ICELL8® scTCR Analyzer User Guide

I. Introduction

The ICELL8 scTCR Analyzer software is only for use with the ICELL8 Human TCR a/b Profiling and ICELL8 cx Human TCR a/b Profiling workflows.

This software uses paired FASTQ files directly generated from an Illumina® MiSeq® sequencer as input. The FASTQ files are then demultiplexed, based on barcodes included in the Read 1 sequences (the first 10 nucleotides), to generate cell-level FASTQs. The ICELL8 scTCR Analyzer then conducts a two-step procedure:

- The primary step:
  - Aligns reads to reference V, D, J and C gene segments of T-cell receptor
  - Assembles the aligned reads to extract *CDR3* gene region
  - Generates a whole-chip clonotype summary report
- The secondary step is a quality control step:
  - Filters out clonotypes with relatively poor quality due to deletion and/or frameshift in *CDR3* amino acid sequences
  - Based on negative controls included on the chip

The analyzer reports filtering statistics and produces tables extracting the top three clonotypes. The analyzer also generates files that could be used as input for data visualization—for example, a VJ join chord diagram.

The ICELL8 scTCR Analyzer is written in Python and Java, with sequence alignment and clonotype assembling conducted by MiXCR. A graphical user interface (GUI) is provided for Windows and MacOS users. Linux users launch the analyzer with a simple configuration file.

II. Before You Begin

A. Supported Operating Systems

The ICELL8 scTCR Analyzer software has been tested and supported on the following operating systems:

- Windows: Windows 7 and Windows 10
- Linux: Centos 6, Centos 7
- Mac: OS X El Capitan (version 10.11 and up)

B. Hardware Requirements

The ICELL8 scTCR Analyzer software requires an Apple Mac, PC, or Linux computer with 16 GB RAM and 100 GB of hard drive free space.

C. User Account Requirements

When installing the software, the account being used should have full administrator privileges for the installation directory. If other user accounts need to have access to run the software, install so all users can use it and not just the installing account.

D. Additional Hardware and Software Dependencies and Recommendations

- **ICELL8 cx Single-Cell System or ICELL8 Single-Cell System (required)**
- **MiXCR (third party dependency)**

  The primary analysis step for sequence alignment and clonotype assembling is conducted by MiXCR. MiXCR is available per the terms of use at Github: https://github.com/milaboratory/mixcr/releases.

  NOTE: Please make sure to download the .zip file or .tar.gz file corresponding to the operating system the analyzer software is installed on. The ICELL8 scTCR Analyzer is compatible with MiXCR version v2.1.8. Other versions may work with the analyzer or have unexpected results.

- **VDJviz Browser (optional)**

  VDJviz is a web-based application that allows graphical analysis of Rep-seq data. The ICELL8 scTCR Analyzer software creates output data files suitable for import into the tool. VDJviz is open source and can be accessed at https://vdjviz.cdr3.net/.

E. Required Input Files

- Paired-read FASTQ files
- CellSelect Well List File—A well selection text file derived from the ICELL8 CellSelect® software, it contains well-level sample information. A typical well selection text file generated by CellSelect should have no more than 1,728 wells.

F. Verifying Python & Java installation

The ICELL8 scTCR Analyzer is written in Python 3 and Java and is compatible with any operating system where Python 3.6 and Java 1.7 or newer is installed.

To check the installed versions of Python and Java:

1. Open a terminal window on the computer on which the ICELL8 scTCR Analyzer software will be installed:

   a. **Windows:** Click the **[Windows]+[R]** buttons and type `cmd` in the Run box.

   Alternatively, select the Command Prompt from **Start > Command Prompt**.

   b. **Linux (Ubuntu):** Use the keyboard shortcut **[Ctrl]+[Alt]+[T].**

   Alternatively, you can find the Terminal by opening the Dash (upper left on most desktops), typing `terminal` and selecting the Terminal application.

   c. **Mac:** The Terminal application is typically found under **Applications > Utilities > Terminal.** Alternatively, search for `terminal` in Spotlight search.

2. After opening a terminal window, verify the version of Java installed by typing:

   ```
   java -version
   ```

   into the terminal window. Text similar to Figure1 should display.

Figure 1. How to verify the Java version of your OS. After typing java -version at the prompt, you should see a version number displayed. Verify that the first two numbers are 1.7 or higher.

3. In the same terminal window, verify the version of Python installed by typing

   ```
   python -V
   ```

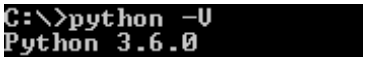


Figure 2. How to verify the Python version of your OS. After typing `python -V` into your terminal, you should see a version number displayed. Verify that it is 3.6.x or higher.

4. The ICELL8 scTCR Analyzer also requires the following Python modules be installed on the computer:

Table 1. Required Python modules.

collections	csv	getopt	glob
logging	numpy	os	pandas
platform	re	subprocess	sys

   To verify if a certain module is installed, type **`python`** in the terminal window, then

   ```
   python [module name]
   ```

   If a module is not installed, a `ModuleNotFoundError` will be reported.

   For example:

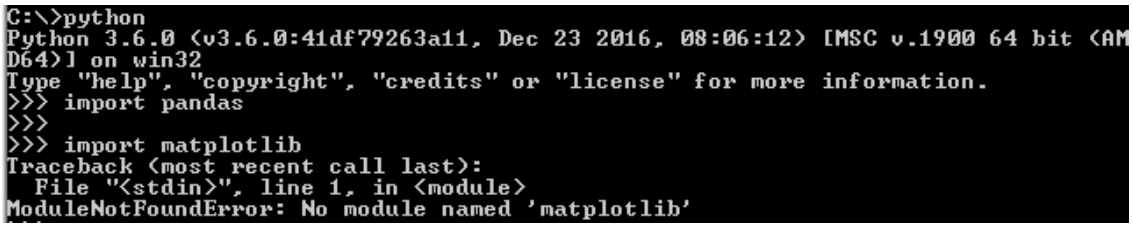


Figure 3. How to verify a Python module is installed. After typing `python [module name]` at the prompt, a module that is not installed will report "ModuleNotFoundError". Here, the module "pandas" is installed; the module "matplotlib" is missing ("matplotlib is not used by the ICELL8 scTCR Analyzer).

   Most packages listed in Table 1 are bundled with the Python distribution, although pandas and numpy will likely need to be installed. For assistance installing these modules individually, please refer to the website https://docs.python.org/3/installing/.

III. Install the ICELL8 scTCR Analyzer

The ICELL8 scTCR Analyzer software is available for download as a compressed file from the ICELL8 software portal takarabio.com/ICELL8-software.

NOTE: Please make sure to specify and download the `.zip` file or `.tar.gz` file corresponding to the operating system the analyzer software will be installed on.

To install:

1. Download or copy the compressed file (`.zip` file or `.tar.gz` file) to the computer on which the software will be installed into the directory you want to install it.
2. Unzip the archive in the installation directory.

   After unzipping, there will be a directory `ICELL8_scTCR_Analyzer\` and four sub-folders within it: `bin\`, `config\`, `test\` and `tools\`.

   NOTE: Windows and MacOS versions will also have an executable GUI launcher, `ICELL8_scTCR_Analyzer.exe` (Windows) or `ICELL8_scTCR_Analyzer` (Mac), and a GUI icon image, `config_icon.png`, in this folder.

