Inference of CRISPR Edits (ICE)

Knockout Analysis

Developed by EditCo

Introduction

CRISPR can be used to knock out a gene of interest so that the corresponding functional protein is no longer produced. These experiments involve transfecting cells with a target-specific guide RNA (gRNA) and a Cas9 nuclease. These components form ribonucleoproteins, which subsequently bind to the targeted genomic locus and induce a double-strand break in the DNA. Non-homologous end joining (NHEJ), a common DNA repair mechanism, is then exploited to generate insertion and/ or deletions (indels) at the cut site. Indels that cause frameshift mutations or nonsense mutations will likely disrupt gene function.

A CRISPR-edited cell population contains a heterogeneous mix of indels and wild type sequences. The percentage of edited sequences is represented as the editing efficiency. If genotypically identical cells are needed, clonal populations can be generated from the knockout cell pool. Monoclonal populations should contain 100% edited cells with the same type of edit.

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) samples. Potential editing outcomes are proposed and fitted to the observed data using linear regression. The ICE report outputs both the total indel percentage and the Knockout Score (the percentage of outcomes that lead to a putative knockout).

This protocol provides instructions for how to run the ICE analysis using Sanger data (.ab1 files) and interpret the results. For additional resources, EditCo's Genotyping protocol and ICE Knockin 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 editing when up to three gRNAs are used. The software can detect indels of up to ~40 bp when 1 gRNA is used, and up to ~150 bp when up to 2-3 gRNAs are used (e.g., Synthego's multi-guide sgRNA). Please see the "Multi-guide Samples" sections throughout the protocol for information relevant to editing with multiple gRNAs or Synthego's multi-guide sgRNA.
- The ICE tool is only compatible with experiments using Cas9 nuclease from Streptococcus pyogenes (SpCas9).
- ICE can be used to analyze knockout cell pools, as well as clones. Please see the "*Knockout Cell Pools*" and "*Knockout Cell Clones*" sections throughout the protocol for information pertaining to each type of sample.
- The Knockout (KO) Score (p. 6) and the list of individual contributions (p. 7) are useful information for quickly understanding the knockout efficiency and types of indels present in a sample.
- ICE is a helpful screening tool that can be used for genotyping analysis, but it is recommended that a gene knockout be confirmed via protein expression or a functional assay.

Before You Start

- Use the latest version of Google Chrome for launching the ICE online tool.
- Ensure wild type and the edited sequencing samples were generated with the same PCR/sequencing primers. Please see the EditCo's Genotyping protocol for general primer design recommendations.
- Be sure to enter the correct gRNA sequence(s) in DNA or RNA format that was used in the CRISPR experiment.
- Open the .ab1 files and check that the traces are of the same length as the PCR product that was sequenced. If they are not, contact the sequencing company and make sure they send unedited .ab1 files.
- Assess 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.
    - ***Control (wild type) sample:*** In the control (wild type) trace, you should see prominent single peaks. See Fig 1 below for an example of what a high quality control sequence should look like.

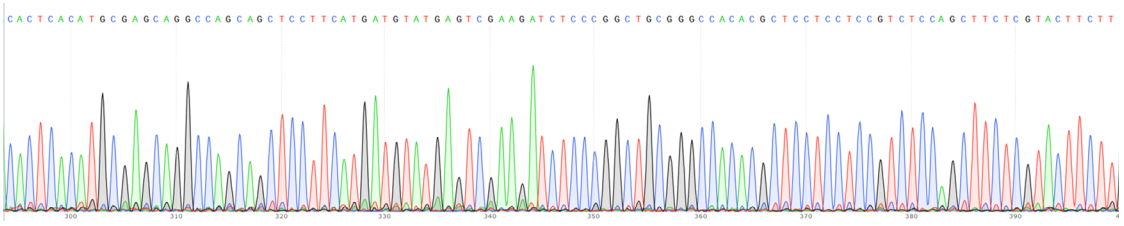

Figure 1.
An example of a high-quality Sanger trace of a control (wild type) sample. Note that mostly a single prominent peak is present for each base call.

