BWH Single Cell Genomics Core Data Output User Guide

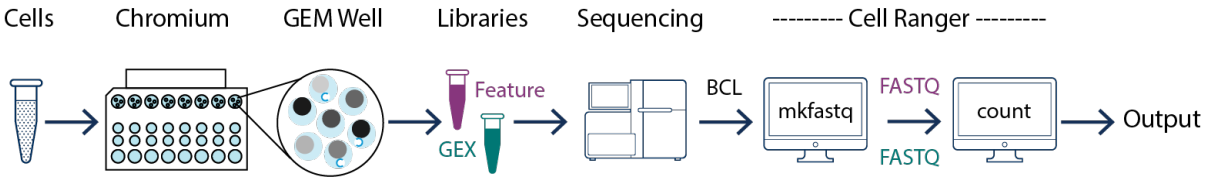
Updated: 06.30.2021

10X Genomics single-cell RNA-seq Protocol

10x Genomics single-cell RNA-seq (scRNA-seq) libraries were prepared according to the 10x Genomics User Guide. Specifically, single cells, reverse transcription (RT) reagents, Gel Beads containing barcoded oligonucleotides, and oil are combined on a microfluidic chip to form reaction vesicles called Gel Beads in Emulsion, or GEMs. Within each GEM reaction vesicle, a single cell is lysed, the Gel Bead is dissolved to free the identically barcoded RT oligonucleotides into solution, and reverse transcription of polyadenylated mRNA occurs. As a result, all cDNAs from a single cell will have the same barcode, allowing the sequencing reads to be mapped back to their original single cells of origin. The preparation of NGS libraries from these barcoded cDNAs is then carried out in a highly efficient bulk reaction.

Summary of Process

The single-cell RNA-seq data are processed by Cell Ranger workflow. The samples are first processed through GEM wells, and then libraries are generated and sequenced. The raw base call files (BCL) generated by Illumina sequencers are then de-multiplexed into FASTQ files using `cellranger mkfastq`. Running `cellranger count` takes FASTQ files and performs alignment, filtering, barcode counting and UMI counting.



Data are processed using Cell Ranger version 6.0.2.

Description of Data

The folder named **cellranger-6.0.2** contains the processed data, including BAM files. BAM files contain all the sequence information.

1) cellranger-6.0.1
    a) GRCh38
        i) BRI-158
            (1) outs
                (a) cloupe.cloupe
                (b) web_summary.html
                (c) filtered_feature_bc_matrix
                    (i) barcodes.tsv.gz
                    (ii) features.tsv.gz
                    (iii) matrix.mtx.gz
                (d) raw_feature_bc_matrix
                (e) possorted_genome_bam.bam
        ii) BRI-159
        iii) …

The processed data are found under the folder named after the reference transcriptome (e.g. **mm10, GRCh38, hg19**) that was used. Inside this directory, the folders are named after their library ID (e.g. **BRI-158, BRI-159**, etc.). The Cell Ranger pipeline outputs a summary HTML file named web_summary.html that contains summary metrics and automated secondary analysis results of the experiment. The **cloupe.cloupe** file can be loaded into the Loupe Browser for visualization and analysis.
The **filtered_feature_bc_matrix** folder stores the filtered gene by cell barcode matrix **matrix.mtx.gz** that excludes barcodes corresponding to background noise from GEMs. Each element of the matrix is the number of UMI associated with a feature (row) and a barcode (column). In the directory, there is the **features.tsv.gz** file that stores information of each gene (feature ID, name, and type). The **barcodes.tsv.gz** file contains barcode sequences corresponding to column indices of the matrix. The unfiltered feature-barcode matrix is also in the output folder (**raw_feature_bc_matrix**), all barcodes from the fixed list of known-good barcode sequences that has at least 1 read are included, they can be background and cell associated barcodes.
These matrices can then be loaded into R or Python for subsequent analyses. We recommend users to further apply additional filters to remove bad quality cells.
For each library, the cellranger pipeline also outputs an indexed BAM file (**possorted_genome_bam.bam**) containing position-sorted reads aligned to the genome and transcriptome. Each read in the BAM file has Chromium cellular and molecular barcode

information attached. More details on the BAM format are available online. Please check here for more information.
The BAM files can also be converted to FASTQ files using the cellranger function `bamtofastq`. Click here for downloading, installing, and running the tool.

