Inference of CRISPR Edits (ICE)

Knock-in Analysis

Developed by EditCo

Introduction

CRISPR can be used to make precise nucleotide changes to the genome or knock in a sequence of interest. These experiments involve introducing a donor DNA template (containing the desired insertion) along with CRISPR components into cells. After CRISPR induces a double-strand break, the desired sequence is inserted into the targeted site via homology-directed repair (HDR). The process of HDR is rather inefficient, so only a fraction of the targeted cells will have the desired insert. Thus, the resulting knock-in cell pool will contain a mix of knock-in and wild type sequences, as well as sequences containing insertions/deletions (indels) resulting from non-homologous end joining (NHEJ). If genotypically identical cells are needed, clonal populations can be generated from the knock-in cell pool.

Inference of CRISPR Edits (ICE) is a free online tool that provides an easy quantitative assessment of indels generated by CRISPR in pools and clones. ICE compares the Sanger sequencing traces of amplicons generated from genomic DNA isolated from both the edited and control (wild type). Potential editing outcomes are proposed and fitted to the observed data using linear regression. ICE is compatible with experiments that utilize Cas9 from *Streptococcus pyogenes* (SpCas9) and can analyze knock-ins of single nucleotide variants (SNVs) and knock-ins of sequences of up to 270 base pairs.

This protocol provides instructions for how to run the ICE analysis using Sanger sequencing data (.ab1 files) and interprets the results.

For additional resources, EditCo's Genotyping protocol and ICE Knockout Analysis protocol are also available.

Abbreviations:
CRISPR: clustered regularly interspaced short palindromic repeats
Cas9: CRISPR associated protein 9
ICE: Inference of CRISPR Edits
PCR: polymerase chain reaction
Indel: insertion or deletion
HDR: homology-directed repair
NHEJ: Non-homologous end joining

Important Considerations

- ICE can be used to analyze knock-in cell pools (cell populations containing the desired insert, other indels, and wild type sequences), as well as clones (genetically identical cells). Please see the “Knock-in Cell Pools” and “Knock-in Cell Clones” sections throughout the protocol for information pertaining to each type of sample.
- The Knock-in (KI) Score (pg 6) and the list of individual contributions (pg 7) are useful information for quickly understanding the knock-in efficiency and any other sequence mutations in a sample.

Before You Start

- Using the latest version of Google Chrome for launching the ICE online tool.
- Uploading high-quality .ab1 files for the wild type and the edited samples.
- Entering the correct gRNA sequence(s) in DNA or RNA format that was used in the CRISPR experiment.
- Entering a donor template sequence (maximum length is 300 bp with 15 bp of homology arms on both ends).
- Making sure that the same PCR/sequencing primers were used for the control (wild type) and edited sample. Please see EditCo’s Genotyping protocol for general primer design recommendations.
- Opening the .ab1 files and checking that the traces are of the same length as the PCR product (i.e., sequences are untrimmed). If they are not, contact the sequencing company and make sure they send unedited .ab1 files.
- Assessing sequencing quality:
    - Open the .ab1 files that you received from your Sanger sequencing provider. Please note that you may need to download a genome browser software (i.e. 4Peaks and SnapGene) to open these files, if you don’t already have one installed on your computer.
    - You should see single peaks for the wild type trace. See our Help Center Article on sequencing quality for an example of what a good quality wild type sequence should look like.
    - For the edited sample, it is crucial that the **trace to the left of the cut site consists of clean single peaks with minimal background**, so that good alignment between the control and edited trace can be found. Visit EditCo.com/help/sanger-sequencing-quality for examples of high and low quality traces. The cut site is predicted to occur 3 bp upstream of the “NGG” PAM site. Note that the region after the sgRNA cut site is expected to contain multiple overlapping peaks. This is a good indication that the genomic locus was edited and that the cells in the cell population that is being analyzed all have slightly different indels (insertions/deletions), which is expected from non-homologous end joining (NHEJ) repair.