    - ***Edited sample:*** 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 traces can be found. 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 editing experiments)

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. 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.

Note: To gain a better understanding of ICE or to practice using the software, you can run a sample-by-sample or batch analysis using our example data. Click “Download Example Files” (on lower left of the screen). Within the “ice_example_files” download folder, the “instructions.txt” file has instructions on how to run the example data. The sample-by-sample examples include low quality, low editing efficiency, and low knockout experimental and control files.

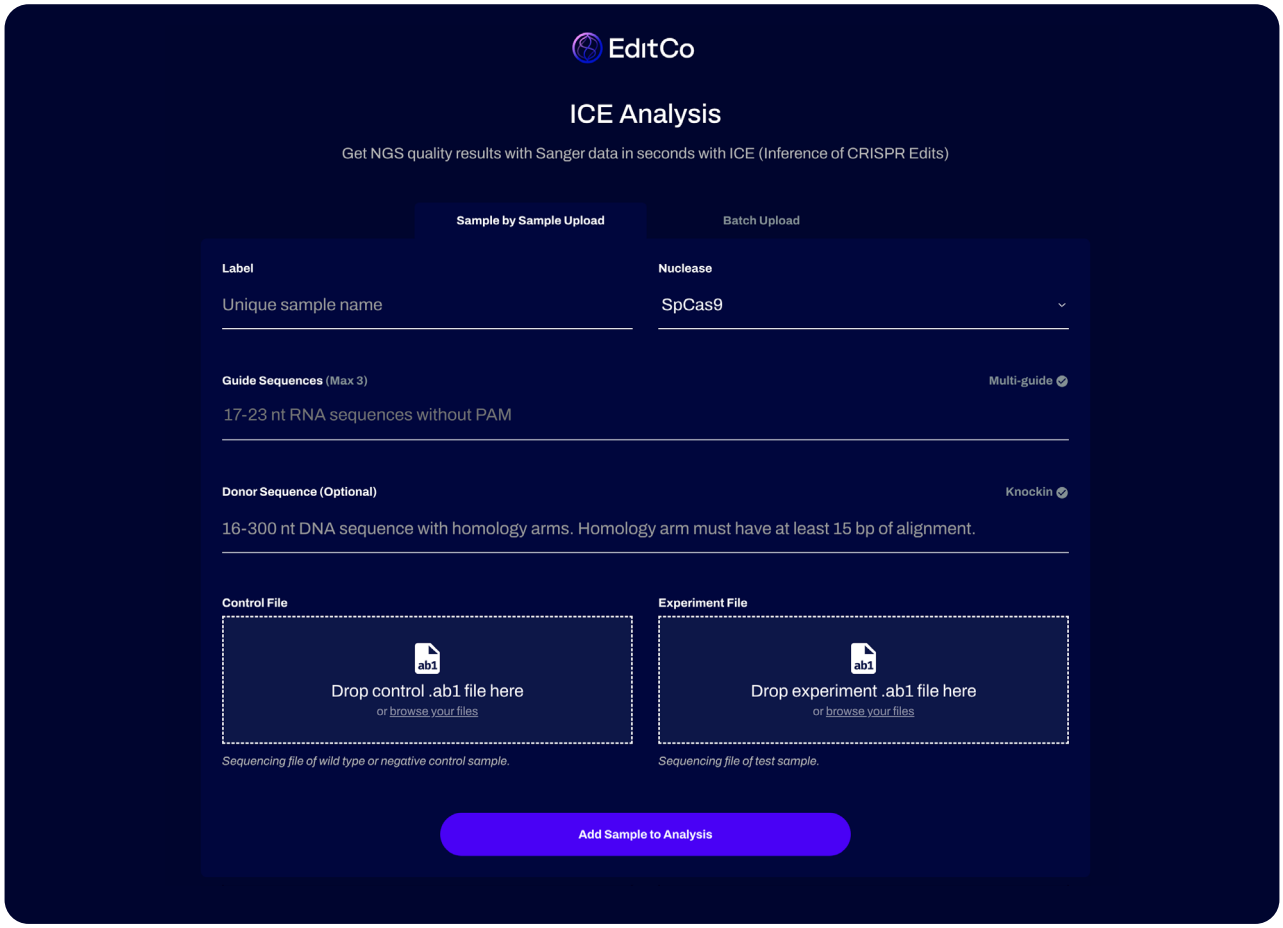

Figure 1. The ICE tool interface.
For knock-ins, the ICE tool has fields for the upload of .ab1 control and experimental files, a sample label, the guide sequence, and the DNA donor sequence. Sample-by-sample and batch uploads are possible.

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. 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. This file must be included in the .zip file.

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 and is therefore optimized for this application only. If gRNA sequences for other nucleases are entered, ICE will give a warning since it cannot find the SpCas9 PAM site (NGG) where it is expected to be (3 bp upstream of the 3’end of the input gRNA sequence). The analysis will proceed despite the error message, but the expected repair outcomes are tailored to SpCas9 and may not be accurate for other nucleases. Please see our help center article for information.

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, a “failed” status message will appear in red. An error message will also pop up with a reason for the failure (see p. 11 for an explanation of error messages). For suboptimal errors, ICE will proceed with analyzing your CRISPR edited samples but can still produce an error message indicated by a status of “succeeded” in orange. If the analysis completes but a warning message appears, see p. 12 for an explanation. See our help center article for more information.

Checking the Reliabilty of Results: If the ICE analysis succeeds without any error or warning message (a status of “succeeded” in green), 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 the R^2 is less than 0.8, the confidence in the analysis is low. Consider troubleshooting the issue (see p. 10 for a Troubleshooting Guide) before proceeding.

2. Click on the Traces tab and make sure the PAM sequence is 5'-NGG-3'. 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.

