png images when creating the report., in addition to .pdf files.

Supplementary Note 3 - Assessment of CRISPResso performance and limitations using simulated sequencing dataset

To assess the performance and limitations of CRISPResso, we generated several simulated datasets with predefined editing efficiency and mutagenesis profiles. These datasets were designed to simulate paired-end reads (150 bp). For simplicity, we first assumed the sequencing step to have 100% accuracy. The datasets are:

1. 1000 unmodified reads
2. 1000 unmodified reads, 1000 reads with 1 substitution
3. 1000 unmodified reads, 1000 reads with 2 substitutions
4. 1000 unmodified reads, 1000 reads with 3 substitutions
5. 1000 unmodified reads, 1000 reads with an insertion of 5 bp
6. 1000 unmodified reads, 1000 reads with an insertion of 10 bp
7. 1000 unmodified reads, 1000 reads with an insertion of 50 bp
8. 1000 unmodified reads, 1000 reads with a deletion of 5 bp
9. 1000 unmodified reads, 1000 reads with a deletion of 10 bp
10. 1000 unmodified reads, 1000 reads with a deletion of 50 bp

As expected, CRISPResso detected no mutation for Dataset 1, while identifying an equal number of mutated and unmodified reads for all other datasets (Supplementary Fig. 6a). Closer examination indicates that CRISPResso perfectly recovered the correct indel size and number of substitutions (Supplementary Fig. 7a and Supplementary Fig. 8a).

Next, we modified each dataset by further taking into account possible sequencing errors. To this end, we applied ART [6], a widely-used software for simulating sequencing errors in different sequencing platforms. We selected the Illumina Miseq profile for analysis, since it is one of the most commonly used platforms for amplicon sequencing and assessment of genome editing modifications [6-8]. We find that the sequencing errors can generate up to 13% false positives in the quantification of mutations (Supplementary Fig.6b), including 6-7% for each end of a read. This is likely to be an overestimate, since the model does not reflect the latest technical improvements by Illumina in terms of chemistry and sequencing. In comparison, the impact of sequencing errors on the estimation of mutation locations and average sizes is significantly lower (Supplementary Fig.7b and Supplementary Fig.8b).

Our above analysis suggests the importance of including unedited controls in experiment design. Such controls are extremely useful for sequencing quality evaluation, as a safeguard for false positive errors. However, if such controls are unavailable and a single cleavage is expected, an alternative strategy is to count only events overlapping a small window around the cleavage site, where mutations are expected. We implemented this strategy on CRISPResso and repeated our experiment with a window of 10 bp around the cleavage site (5 bp on either side). This strategy greatly improves the quantification accuracy (Supplementary Figs. 6-8c), although a small number of false positive calls were still observed. We note the window size may be adjusted by the user (minimum 2 bp window with 1 bp on either side of the cleavage site). Taken together, these results strongly indicate that CRISPResso can recover the location and frequency of mutations with high accuracy even in presence of sequencing errors.

The reports, data files and code (IPython Notebook) used for this simulation are saved in Supplementary File 4 and accessible to the general public.

Supplementary Note 4 - Production of sequencing libraries

A subclone of K562 cells with genotype rs713040-A/A and rs334-A/A (i.e. modified to possess the sickle cell anemia mutation) were grown in RPMI (Cellgro) with 10% fetal bovine serum (Gemini Bio-Products) and 1x penicillin/streptomycin/L-glutamine (Gemini Bio-Products). CRISPR guides targeting the *HBB* locus were designed using the Optimized CRISPR Design Tool [9]. Oligonucleotides for the guide strands were annealed and ligated into the pX330 plasmid (Addgene) as described [5]. A homologous donor template containing about 1 kb of the *HBB* locus was cloned as described [10] into a TOPO vector (Invitrogen). Site directed mutagenesis was completed to introduce four specific changes: a G in the location of the sickle-associated SNP rs713040, a silent mutation two bases upstream of the sickle mutation to introduce an MspI restriction fragment length polymorphism (RFLP) site, the sickle cell mutation correction, and a silent mutation 22 bases downstream of the sickle mutation to introduce a HhaI RFLP site. The expected HDR reference amplicon included each of these four substitutions.

The CRISPR-Cas9 plasmid and donor template plasmid were delivered to K562 cells by nucleofection using the Amaxa 4D Nucleofector (Lonza) as per the manufacturer's instructions. Briefly, 2 x 10^5 K562 subclone cells per sample were spun at 90g for 10 min before being resuspended in 20 mcl of SF solution per sample. 20 mcl of cells was mixed with 500 ng of Cas9-sgRNA encoding plasmid and 1500 ng of donor template plasmid before being added to the cuvette and pulsed with the FF-120 program as described by the manufacturer. Cells were rested at room temperature for 10 min after which 80 mcl of RPMI supplemented with 10% FBS and penicillin/streptomycin/L-glutamine was added to each cuvette before being transferred to a 24 well plate containing 400 mcl of medium. Three days following nucleofection, cells were harvested and genomic DNA was extracted for downstream analyses.

For library preparation, an initial PCR was completed outside of the donor template region using primers as described [6]. An inner PCR was completed using primers HBB Fwd 5' – ACACGACGCTCTTCCGATCTNNNNGGCAGAGCCATCTATTGCTT – 3' and HBB Rev 5' – GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGTCTCCTTAAACCTGTCTTG – 3' to amplify the specific 198 bp region of *HBB* surrounding the sickle mutation along with primers to add Illumina adapters and indexes (Supplementary Figure 9). Samples were quantified, pooled, purified, and the final library was quantified and run on an Illumina HiSeq machine as paired end 150. The amplicon corresponds to the region hg19 chr11:5,248,322-5,248,123 inclusive of the primers described above. The K562 cells contained the hg19 reference genotype with the exception of the rs334-A/A per above. Primer sequences were trimmed before being run in CRISPResso such that the reference amplicon begins at position 1 immediately following the primer (Supplementary Figure 9).

Supplementary Note 5 - Installation and usage of CRISPRessoPooled

CRISPRessoPooled is a utility to analyze and quantify targeted sequencing CRISPR/Cas9 experiments involving sequencing libraries with pooled amplicons. One common experimental strategy is to pool multiple amplicons (e.g. a single on-target site plus a set of potential off-target sites) into a single deep sequencing reaction. (Briefly, genomic DNA samples for pooled applications can be prepared by first amplifying the target regions for each gene/target of interest with typical amplicons of ~150-400 bp. In a second round of PCR, with minimized cycle numbers, barcode and adaptors are added. With optimization, these two rounds of PCR can be merged into a single reaction. These reactions are then quantified, normalized, pooled, and undergo quality control before being sequenced.) CRISPRessoPooled demultiplexes reads from multiple amplicons and runs the CRISPResso utility with appropriate reads for each amplicon separately.

Installation

CRISPRessoPooled is installed automatically during the installation of CRISPResso, but to use it two additional programs must be installed:

- samtools: http://samtools.sourceforge.net/
- bowtie2: http://bowtie-bio.sourceforge.net/bowtie2

To install these tools please refer to their documentation.

Usage

This tool can run in 3 different modes:

Amplicons mode: Given a set of amplicon sequences, in this mode the tool demultiplexes the reads, aligning each read to the amplicon with best alignment, and creates separate compressed FASTQ files, one for each amplicon. Reads that do not align to any amplicon are discarded. After this preprocessing, CRISPResso is run for each FASTQ file, and separated reports are generated, one for each amplicon.

To run the tool in this mode the user must provide:

1. Paired-end reads (two files) or single-end reads (single file) in FASTQ format (fastq.gz files are also accepted)

2. A description file containing the amplicon sequences used to enrich regions in the genome and some additional information. In particular, this file, is a tab delimited text file with up to 5 columns (first 2 columns required):
   - *AMPLICON_NAME*: an identifier for the amplicon (must be unique).
   - *AMPLICON_SEQUENCE*: amplicon sequence used in the design of the experiment.
   - *sgRNA_SEQUENCE (OPTIONAL)*: sgRNA sequence used for this amplicon without the PAM sequence. If not available, enter *NA.*
   - *EXPECTED_AMPLICON_AFTER_HDR (OPTIONAL)*: expected amplicon sequence in case of HDR. If more than one, separate by commas and not spaces. If not available, enter *NA.*
   - *CODING_SEQUENCE (OPTIONAL)*: Subsequence(s) of the amplicon corresponding to coding sequences. If more than one, separate by commas and not spaces. If not available, enter *NA.*

   A file in the right format should look like this:

   Site1 CACACTGTGGCCCCTGTGCCCAGCCC**TGG**GCTCTCTGTACATGAAGCAAC CCCTGTGCCCAGCCC
   NA NA
   Site2 GTCCTGGTTTTTGGTTTGGGAAATATAGTCATC NA
   GTCCTGGTTTTTGGTTTAAAAAAATATAGTCATC NA
   Site 3 TTTCTGGTTTTTGGTTTGGGAAATATAGTCATC NA NA GGAAATATA

   Note: no column titles should be entered. Also the colors here are used only for illustrative purposes and in a plain text file will be not be present and saved.

The user can easily create this file with any text editor or with spreadsheet software like Excel (Microsoft), Numbers (Apple) or Sheets (Google Docs) and then save it as tab delimited file.

Example:

```
CRISPRessoPooled -r1 SRR1046762_1.fastq.gz -r2 SRR1046762_2.fastq.gz -f AMPLICONS_FILE.txt --name
ONLY_AMPLICONS_SRR1046762 --gene_annotations gencode_v19.gz
```

The output of CRISPRessoPooled Amplicons mode consists of:

1. REPORT_READS_ALIGNED_TO_AMPLICONS.txt: this file contains the same information provided in the input description file, plus some additional columns:
   a. *Demultiplexed_fastq.gz_filename*: name of the files containing the raw reads for each amplicon.
   b. *n_reads*: number of reads recovered for each amplicon.
2. A set of fastq.gz files, one for each amplicon.
3. A set of folders, one for each amplicon, containing a full CRISPResso report.
4. *CRISPRessoPooled_RUNNING_LOG.txt*: execution log and messages for the external utilities called.

Genome mode: In this mode the tool aligns each read to the best location in the genome. Then potential amplicons are discovered looking for regions with enough reads (the default setting is to have at least 1000 reads, but the parameter can be adjusted with the option *--min_reads_to_use_region*). If a gene annotation file from UCSC is provided, the tool also reports the overlapping gene/s to the region. In this way it is possible to check if the amplified regions map to expected genomic locations and/or also to pseudogenes or other problematic regions. Finally CRISPResso is run in each region discovered.