1. Conducting an ICE Analysis

1.1. Launch the ICE Tool & Select Analysis Type

1. Go to the online ICE analysis tool webpage at editco.bio/crispr-analysis
2. Select analysis type:
   - Sample-by-sample: intended for a small number of samples (≤5)
   - Batch: intended for a larger number of samples (>5)

1.2. Upload Files

Sample-by-sample analysis (recommended for up to five control & edited sample pairs)

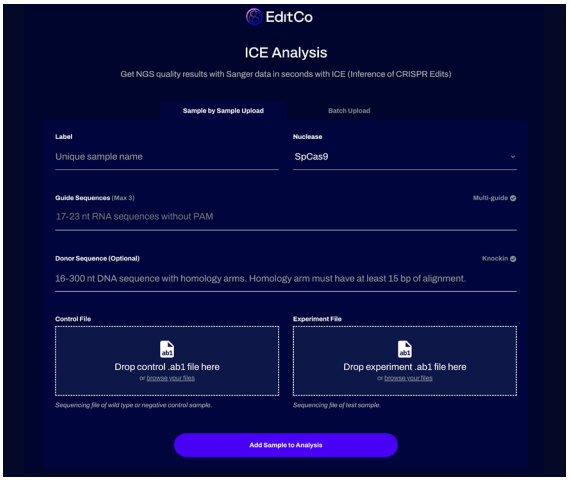
1. Click on the “Sample by Sample Upload” tab.
2. Add .ab1 files by dragging and dropping into the appropriate upload space, or by clicking “browse your files” to select from a list of files on your computer. If the upload file type is correct, the perimeter of the upload space will turn green. If it is not correct, the perimeter will turn red.

   Control File (left): control Sanger sequence file (.ab1 format)
   Experiment File (right): experimental Sanger sequence file (.ab1 format)

3. **Guide sequence:** the 17-23 nucleotide sequence of the DNA-targeting region of the guide RNA excluding the PAM. This can be provided as either DNA or RNA sequence. For knock-in experiments, enter only 1 guide sequence.

   Note: The guide sequence(s) of the gRNAs included in EditCo's products are found on the Quality Control (QC) document shipped with the product.

4. **Label:** a unique name will be automatically generated for each sample using the name of the experimental .ab1 file. These names can be modified to any unique sample name (255 character limit).



5. **Donor Sequence (up to 300 bp):** enter the donor DNA sequence including both homology arms (up to 300 bp). Each homology arm must have at least 15 bp of alignment with the targeted genomic sequence and 50+ bp of alignment is recommended.
6. Click “Add Sample to Analysis”. The files will display in the table called “Your Upload”, which is a running summary of all your uploads.

Additional samples can be added one at a time by filling in the form again with the above information. Each new sample will be added as a new row in the “Your Upload” summary table. Note that knock-in and knock out samples may be analyzed together.

Figure 1. The ICE tool interface. For knock-ins, the ICE tool has fields for the upload of .ab1 control and experimental files, a sample label, the guide sequence, and the DNA donor sequence. Sample-by-sample and batch uploads are possible.

Batch Analysis (recommended for over five editing experiments)

1. On the ICE analysis tool webpage, select the “Batch Upload” tab.

2. Add the Zip archive (containing .ab1 files) and Excel file by dragging and dropping into the appropriate upload space, or by clicking “browse your files” to select from a list of files on your computer. If the upload file type is correct, the dashed line perimeter of the upload space will turn green. If it is not correct, the perimeter will turn red.

   Zip archive containing .ab1 files (left): upload a single .zip file containing experimental and negative control (wild type) Sanger sequence files (.ab1 format). More than one control sequence can be used. Up to 700 samples can be included in the .zip file and analyzed at once (file size limit is 225 MB).

   Definition File (right): a single Microsoft Excel file (.xlsx format only) that lists a unique label, the control file name, the experimental file name, the guide sequence, and the donor sequence. To download a template of the definition file (.xlsx), click “Download Template/Example Files” on the lower left side of the screen. Do not modify the column headers in the template_definitions.xlsx.

   Do not modify the column headers in the template_definitions.xlsx. Follow the instructions below to fill in the template with your file information:

   1. The Label column is used for labeling your samples with a unique name that has a 255 character limit

   2. The Control Files column should contain the name of the AB1 file containing the Sanger sequence for each negative control. The control .ab1 files must be included in the Zip archive

   3. The Experiment Files column should contain the name of the .ab1 file containing the Sanger sequence for each experimental sample. This file must be included in the .zip file.