4. 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 and that the edited graph increases after the cut site.

2. Interpretation of ICE Results

2.1 Summary Analysis Window

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 p. 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 yellow check mark indicates that ICE needed to adjust a particular parameter in order to generate results and a red exclamation point indicates that there was a processing error (no results generated). If yellow or red marks are displayed, hover over them for further information. 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).

4. **Indel Percentage:** 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 %, KO Score, and KI 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 KO 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. **Knockout (KO) Score:** The proportion of sequences that are likely to result in functional protein knockout (frameshift mutations or indels of 21+ bp). The higher the KO Score, the higher the percentage of sequences that result in a knockout of the target gene.

KO Score for Pools vs Clones: Pools contain a mix of edited and unedited cells. Of the edited cells, different alleles of each cell may contain different indels resulting from NHEJ. The KO Score represents the total percentage of sequences in the pool predicted to cause a knockout. Clones are genotypically identical cells. The expected indel frequency for a KO clone (diploid cell line) are as follows:

- Homozygous knockout: ~100%
- Heterozygous knockout: ~50%
- Heterozygous compound knockout (different alleles have different indels): ~100%

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. You will be prompted to enter your email address and a zip file with the ICE results will be emailed to you. Please note that the graphs in the zip file are not identical to the graphs presented in the online analysis, though the results will be the same.

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 three tabs called “Contributions,” “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 Knockout Score are all shown for a particular edited sample (Fig 3). Below this information, the inferred sequences present in your edited population (“Sequences”) and their relative representation in the population (“Contribution”) are shown. For knockout cell pools (non-clonal edited cell populations), there is typically a mix of different indels present in the sample as a result of NHEJ repair. From the list of relative contributions, one can identify the major sequences detected by ICE. 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.

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

   Clone Samples: 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.

