CeGaT

Research &
Pharma Solutions

Single-Cell RNA Sequencing

Bioinformatic Note

Single-Cell RNA Sequencing

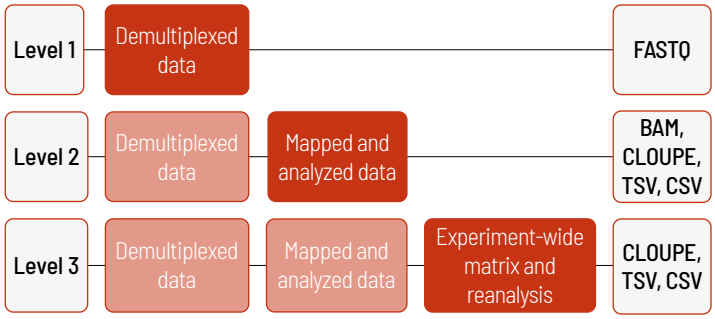
The human body consists of more than 30 trillion cells. Within an organ or tissue, different cell types and cell populations occur. These cells are highly heterogeneous regarding their transcriptional state. Cellular heterogeneity is an important feature of normal physiological processes. Different cell types and populations have different cellular functions, which are all required for the tissue's or organ's correct functioning. During disease progression, the composition of the different cell populations and their transcriptome may change or even contribute to the disease.

Traditional bulk RNA-sequencing methods cannot resolve the heterogeneity of samples of different cell populations. However, single-cell RNA sequencing technologies now enable the analysis of individual transcriptomes from thousands of cells in a single sample. Thus, this technology can help resolve the transcriptional change on a cell type-specific level that may lead to the identification of new biomarkers or help further understand the biology of tissues or diseases.

Application areas of single-cell RNA sequencing:

- Detection of tumor heterogeneity
- Cell differentiation and lineage tracing
- Response to therapeutic interventions
- Biomarker discovery

Different levels of bioinformatic data analysis are available:



With increasing bioinformatic levels, more data is delivered. All higher levels include the data from the lower levels, e.g., in Level 2, trimmed data as well as mapped and analyzed data are provided. In addition to the data, and independent of the analysis level, a project report is generated.

Level 1

If you wish to analyze your data yourself, we recommend Level 1. At this level, demultiplexed reads in FASTQ format are delivered. In this level, the sequencing data are demultiplexed. Additionally, the index reads are delivered. This level is provided for every project, regardless of additional purchased bioinformatic analyses.

The additionally generated project report provides information for every sample about the laboratory protocol, including quality control data of the starting material, library preparation, sequencing parameters, and the Q30 value of the sequencing.

Level 2

In Level 2, the trimmed reads are processed. In this step, the alignment, filtering, barcode counting, UMI counting of the reads, and cell calling is performed. In addition, cell clusters are determined, and differential gene expression analysis is performed.

Several output files and folders are generated:

[1] Pos(ition)sorted genome files (BAM and BAI format),

[2] a web summary file (HTML format),

[3] a metrics summary file (CSV format),

[4] a cloupe file (CLOUPE format),

[5] a molecule info file (H5 format),

[6] a raw and a filtered feature barcode matrix folder (with MEX and TSV files),

[7] a raw and a filtered feature barcode matrix file (H5 format),

[8] and an analysis folder.

Every project receives a unique S-number (SXXXX), and every sample a unique identifier. In this example, the S-number is S1163.

[1] Pos(ition)-sorted Genome Files

The mapped data are provided as BAM files and contain position-sorted reads aligned to the genome and transcriptome. Each read in this BAM file has 10x cellular and molecular barcode information attached.

[2] Web Summary File

The web summary file in HTML format contains detailed information on the sample, QC metrics, and a basic interactive view of cell clusters as t-SNE plot (t-distributed stochastic neighbor embedding), which is a statistical method to visualize high-dimensional data in a two or three-dimensional map. The HTML file can be opened with a web browser. It is divided into a summary part and an analysis part.

The summary part contains information about:

- Sequencing QC metrics, such as number of total reads, sequencing saturation, valid barcodes, and Q30 values;
- Mapping metrics, such as percentage of reads mapped to the genome, transcriptome, exonic, and intronic regions;
- Cell metrics, such as the estimated number of cells, fraction reads in cells, mean reads and median genes per cell;
- Sample information, such as the sample ID, software version used, and reference genome.

The analysis part consists of:

- a t-SNE projection of cells colored by UMI counts and a t-SNE projection of cells colored by corresponding clusters. Here, the clustering type can be changed between graph-based and K-means. The choice of the clustering type affects the number of cell clusters.
- a differential gene expression analysis with a list of top features, including the log2 fold-change of the gene expression and corresponding p-value in each cluster. When the clustering type is changed, the top feature list displays the current cluster-specific differential gene expression and p-values.
- a plot of the sequencing saturation,
- and a plot of the median number of genes per cell.