      Note: With a Mac computer, the file names can simply be selected and copy and pasted into the excel spreadsheet. For copying the file names into an excel spreadsheet in Windows, do the following: 1) Select all the files you want to use for batch analysis 2) Click "Copy path," you can find this button in the Home bar for Windows 10 3) Paste into excel spreadsheet 4) Use “Find and Replace” to delete all the folder names, leaving behind only the "filename.ab1"

   4. The Guide Sequence column should contain the 17-23 nucleotide sequence of the DNA- targeting region of the guide RNA (excluding the PAM) for each sample. This can be provided as either DNA or RNA sequence. ICE was designed assuming that SpCas9 is used as the nuclease. However, ICE does not actually check if the PAM site is NGG, but instead uses the input guide sequence to place the predicted cut site 3 bp upstream of the end of the input sequence. If you wish to analyze data that uses other nucleases, input 17-23 bp of the genomic DNA sequence in the “Guide Sequence” field such that the expected cut site is 3 bp upstream from the end of the sequence.

   5. The Donor Sequence column should contain the donor DNA sequence including both homology arms (up to 300 bp). Each homology arm must have at least 15 bp of alignment with the targeted genomic sequence.

3. To complete the analysis, click “Analyze Experiment.”

1.3. Checking for Errors & Ensuring Reliability

After an ICE analysis is complete the results will load automatically. It is important to make sure that the analysis is completed without any errors and that the results are of good quality.

Failure/ warnings associated with an ICE analysis: If the ICE analysis fails, an error message will pop up with a reason for the failure (refer to Table 2 for an explanation of error messages). If the analysis completes but a warning message appears, refer to Table 3 for an explanation.

Failure/ warnings associated with an ICE analysis: If the ICE analysis succeeds without any error or warning message, follow these guidelines to determine if the results are reliable. If you run multiple samples (edited and control pairs), you will first see a summary window. Click on each sample and check the following:

1. Check the Model Fit R^2 value. If $R^2 < 0.8$, see Table 1 for troubleshooting tips. We generally consider R^2 values of up 0.8 as acceptable. For lower R^2 values, the confidence in the analysis is lower, though some useful information may still be retrieved.
2. Click on the Traces tab and make sure the PAM sequence is "NGG." Make sure the cut site (depicted as a vertical dashed line) is 3 bp upstream of the PAM site.
3. Click on the Contributions tab and make sure that you see some contributions that are greater than 5%. If only small percentages are listed (<5%), they may not be real contributions, and re-sequencing may be needed.
4. Check that the donor template aligns well with the control sample.
5. Click on the Indel Distribution tab and look at the discordance plot on the right side. Make sure the control (orange) and edited (green) graphs are close together prior to the cut site. Also, make sure that the edited graph goes up after the cut site.

For an explanation of analysis results, see the Interpretation of ICE Results section below.

2. Interpretation of ICE Results

2.1 Summary Analysis Window

After an ICE analysis is complete the results will load automatically. For analyses that include multiple samples, a summary window will appear with bar graphs of the results and a list of the analyzed samples (Fig 2). The analysis can be sorted by any of the parameters displayed on the summary table. In order to search for a particular sequence or name, your browser's "Control F" functionality can be used to find a guide or name. Note that the control sequence is not listed in the summary table.

If only one sample was included in your analysis, then ICE will skip the summary and navigate directly to the analysis of that sample (see "Individual Samples" on pg 7). If a sample run has no issues, the summary window will show a green checked circle to the left of the sample name. A orange 'Succeeded' check mark indicates that ICE needed to adjust a particular parameter in order to generate results, and further details may be provided in the form of an error message. If there are no results or there was a processing error, a red 'failed' message will be displayed alongside a message with details on the error type.

Successfully analyzed samples will display the following parameters:

1. **Sample:** The unique label name for each sample.
2. **Guide Target:** The 17-23 bp sequence of the guide RNA(s) that bind to the genomic DNA, excluding the PAM sequence.
3. **PAM Sequence:** The Protospacer Adjacent Motif (PAM) sequence for the nuclease used. Currently, ICE is

configured for the Cas9 nuclease from *Streptococcus pyogenes* (SpCas9). SpCas9 PAM sequence is NGG.

4. **Indel %:** The percentage of sequences that contain an insertion or deletion (indel) in the sample. It includes all sequences that are not wild type, whether they represent a knockout or knock-in mutation. The indel percentage is equivalent to the editing efficiency in a mixed population of cells and a general indication of how well a given gRNA cuts under the present transfection conditions.