   Multi-guide Samples: For samples edited by multi-guide sgRNA, the vertical dotted line indicates the most upstream cut site. In the “Indel” column, the type and size of indel are indicated, along with information about insertions or deletions generated by the different sgRNAs (e.g., g1, g2, g3). Large deletions (21+ bp) are labeled with the word, “fragment deletion” along with the number of nucleotides deleted.

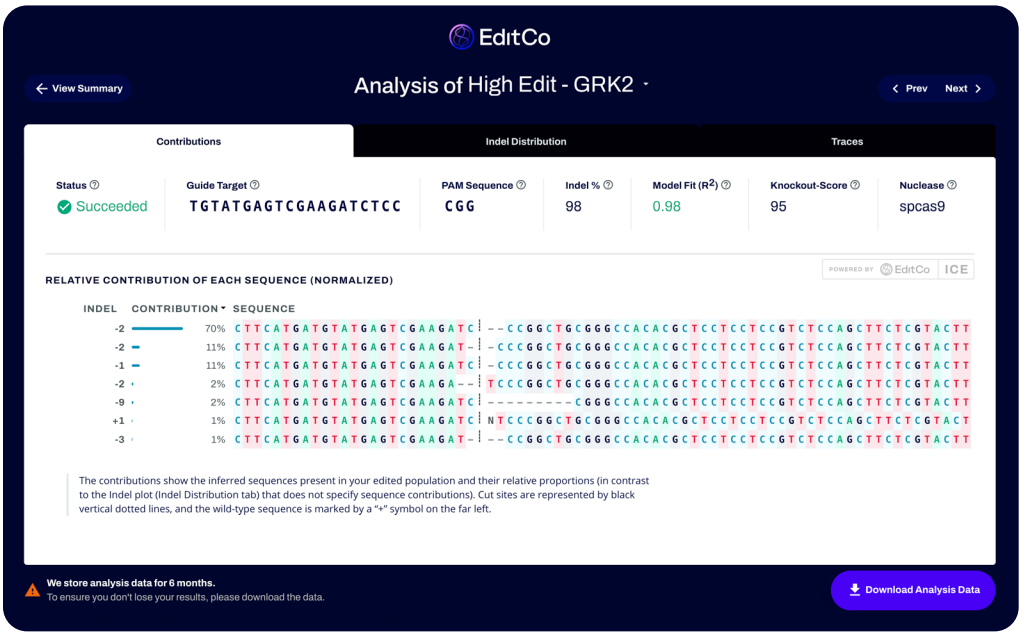

Figure 3. Contributions tab. The contributions tab displays all of the indels and their associated frequencies (% contribution) of a particular sample. The guide sequence, indel %, and Knockout Score, and other information about the sample are also presented. If the unedited control (wild type) sequence is present at a high frequency in the sample, it will be demarcated by an orange “+” sign on the far left. However, if the wild type sequence is present at a very low frequency or absent, then it will not appear as a contribution (as in this example).

2. **Indel Distributions Tab**

In this tab, you will find a histogram that displays the inferred distribution of indel sizes in the entire edited population (Fig 4, left). 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 4, 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.

Multi-guide Samples: Multiple cut sites (vertical dotted lines) will be displayed on the discordance plot for multi-guide samples.

Figure 4. Indel distributions tab.
The histogram on the left shows the percentage of each indel type in 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.

3. **Traces Tab**

The Traces tab shows the edited (top) and unedited control (bottom) Sanger traces in the region around the guide RNA binding site. The sequence base calls from the .ab1 file are also shown above each trace (Fig 5). In the control sample, the guide sequence is marked by a solid black underline and the PAM site marked by a dashed red underline. The vertical black dashed line on both traces represents the cut site. Non-homologous end-joining (NHEJ) repair typically results in a heterogeneous mix of bases downstream of the cut.

Clone Samples: It is best practice to compare the .ab1 traces of clones to the ICE traces on this tab to make sure that they agree well with one another.

Multi-guide Samples: Multi-guide samples have multiple guides, PAM sites, and cut sites displayed in this tab.

Figure 5. 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 cut 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.