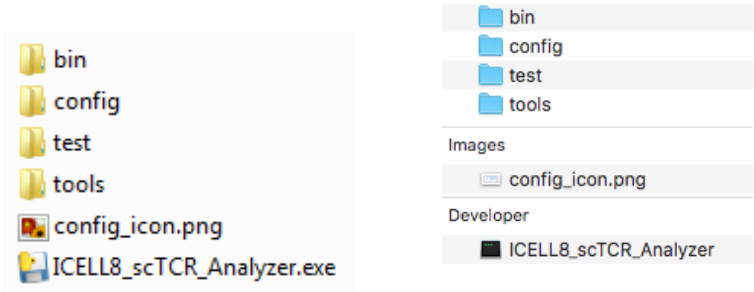


Figure 4. The sub-folder and files list of the `ICELL8_scTCR_Analyzer\` folder for Windows (left) and Macintosh (right).

3. Install MiXCR into the `tools\` directory.

   The resulting folder will be named `mixcr-[version]\` (E.g., `mixcr-2.1.8\`) and will contain the folder `libraries\` and three files: `LICENSE`, `mixcr`, and `mixcr.jar`.

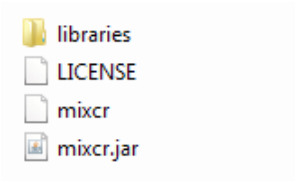


Figure 5. The `mixcr-[version]\` directory sub-folder and files list

Reference Section II.D for more information about obtaining MiXCR.

IV. Test Run with Sample Data

Sample data is provided in the `test\sample_input\` folder and should be used to confirm installation and operation of the ICELL8 scTCR Analyzer prior to usage with customer data.

The sample data directory includes the following files:

1. **`mini_read1.fastq`** & **`mini_read2.fastq`**: Paired-read FASTQ files
2. **`mini_well_list.txt`**: A well selection text file derived from the CellSelect software, it contains well level sample information. A typical well selection text file generated by CellSelect Software should have no more than 1,728 wells.

NOTE: The example data set contains sequences from 20 wells, with 6,876 total read pairs. This representative dataset was a small subset acquired from a library synthesized using the ICELL8 ICELL8 scTCR a/b Profiling Kit and sequenced on Illumina MiSeq.

A. For Windows and MacOS users

The ICELL8 scTCR Analyzer runs from a Graphic User Interface (GUI). Users can do a test run of the sample data with the following steps:

1. Double click the ICELL8 scTCR Analyzer executable file in the `ICELL8_scTCR_Analyzer\` folder, which brings up the GUI.

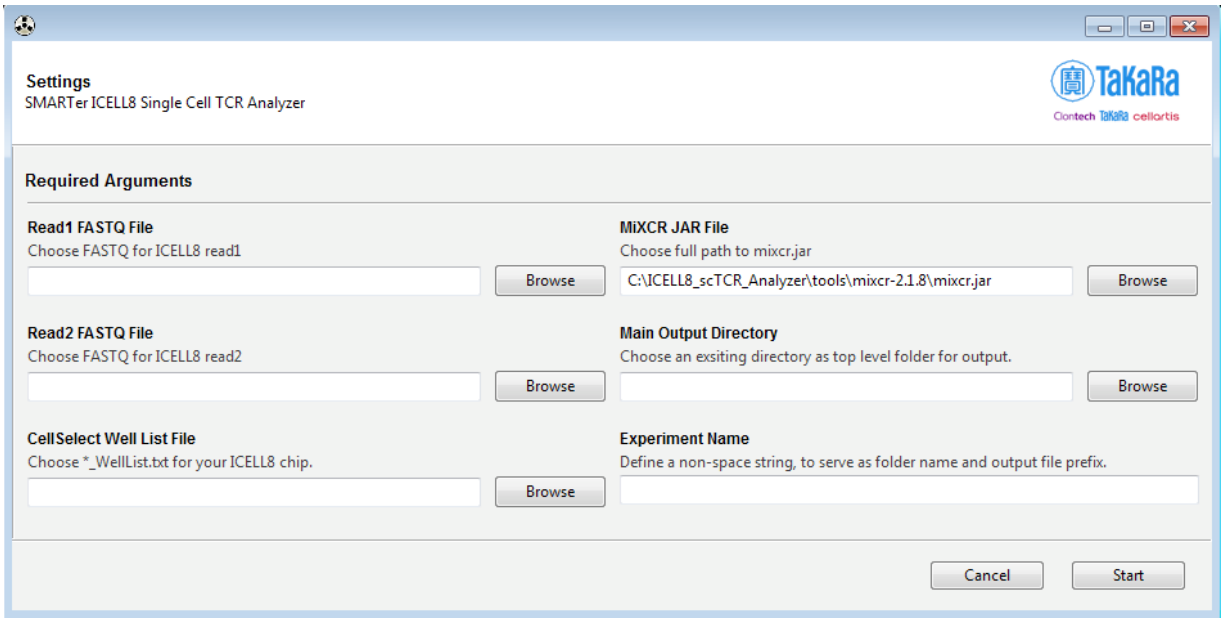


Figure 6. Initial *Settings* GUI page of the ICELL8 scTCR Analyzer.

2. In the *Settings* page (Figure 6), use the [Browse] button of each field to configure the following information.

   NOTE: The configuration field text values should only include alphanumeric characters, hyphens, and/or underscores. Spaces or other special characters should not be used; they may cause the analyzer to behave unreliably or crash.

   a. Read 1 FASTQ File: Browse and choose **`mini_read1.fastq`**
   b. Read 2 FASTQ File: Browse and choose **`mini_read2.fastq`**
   c. CellSelect Well List File: Browse and choose **`mini_well_list.txt`**

d. MiXCR JAR File: After installation, this text box would be auto-filled with the full path of the **`mixcr.jar`** file in the path `ICELL8_scTCR_Analyzer\tools\mixcr-2.1.8\` folder; if another version is being used, browse and choose that **`mixcr.jar`** file

e. Main Output Directory: Browse and choose an existing directory to use as the top-level folder for the analyzer output files

f. Experiment Name: A character string without spaces. This string will name a sub-folder created in the Main Output Directory and is prepended to many of the report/output file names generated by ICELL8 scTCR Analyzer

E.g., The Experiment Name "analysis20180531" will become the name of a new folder `analysis20180531\` and the prefix for a file `analysis20180531_simplified_top3.csv`

NOTE: In order to prevent output files from being overwritten, a unique Experiment Name identifier must be specified for each analysis run (i.e., the text in the Experiment Name field does not match any of the existing sub-folder names in the main output directory). If a duplicate name is given as input, the ICELL8 scTCR Analyzer will display the error `Analysis dir already exists`, and the program will exit without performing the analysis.