5. **Model Fit (R^2):** The R^2 value (Pearson correlation coefficient) is a measure of how well the proposed indel distribution fits the Sanger sequence data of the edited sample. The maximum R^2 value is 1.0 and the sum of all individual contributions will be equal to the R^2 value. For example, if the R^2 is 0.84, then all of the contributions will add up to 84%. The difference between 1.0 and the R^2 (e.g., 100% - 84% = 16%) represents the percentage of Sanger sequencing data that is unexplained and does not match the expected outcomes. The R^2 value is critical for assessing the indel %, KI Score, and KO Score (when applicable), as it sets the maximum value for these metrics. The higher the R^2 value, the higher the confidence in the indel percentage and KI Score. An R^2 value of 0.8 or more indicates a robust analysis, but R^2 values below 0.8 should be considered with caution.

6. **Knock-in Score:** the Knock-in Score is the percentage of sequences in the sample that contains the desired knock-in sequence.

    - *Knock-in Cell Pools:* Knock-in cell pools contain a mix of edited and unedited cells. Of the edited cells, some will contain the desired knock-in fragment, while others will contain indels resulting from NHEJ. The Knock-in Score for pools is typically lower than the indel %.

    - *Knock-in Cell Clones:* Knock-in cell clones are genotypically identical cells. The expected KI Scores for a successfully edited clone (diploid cell line) are as follows:

        - Homozygous knock-in Score: ~100%
        - Heterozygous knock-in Score: ~50%

Note: Lower frequency (<5%) mutations cannot be accurately measured and are either background noise or minor sequences in the PCR pool. The KI Scores listed above are approximations.

Downloading Results: The entire analysis can be downloaded as a .zip file by clicking “Download Analysis Data” on the bottom right of the analysis screen. Your results will also be sent to you by email

Figure 2. ICE summary window. The summary window appears when multiple samples are analyzed simultaneously. The window shows basic information about your edited samples, including the Indel % (indel frequency) and Knockin-Score (% sequences that have desired insert). If knockout samples are also included, then the summary window will display the Knockout-Score (% sequences that are putative knockouts) for those samples.

2.2. Individual Samples

Each sample can be individually inspected in greater detail by clicking on the sample name or on its corresponding bar graph entry. This will open up a new window with four tabs called "Contributions", "Donor Alignment", "Indel Distribution", and "Traces." To return to the summary window, hit the “View Summary” button on the top left of the screen. Below are descriptions of each tab.

1. Contributions Tab

At the top of this tab, the analysis status, guide target, PAM sequence, indel %, model fit (R^2), and Knockin Score are all shown for a particular edited sample. Below this information, the inferred sequences present in your edited population (“Sequences”) and their relative representation in the population (“Contribution”) are shown. The knock-in sequence is highlighted in orange and labeled “HDR” under the “Indel” column.

For each sequence in the population, the number of nucleotides inserted (+) or deleted (-) is indicated in the “Indel” column. An orange “+” symbol on the far left marks the unedited sequence (wild type) if it is present above background frequency. The black vertical dotted line represents the cut site.

- *Knock-in Cell Pools:* For knock-in cell pools (non-clonal populations), there are typically other indels present in the population (besides the desired insert) as a result of NHEJ repair. Thus, the contributions tab will display the sequence with the desired insert, as well as other indels detected in the pool (Fig 3). Keep in mind that ICE cannot accurately detect mutations that are present at 5% or less because Sanger sequencing has some inherent noise. It is not possible to determine whether mutations below this threshold are truly present in the sample.