To run the tool in this mode the user must provide:

1. Paired-end reads (two files) or single-end reads (single file) in FASTQ format (fastq.gz files are also accepted)
2. The full path of the reference genome in bowtie2 format (e.g. /homes/luca/genomes/human_hg19/hg19). Instructions on how to build a custom index or precomputed index for human and mouse genome assembly can be downloaded from the bowtie2 website: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml.
3. Optionally the full path of a gene annotations file from UCSC. The user can download this file from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start) selecting as table "knowGene",

as output format "all fields from selected table" and as file returned "gzip compressed". (e.g. like: homes/luca/genomes/human_hg19/gencode_v19.gz)

Example:

```
CRISPRessoPooled -r1 SRR1046762_1.fastq.gz -r2 SRR1046762_2.fastq.gz -x
/gcdata/gcproj/Luca/GENOMES/hg19/hg19 --name ONLY_GENOME_SRR1046762 --gene_annotations gencode_v19.gz
```

The output of CRISPRessoPooled Genome mode consists of:

1. REPORT_READS_ALIGNED_TO_GENOME_ONLY.txt: this file contains the list of all the regions discovered, one per line with the following information:
   - chr_id: chromosome of the region in the reference genome.
   - bpstart: start coordinate of the region in the reference genome.
   - bpend: end coordinate of the region in the reference genome.
   - fastq_file: location of the fastq.gz file containing the reads mapped to the region.
   - n_reads: number of reads mapped to the region.
   - sequence: the sequence, on the reference genome for the region.
2. MAPPED_REGIONS (folder): this folder contains all the fastq.gz files for the discovered regions.
3. A set of folders with the CRISPResso report on the regions with enough reads.
4. *CRISPRessoPooled_RUNNING_LOG.txt*: execution log and messages for the external utilities called.

This running mode is particular useful to check if there are mapping artifacts or contaminations in the library. In an optimal experiment, the list of the regions discovered should contain only the regions for which amplicons were designed.

Mixed mode (Amplicons + Genome): in this mode, the tool first aligns reads to the genome and, as in the **Genome mode**, discovers aligning regions with reads exceeding a tunable threshold. Next it will align the amplicon sequences to the reference genome and will use only the reads that match both the amplicon locations and the discovered genomic locations, excluding spurious reads coming from other regions, or reads not properly trimmed. Finally CRISPResso is run using each of the surviving regions.

To run the tool in this mode the user must provide:

- Paired-end reads (two files) or single-end reads (single file) in FASTQ format (fastq.gz files are also accepted)
- A description file containing the amplicon sequences used to enrich regions in the genome and some additional information (as described in the Amplicons mode section).
- The reference genome in bowtie2 format (as described in Genome mode section).
- Optionally the gene annotations from UCSC (as described in Genome mode section).

Example:

```
CRISPRessoPooled -r1 SRR1046762_1.fastq.gz -r2 SRR1046762_2.fastq.gz  -f AMPLICONS_FILE.txt  -x
/gcdata/gcproj/Luca/GENOMES/hg19/hg19 --name AMPLICONS_AND_GENOME_SRR1046762 --gene_annotations
gencode_v19.gz
```

The output of CRISPRessoPooled Mixed Amplicons + Genome mode consists of these files:

1. REPORT_READS_ALIGNED_TO_GENOME_AND_AMPLICONS.txt: this file contains the same information provided in the input description file, plus some additional columns:
   a. Amplicon_Specific_fastq.gz_filename: name of the file containing the raw reads recovered for the amplicon.
   b. *n_reads*: number of reads recovered for the amplicon.
   c. *Gene_overlapping:* gene/s overlapping the amplicon region.
   d. chr_id: chromosome of the amplicon in the reference genome.
   e. bpstart: start coordinate of the amplicon in the reference genome.
   f. bpend: end coordinate of the amplicon in the reference genome.
   g. Reference_Sequence: sequence in the reference genome for the region mapped for the amplicon.

2. MAPPED_REGIONS (folder): this folder contains all the fastq.gz files for the discovered regions.
3. A set of folders with the CRISPResso report on the amplicons with enough reads.
4. *CRISPRessoPooled_RUNNING_LOG.txt*: execution log and messages for the external utilities called.

The Mixed mode combines the benefits of the two previous running modes. In this mode it is possible to recover in an unbiased way all the genomic regions contained in the library, and hence discover contaminations or mapping artifacts. In addition, by knowing the location of the amplicon with respect to the reference genome, reads not properly trimmed or mapped to pseudogenes or other problematic regions will be automatically discarded, providing the cleanest set of reads to quantify the mutations in the target regions with CRISPResso.

If the focus of the analysis is to obtain the best quantification of editing efficiency for a set of amplicons, we suggest running the tool in the Mixed mode. The Genome mode is instead suggested to check problematic libraries, since a report is generated for each region discovered, even if the region is not mappable to any amplicon (however, this may be time consuming). Finally the Amplicon mode is the fastest, although the least reliable in terms of quantification accuracy.

As an illustrative example we reanalyzed data from a recent paper comparing the on-target and off-target efficiency of Cas9 [8]. We downloaded the raw data (SRX380876 from the SRA archive) from one of the pooled experiments containing two target sites for the *VEGFA* gene, one target site for *EMX1* and several off-target sites. Using CRISPRessoPooled we were able to automatically recover reads for the different amplicons pooled and to generate CRISPResso reports and mapping statistics for the amplicon with adequate reads and for which the amplifying primers were reported (Supplementary Figure 13 and Supplementary Table 1-2). The full CRISPREssoPooled reports for this experiment can be downloaded as Supplementary File 5.

List of all the command line parameters

```
  -r1 FASTQ_R1, --fastq_r1 FASTQ_R1
                        First fastq file (default: Fastq filename)
  -r2 FASTQ_R2, --fastq_r2 FASTQ_R2
                        Second fastq file for paired end reads (default: )
  -f AMPLICONS_FILE, --amplicons_file AMPLICONS_FILE
                        Amplicons description file. In particular, this file,
                        is a tab delimited text file with up to 5 columns (2
                        required): AMPLICON_NAME: an identifier for the
                        amplicon (must be unique) AMPLICON_SEQUENCE: amplicon
                        sequence used in the design of the experiment
                        sgRNA_SEQUENCE (OPTIONAL): sgRNA sequence used for
                        this amplicon without the PAM sequence. If more than
                        one separate them by commas and not spaces. If not
                        available enter NA. EXPECTED_AMPLICON_AFTER_HDR
                        (OPTIONAL): expected amplicon sequence in case of HDR.
                        If not available enter NA. CODING_SEQUENCE (OPTIONAL):
                        Subsequence(s) of the amplicon corresponding to coding
                        sequences. If more than one separate them by commas
                        and not spaces. If not available enter NA. (default: )
  -x BOWTIE2_INDEX, --bowtie2_index BOWTIE2_INDEX
                        Basename of Bowtie2 index for the reference genome
                        (default: )
  --gene_annotations GENE_ANNOTATIONS
                        Gene Annotation Table from UCSC Genome Browser Tables
                        (http://genome.ucsc.edu/cgi-
                        bin/hgTables?command=start), please select as table
                        "knowGene", as output format "all fields from selected
                        table" and as file returned "gzip compressed"
                        (default: )
  -p N_PROCESSES, --n_processes N_PROCESSES
                        Number of processes to use for the bowtie2 alignment
                        (default: 8)
```

```
--botwie2_options_string BOTWIE2_OPTIONS_STRING
                      Override options for the Bowtie2 alignment command
                      (default: -k 1 --end-to-end -N 0 --np 0 )
--min_reads_to_use_region MIN_READS_TO_USE_REGION
                      Minimum number of reads that align to a region to
                      perform the CRISPResso analysis (default: 1000)
-q MIN_AVERAGE_READ_QUALITY, --min_average_read_quality MIN_AVERAGE_READ_QUALITY
                      Minimum average quality score (phred33) to keep a read
                      (default: 0)
-s MIN_SINGLE_BP_QUALITY, --min_single_bp_quality MIN_SINGLE_BP_QUALITY
                      Minimum single bp score (phred33) to keep a read
                      (default: 0)
--min_identity_score MIN_IDENTITY_SCORE
                      Min identity score for the alignment (default: 60.0)
-n NAME, --name NAME  Output name (default: )
-o OUTPUT_FOLDER, --output_folder OUTPUT_FOLDER
--trim_sequences      Enable the trimming of Illumina adapters with
                      Trimmomatic (default: False)
--trimmomatic_options_string TRIMMOMATIC_OPTIONS_STRING
                      Override options for Trimmomatic (default: ILLUMINACLI
                      P:/Users/luca/Projects/CRISPResso/CRISPResso/data
                      /NexteraPE-PE.fa:0:90:10:0:true MINLEN:40)
--min_paired_end_reads_overlap MIN_PAIRED_END_READS_OVERLAP
                      Minimum required overlap length between two reads to
                      provide a confident overlap. (default: 4)
--max_paired_end_reads_overlap MAX_PAIRED_END_READS_OVERLAP
                      parameter for the flash merging step, this parameter
                      is the maximum overlap length expected in
                      approximately 90% of read pairs. Please see the flash
                      manual for more information. (default: 80)
-w WINDOW_AROUND_SGRNA, --window_around_sgrna WINDOW_AROUND_SGRNA
                      Window(s) in bp around each sgRNA to quantify the
                      indels. Any indels outside this window is excluded. A
                      value of -1 disable this filter. (default: -1)
--cleavage_offset CLEAVAGE_OFFSET
                      Cleavage offset to use within respect to the provided
                      sgRNA sequence. Remember that the sgRNA sequence must
                      be entered without the PAM. The default is -3 and is
                      suitable for the SpCas9 system. For alternate
                      nucleases, other cleavage offsets may be appropriate,
                      for example, if using Cpf1 set this parameter to 1.
                      (default: -3)
--exclude_bp_from_left EXCLUDE_BP_FROM_LEFT
                      Exclude bp from the left side of the amplicon sequence
                      for the quantification of the indels (default: 5)
--exclude_bp_from_right EXCLUDE_BP_FROM_RIGHT
                      Exclude bp from the right side of the amplicon
                      sequence for the quantification of the indels
                      (default: 5)
--hdr_perfect_alignment_threshold HDR_PERFECT_ALIGNMENT_THRESHOLD
                      Sequence homology % for an HDR occurrence (default:
                      98.0)
--needle_options_string NEEDLE_OPTIONS_STRING
                      Override options for the Needle aligner (default:
                      -gapopen=10 -gapextend=0.5 -awidth3=5000)
--keep_intermediate   Keep all the intermediate files (default: False)
--dump                Dump numpy arrays and pandas dataframes to file for
                      debugging purposes (default: False)
--save_also_png       Save also .png images additionally to .pdf files
                      (default: False)
```

Supplementary Note 6 - Installation and usage of CRISPRessoWGS

CRISPRessoWGS is a utility for the analysis of genome editing experiment from whole genome sequencing (WGS) data. CRISPRessoWGS allows exploring any region of the genome to quantify targeted editing or potentially off-target effects.

Installation

CRISPRessoWGS is installed automatically during the installation of CRISPResso, but to use it two additional programs must be installed:

- samtools: http://samtools.sourceforge.net/
- bowtie2: http://bowtie-bio.sourceforge.net/bowtie2

To install these tools please refer to their documentation.