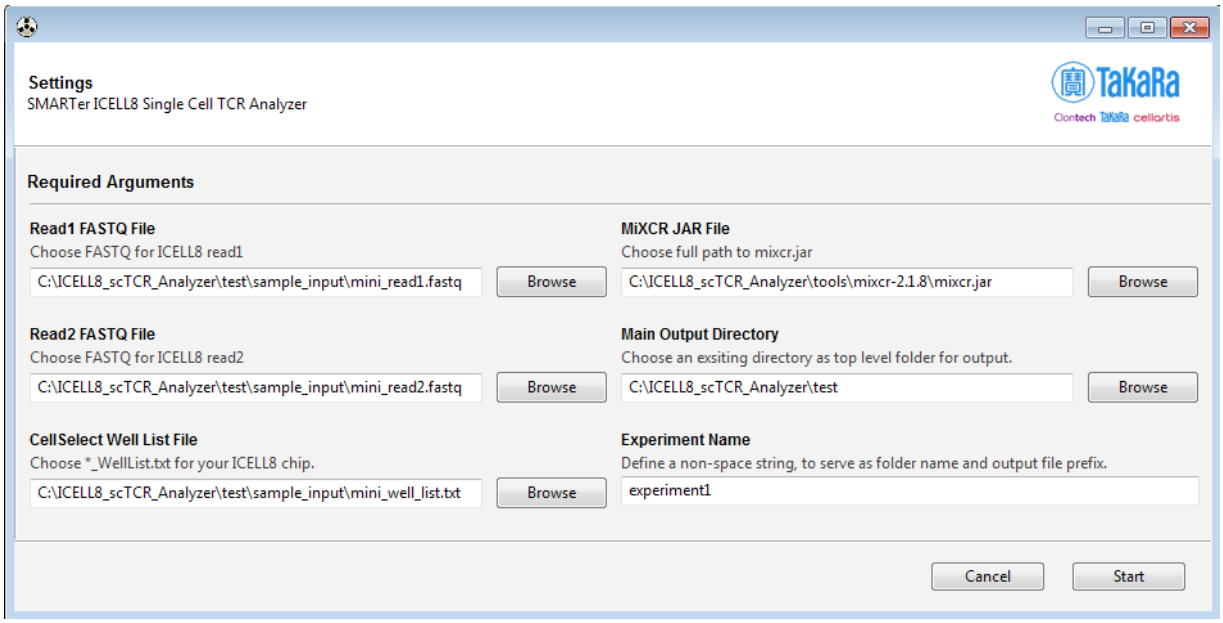


Figure 7. Example of the ICELL8 scTCR Analyzer GUI selection.

3. Once all parameters are populated, click **[Start]** to begin the analysis procedure.

   The ICELL8 scTCR Analyzer reports progress to the GUI.

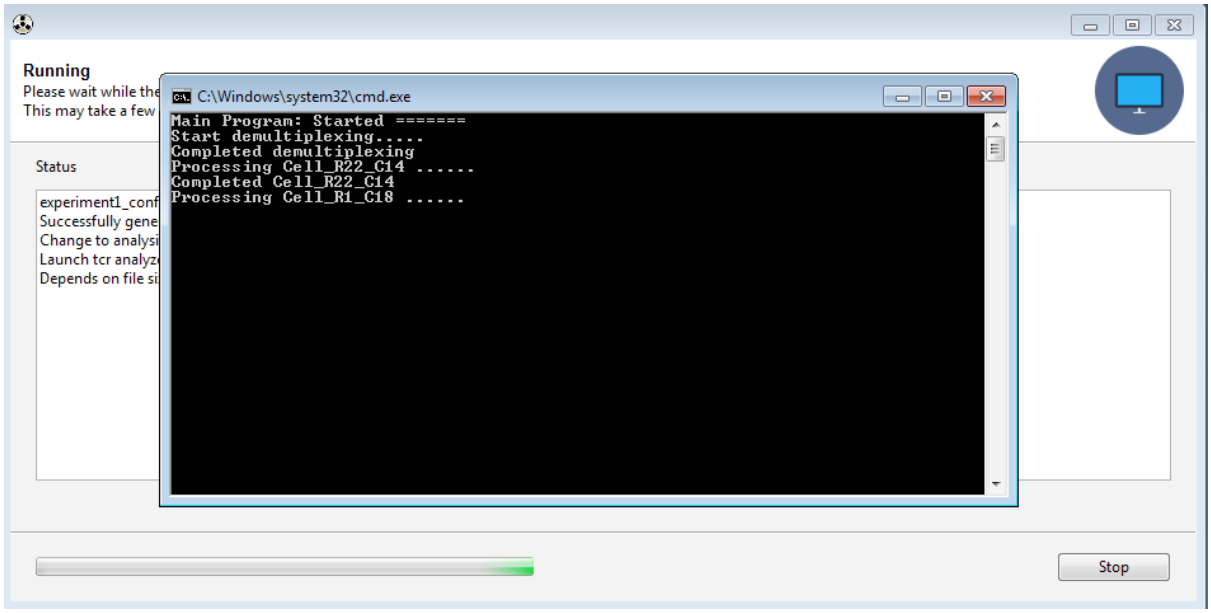


Figure 8. Example of the ICELL8 scTCR Analyzer GUI report, analysis in progress. Windows users will see two progress windows on top of each other, as shown. MacOS users will only see a single window.

4. Once the analysis is finished, a pop-up window will display the message "Program completed successfully!

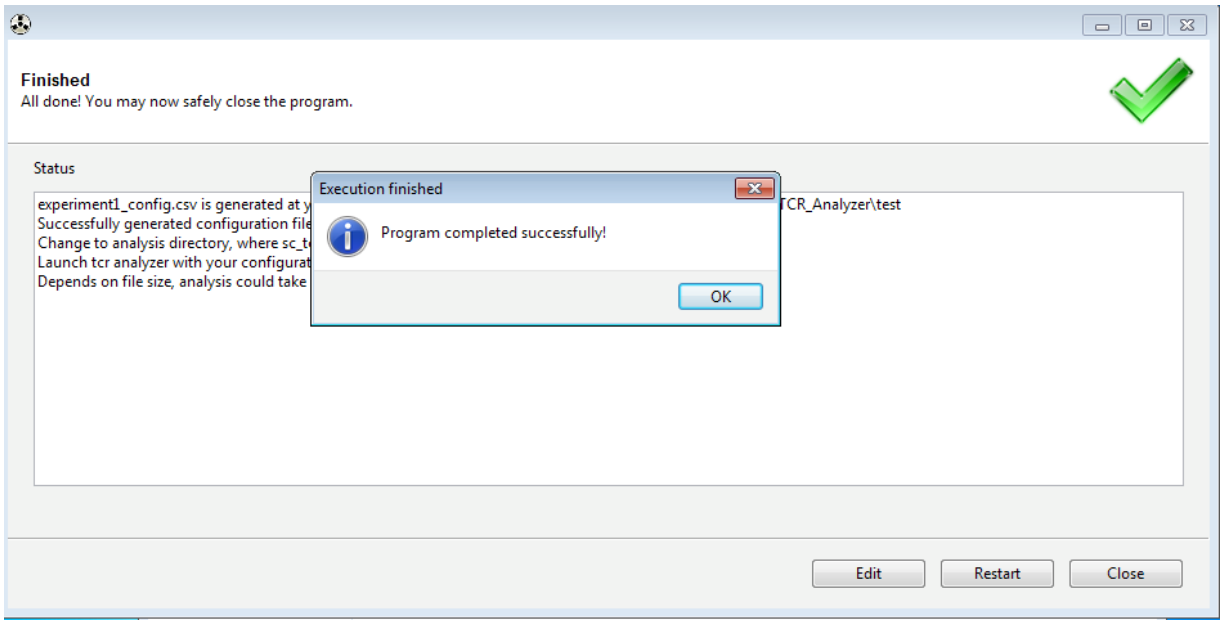


Figure 9. Example of the ICELL8 scTCR Analyzer GUI report, execution finished.

B. For Linux Users

With Linux, the ICELL8 scTCR Analyzer is launched via command line with a call to a comma-separated (CSV) configuration file which specifies the parameters for the analysis.

The configuration file contains three sections:

1. INPUT
2. OUTPUT
3. REFERENCES and TOOLS

Each section contains predefined keywords and their values in two columns. The values should be modified to specify file locations or parameters for the user's analysis; the predefined keywords should not be changed. Any line starting with # is ignored, as in many programming languages, so custom comments can be added to the file.

The ICELL8 scTCR Analyzer bundle includes a configuration template file, `sample_config.csv`, in the `test/` folder that will need to be customized before use. Details of keywords and values in the configuration file are listed in Table II.