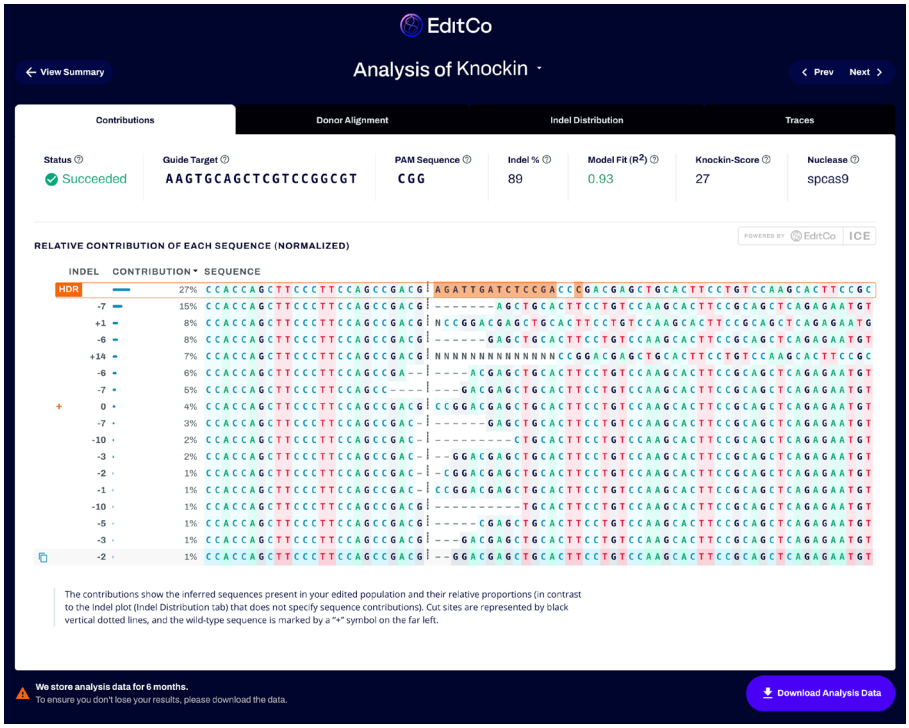
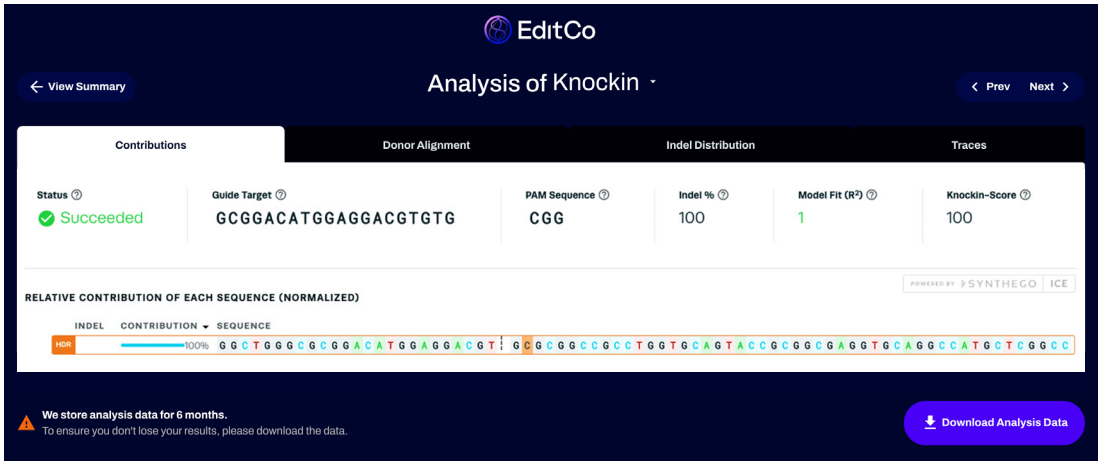


Figure 3. ICE contributions tab for knock-in cell pools. The contributions tab displays the inserted sequence (highlighted in orange and marked by “HDR” on the left), its associated frequency (% contribution). The % contribution of the inserted sequence equals the Knock-in Score. In this example, a sequence is inserted (AGATTGATCTCCGA) and a nucleotide is changed from G to C. The Knock-in Score indicates that 33% of sequences contain this alteration. Other indel outcomes present in the sample are indicated in the “Indel” column. If the unedited (wild type) sequence is present as a contribution, it will be demarcated by an orange “+” sign on the far left.

- *Knock-in Cell Clones:* For knock-in cell clones, the edited cells are genotypically identical. Thus, for a diploid cell line one contribution is expected if the knock-in is homozygous (Fig 4a) and two contributions (one contribution being the knock-in and the other contribution being the wildtype or indel) are expected if the knock-in is heterozygous (Fig 4b).