To run CRISPRessoWGS you must provide:

1. A genome aligned BAM file. To align reads from a WGS experiment to the genome there are many options available, we suggest using either **Bowtie2 (**http://bowtie-bio.sourceforge.net/bowtie2/) or **BWA (**http://bio-bwa.sourceforge.net/**).**
2. A FASTA file containing the reference sequence used to align the reads and create the BAM file (the reference files for the most common organism can be download from UCSC: http://hgdownload.soe.ucsc.edu/downloads.html. Download and uncompress only the file ending with .fa.gz, for example for the last version of the human genome download and uncompress the file hg38.fa.gz)
3. Descriptions file containing the coordinates of the regions to analyze and some additional information. In particular, this file is a tab delimited text file with up to 7 columns (4 required):
   - chr_id: chromosome of the region in the reference genome.
   - bpstart: start coordinate of the region in the reference genome.
   - bpend: end coordinate of the region in the reference genome.
   - *REGION_NAME*: an identifier for the region (must be unique).
   - *sgRNA_SEQUENCE (OPTIONAL)*: sgRNA sequence used for this genomic segment without the PAM sequence. If not available, enter *NA.*
   - *EXPECTED_SEGMENT_AFTER_HDR (OPTIONAL)*: expected genomic segment sequence in case of HDR. If more than one, separate by commas and not spaces. If not available, enter *NA.*
   - *CODING_SEQUENCE (OPTIONAL)*: Subsequence(s) of the genomic segment corresponding to coding sequences. If more than one, separate by commas and not spaces. If not available, enter *NA.*

   A file in the right format should look like this:

   ```
   chr1    65118211        65118261        R1      CTACAGAGCCCCAGTCCTGG NA       NA
   chr6    51002798        51002820        R2      NA       NA       NA
   ```

   Note: no column titles should be entered. As you may have noticed this file is just a BED file with extra columns. For this reason a normal BED file with 4 columns, is also **accepted** by this utility.

4. Optionally the full path of a gene annotations file from UCSC. You can download the this file from the UCSC Genome Browser (http://genome.ucsc.edu/cgi-bin/hgTables?command=start) selecting as table "knowGene", as output format "all fields from selected table" and as file returned "gzip compressed". (something like: homes/luca/genomes/human_hg19/gencode_v19.gz)

Example:

```
CRISPRessoWGS -b WGS/50/50_sorted_rmdup_fixed_groups.bam -f WGS_TEST.txt -r
/gcdata/gcproj/Luca/GENOMES/mm9/mm9.fa --gene_annotations ensemble_mm9.txt.gz  --name
CRISPR_WGS_SRR1542350
```

The output from these files will consist of:

1. REPORT_READS_ALIGNED_TO_SELECTED_REGIONS_WGS.txt: this file contains the same information provided in the input description file, plus some additional columns:
    a. sequence: sequence in the reference genome for the region specified.
    b. *gene_overlapping:* gene/s overlapping the region specified.
    c. *n_reads*: number of reads recovered for the region.
    d. bam_file_with_reads_in_region: file containing only the subset of the reads that overlap, also partially, with the region. This file is indexed and can be easily loaded for example on IGV for visualization of single reads or for the comparison of two conditions. For example, in the figure below (fig X) we show reads mapped to a region inside the coding sequence of the gene Crygc subjected to NHEJ (CRISPR_WGS_SRR1542350) vs reads from a control experiment (CONTROL_WGS_SRR1542349).
    e. fastq.gz_file_trimmed_reads_in_region: file containing only the subset of reads fully covering the specified regions, and trimmed to match the sequence in that region. These reads are used for the subsequent analysis with CRISPResso.
2. ANALYZED_REGIONS (folder): this folder contains all the BAM and FASTQ files, one for each region analyzed.
3. A set of folders with the CRISPResso report on the regions provided in input with enough reads (the default setting is to have at least 10 reads, but the parameter can be adjusted with the option *--min_reads_to_use_region*).
4. *CRISPRessoPooled_RUNNING_LOG.txt*: execution log and messages for the external utilities called.

This utility is particular useful to investigate and quantify mutation frequency in a list of potential target or off-target sites, coming for example from prediction tools, or from other orthogonal assays.

As an illustrative example we reanalyzed data from a recent study that uses WGS data for gene correction by CRISPR/Cas9 in mouse spermatogonial stem cells [11]. In this study, the investigators first created mutant mouse stem cells with a biallelic disease-causing deletion of 1 bp in the coding sequence of the gene *Crygc* (CONTROL_SRR15423349). Subsequently, they tried to repair this gene by HDR in mutant stem cells, and isolated various clones with HDR-mediated bialleleic correction, biallelic NHEJ indel events, and compound heterozygous HDR and NHEJ alleles. They subjected to WGS one clone which had compound heterozygous NHEJ alleles (one with a 1 bp insertion 6 bp from the disease causing deletion and the other with a 1 bp insertion 7 bp from the disease-causing deletion; CRISPR_SRR1542350). After downloading and realigning the raw data (we used bowtie2 and mm9 reference genome from UCSC) to create a .bam file we run CRISPRessoWGS in the *Crygc* region to see if we can recapitulate the two events: the 1 bp deletion in the control cells, and each of the additional 1 bp insertions in the edited clone. As shown in Supplementary Fig. 14 (a,b), CRISPRessoWGS correctly recovers these events and demonstrates their frequency. In addition, the reads fully mapped to this region are extracted and a .bam file is created for a more detailed visualization for example with IGV (Supplementary Figure 14 c). The full CRISPREssoWGS reports for this experiment can be downloaded as Supplementary File 6. (This report includes also a control genomic region, named R2, that as expected, does not show any mutations in either sample.)

List of all the command line parameters

```
  -b BAM_FILE, --bam_file BAM_FILE
                        WGS aligned bam file (default: bam filename)
  -f REGION_FILE, --region_file REGION_FILE
                        Regions description file. A BED format file containing
                        the regions to analyze, one per line. The REQUIRED
                        columns are: chr_id(chromosome name), bpstart(start
                        position), bpend(end position), the optional columns
                        are:name (an unique indentifier for the region),
                        guide_seq, expected_hdr_amplicon_seq,coding_seq, see
                        CRISPResso help for more details on these last 3
                        parameters) (default: )
  -r REFERENCE_FILE, --reference_file REFERENCE_FILE
                        A FASTA format reference file (for example hg19.fa for
                        the human genome) (default: )
  --min_reads_to_use_region MIN_READS_TO_USE_REGION
```

```
                      Minimum number of reads that align to a region to
                      perform the CRISPResso analysis (default: 10)
--gene_annotations GENE_ANNOTATIONS
                      Gene Annotation Table from UCSC Genome Browser Tables
                      (http://genome.ucsc.edu/cgi-
                      bin/hgTables?command=start), please select as table
                      "knowGene", as output format "all fields from selected
                      table" and as file returned "gzip compressed"
                      (default: )
-q MIN_AVERAGE_READ_QUALITY, --min_average_read_quality MIN_AVERAGE_READ_QUALITY
                      Minimum average quality score (phred33) to keep a read
                      (default: 0)
-s MIN_SINGLE_BP_QUALITY, --min_single_bp_quality MIN_SINGLE_BP_QUALITY
                      Minimum single bp score (phred33) to keep a read
                      (default: 0)
--min_identity_score MIN_IDENTITY_SCORE
                      Min identity score for the alignment (default: 60.0)
-n NAME, --name NAME  Output name (default: )
-o OUTPUT_FOLDER, --output_folder OUTPUT_FOLDER
--trim_sequences      Enable the trimming of Illumina adapters with
                      Trimmomatic (default: False)
--trimmomatic_options_string TRIMMOMATIC_OPTIONS_STRING
                      Override options for Trimmomatic (default:
                      ILLUMINACLIP:/Users/luca/anaconda/lib/python2.7/site-
                      packages/CRISPResso-0.8.2-py2.7.egg/CRISPResso/data
                      /NexteraPE-PE.fa:0:90:10:0:true MINLEN:40)
--min_paired_end_reads_overlap MIN_PAIRED_END_READS_OVERLAP
                      Minimum required overlap length between two reads to
                      provide a confident overlap. (default: 4)
-w _AROUND_SGRNA, --window_around_sgrna WINDOW_AROUND_SGRNA
                      Window(s) in bp around the cleavage position (half on
                      on each side) as determined by the provide guide RNA
                      sequence to quantify the indels. Any indels outside
                      this window are excluded. A value of -1 disables this
                      filter. (default: -1)
--cleavage_offset CLEAVAGE_OFFSET
                      Cleavage offset to use within respect to the 3' end of
                      the provided sgRNA sequence. Remember that the sgRNA
                      sequence must be entered without the PAM. The default
                      is -3 and is suitable for the SpCas9 system. For
                      alternate nucleases, other cleavage offsets may be
                      appropriate, for example, if using Cpf1 this parameter
                      would be set to 1. (default: -3)
--exclude_bp_from_left EXCLUDE_BP_FROM_LEFT
                      Exclude bp from the left side of the amplicon sequence
                      for the quantification of the indels (default: 5)
--exclude_bp_from_right EXCLUDE_BP_FROM_RIGHT
                      Exclude bp from the right side of the amplicon
                      sequence for the quantification of the indels
                      (default: 5)
--hdr_perfect_alignment_threshold HDR_PERFECT_ALIGNMENT_THRESHOLD
                      Sequence homology % for an HDR occurrence (default:
                      98.0)
--needle_options_string NEEDLE_OPTIONS_STRING
                      Override options for the Needle aligner (default:
                      -gapopen=10 -gapextend=0.5 -awidth3=5000)
--keep_intermediate   Keep all the intermediate files (default: False)
--dump                Dump numpy arrays and pandas dataframes to file for
                      debugging purposes (default: False)
--save_also_png       Save also .png images additionally to .pdf files (default: False)
```

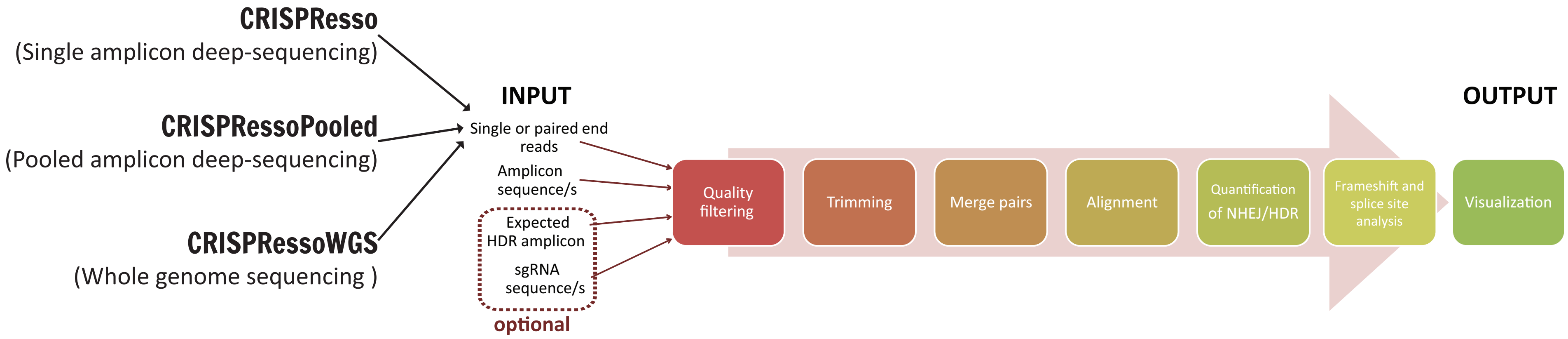
Acknowledgements

We are grateful to Feng Zhang, David Scott and Samuel Lessard for useful feedback and suggestions; the FAS Research Computing Team, in particular Daniel Kelleher, for great support in hosting the web application of CRISPResso; and Sorel Fitz-Gibbon from UCLA for help in sharing data. Finally, we thank all members of the Guo-Cheng Yuan lab for testing the software. L.P. is supported by NHGRI Career Development Award K99HG008399. M.C.C. is supported by F30DK103359-01A1. S.H.O. is supported by P01HL032262 and P30DK049216 (Center of Excellence in Molecular Hematology). D.E.B. is supported by NIDDK Career Development Award K08DK093705, Doris Duke Charitable Foundation Innovations in Clinical Research Award (2013137), and Charles H. Hood Foundation Child Health Research Award. G.C.Y. is supported by NIH R01HL119099 and R01HG005085.