Table 2. Summary of keywords and values in configuration file.

Section	Keyword	Value
#INPUT	read1_fastq	The full path to the ICELL8 Read 1 FASTQ file
	read2_fastq	The full path to the ICELL8 Read 2 FASTQ file
	selected_barcode_list	The full path to the well selection text file of the current experiment
	printed_barcode_list	The full path to `printed_list_1728.txt`, located in `ICELL8_scTCR_Analyzer\config\`
#OUTPUT	analysis_dir	The full path to a new output folder to be created which will store analysis outputs.
	experiment_name	A name string defined by the user which will be used as a prefix prepended to output files. The name string can include alphanumeric characters, hyphens, and/or underscores; no spaces or other special characters.
#REFERENCES and TOOLS	species	`hsa`
	demux_tool_loc	The full path to `ICELL8TCR_Demux.jar`
	mixcr_loc	The full path to `mixcr.jar`

For the purposes of this section, the following items are defined:

- `[install_dir]`—The top-level installation location for the ICELL8 scTCR Analyzer software
- `[main_dir]`—The directory in which the user will perform the test run with sample data
- `[output_folder]`—A folder defined by the user which will be created under `[main_dir]` during the analysis process

NOTE: The path defined by `[main_dir]/[output_folder]` is equivalent to the value of the keyword `analysis_dir` in the configuration file (see Table 2).

In order to prevent output files from being overwritten, the `[output_folder]` name must be unique for each analysis run (i.e., a sub-folder matching that name must not exist in `[main_dir]` prior to the analysis run). If a duplicate name is given as input, the ICELL8 scTCR Analyzer will display the error `Analysis dir already exists`, and the program will exit without analyzing.

Linux users can run the sample data using the steps below:

1. Create a local copy of the configuration file for the test run

   ```
   cp -i [install_dir]/test/sample_config.csv [analysis_dir]/
   ```

   NOTE: The configuration file should be edited so that all entries in it which contain a file path are updated to refer to the appropriate locations in the user's installation environment.

2. Change the working directory at the terminal prompt to the `[analysis_dir]`.

3. Execute the following command to run the Analyzer on the test data

   ```
   [install_dir]/bin/sc_tcr_analysis.py -c sample_config.csv
   ```

4. Use the following command to compare the CSV report file results from Step 3. with the expected regression test results

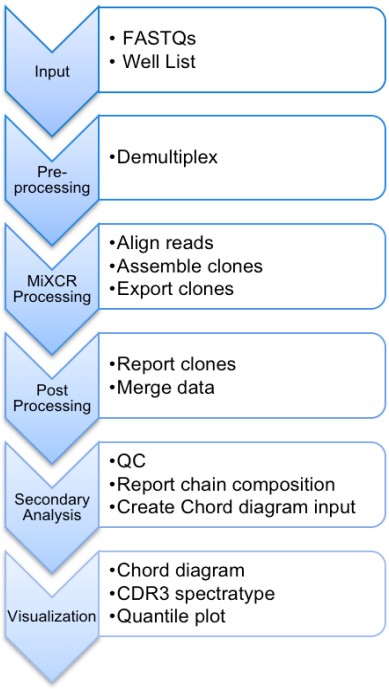
   ```
   diff [analysis_dir]/[output_folder]/[experiment_name]_summary.csv \
   [install_dir]/test/sample_output/sample_output_summary.csv
   ```

   - If these files match exactly, the regression test was successful. If not, begin troubleshooting by looking at the analysis log file in `[analysis_dir]/sc_tcr_analysis.log`
   - Further troubleshooting is beyond the scope of this document.

V. ICELL8 scTCR Analyzer Workflow

The ICELL8 scTCR Analyzer processes data via primary and secondary analysis.

A. Primary Analysis Steps



The primary analysis includes three steps:

1. Pre-processing:
   a. Demultiplex paired-end read FASTQs into cell-level paired FASTQs
2. MiXCR processing, which uses MiXCR to:
   a. Align reads to reference V, D, J, and C gene segments of T-cell receptor
   b. Assemble the aligned reads to extract *CDR3* gene region
   c. Export clonotypes with nucleotide sequences and amino acid sequences
3. Post-processing:
   a. Converts the well-level clonotype report to a format which can be uploaded to VDJviz (https://vdjviz.cdr3.net/) for visualization (chord diagram, *CDR3* spectra type, and quantile plot)
   b. Merges well-level data and generates a clonotype summary report

Figure 10. The ICELL8 scTCR Analyzer data processing workflow.

B. Secondary Analysis Steps

Quality control occurs during the secondary analysis step, filtering samples on the chip based on the negative controls used, if any.

- If negative controls were included on the chip:
  1. The analyzer calculates mean counts and standard deviation of the top clonotype of all negative control samples and uses the resulting mean + 3x standard deviation as a cutoff threshold for clonotype filtering.
  2. For every single-cell sample on chip, the analyzer scans its top 3 clonotypes to check if any of them have counts less than the cutoff calculated in Step 1. If yes, the corresponding clonotype is removed, and its count number is reset to 0.
  3. The analyzer then scans all remaining top three clonotypes for their amino acid sequence. If there are any deletions or frame shifts identified, the corresponding clonotype is also removed, and its count is reset to 0.
  4. If all of the top 3 clonotypes are removed in Step 3, the sample is dropped.
  5. The analyzer summarizes all remaining cells into a simplified summary file, records chain composition of the simplified chip-wise result, and records filtering statistics.
- If there are no negative controls on the chip, then the cutoff threshold for clonotype filtering (result of Step 1) is set to 0. Steps 2-5 remain the same.

C. Additional Data Processing

The analyzer also subsets results by cell type and chain type (alpha chain or beta chain) and generates chip-wise tables that could be uploaded to VDJviz for visualization.

For every possible cell type/chain type combination detected, a gene VJ join cell count table is produced to summarize the number of cells with various VJ usage.

VI. ICELL8 scTCR Analyzer Output

The following section lists the output files of the ICELL8 scTCR Analyzer for the Primary and Secondary Analysis steps in Section V. For more information about the information contained in these files, please refer to Appendix A.

A. Primary Analysis Output

- `sc_tcr_analysis.log`—Log file containing the ICELL8 scTCR Analyzer's processing information
- `[experiment_name]_config.csv`—Configuration file recording all the parameter settings the user selected via the GUI (only for Windows/MacOS; Linux OS uses this file as input)
- `[experiment_name]\`—Folder named by Windows/MacOS user in the GUI configuration steps, or LinuxOS users in the configuration file. It is created in the Main Output Directory and stores the results output files of the ICELL8 scTCR Analyzer, including five sub-folders and four files.

  Folders:

  - `split_fastqs\`—Contains cell level pair read FASTQs after demultiplexing and a log file (`*_demultiplexing.log`) which records error messages (if any).
  - `align\`: Contains cell level alignment reports (`*_align_report.txt`) and MiXCR alignment file (`*.vdjca`)
  - `assemble_clones\`—Contains cell-level clone assemble reports (`*_clones_report.txt`) and MiXCR assemble file (`*.clns`)
  - `export_clones\`—Contains cell level clone summary files (`*_clones_all.txt`)
  - `vdjviz_input\`—Contains cell level clonotype files (`*vdjviz.txt`) that could be loaded to VDJviz for visualization (see Section VII for details).