General Recommendations for Downstream Analysis

cloupe.cloupe file

To load the cloupe.cloupe file in the Loupe Browser, please first decompress the .tar.gz file by just clicking on the file or through command line, with the following command: `tar -xvzf yourFile.tar.gz`. This will open a directory with more folders and files inside. Navigate to the **outs** folder where the cloupe.cloupe file is stored and consult the Loupe Browser tutorials for more details on how this works.

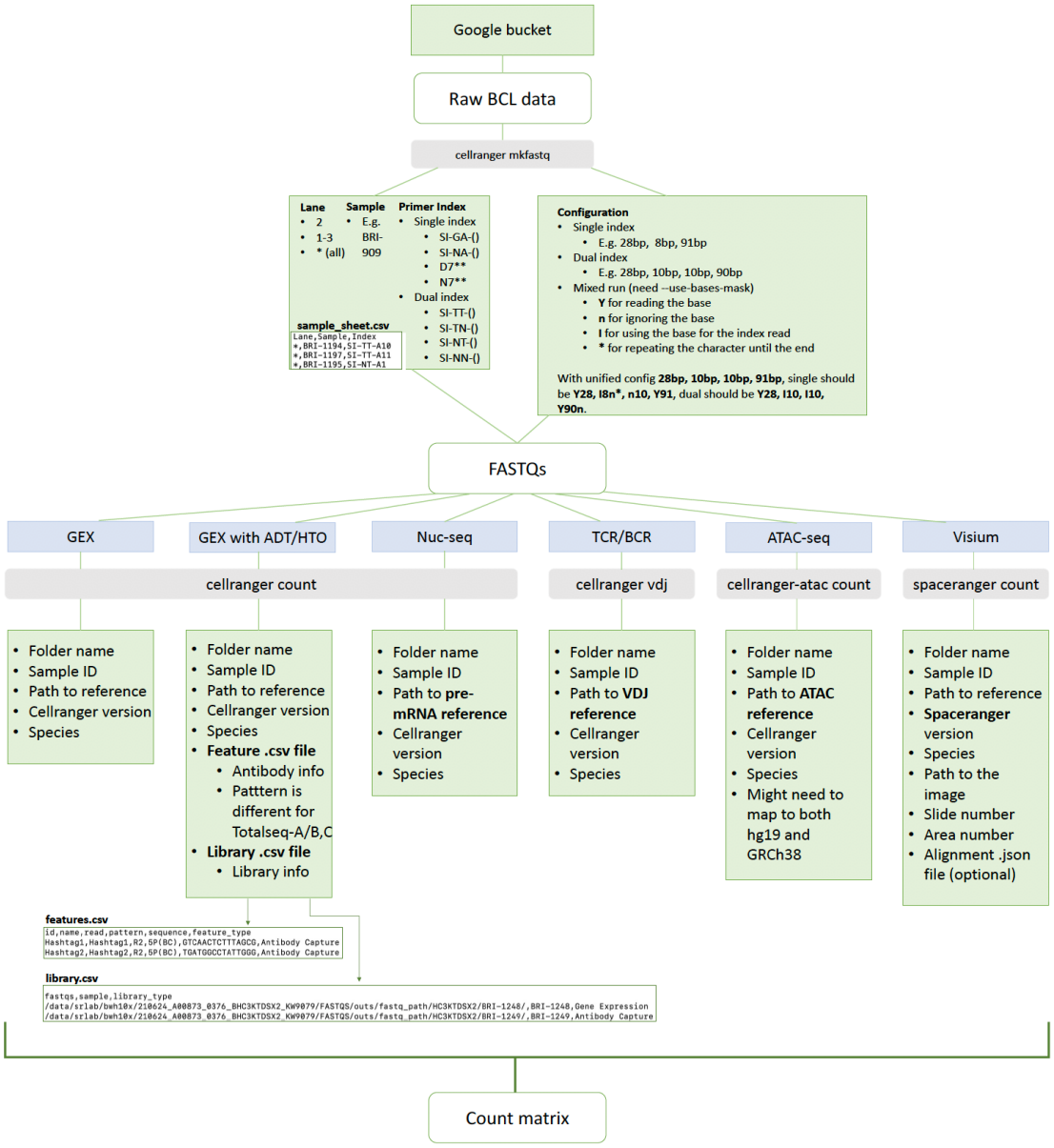
QC Summary Statistics

To see the general QC summary statistics of the sequencing data, please consult the **web_summary.html** file under the **outs** directory.

Processed Data Analysis

The processed matrices of filtered gene expression can be read using a single command with the R package **Seurat** (Tutorials).
For differential expression of scRNA-seq data, our lab uses the `wilcoxauc` function from the **presto** (GitHub Page) package developed by Ilya Korsunsky in our lab to quickly find the main markers that characterize each cluster, and other methods that use linear regression that can include covariates are recommended for more formal differential expression analyses (Limma, DESeq2, Langefeld).
For pathway/gene set enrichment analyses, our lab uses GSEA and liger. The gene sets that we find most useful are from MSigDB, particularly C5 (GO gene sets) and potentially C7 (immunologic signatures).

You can also consult the 10X genomics training modules for more information, and this free book on "Orchestrating Single Cell Analysis" targeted for experimental biologists.


Google bucket