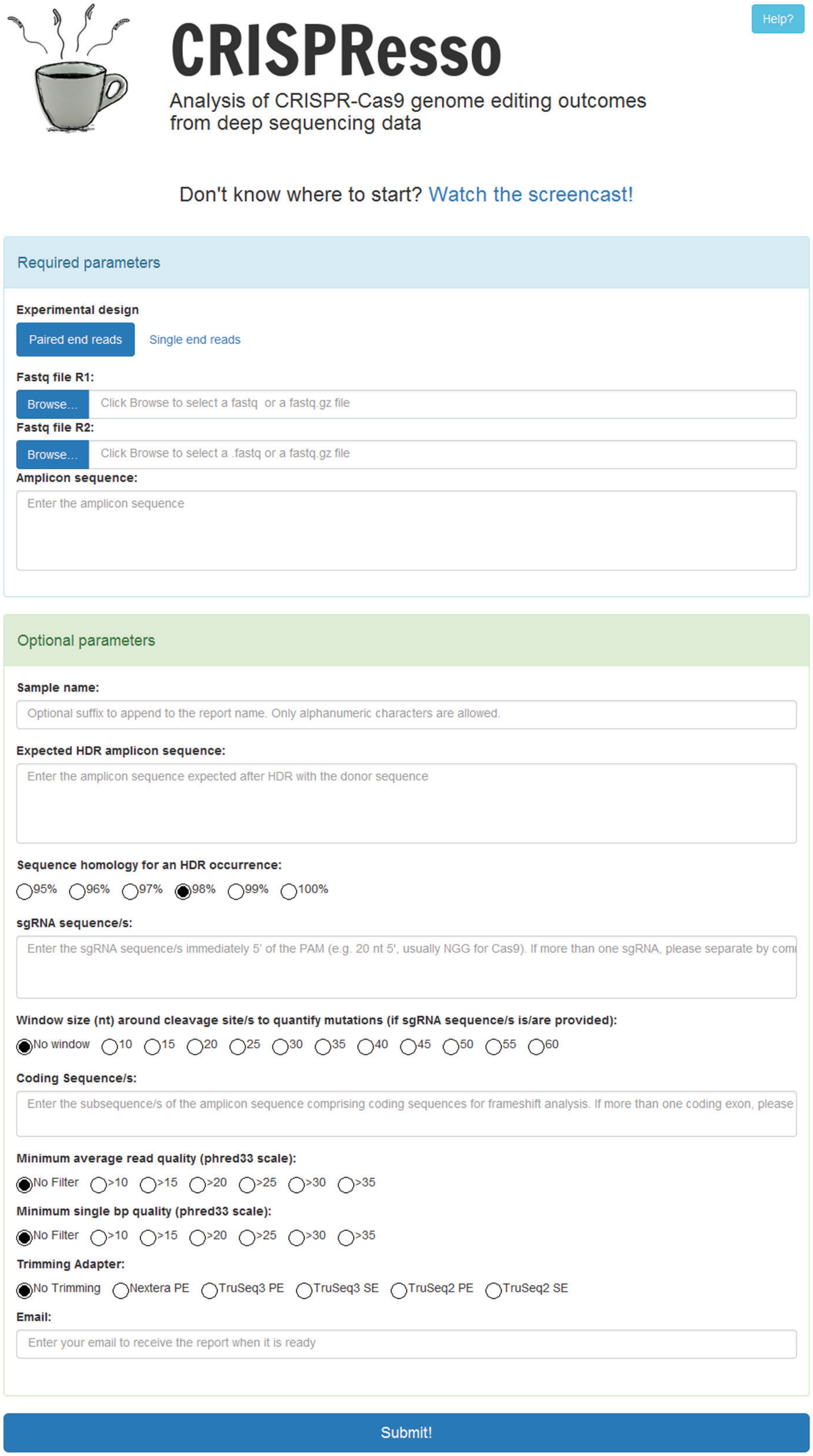
Supplementary References

1. Ewing, B. and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res*, **8**, 186-194.
2. Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114-2120.
3. Magoc, T. and Salzberg, S.L. (2011) FLASH: fast length adjustment of short reads to improve genome assemblies, *Bioinformatics*, **27**, 2957-2963.
4. Rice, P., Longden, I. and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet*, **16**, 276-277.
5. Huang, W. et al (2012) ART: a next-generation sequencing read simulator, Bioinformatics (2012) **28** (4): 593-594.
6. Güell M, Yang L, Church GM: Genome editing assessment using CRISPR Genome Analyzer (CRISPR-GA). Bioinformatics 2014, 30:1–3.
7. Canver, M.C. et al. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature (2015).
8. Fu, Y., Sander, J.D., Reyon, D., Cascio, V.M. & Joung, J.K. Improving CRISPR-Cas nuclease specificity using truncated guide RNAs. Nature biotechnology 32, 279-284 (2014).
9. Ran, F.A.*, et al.* (2013) Genome engineering using the CRISPR-Cas9 system, *Nat Protoc*, **8**, 2281-2308.
10. Hoban, M.D.*, et al.* (2015) Correction of the sickle cell disease mutation in human hematopoietic stem/progenitor cells, *Blood*, **125**, 2597-2604
11. Wu, Y. et al. Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells. Cell research 25, 67-79 (2015).

Supplementary Figure 1


CRISPResso
(Single amplicon deep-sequencing)
CRISPRessoPooled
(Pooled amplicon deep-sequencing)
CRISPRessoWGS
(Whole genome sequencing)
INPUT
Single or paired end reads
Amplicon sequence/s
Expected HDR amplicon
sgRNA sequence/s
optional
Quality filtering
Trimming
Merge pairs
Alignment
Quantification of NHEJ/HDR
Frameshift and splice site analysis
OUTPUT
Visualization

Supplementary Figure 2



Help?

CRISPResso

Analysis of CRISPR-Cas9 genome editing outcomes from deep sequencing data

Don't know where to start? Watch the screencast!

Required parameters

Experimental design

[✓ Paired end reads] Single end reads

Fastq file R1:

Browse... Click Browse to select a fastq or a fastq.gz file

Fastq file R2:

Browse... Click Browse to select a .fastq or a fastq.gz file

Amplicon sequence:

Enter the amplicon sequence

Optional parameters

Sample name:

Optional suffix to append to the report name. Only alphanumeric characters are allowed.

Expected HDR amplicon sequence:

Enter the amplicon sequence expected after HDR with the donor sequence

Sequence homology for an HDR occurrence:

○95% ○96% ○97% ●98% ○99% ○100%

sgRNA sequence/s:

Enter the sgRNA sequence/s immediately 5' of the PAM (e.g. 20 nt 5', usually NGG for Cas9). If more than one sgRNA, please separate by com

Window size (nt) around cleavage site/s to quantify mutations (if sgRNA sequence/s is/are provided):

●No window ○10 ○15 ○20 ○25 ○30 ○35 ○40 ○45 ○50 ○55 ○60

Coding Sequence/s:

Enter the subsequence/s of the amplicon sequence comprising coding sequences for frameshift analysis. If more than one coding exon, please

Minimum average read quality (phred33 scale):

●No Filter ○>10 ○>15 ○>20 ○>25 ○>30 ○>35

Minimum single bp quality (phred33 scale):

●No Filter ○>10 ○>15 ○>20 ○>25 ○>30 ○>35

Trimming Adapter:

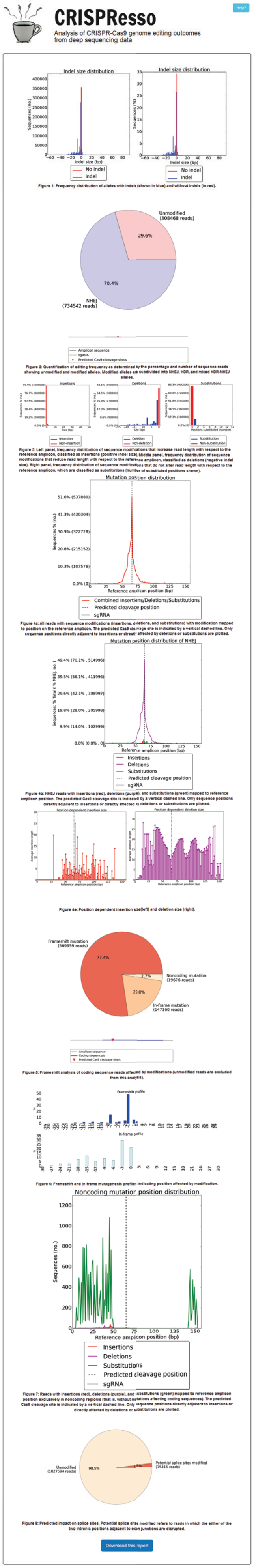
●No Trimming ○Nextera PE ○TruSeq3 PE ○TruSeq3 SE ○TruSeq2 PE ○TruSeq2 SE

Email:

Enter your email to receive the report when it is ready

Submit!

Supplementary Figure 3



Help?
CRISPResso
Analysis of CRISPR-Cas9 genome editing outcomes from deep sequencing data
Indel size distribution
Sequences (no.)
Sequences (%)
Indel size (bp)
No indel
Indel
Unmodified (308468 reads)
29.6%
70.4%
NHEJ (734542 reads)
Amplicon sequence
sgRNA
Predicted Cas9 cleavage site/s
Insertions
Deletions
Substitutions
Sequences % (no.)
Size (bp)
Positions substituted (number)
Insertion
Non-insertion
Deletion
Non-deletion
Substitution
Non-substitution
Mutation position distribution
51.6% (537880)
41.3% (430304)
30.9% (322728)
20.6% (215152)
10.3% (107576)
0.0% (0)
Reference amplicon position (bp)
Combined Insertions/Deletions/Substitutions
Predicted cleavage position
sgRNA
Mutation position distribution of NHEJ
Sequences: % Total (% NHEJ, no.)
49.4% (70.1% , 514996)
39.5% (56.1% , 411996)
29.6% (42.1% , 308997)
19.8% (28.0% , 205998)
9.9% (14.0% , 102999)
0.0% (0.0% , 0)
Insertions
Deletions
Substitutions
Predicted cleavage position
sgRNA
Position dependent insertion size
Average insertion length
Position dependent deletion size
Average deletion length
Reference amplicon position (bp)
Frameshift mutation (569959 reads)
77.4%
2.7%
Noncoding mutation (19676 reads)
20.0%
In-frame mutation (147160 reads)
Amplicon sequence
Coding sequence/s
Predicted Cas9 cleavage site/s
Frameshift profile
In-frame profile
%
Noncoding mutation position distribution
Sequences (no.)
Reference amplicon position (bp)
Insertions
Deletions
Substitutions
Predicted cleavage position
sgRNA
Unmodified (1027594 reads)
98.5%
1.5%
Potential splice sites modified (15416 reads)
Download this report

Figure 1: Frequency distribution of alleles with indels (shown in blue) and without indels (in red).