  Files:

  - `[experiment_name]_summary.csv`— File summarizing all cells and their top 3 clonotypes, aggregating output from MiXCR.
  - `[experiment_name]_mixcr.stdout`—Log file containing details of the MiXCR procedures
  - `[experiment_name]_mixcr.stderr`—Text file containing all error messages produced by MiXCR (if any). For a successful analysis, this file will be empty.
  - `config.bak`—Text file which stores all parameter settings the user selected via the GUI (i.e., a configuration file)

B. Secondary Analysis Outputs

The ICELL8 scTCR Analyzer's secondary analysis outputs are all in the `[experiment_name]folder\` named by the user and includes:

- `[experiment_name]_simplified_top3.csv`—Simplified top3 clonotype summary file.
- `[experiment_name]_cell_filter_table.csv`—File containing step by step cell quality control filtering statistics.
- `[experiment_name]_chain_composition_tables.csv`—Summarizes the overall chain composition; if there are no samples remaining after the secondary analysis quality control filtering (Section V. B Step 4, above), this file will not exist.
- `[cell_type]_[chain_type]_vdjviz.txt`—A VDJ gene usage summary table which can be uploaded to VDJviz for whole chip overall chain composition visualization.
- `[cell_type]_[chain_type]_vj.txt`—A by cell type, by chain type cell count table, summarizing number of cells with certain VJ gene composition.

VII. Data Visualization

The ICELL8 scTCR Analyzer generates two levels of VDJ gene summary tables for data visualization:

- The chip level table is generated under the `[experiment_name]` folder and named as `[cell_type]_[chain_type]_vdjviz.txt`
- The single cell level table can be found in the `vdjviz_input\` folder inside the `[experiment_name]` folder

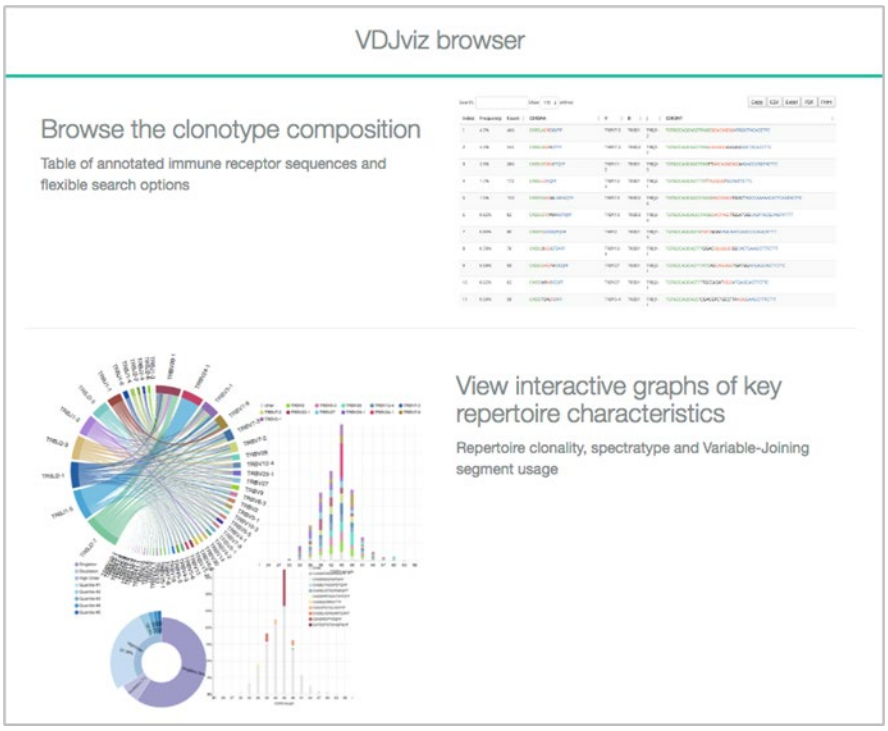


Figure 11. A sample view of VDJviz Browser

New users can register and create an account to use the VDJviz browser:

1. At the website, click the "Browser" button on top-left corner, which leads to a login page.
2. Follow the directions on the page to register for an account.
3. Once registration is completed, login in with your username and password
4. After login, on the left panel, choose "Upload new samples".
5. Select and upload any `[cell_type]_[chain_type]_vdjviz.txt` to start visualization.