Raw BCL data
cellranger mkfastq
Lane
2
1-3
* (all)
Sample
E.g.
BRI-
909
Primer Index
Single index
SI-GA-()
SI-NA-()
D7**
N7**
Dual index
SI-TT-()
SI-TN-()
SI-NT-()
SI-NN-()
sample_sheet.csv
Lane,Sample,Index
*,BRI-1194,SI-TT-A10
*,BRI-1197,SI-TT-A11
*,BRI-1195,SI-NT-A1
Configuration
Single index
E.g. 28bp, 8bp, 91bp
Dual index
E.g. 28bp, 10bp, 10bp, 90bp
Mixed run (need --use-bases-mask)
Y for reading the base
n for ignoring the base
I for using the base for the index read
* for repeating the character until the end
With unified config 28bp, 10bp, 10bp, 91bp, single should be Y28, I8n*, n10, Y91, dual should be Y28, I10, I10, Y90n.
FASTQs
GEX
GEX with ADT/HTO
Nuc-seq
TCR/BCR
ATAC-seq
Visium
cellranger count
cellranger vdj
cellranger-atac count
spaceranger count
Folder name
Sample ID
Path to reference
Cellranger version
Species
Feature .csv file
Antibody info
Patttern is different for Totalseq-A/B,C
Library .csv file
Library info
Path to pre-mRNA reference
Path to VDJ reference
Path to ATAC reference
Might need to map to both hg19 and GRCh38
Spaceranger version
Path to the image
Slide number
Area number
Alignment .json file (optional)
features.csv
id,name,read,pattern,sequence,feature_type
Hashtag1,Hashtag1,R2,5P(BC),GTCAACTCTTTAGCG,Antibody Capture
Hashtag2,Hashtag2,R2,5P(BC),TGATGGCCTATTGGG,Antibody Capture
library.csv
fastqs,sample,library_type
/data/srlab/bwh10x/210624_A00873_0376_BHC3KTDSX2_KW9079/FASTQS/outs/fastq_path/HC3KTDSX2/BRI-1248/,BRI-1248,Gene Expression
/data/srlab/bwh10x/210624_A00873_0376_BHC3KTDSX2_KW9079/FASTQS/outs/fastq_path/HC3KTDSX2/BRI-1249/,BRI-1249,Antibody Capture
Count matrix