Additional information about the summary and analysis parts are available in the interactive HTML file.

[3] Metrics Summary File

The information in the summary section of the web summary file is also provided in a metrics summary file in CSV format. An excerpt of the file is shown in table 1. Selected QC metrices, such as estimated number of cells, mean reads per cell, and median genes per cell are additionally provided in the project report.

[4] Cloupe File

A Loupe Browser file in CLOUPE format is generated that enables the interactive visualization and analysis of single-cell data with the Loupe Browser software, a desktop application provided by 10x Genomics.

[5] Molecule Info File

A molecule information file in HDF5 format contains per-molecule information for all molecules with a valid barcode, valid UMI, and a confident assignment to a gene or feature barcode. This HDF5 file contains data corresponding to the observed molecules, and data about the libraries used. HDF5 stands for the Hierarchical Data Format version 5 and is an open source file format for large, complex, and heterogeneous data. With this file format, data can be organized in structured ways. A HDF5 file can consist of groups, which are folder-like elements, and datasets, which contain the actual data. As the HDF5 format is an open source format, supporting libraries and viewers can be downloaded. HDF5 files can also be used with open source programming languages, such as R or Python.

[6] Raw and Filtered Feature Barcode Matrix Folders

Raw (or unfiltered) and filtered feature-barcode matrices are generated. The raw feature barcode matrix contains all detected barcodes, while the filtered feature barcode matrix contains only detected cell-associated barcodes. For both, the raw and the filtered feature-barcode matrices, a folder with three files is created:

- a matrix file in MEX format (compressed),
- a features table in TSV format (compressed),
- and a barcodes table in TSV format (compressed).

Matrix file in MEX format

The matrix is stored in the Market Exchange Format (MEX) for sparse matrices (figure 1). The first three lines of the matrix file contain headers: the file title (%%MatrixMarket matrix coordinate integer general), metadata (e.g., software and format version), and the total number of rows of the features.tsv file, barcodes.tsv file, and matrix.mtx file. In the following lines, the features are listed with feature index, cell index, and corresponding UMI count in columns. The feature and cell indices correspond to the lines in the features and barcodes files, respectively. As seen in line 4 of figure 1, the feature in line 47 of the features file and the cell barcode in line 1 of the barcodes file has an UMI count of 1.

```
%%MatrixMarket matrix coordinate integer general
%metadata_json: {"software_version": "cellranger-7.1.0", "format_version": 2}
32738 4898 13388388
47 1 1
95 1 1
```

Figure 1 | First five lines of the matrix file of sample S1163Nr606 in MEX format.

Table 1 | Excerpt of the metrics summary file of sample S1163Nr606 in CSV format.

Estimated Number of Cells	Mean Reads per Cell	Median Genes per Cell	Number of Reads	Valid Barcodes	Sequencing Saturation	Q30 Bases in Barcode
4,898	58,666	2,665	287,347,593	96.6%	66.8%	96.4%

Features file in TSV format

The features file in gzipped TSV format contains all annotated features with feature identifier (ID), feature name, and feature type (Gene Expression, Antibody Capture, CRISPR Guide Capture, Multiplexing Capture, CUSTOM). An excerpt of the features file is shown in table 3.

Table 3 | Excerpt of the features file of sample S1163Nr606 in TSV format (header line added for the sake of clarity).

Feature ID	Feature name	Feature type
ENSG00000243485	MIR1302-10	Gene Expression
ENSG00000237613	FAM138A	Gene Expression
ENSG00000186092	OR4F5	Gene Expression
ENSG00000238009	RP11-34P13.7	Gene Expression
ENSG00000239945	RP11-34P13.8	Gene Expression
ENSG00000237683	AL627309.1	Gene Expression

Barcodes file in TSV format

The barcodes file in gzipped TSV format contains the barcode sequences. The number attached to the barcode sequence refers to the GEM well (Gel Beads-in-emulsion) and is used to achieve a higher effective barcode diversity when combining samples generated from separate GEM wells. An excerpt is shown in table 4.

The filtered feature-barcode matrices can be used as a starting point to analyze the data with alternative software packages (e.g., Seurat). For additional information on these three files or for instructions on how to import these matrices to R or Python or how to convert the matrices to CSV format, the reader is referred to the 10x support page.

Table 4 | Excerpt of the barcodes file of sample S1163Nr606 in TSV format (header line added for the sake of clarity)

Barcodes
AAACCCAAGCGGGTAT-1
AAACCCAAGTCCCTAA-1
AAACCCAGTCCATACA-1
AAACCCAGTGACTCGC-1
AAACCCAGTTATCTGG-1

[7] Raw and Filtered Feature Barcode Matrix Files

In addition to the MEX format, matrices are also provided in the Hierarchical Data Format version 5 (HDF5).