Figure 2: Quantification of editing frequency as determined by the percentage and number of sequence reads showing unmodified and modified alleles. Modified alleles are subdivided into NHEJ, HDR, and mixed HDR-NHEJ alleles.

Figure 3: Left panel, frequency distribution of sequence modifications that increase read length with respect to the reference amplicon, classified as insertions (positive indel size). Middle panel, frequency distribution of sequence modifications that reduce read length with respect to the reference amplicon, classified as deletions (negative indel size). Right panel, frequency distribution of sequence modifications that do not alter read length with respect to the reference amplicon, which are classified as substitutions (number of substituted positions shown).

Figure 4a: All reads with sequence modifications (insertions, deletions, and substitutions) with modification mapped to position on the reference amplicon. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4b: NHEJ reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4e: Position dependent insertion size(left) and deletion size (right).

Figure 5: Frameshift analysis of coding sequence reads affected by modifications (unmodified reads are excluded from this analysis).

Figure 6: Frameshift and in-frame mutagenesis profiles indicating position affected by modification.

Figure 7: Reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position exclusively in noncoding region/s (that is, without mutations affecting coding sequences). The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 8: Predicted impact on splice sites. Potential splice sites modified refers to reads in which the either of the two intronic positions adjacent to exon junctions are disrupted.

Supplementary Figure 4

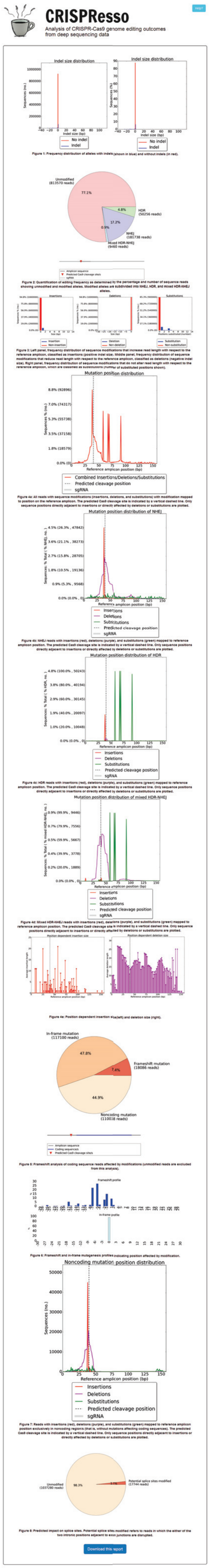


Figure 1: Frequency distribution of alleles with indels (shown in blue) and without indels (in red).

Figure 2: Quantification of editing frequency as determined by the percentage and number of sequence reads showing unmodified and modified alleles. Modified alleles are subdivided into NHEJ, HDR, and mixed HDR-NHEJ alleles.

Figure 3: Left panel, frequency distribution of sequence modifications that increase read length with respect to the reference amplicon, classified as insertions (positive indel size). Middle panel, frequency distribution of sequence modifications that reduce read length with respect to the reference amplicon, classified as deletions (negative indel size). Right panel, frequency distribution of sequence modifications that do not alter read length with respect to the reference amplicon, which are classified as substitutions (number of substituted positions shown).

Figure 4a: All reads with sequence modifications (insertions, deletions, and substitutions) with modification mapped to position on the reference amplicon. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4b: NHEJ reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4c: HDR reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4d: Mixed HDR-NHEJ reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4e: Position dependent insertion size(left) and deletion size (right).

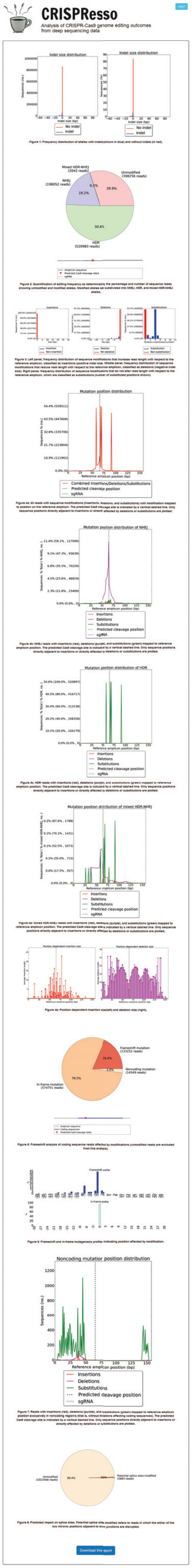
Figure 5: Frameshift analysis of coding sequence reads affected by modifications (unmodified reads are excluded from this analysis).

Figure 6: Frameshift and in-frame mutagenesis profiles indicating position affected by modification.

Figure 7: Reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position exclusively in noncoding region/s (that is, without mutations affecting coding sequences). The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 8: Predicted impact on splice sites. Potential splice sites modified refers to reads in which the either of the two intronic positions adjacent to exon junctions are disrupted.

Supplementary Figure 5


Help?
Indel size distribution
Sequences (no.)
Indel size (bp)
Sequences (%)
No indel
Indel
Mixed HDR-NHEJ (2042 reads)
Unmodified (308256 reads)
NHEJ (198052 reads)
0.2%
29.9%
19.2%
50.6%
HDR (520983 reads)
Amplicon sequence
Predicted Cas9 cleavage site/s
sgRNA
Insertions
Deletions
Substitutions
Insertion
Non-insertion
Deletion
Non-deletion
Substitution
Non-substitution
Positions substituted (number)
Mutation position distribution
Sequences % (no.)
Reference amplicon position (bp)
Combined Insertions/Deletions/Substitutions
Predicted cleavage position
Mutation position distribution of NHEJ
Sequences: % Total (% NHEJ, no.)
Mutation position distribution of HDR
Sequences: % Total (% HDR, no.)
Mutation position distribution of mixed HDR-NHEJ
Sequences: % Total (% mixed HDR-NHEJ, no.)
Position dependent insertion size
Average insertion length
Position dependent deletion size
Average deletion length
Frameshift mutation (133252 reads)
18.4%
Noncoding mutation (14549 reads)
2.0%
79.5%
In-frame mutation (574751 reads)
Coding sequence/s
Frameshift profile
In-frame profile
%
Noncoding mutation position distribution
Sequences (no.)
Unmodified (1023468 reads)
99.4%
0.6%
Potential splice sites modified (5865 reads)
Download this report

CRISPResso

Analysis of CRISPR-Cas9 genome editing outcomes from deep sequencing data

Figure 1: Frequency distribution of alleles with indels(shown in blue) and without indels (in red).

Figure 2: Quantification of editing frequency as determined by the percentage and number of sequence reads showing unmodified and modified alleles. Modified alleles are subdivided into NHEJ, HDR, and mixed HDR-NHEJ alleles.

Figure 3: Left panel, frequency distribution of sequence modifictions that increase read length with respect to the reference amplicon, classified as insertions (positive indel size). Middle panel, frequency distribution of sequence modifications that reduce read length with respect to the reference amplicon, classified as deletions (negative indel size). Right panel, frequency distribution of sequence modificaions that do not alter read length with respect to the reference amplicon, which are classified as substitutions (numer of substituted positions shown).

Figure 4a: All reads with sequence modifications (insertions, deletions, and substitutions) with modification mapped to position on the reference amplicon. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4b: NHEJ reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4c: HDR reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4d: Mixed HDR-NHEJ reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position. The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 4e: Position dependent insertion size(left) and deletion size (right).

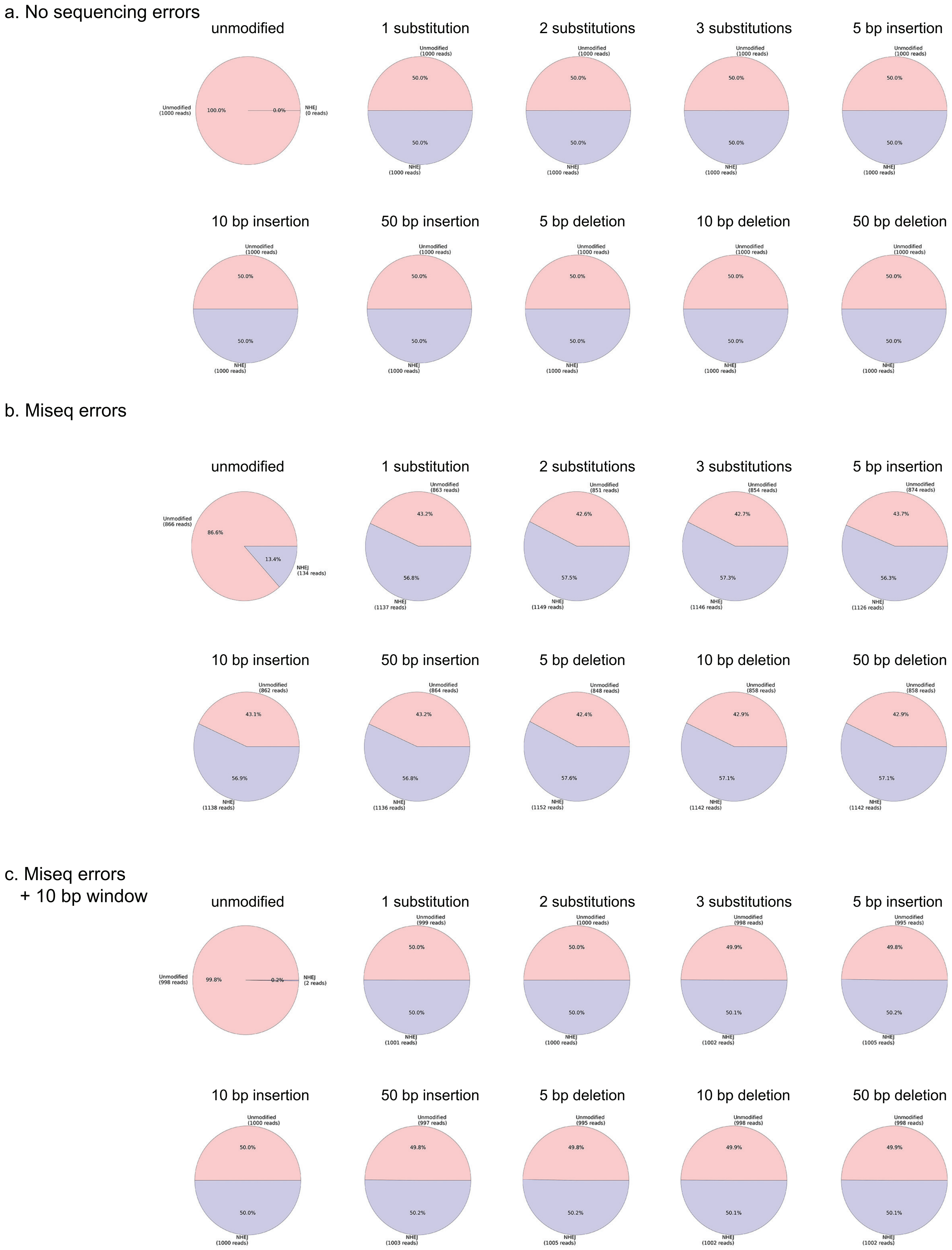
Figure 5: Frameshift analysis of coding sequence reads affected by modifications (unmodified reads are excluded from this analysis).