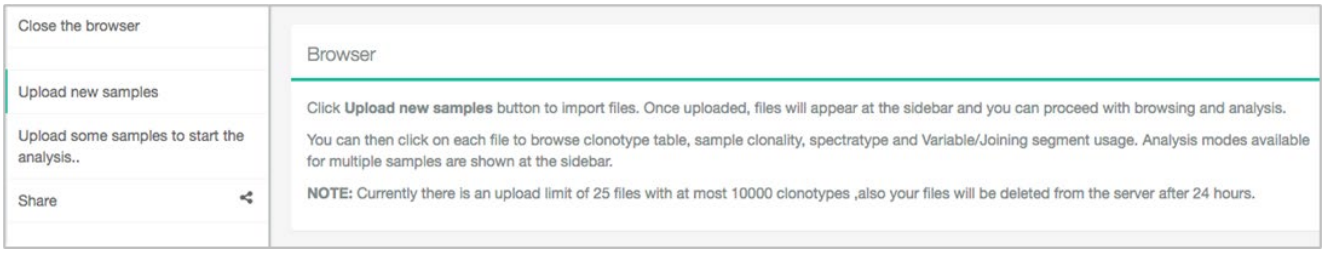


Figure 12. VDJviz Browser sample upload page

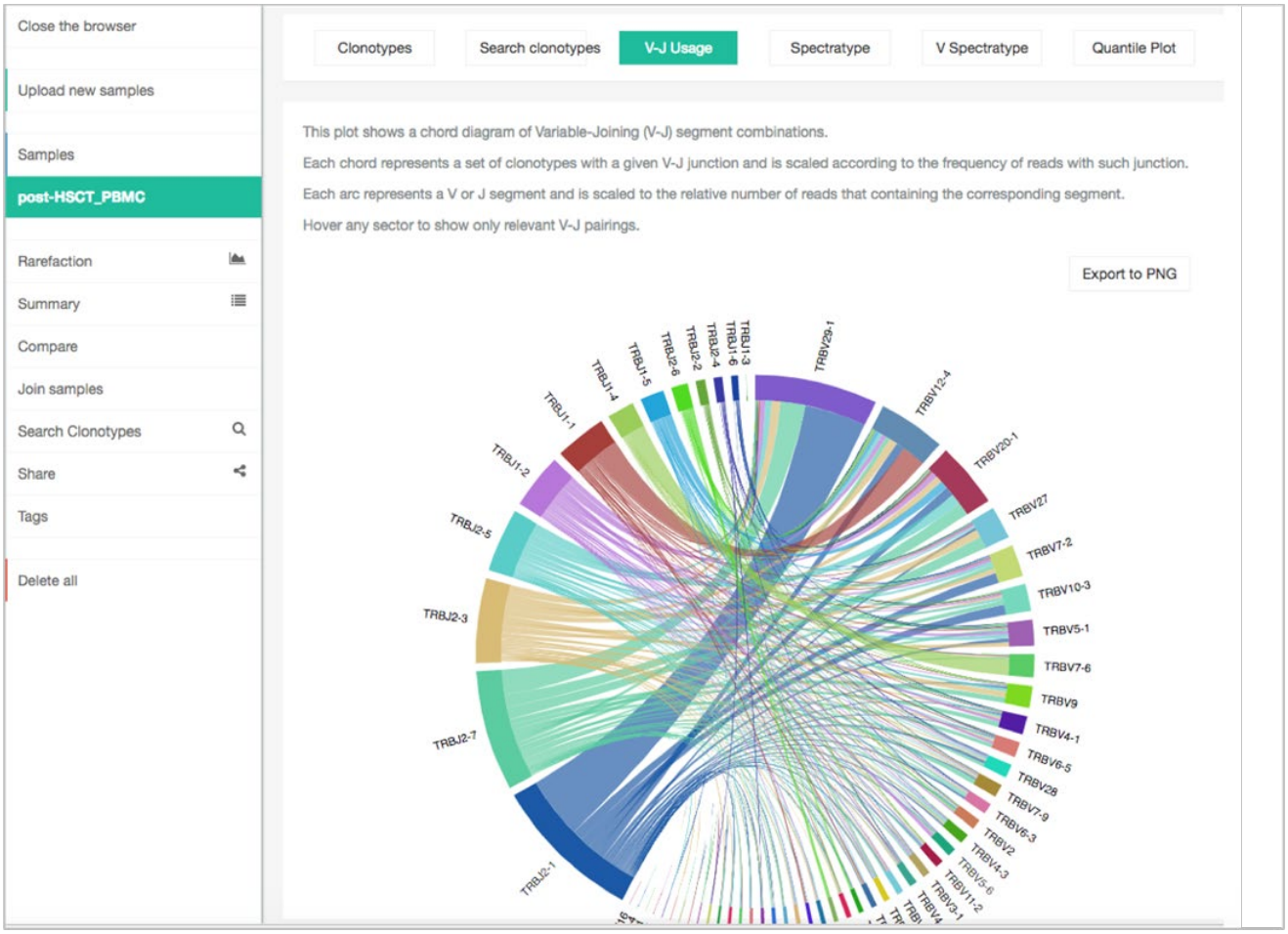


Figure 13. Example Variable-Joining Chord diagram generated by VDJviz

VIII. Technical Support

For any technical support regarding this protocol, please visit: takarabio.com/support

IX. Related Products

Cat. #	Product	Size
640182	ICELL8® Human TCR a/b Profiling Reagent Kit	Each
640179	ICELL8® Human TCR a/b Profiling - Indexing Primer Set	1 Chip
640180		5 Chips
640181		10 Chips
640200	ICELL8® cx TCR Chip*	1 Chip
640178	ICELL8® TCR Chip†	1 Chip

*For use with the ICELL8 cx Single-Cell System (Cat. # 640188, 640189)
†For use with the ICELL8 Single-Cell System (Cat. # 640000)

Appendix. File Output Details

A. Demultiplexing Output

The ICELL8 scTCR Analyzer uses ICELL8 paired read fastq files directly generated from the Illumina MiSeq sequencer as input. It demultiplexes with barcodes incorporated into Read 1 sequences and generate well level paired fastq files containing sorted read data.

1. Files directory location

- For Windows or MacOS users, the output files can be found in

  `[Main_Output_Directory]\[experiment_name]\split_fastqs`

- For Linux users. the output files can be found in `[analysis_dir]/split_fastqs`

NOTE: `[analysis_dir]` is the alias for the path defined by `[main_dir]/[outputfolder]`. For more information, see Table 2.

2. File names and information

For each ICELL8 well, the demultiplexing step produces paired fastq files with file names with the following naming schema:

- `[cell_type]-R[chip_rowID]-C[chip_columnID]_[fastq_prefix]_read1.fastq`
- `[cell_type]-R[chip_rowID]-C[chip_columnID]_[fastq_prefix]_read1.fastq`

`[cell_type]`, `[chip_rowID]`, and `[chip_columnID]` correspond to information in the CellSelect well list for the current experiment.

Example:

If you are looking for the information for Cell 1 on chip Row 20, Column 35 and the ICELL8 paired read fastq files are named `experiment1_read1.fastq` and `experiment1_read2.fastq` (`experiment1` being the fastq_prefix value), the names of the fastq files you would be interested in are named:

- `Cell1_R20_C35_experiment1_read1.fastq`
- `Cell1_R20_C35_experiment1_read2.fastq`

There is also a log file stored in the directory named `[experiment_name]_demultiplexing.log`. If the demultiplexing step completes successfully, the file will be empty. If there is an issue during this step, the log will contain error messages, such as unpaired reads detected during processing.

B. Alignment Output

The ICELL8 scTCR Analyzer uses the MiXCR default parameters for alignment; it requires the MiXCR "-chains" parameter to be set as "TRA, TRB".

1. Files directory location

- For Windows or MacOS users, these files can be found in

  `[Main_Output_Directory]\[experiment_name]\align`

- For Linux users. these alignment outputs can be found in

  `[analysis_dir]/align`

2. File names and information

For each ICELL8 well, the build alignment step produces two files:

- An alignment report file

  `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_align_report.txt`

- A MiXCR specific binary output file

  `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode].vdjca`

The MiXCR manual provides an in-depth description of these files, which is available at

https://media.readthedocs.org/pdf/mixcr/latest/mixcr.pdf

A brief description of each of the output files and their contents is provided for convenience.

- `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_align_report.txt` is a plain text file that records details about alignment statistics. The columns have the following headers:
  - Analysis date
  - Input files
  - Input command line argument
  - Analysis time
  - Total sequencing reads
  - Successfully aligned reads
  - Alignment failed, no hits (not TCR/IG?)*
  - Alignment failed because of absence of J hits
  - Overlapped
  - Overlapped and aligned
  - Alignment-aided overlaps
  - Overlapped and not aligned

* - This is exactly how the column header reads

- `[cell_type]-[chip_rowID]-[chip_columnID].vdjca` is a MiXCR-specific binary output file. This file serves as input for alignment manual inspection. More information about the file is available at

  http://mixcr.readthedocs.io/en/latest/export.html#exporting-well-formatted-alignments-for-manual-inspection

C. Assemble Clones Output

The ICELL8 scTCR Analyzer uses MiXCR default parameter for clone assembling, which builds a set of clones using alignments to extract *CDR3* region.

1. Files directory location

- For Windows or MacOS users, these files can be found in

  `[Main_Output_Directory]\[experiment_name]\assemble_clones`

- For Linux users, assembling outputs can be found in

  `[analysis_dir]/assemble_clones`

2. File names and information

For each ICELL8 well, this step produces two files:

- An alignment report file

  `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_clones_report.txt`

- A MiXCR specific binary output file

  `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode].clns`

  The MiXCR manual provides an in-depth description of these files, which is available at

  https://media.readthedocs.org/pdf/mixcr/latest/mixcr.pdf

A brief description of each of the output files and their contents is provided for convenience.

- `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_clones_report.txt` is a plain text file with details about assembling statistics:
  - Analysis date
  - Input files
  - Input command line argument
  - Final clonotype count
  - Average number of reads per clonotype
  - Reads used in clonotypes, ercent of total
  - Reads used in clonotypes before clustering, percent of total
  - Number of reads used as a core, percent of used
  - Mapped low quality reads, percent of used
  - Reads clustered in PCR error correction, percent of used
  - Reads pre-clustered due to the similar VJC-lists, percent of used
  - Reads dropped due to the lack of a clone sequence
  - Reads dropped due to low quality

- Reads dropped due to failed mapping
- Reads dropped with low quality clones
- Clonotypes eliminated by PCR error correction
- Clonotypes dropped as low quality
- Clonotypes pre-clustered due to the similar VJC-lists

Depending on input data, not all of the statistics listed above will be reported.

- `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode].clns` is a MiXCR specific binary output file which contains comprehensive information about clones. Information stored in it is extracted at the "export clone" step and reported in the final clone report.