[8] Analysis Folder

This folder contains the results of the different cluster analyses, the differential gene expression analysis between clusters, and PCA, t-SNE, and UMAP dimensionality reduction in CSV format.

Level 3

If more than one sample from the same experiment is sequenced, the data from multiple samples can be combined into an experiment-wide feature-barcode matrix and analysis. All samples are normalized to the same sequencing depth, and then the feature-barcode matrices are recomputed and the combined data are analyzed

A set of output files for all samples combined is generated, which is similar to the output for each individual sample, except for FASTQ files and alignments. Thus, these files will not be explained again.

A web summary file in HTML format is generated. Similar to the previously described web summary file, the HTML file can be browsed interactively. The aggregation summary file contains information about all aggregated samples, including the total number of reads and mean reads before and after normalization. A t-SNE projection colored by the library ID is shown, as well as the estimated number of cells, fraction reads in cells, and median gene number and UMI counts per cell. In the analysis part of the interactive web summary, the clustering type can be individually chosen. Different t-SNE projections are shown, and the top features by clusters are provided.

An aggregation file is provided in CSV format (see table 5). The file provides information about each sample, its mixture status, and the sample types.

Table 5 | Aggregation file in CSV format.

sample_id	molecule_h5	mixture	mixed_sample	sample_type
S1163Nr606	S1163Nr606/outs/molecule_info.h5	100:0	no	PBMC
S1163Nr607	S1163Nr607/outs/molecule_info.h5	90:10	yes	PBMC-HEK

10X Genomics Resources

For further information, a few 10x Genomics resources are listed below:

- 10x Genomics website: www.10xgenomics.com/
- 10x Genomics software downloads (Cell Ranger and Loupe Browser): https://support.10xgenomics.com/single-cell-gene-expression/software/downloads/latest
- 10x Genomics, explanation of outputs: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/output/overview
- 10x Genomics, detailed description of the software Cell Ranger: https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/what-is-cell-ranger
- 10x Genomics, detailed description of the software Loupe Browser: https://support.10xgenomics.

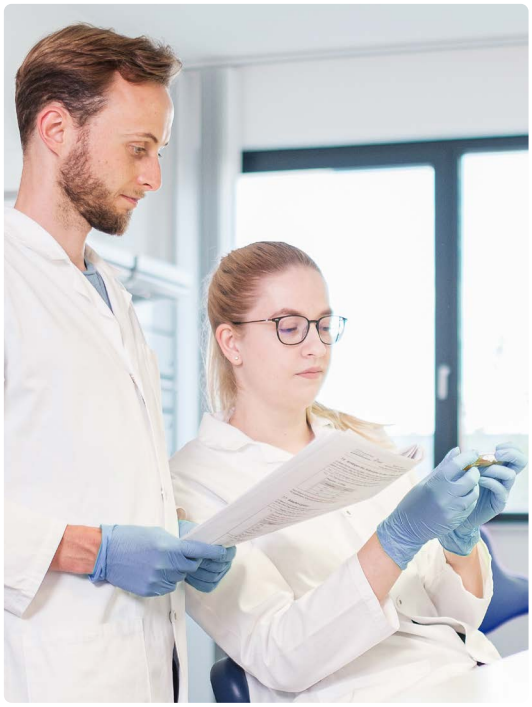
About Us

CeGaT was founded in 2009 in Tübingen, Germany. Our scientists are specialized in next-generation sequencing (NGS) for genetic diagnostics, and we also provide a variety of sequencing services for research purposes and pharma solutions. Our sequencing service portfolio is complemented by analyses suited for microbiome, immunology, and translational oncology studies.

Our dedicated project management team of scientists and bio-informaticians works closely with you to develop the best strategy to realize your project. Depending on its scope, we select the most suitable library preparation and conditions on our sequencing platforms.

We would be pleased to provide you with our excellent service.
Contact us today to start planning your next project.

For more details please visit
www.cegat.com/rps



CeGaT GmbH
Research & Pharma Solutions
Paul-Ehrlich-Str. 23
72076 Tübingen
Germany

Phone: +49 7071 56544-333
Fax: +49 7071 56544-56
Email: rps@cegat.com

CAP ACCREDITED
COLLEGE of AMERICAN PATHOLOGISTS
CLIA CERTIFIED ID: 99D2130225

DAkkS Deutsche Akkreditierungsstelle D-PL-13206-01-00
Accredited by DAkkS according to DIN EN ISO/IEC 17025:2018

2023/12