Figure 6: Frameshift and in-frame mutagenesis profiles indicating position affected by modification.

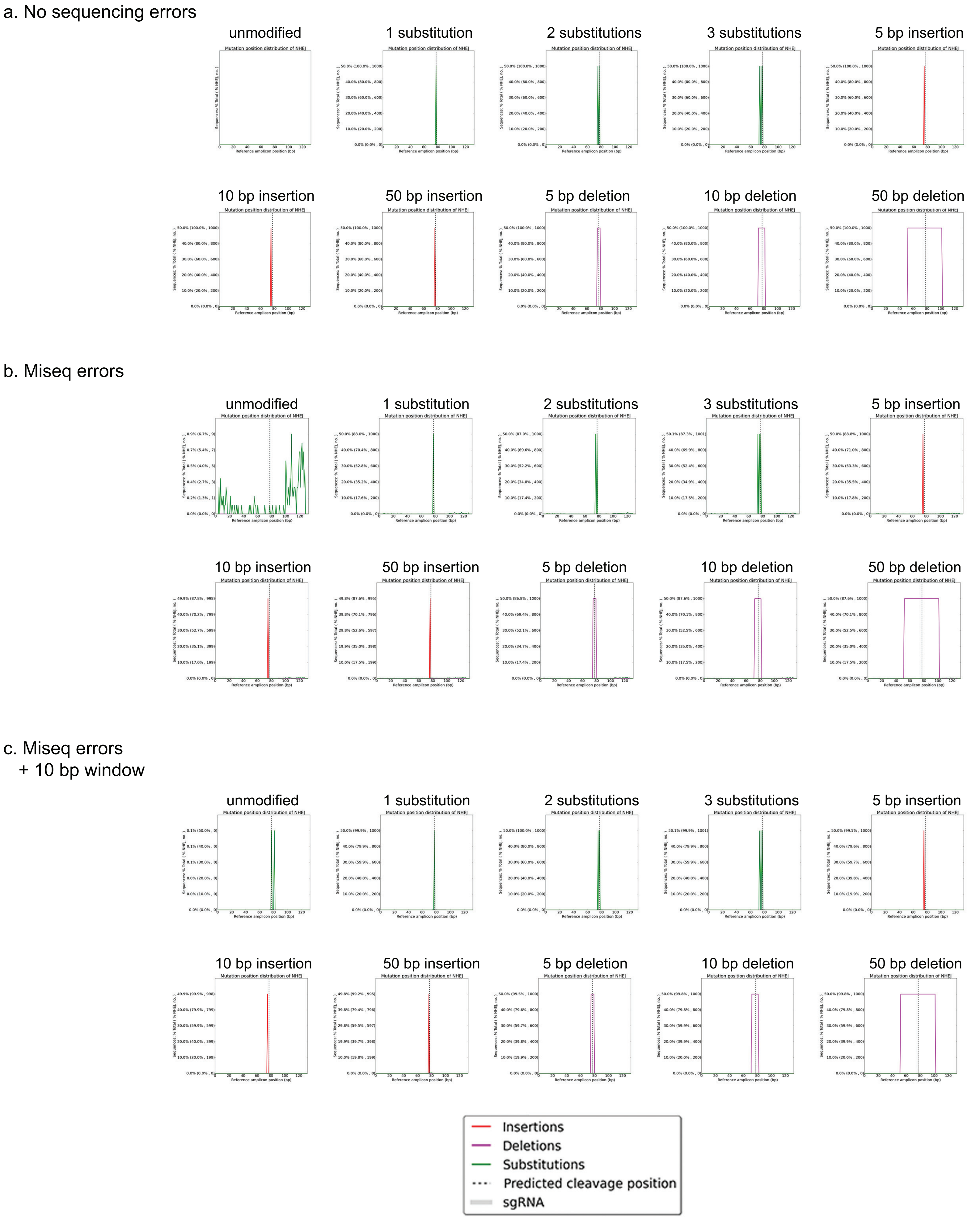
Figure 7: Reads with insertions (red), deletions (purple), and substitutions (green) mapped to reference amplicon position exclusively in noncoding region/s (that is, without mutations affecting coding sequences). The predicted Cas9 cleavage site is indicated by a vertical dashed line. Only sequence positions directly adjacent to insertions or directly affected by deletions or substitutions are plotted.

Figure 8: Predicted impact on splice sites. Potential splice sites modified refers to reads in which the either of the two intronic positions adjacent to exon junctions are disrupted.