D. Export Clones Output

The ICELL8 scTCR Analyzer uses the MiXCR default parameters for clone export, which exports all clones from a binary clone file, `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode].clns` and converts the information into human-readable text format.

1. File directory location

- For Windows or MacOS users, the output file can be found in

  `[Main_Output_Directory]\[experiment_name]\export_clones`

- For Linux users, the export report can be found in

  `[analysis_dir]/export_clones`

2. File name and information

For each ICELL8 well, this step produces one file, a clone report file

`[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_clones_all.txt`

The MiXCR manual provides an in-depth description of these files, which is available at

https://media.readthedocs.org/pdf/mixcr/latest/mixcr.pdf

A brief description of the output file and its contents is provided for convenience.

`[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_clones_all.txt` is a tab-delimited, plain text file listing detailed clonotype information. The file lists clonotypes by rows; each clonotype has 35 features reported:

Table 3. Reported clonotype features.

cloneId	cloneCount	cloneFraction	clonalSequence
clonalSequenceQuality	allVHitsWithScore	allDHitsWithScore	allJHitsWithScore
allCHitsWithScore	allVAlignments	allDAlignments	allJAlignments
allCAlignments	nSeqFR1	minQualFR1	nSeqCDR1
minQualCDR1	nSeqFR2	minQualFR2	nSeqCDR2
minQualCDR2	nSeqFR3	minQualFR3	nSeqCDR3
minQualCDR3	nSeqFR4	minQualFR4	aaSeqFR1
aaSeqCDR1	aaSeqFR2	aaSeqCDR2	aaSeqFR3
aaSeqCDR3	aaSeqFR4	refPoints	

E. Well-level Files for VDJviz Browser

For visualization purpose, the ICELL8 scTCR Analyzer converts well-level clonotype files into another file that can be uploaded to VDJviz.

1. File directory location

- For Windows or MacOS users, VDJviz input files can found in

  `[Main_Output_Directory]\[experiment_name]\vdjviz_input`

- For Linux users, VDJviz input files can be found in

  `[analysis_dir]/vdjviz_input`

2. File name and information

These VDJviz input files are named as

`[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_vdjviz.txt`

This is a tab-delimited, plain text file with a header line starting with "#". This file has six columns, with each row representing a clonotype:

- count—Number of reads for the current clonotype
- freq—The fraction of reads of the current clonotype as part of the total reads of all clonotypes
- cdr3nt—*CDR3* region nucleotide sequence
- cdr3aa—*CDR3* region amino acid sequence
- v—V region segment ID
- d—D region segment ID
- j—J region segment ID

NOTE: A valid VDJviz input file should have more than one clonotype. If only a single clonotype is detected in a given ICELL8 well, the analyzer would not generate a VDJviz input file for those wells.

F. Chip-level Summary File

After individual analysis at well-level is completed, the ICELL8 scTCR Analyzer merges results from all wells on ICELL8 chip to generate a chip-level summary file.

1. File directory location

- For Windows or MacOS users, this summary file is in

  `[Main_Output_Directory]\[experiment_name]`

- For Linux users, this summary file is in

  `[analysis_dir]/[experiment_name]`

2. File name and information

The chip-level summary information is stored in a comma-separated value (CSV) file named

`[experiment_name]_summary.csv`

The last three rows of the file record the [*experiment name*], total number of reads of all selected barcodes (barcode selected in CellSelect Well list), and the total number of aligned reads.

The rest of the information is in data-matrix format, with individual wells listed in rows and 83 corresponding features listed in columns. The details are summarized in Table 4 below.

Table 4. Summary of `[experiment_name]_summary.csv` well feature columns.

Column	Feature name	Description
1	All_name	Full name, `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]`
2	Sample	Partial name, `[cell_type]-[chip_rowID]-[chip_columnID]`
3	Barcode	Barcode sequence
4	Total_reads	Number of total reads
5	Aligned_reads	Number of aligned reads
6	Aligned (percent of total)	Fraction of alignment
7	Final clonotype count	Total number of clonotypes
8	Reads_used_in_clonotypes	Number of reads used for clonotype assembling
9	Reads_used (percent of total)	Fraction of reads used for clonotype assembling out of total reads
10	Top1_clonotype	The most numerous clonotype identified for current sample
11	Top1_clonotype (fraction)	Fraction of top clonotype out of all clonotypes
12	Top2_clonotype	The second-most numerous clonotype identified for current sample
13	Top2_clonotype (fraction)	Fraction of top2 clonotype out of all clonotypes
14	Top3_clonotype	The third-most numerous clonotype identified for current sample
15	Top3_clonotype (fraction)	Fraction of top3 clonotype out of all clonotypes
16	Clonotype Count (>= 0.01%)	Number of clonotypes which comprises more than 0.01% of the total number of clonotype
17	Clonotype Count (>=0.01% and read 10)	Number of clonotypes which comprises more than 0.01% of the total number of clonotype, assembled based on more than 10 reads (read used in clonotype)
18,34,50	Clone ID	Clonotype information for the top 3 clonotypes identified for corresponding well. Information extracted from the `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_clones_report.txt` file for the individual well. For example: column 18 is Clone ID for the most numerous clonotype; column 34 is Clone ID for the second-most numerous clonotype; column 50 is Clone ID for the third-most numerous clonotype.
19,35,51	Clone count	
20,36,52	Clone fraction	
21,37,53	Clonal sequence(s)	
22,38,54	Clonal sequence quality(s)	
23,39,55	All V hits with score	
24,40,56	All D hits with score	
25,41,57	All J hits with score	
26,42,58	All C hits with score	
27,43,59	All V alignments	
28,44,60	All D alignments	
29,45,61	All J alignments	
30,46,62	N. Seq. CDR3	
31,47,63	Min. qual. CDR3	

Column	Feature name	Description
32,48,64	AA. Seq. CDR3	
33,49,65	Ref. points	
66	Alignment failed no hits (not TCR/IG?)	Alignment information extracted from `[cell_type]-[chip_rowID]-[chip_columnID]-[barcode]_align_report.txt` file for the individual well.
67	Alignment failed because of absence of V hits	
68	Alignment failed because of absence of J hits	
69	Alignment failed because of low total score	
70	Overlapped	
71	Overlapped and aligned	
72	Overlapped and not aligned	
73	Reads used as core (percent of used)	
74	Mapped low quality reads (percent of used)	
75	Reads clustered in PCR error correction (percent of used)	
76	Clonotypes eliminated by PCR error correction	
77	Percent of reads dropped (lack of clonal sequence)	
78	Percent of reads dropped (low quality)	
79	Percent of reads dropped (failed mapping)	
80	Total V hits	All V segments identified and their clone counts, in the format of ID: count, and concatenated by &. In short, this summarized all V segments identified in cell type]-`[chip_rowID]-[chip_columnID]-[barcode]_clones_all.txt`
81	Total J hits	All J segments identified and their clone counts, in the format of ID: count, and concatenated by & In short, this summarized all J segments identified in cell type]-`[chip_rowID]-[chip_columnID]-[barcode]_clones_all.txt`

G. File for Simplified Top3 Clonotypes

The ICELL8 scTCR Analyzer conducts secondary analysis for quality control purpose with sample filtering on the chip based on negative controls used, if any. It generates a simplified top3 clonotype summary file: `[experiment_name]_simplified_top3.csv`. This a comma-separated list file with wells listed in rows and their features listed in columns, as detailed in Table 5 below.

Table 5. Summary of `[experiment_name]_summary.csv` well feature columns.