  Note: If there are more than two main contributions present or the sequences have frequencies that are very different from the expected 50% (heterozygous edit) or 100% (homozygous edit), the cells are either not clonal or they may not be diploid.

a. Homozygous knock-in clone



b. Heterozygous knock-in clone

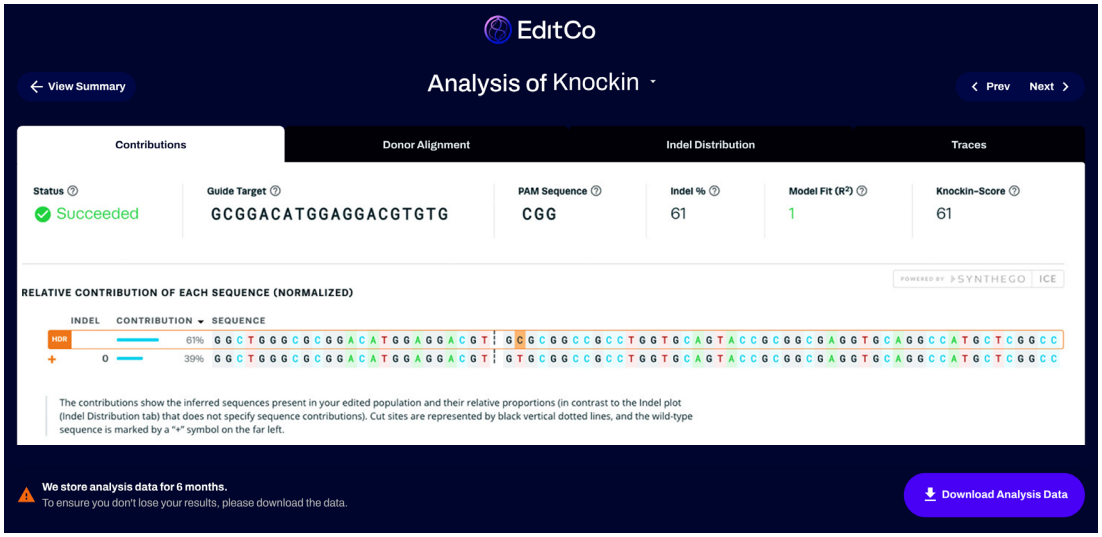


Figure 4. ICE contributions tab for knock-in cell clones.
Because knock-in cell clones are genotypically identical cells, the contributions will reflect whether the clone is homozygous, heterozygous, or compound heterozygous for the inserted sequence or single nucleotide variant (SNV). For a diploid cell line, a homozygous knock-in clone (Figure 4a) has 100% contributions with the desired SNV and a KI Score of 100%. A heterozygous knock-in clone (Figure 4b) has contributions with the desired SNV and wild type sequence. While this example shows a 61% KI Score, typically a heterozygous KI clone has a KI Score of ~50%

2. Donor Alignment Tab

The Donor Alignment tab displays the donor sequence in relation to the control (wild type) sequence. There are two views of this alignment. On the top, the donor and control nucleotide sequences are shown with bases in common linked by short vertical lines. The insertion in the donor sequence is absent in the control sequence (missing nucleotides are indicated by horizontal dashes).

Directly below the donor and control sequences is a “zoomed out” view of the alignment. A rectangular box is situated on this graphic so that one can easily scroll across the aligned sequences.

