Summary of files stored in CZI Nasal Mucosa Terra Workspace

Last updated 5/31/2023

* Upstream Processing Files
  + Custom Reference Genomes
    - 20230216\_GRCh38\_SARSCoV2\_RSV
      * Joint human, SARSCoV2, and RSV genome
      * fastas and gtfs used to generate this genome are stored in this folder
      * Used during Cellranger alignment by specifying path to "GRCH38\_SARSCoV2\_RSV\_A\_B.tar.gz" file under the Reference column
  + scRNA-seq
    - BCL files
      * This is a temporary storage space before Cellranger alignment is run. BCL files are permanently stored in the google drive folder: JOMLab/projectsCollaborative/CZI\_Pediatric\_Nasal\_Swabs/scRNAseq/raw\_data/BCL\_files
    - FASTQs
      * These are generated during the "mkfastq" part of the cellranger alignment. They only need to be generated from the BCL files once, and then can be used for multiple alignments.
    - Cellranger Alignments
      * GRCh38
        + This is necessary to run Freemuxlet for demultiplexing
      * GRCH38\_SARSCoV2\_RSV
        + These are the files used for the majority of analysis.
        + The "Alignment\_h5files" folder has a subfolder for each pool with only the filtered\_feature\_bc\_matrix.h5 file in it. This is used for the input to create the Seurat object
      * GRCh38\_SARSCoV2\_RSV\_Force30k
        + These files are used for input to cellbender. They are forced to have 30,000 "cells" so we can estimate viral RNA in ambient droplets.
    - Cellbender corrections
      * These are the post-cellbender files used for viral positive assignment
  + bulkRNA-seq
    - BCL files
      * This is also temporary storage for bulk BCL files
    - FASTQs
      * Results from bcl -> fastq workflow. These are used as input for the demultiplexing.
    - Alignments
      * Smartseq2 aligned files. The count\_matrix can be used for analysis.
  + Demultiplexing results
    - Freemuxlet
      * Output from running Freemuxlet on each pool with a designated number of samples
    - bulk RNA-seq merged vcf
      * After the FASTQs for each individual participant are converted through multiple steps into a VCF, the VCFs for each pool are aggregated into one file. The aggregated files for each pool are stored in this folder
    - vcf match sample id results
      * Sample\_Fmux\_IDs table contains the translation between the Freemuxlet output and sample identification
* Analysis Files
  + Code - RMarkdown files used for analysis
    - Create objects, merge objects for each pool, initial filtering
    - Preliminary clustering
  + Seurat Objects
    - 20230523\_Batch2-12\_merged\_filtered.RData - this object has the compilation of all of the cells from all of the sequenced pools. Filters used.
      * Singlets (Freemuxlet)
      * Percent mitochondrial < 35%
      * Number of unique genes: > 300
      * Number of UMIs: > 500
    - 20230523\_Batch2-12\_annotated.RData - this object contains the same cells as the other object, was saved after doing SCTransform normalization, clustering, and annotation of preliminary clusters.
      * Cluster labels are stored in the “cell.type.broad” metadata variable
      * This object should be used for all further analysis
  + Gene signature lists
    - These are the input used for gene module scoring
  + Output
    - Plots and marker gene lists