Column	Feature Name	Description
1	Sample	Sample name, `[cell_type]-[chip_rowID]-[chip_columnID]`
2	Barcode	Barcode sequence
3	Type	Cell Type
4	Chains	chain composition of the top3 clonotypes
5	Total_Reads	Total reads for sample
6	Aligned_Reads	Aligned reads for sample
7	Reads_Used_in_Clonotypes	Number of reads used for assembling clonotype
8	Total_Clonotypes	Number of total clonotypes identified
9	Clono1	Chain type of the most numerous clonotype, TCRa or TCRb
10	Clono1_Freq	Frequency of the most numerous overall clonotype
11	Clono1_Count	Total reads of the most numerous clonotype
12	Clono1_N_Seq	Nucleotide sequence of the most numerous clonotype
13	Clono1_AA_Seq	Amino acid sequence of the most numerous clonotype
14	Clono1_V	V region ID of the most numerous clonotype
15	Clono1_D	D region ID of the most numerous clonotype
16	Clono1_J	J region ID of the most numerous clonotype
17	Clono2	Chain type of the second-most numerous clonotype, TCRa or TCRb
18	Clono2_Freq	Frequency of the second-most numerous clonotype
19	Clono2_Count	Total reads of the second-most numerous clonotype
20	Clono2_N_Seq	Nucleotide sequence of the second-most numerous clonotype
21	Clono2_AA_Seq	Amino acid sequence of the second-most numerous clonotype
22	Clono2_V	V region ID of the second-most numerous clonotype
23	Clono2_D	D region ID of the second-most numerous clonotype
24	Clono2_J	J region ID of the second-most numerous clonotype
25	Clono3	Chain type of the third-most numerous clonotype, TCRa or TCRb
26	Clono3_Freq	Frequency of the third-most numerous clonotype

Column	Feature Name	Description
27	Clono3_Count	Total reads of the third-most numerous clonotype
28	Clono3_N_Seq	Nucleotide sequence of the third-most numerous clonotype
29	Clono3_AA_Seq	Amino acid sequence of the third-most numerous clonotype
30	Clono3_V	V region ID of the third-most numerous clonotype
31	Clono3_D	D region ID of the third-most numerous clonotype
32	Clono3_J	J region ID of the third-most numerous clonotype
33	TrueChains	Chain composition in the same order as their clonotypes. *Example: if the top1 clonotype is TCRa, top2 clonotype is missing, and top3 clonotype is TCRb, the value in the TrueChains column would be "a0b".*

H. File for Cell-level Quality Control Statistics

After secondary analysis, the ICELL8 scTCR Analyzer produces the file `[experiment_name]_cell_filter_table.csv`, with step-by-step cell quality control filtering statistics.

- Filter1 is a clonotype N/A (null) filter—If a well has all of its top3 clonotypes missing, this well is filtered out.
- Filter2 is the negative control threshold filter—Collect read counts of all negative controls' top1 clonotype and use mean + 3x standard deviation as the threshold. If any of the top3 clonotype in a well sample has less reads than the calculated threshold, filter out the corresponding clonotype.
- Filter3 is *CDR3* deletion filter—If there is any deletion or frameshift found in the amino acid sequence of a clonotype, drop the corresponding clonotype.
- After applying each filter, if a well sample has all three top clonotypes filtered out, the ICELL8 scTCR Analyzer eliminates this well sample.

This file contains three tables:

1. Table1. Number of Remaining Cells by filters
   - This table summarizes total number of cells for each cell type, and remaining cell numbers after applying three filters.
2. Table2. Percentage Remaining by filters
   - This percentage table summarizes percentage remaining after filtering. It is produced by divide values in table1 by total number of cells.
3. Table3. Detailed Filtering Stats

Table 6 below shows a detailed breakdown statistics of values in Table1 of the file.

Table 6. Summary of [experiment_name]_cell_filter_table.csv column names in Table1 of the file.

Column	Feature name	Description
1	number_cell	Total number of cells before QC
2	top1_na	Number of cells with missing value in top1 clonotype
3	top2_na	Number of cells with missing value in top2 clonotype
4	top3_na	Number of cells with missing value in top3 clonotype
5	drop_filter1	Number of cells dropped after filter1
6	count_drop_clono1	Number of cells with its top1 clonotype reads less than filter2
7	count_drop_clono2	Number of cells with its top2 clonotype reads less than filter2
8	count_drop_clono3	Number of cells with its top3 clonotype reads less than filter2
9	drop_filter2	Number of cells dropped after filter2
10	AA_deletion_drop_clono1	Number of cells with deletion or frameshift in amino acid sequence of its top1 clonotype
11	AA_deletion_drop_clono2	Number of cells with deletion or frameshift in amino acid sequence of its top2 clonotype
12	AA_deletion_drop_clono3	Number of cells with deletion or frameshift in amino acid sequence of its top3 clonotype
13	drop_filter3	Number of cells dropped after filter3

I. File for Cell-level QC Statistics

After secondary analysis, the ICELL8 scTCR Analyzer also produces the file [experiment_name]_chain_composition_table.csv, which, for certain cell types included on the chip, summarizes the overall α- and β-chain composition of the top three clonotypes. If there are no cells remaining after the quality control step, this file won't be available.

The [experiment_name]_chain_composition_table.csv has two tables:

- Table4. Sample Number by Alpha Beta Chain Composition
- Table5. Sample Fractions by Alpha Beta Chain Composition

All possible chain compositions for the top3 clonotypes are listed below.

Table 7. Possible chain compositions in the Cell-level QC Statistics file .

Chain Composition	Description
a	A single alpha chain found in the top3 clonotypes
b	A single beta chain found in the top3 clonotypes
aa	Two alpha chains found in the top3 clonotypes, no beta chain
bb	Two beta chains found in the top3 clonotypes, no alpha chain
ab	An alpha chain and a beta chain found in the top3 clonotypes
aab	Two alpha chains and a beta chain found in the top3 clonotypes
abb	One alpha chain and two beta chains found in the top3 clonotypes
aaa	The top3 clonotypes all have alpha chains
bbb	The top3 clonotypes all have beta chains

J. Chip-level VDJviz Input File

Chip-level VDJviz input files are generated by the ICELL8 scTCR Analyzer based on simplified Top3 Clonotypes. Files are named as [cell_type]_[chain_type]_vdjviz.txt for each individual cell type and each detected chain type on the chip.

A valid VDJviz input file should have more than one clonotype; if a certain cell type + chain combination only has a single clonotype, the VDJviz input will also not be available.

K. Chip-level Clonotype VJ-join Cell Count File

Based on the `[experiment_name]_simplified_top3.csv` file, the ICELL8 scTCR Analyzer also summarizes chip-level vj-join cell counts – it counts how many cells have a certain vj-join region. Similar to chip-level VDJviz input, files are named with the syntax `[cell_type]_[chain_type]_vj_join.txt`.

The file has three columns:

Table 8. Simplified Top3 clonotype summary data fields

Column	Feature Name	Description
1	cell_count	number of cells
2	v	V region ID
3	j	j region ID

Because summarization is based on all three clonotypes, the sum of the cell_count column could be more than the total cells in `[experiment_name]_simplified_top3.csv`.

Contact Us	
Customer Service/Ordering	Technical Support
tel: 800.662.2566 (toll-free)	tel: 800.662.2566 (toll-free)
fax: 800.424.1350 (toll-free)	fax: 800.424.1350 (toll-free)
web: takarabio.com/service	web: takarabio.com/support
e-mail: ordersUS@takarabio.com	e-mail: technical_support@takarabio.com

Takara Bio USA, Inc.
2560 Orchard Parkway, San Jose, CA 95131, USA
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999

08.21 US