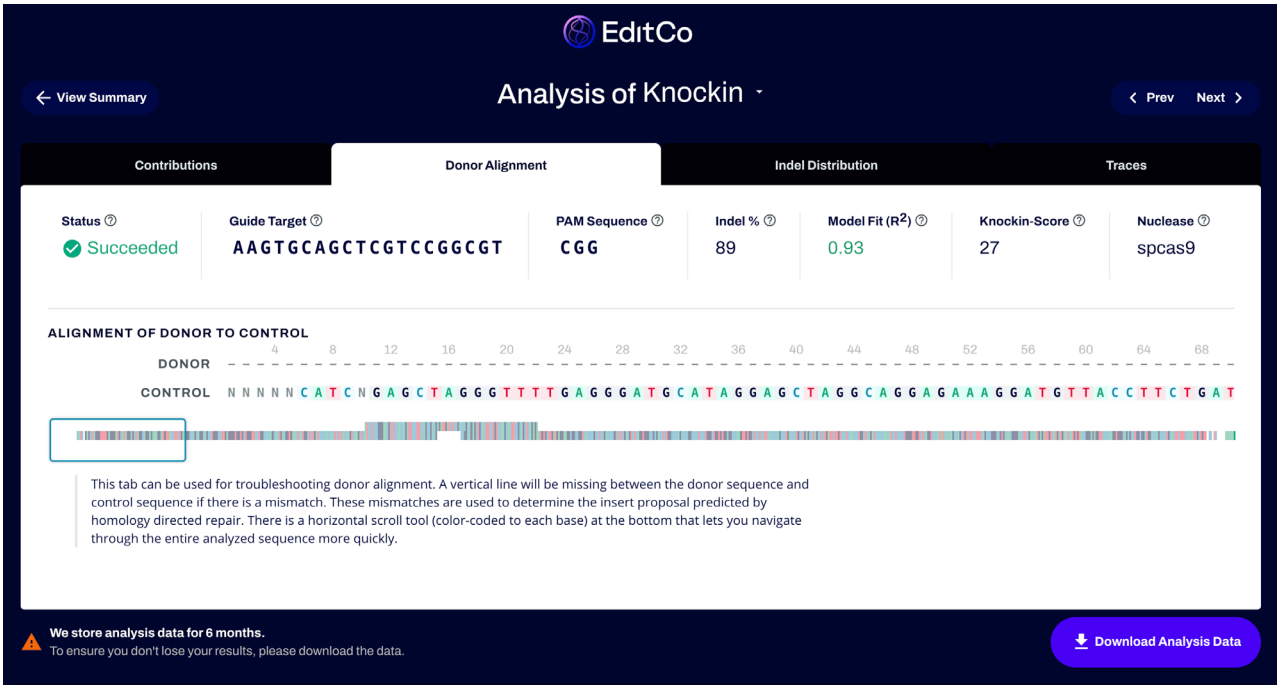


Figure 5. ICE donor alignment tab.
The donor alignment tab portrays the DNA donor sequence (insertion + homology arms) in relation to the control (wild type) sequence. The insert (AGATTGATCTCCGA) is present in the donor sequence but absent in the control sequence. The alignment can be navigated by moving the rectangular box across the graphic at the bottom.

3. Indel Distributions Tab

In this tab, you will find a histogram that displays the inferred knock-in sequence and its frequency in the edited population (Fig 6, left labeled “HDR”). Frequencies of other indels present in the sample (resulting from NHEJ) are also shown. Hovering over each bar of the graph shows the size of the insertion or deletion (+ or -1 or more nucleotides), along with the percentage of sequences that contain it.

The discordance plot shows the level of disagreement between the non-edited wild type (control) and the edited sample in the inference window, i.e., the region around the cut site used for analysis (Fig 6, right). It shows, base-by-base, the average amount of signal that disagrees with the reference sequence derived from the control trace file. On the plot, the green (edited sample) and orange (control sample) lines should mostly overlap before the cut site, and a typical CRISPR edit results in a jump in the discordance near the cut site and continuing after the cut site.

Figure 6. ICE distributions tab.
The histogram on the left shows the percentage of the knock-in insert (labeled HDR) and other indels the sample. On the right, a discordance plot displays the level of alignment between the edited (green) and control (orange) traces before and after the cut site (vertical black dotted line). Typically, the edited and control traces are aligned before the cut site and then become discordant after the cut site.

4. Traces Tab

The Traces tab shows the edited (top) and control (bottom) Sanger traces with the base calls above each trace (Fig 7). The horizontal black underlined region of the control sample represents the guide sequence and the horizontal dotted red-underlined region is the associated PAM site. The vertical black dotted line on both traces represents the cut site. Repair of the double-strand break (via HDR and NHEJ) typically results in a heterogeneous mix of bases downstream of the cut.

Figure 7. ICE traces tab.
The traces tab edited (top) and control (bottom) Sanger traces in the region around the guide RNA binding site (horizontal black line on the control sample). On the control trace, the guide RNA sequence is underlined in black and the PAM site underlined with a red dotted line. The cute site is indicated with a vertical dotted line on both traces.

To return to the summary window containing all the samples, click “View Summary." The “Next” and “Previous” buttons or pressing the arrow keys on your keyboard will take you to the next sample on the summary table.

For additional details on ICE, read our BioRxiv paper.