Supplementary Figure 6



Supplementary Figure 7

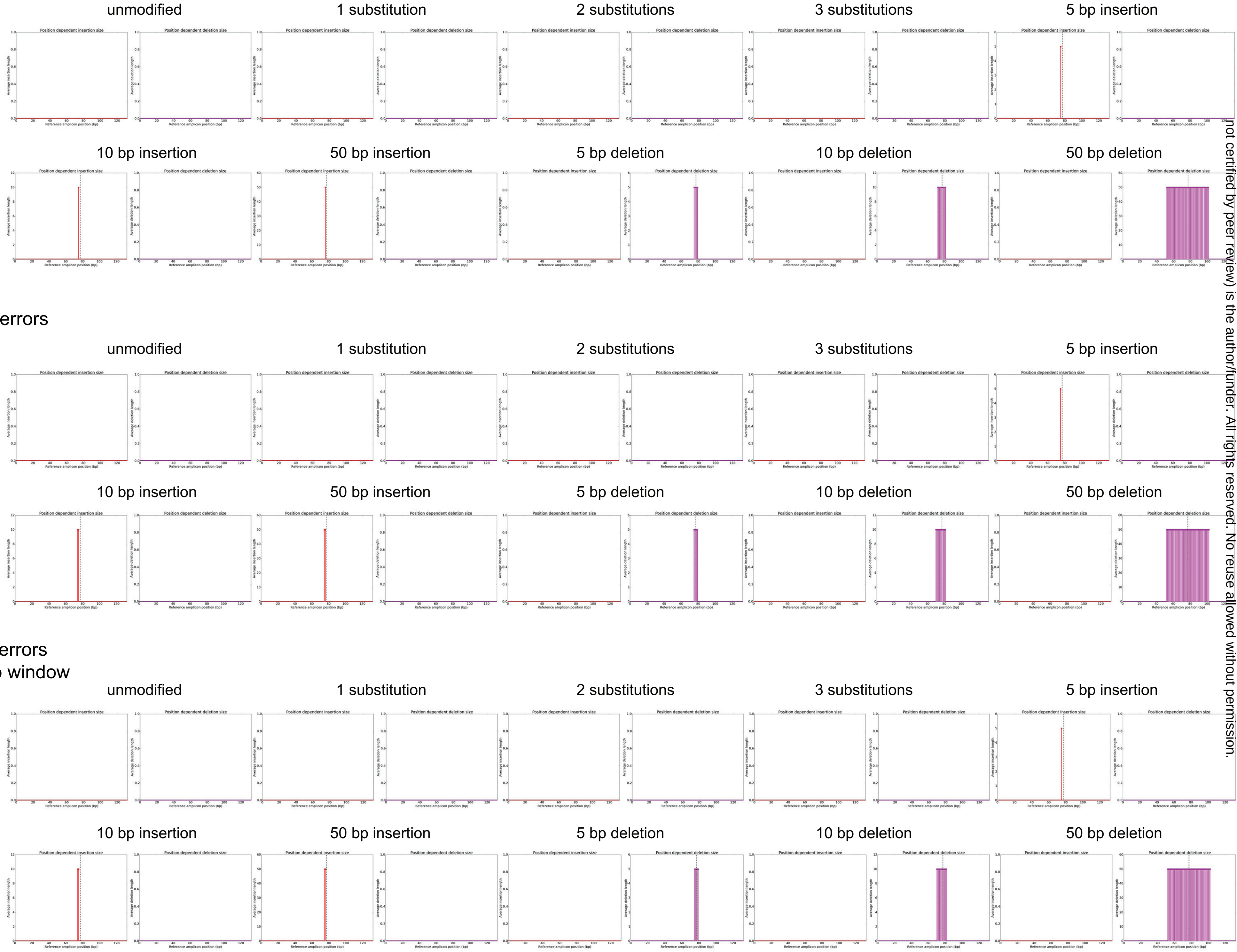


a. No sequencing errors

b. Miseq errors

c. Miseq errors + 10 bp window

Supplementary Figure 8

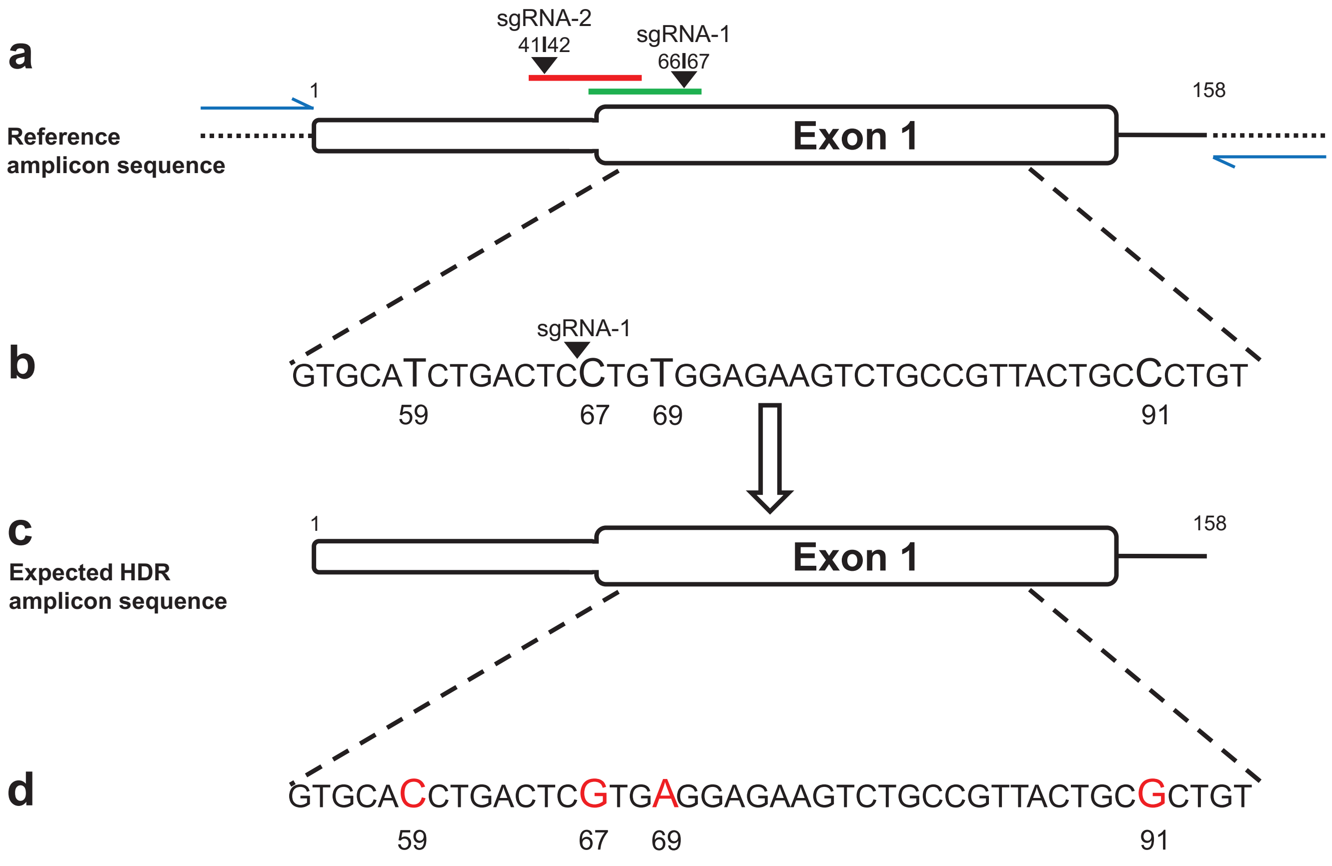


a. No sequencing errors

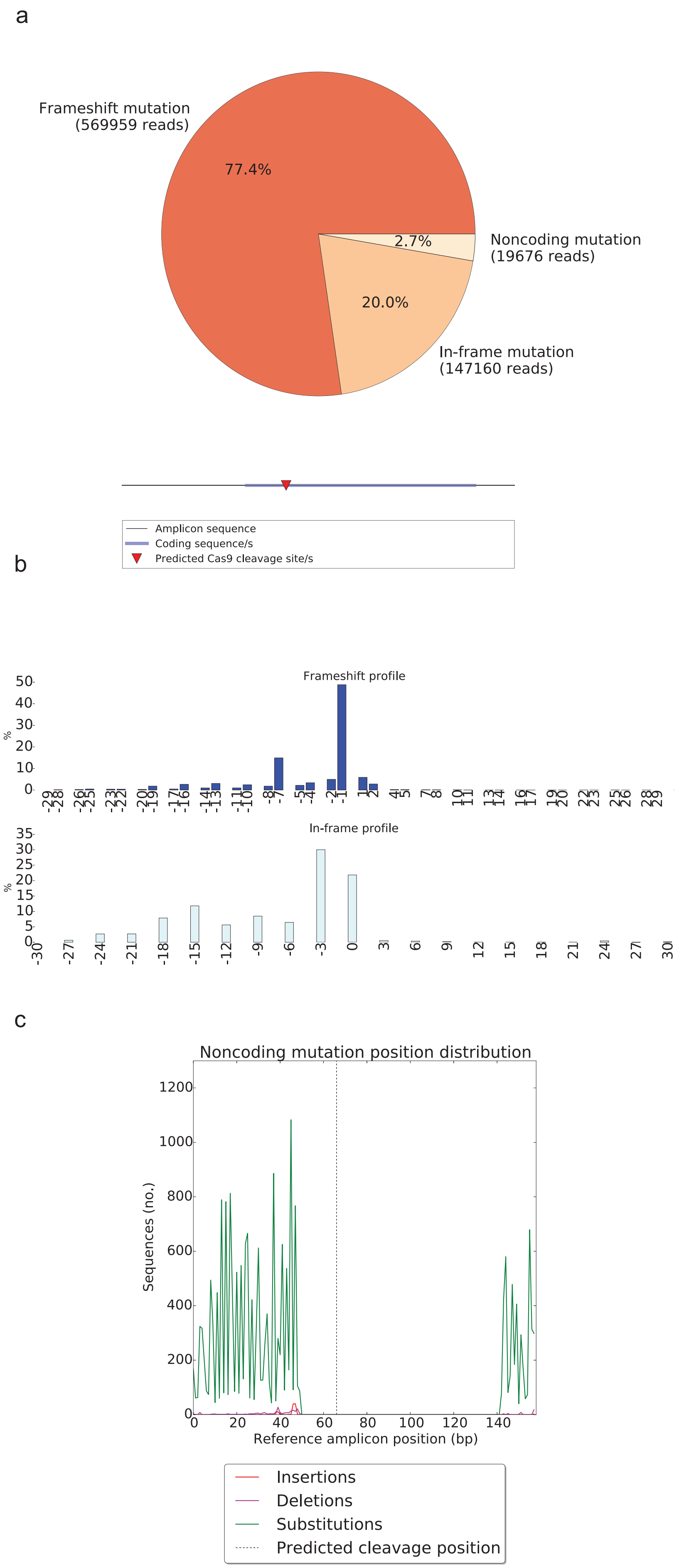
b. Miseq errors

c. Miseq errors
+ 10 bp window

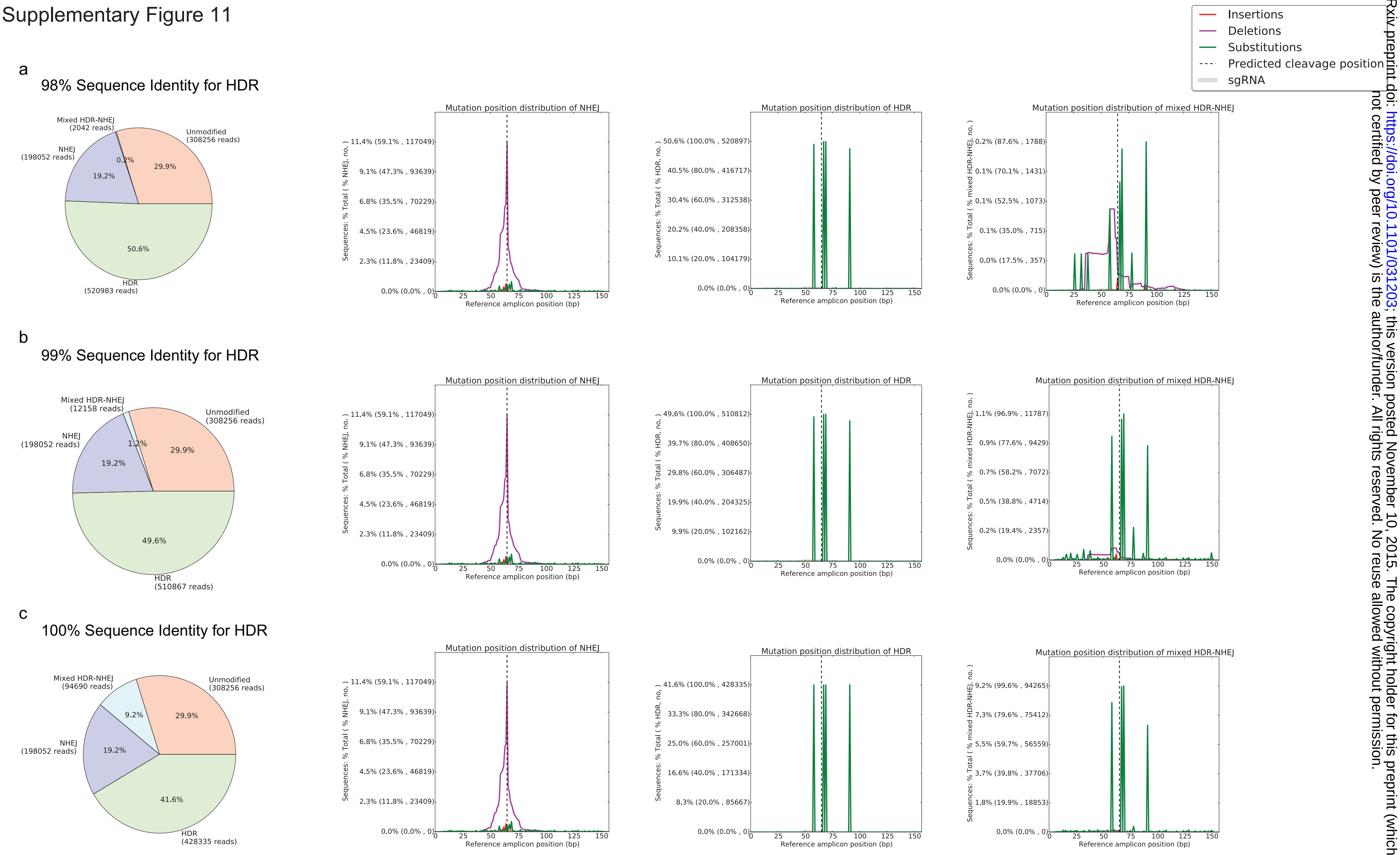
Supplementary Figure 9


a
sgRNA-2
41I42
sgRNA-1
66I67
1
158
Reference amplicon sequence
Exon 1
b
sgRNA-1
GTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGT
59
67
69
91
c
1
158
Expected HDR amplicon sequence
Exon 1
d
GTGCACCTGACTCGTGAGGAGAAGTCTGCCGTTACTGCGCTGT
59
67
69
91