Troubleshooting Guide

Table 1. Troubleshooting Common Problems

Problem	Possible Cause(s)	Recommended Solutions
Low Sanger Sequencing Quality	Nonspecific amplicons for wild type sample	Check the PCR product on an agarose gel to make sure you are getting a single band. If you get multiple bands, you will have to optimize the PCR by trying different annealing temperatures or trying a different PCR primer set.
	Sequencing primer is too close to the cut site	Design a new PCR/sequencing primer that allows the sequencing primer to be at least 150 bp away from the closest cut site
	Sequences contain nucleotide repeats which can cause sequencing slippage	Try a different sequencing/PCR primers if possible to avoid the repeat regions.
	Sequencing primer is not specific	Design a new nested sequencing primer following the recommendations in EditCo's Genotyping protocol.
	PCR product is not pure enough	Make sure that the PCR product was purified with a PCR clean-up kit or gel-purified. If PCR cleanup was outsourced, consult the Sequencing company to see if DNA purity can be increased.
Low R^2 value (<0.8)	Incorrect gRNA sequence was entered	Make sure that the gRNA sequences that were delivered into the cells were entered correctly and that the PAM site was not included. Check that the inferred cutsite is in the correct place.
	Low Sanger Sequencing Quality	See the above troubleshooting section.
	Unexpected repair outcomes (i.e., deletion of > 40 bp for a single gRNA experiment or deletions > 150 bp for a multi-guide gRNA experiment)	ICE will not be the best approach in this case. Use an alternative method like NGS to confirm these results.
Trimmed Sanger sequence files	Sequencing company may have edited the sequences	Reach out to sequencing company and make sure they send unedited .ab1 files

Table 2. ICE Error Messages
These messages occur when the ICE analysis completely fails.

ICE Error Message	Possible Cause(s)	Recommended Solutions
Guide {guide sequence} not found in control sequence	The incorrect guide sequence was entered	Make sure that the gRNA sequences that were delivered into the cells were entered correctly and that the PAM site was not included.
	The incorrect region of the genome was amplified and/or sequenced	Verify that the gRNA sequence you entered is present in the control .ab1 file by opening the .ab1 file and using the "Ctrl/Command+F" function to search for the entered guide sequence. If you cannot find the gRNA sequence, design new PCR primers that are specific for your target gene.
	Low Sanger sequencing quality	See Table 1 for troubleshooting.
Negative dimensions not allowed	Low Sanger sequencing quality	See Table 1 for troubleshooting.
Sample .ab1 {control sample name} quality scores too low	Low Sanger sequencing quality for experimental .ab1 file	See Table 1 for troubleshooting.
Control .ab1 trace quality scores too low	Low Sanger sequencing quality for control .ab1 file.	See Table 1 for troubleshooting
{Provided guide or donor} contains invalid gRNA sequences	Incorrect guide or donor template sequence was entered	Make sure that the correct gRNA sequence(s) was entered. Double-check donor template and make sure it contains 15 bp of homology arms on both ends.
Control .ab1 {control sample name} quality before cut site is too low	Low Sanger sequencing quality.	See Table 1 for troubleshooting
No alignment found between control and edited sample	Different sequencing/PCR primer was used for control and edited samples.	Double-check the primers used. Repeat the PCR/sequencing reaction for control and/or edited sample using the same PCR/sequencing primer.
	Low Sanger sequencing quality.	See Table 1 for troubleshooting

Table 3. ICE Warning Messages
These messages occur when the ICE analysis is suboptimal, but it does not completely fail.

ICE Error Message	Possible Cause(s)	Recommended Solutions
Padding alignment window to be at least 40bp in length or to start of seq	Sequencing primer is too close to the cut site.	Design a new PCR/sequencing primer that allows the sequencing primer to be at least 150 bp away from the closest cut site.
No PAM up/ downstream of guide {guide sequence}	The incorrect guide sequence was entered.	Make sure that the gRNA sequences that were delivered into the cells were entered correctly and that the PAM site was not included.
	Cas nuclease enzyme other than SpCas9 was used.	ICE has only been tested and optimized for SpCas9, so we cannot guarantee the accuracy of ICE results for other nuclease enzymes (especially if they introduce staggered cuts rather than blunt cuts).

Additional Information

For an up-to-date list of all protocols and other resources, please visit this link.

For technical assistance, contact our Scientific Support Team at technicalsupport@editco.bio.

For common FAQs, please visit this link.