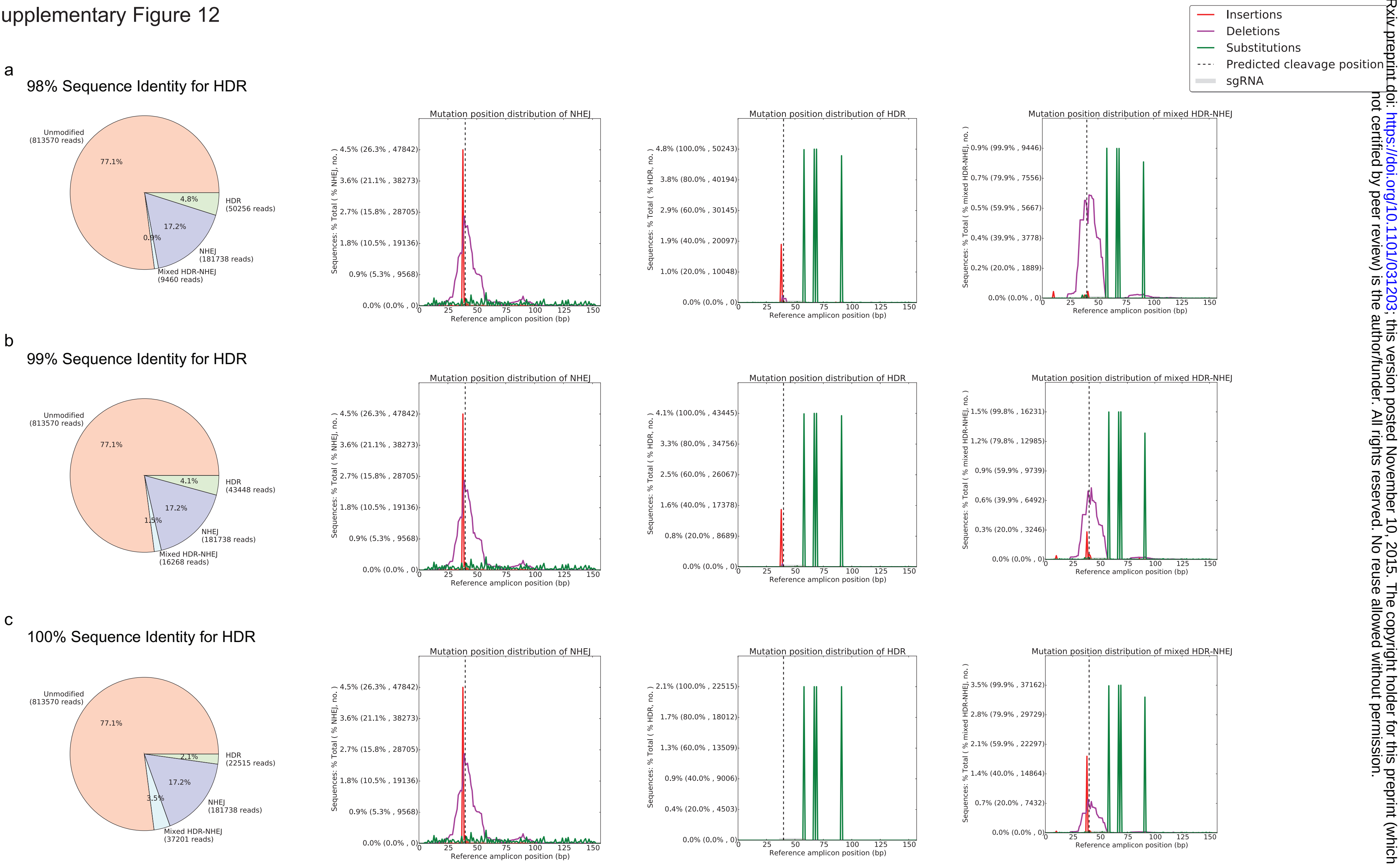
Supplementary Figure 10



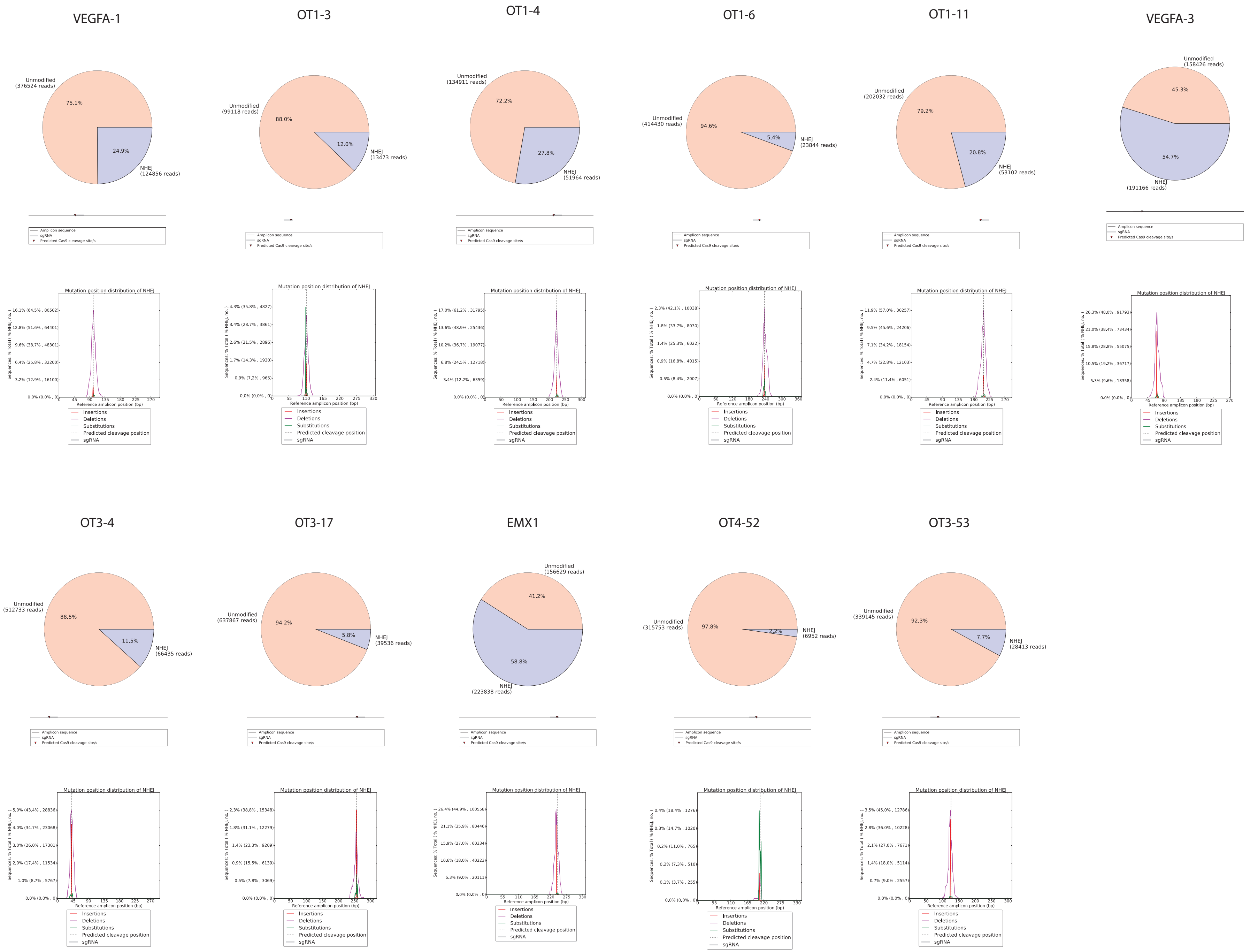
Supplementary Figure 11



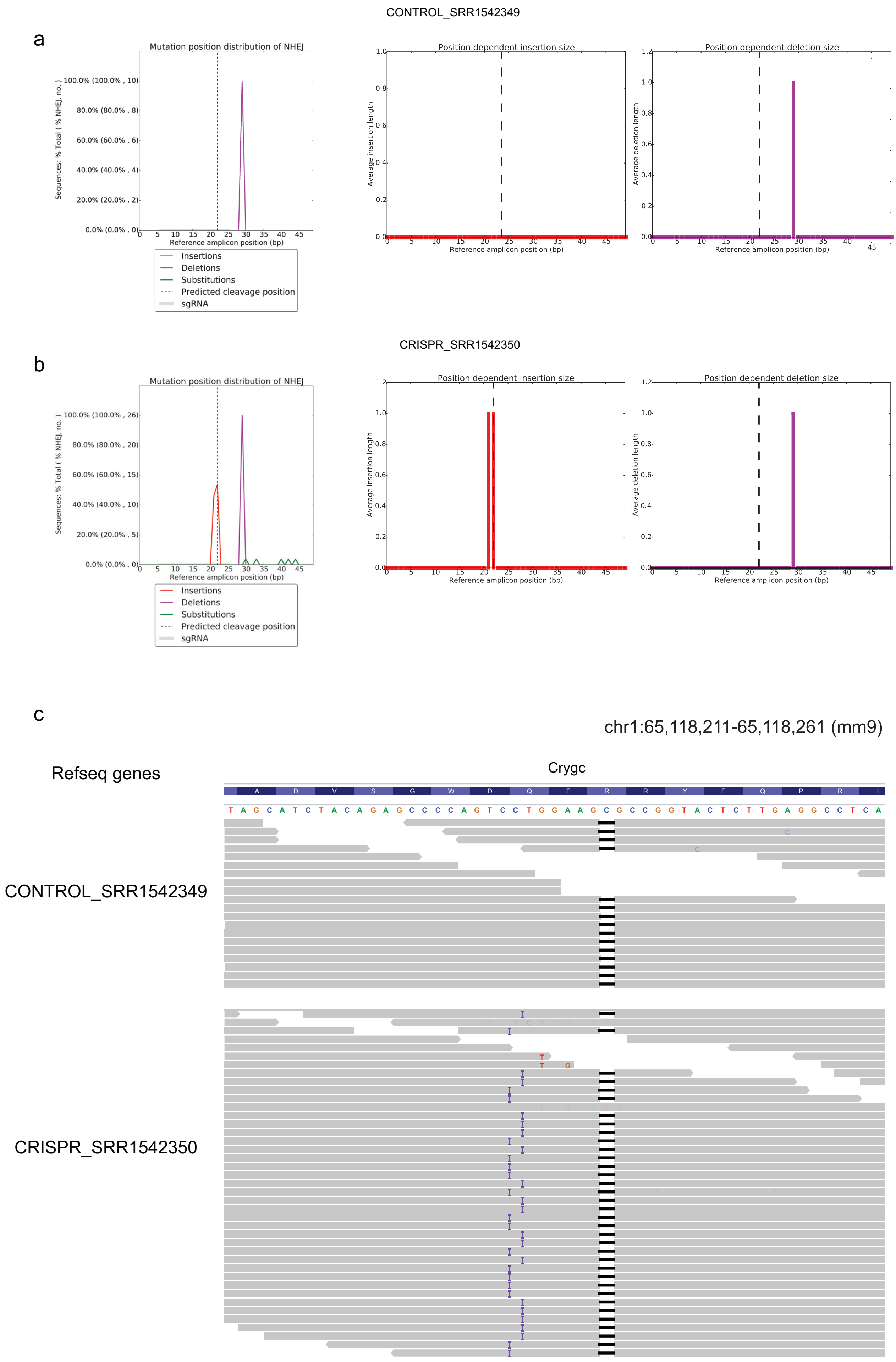
Supplementary Figure 12



Supplementary Figure 13



Supplementary Figure 14


CONTROL_SRR1542349
a
Mutation position distribution of NHEJ
Sequences: % Total (% NHEJ, no.)
100.0% (100.0% , 10)
80.0% (80.0% , 8)
60.0% (60.0% , 6)
40.0% (40.0% , 4)
20.0% (20.0% , 2)
0.0% (0.0% , 0)
Reference amplicon position (bp)
Insertions
Deletions
Substitutions
Predicted cleavage position
sgRNA
Position dependent insertion size
Average insertion length
Reference amplicon position (bp)
Position dependent deletion size
Average deletion length
Reference amplicon position (bp)
CRISPR_SRR1542350
b
Mutation position distribution of NHEJ
Sequences: % Total (% NHEJ, no.)
100.0% (100.0% , 26)
80.0% (80.0% , 20)
60.0% (60.0% , 15)
40.0% (40.0% , 10)
20.0% (20.0% , 5)
0.0% (0.0% , 0)
Reference amplicon position (bp)
Insertions
Deletions
Substitutions
Predicted cleavage position
sgRNA
Position dependent insertion size
Average insertion length
Reference amplicon position (bp)
Position dependent deletion size
Average deletion length
Reference amplicon position (bp)
c
chr1:65,118,211-65,118,261 (mm9)
Refseq genes
Crygc
A D V S G W D Q F R R Y E Q P R L
T A G C A T C T A C A G A G C C C C A G T C C T G G A A G C G C C G G T A C T C T T G A G G C C T C A
CONTROL_SRR1542349
CRISPR_SRR1542350