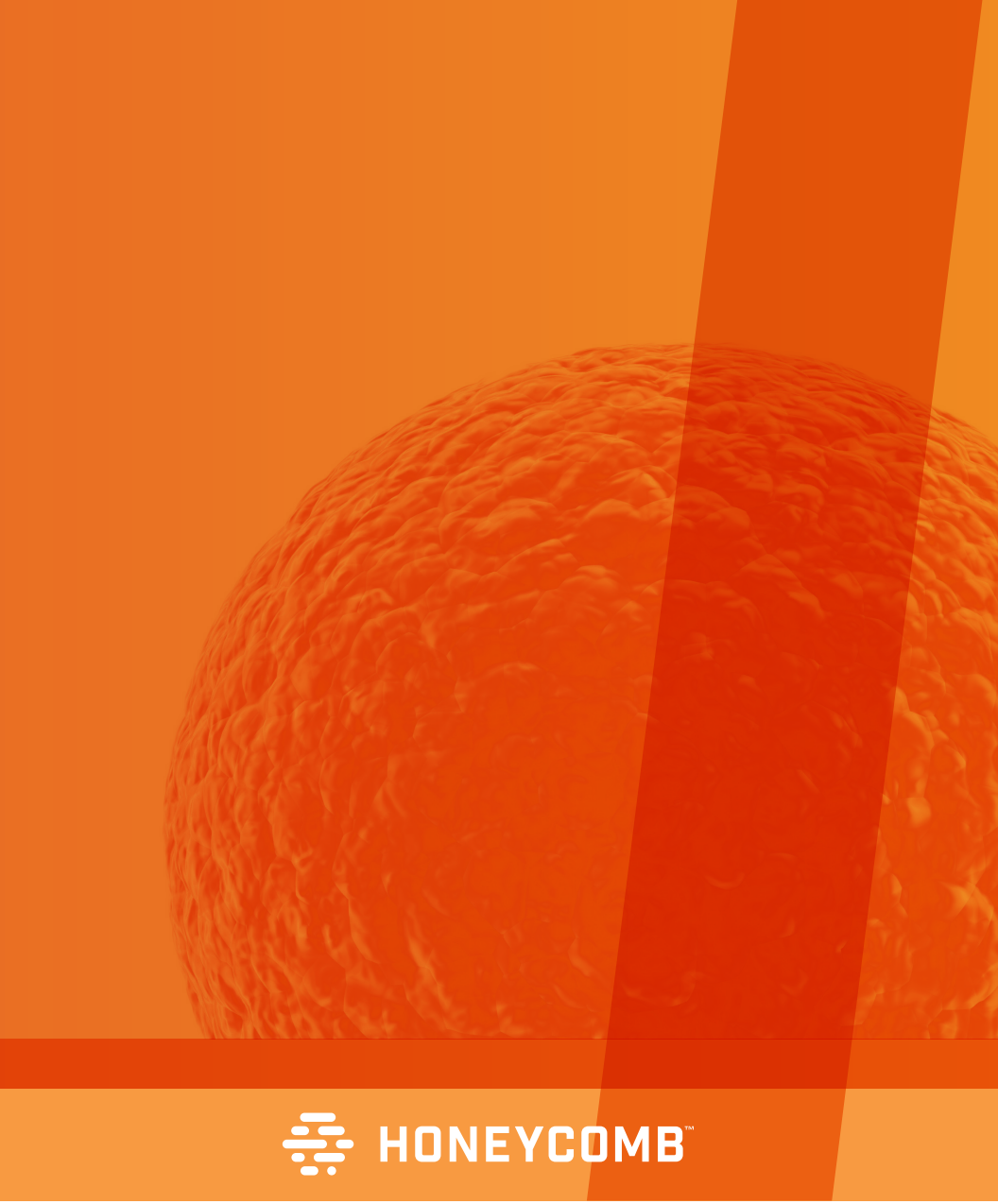


HONEYCOMB™

BeeNetPLUS hosted on Terra.bio v1.0.X Guide

Automated Single Cell RNA Sequencing Analysis

v1.0

This product is for research use only.
Not for use in diagnostic procedures.

BeeNetPLUS scRNAseq Analysis Guide

///// Introduction

This document, along with the accompanying video tutorial **BeeNetPLUS hosted on Terra.bio**, details how to run single cell RNAseq analysis using the BeeNetPLUS Workflow hosted on Terra.bio. The BeeNetPLUS workflow combines BeeNet™ , a custom software designed to process data from HIVE™ single-cell RNA-seq libraries, with a downstream Seurat analysis workflow. This document is oriented to support a user familiar with Terra.bio Workspaces and Google Cloud Platform (GCP) storage buckets.

BeeNet™ is a custom software designed to process data from paired-end Illumina® sequencing of single-cell RNA-seq libraries produced by the HIVE™ scRNAseq Processing Kit. The software consists of a set of programs that receives demultiplexed FASTQ file inputs and yields a transcript and gene count matrix (CM), aligned BAM file, and a quality metric (QC) file.

The Terra.bio BeeNetPLUS pipeline also performs preliminary scRNAseq analysis using a Seurat-based workflow. Count matrices are input into the workflow, which performs basic filtering, clustering, cell-type annotation, and marker gene analysis. A basic analysis report is generated at the end of this workflow that provides quality metrics, plots, and an R object for additional custom analysis. More information about the BeeNetPLUS workflow can be found on the Honeycomb Biotechnologies Support Page.

BeeNet™ is also available as an executable file that can be run only on Linux™ systems. The **BeeNet™ Software User Guide** along with the accompanying video tutorials **BeeNet™ Download** and **BeeNet™ Running Analysis**, detail how to download, install, and use the software to analyze HIVE™ scRNAseq data on a Linux™ system from the command-line. BeeNetPLUS is not available as a downloadable executable file.

///// HIVE™ scRNAseq

HIVE™ scRNAseq is a picowell array technology that enables users to collect, store, and process single cells into NGS libraries without specialized instrumentation. HIVE™ scRNAseq libraries are indexed with a sample-specific identifier and Illumina® adapters. Each molecule within a sample is labeled with a unique cell barcode to delineate the cell of origin.

Paired-end sequencing libraries are generated by Illumina® sequencing with Read 1 yielding the individual cell barcode and Read 2 yielding the mRNA sequence.

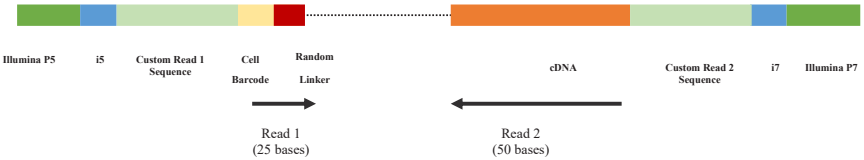


Figure 1: shows the structure of a single library molecule (Illumina P5: Illumina P5 adapter, i5: index 2, i7: index 1, Illumina P7: Illumina P7 adapter)

///// BeeNetPLUS Workflow

Terra.bio is a cloud-native platform for biomedical researchers to access data, run analysis tools, and collaborate. Terra's graphic user interface (GUI) allows users from any background to run and automate workflows without prior knowledge of command-line tools or cloud computing. Terra.bio runs using files placed on Google Cloud Platform (GCP). Therefore, all fastqs must be loaded into a Google Cloud Bucket.

The BeeNetPLUS workflow hosted on Terra.bio is an integrated set of programs that will process Read 1 and Read 2 FASTQ files to create Count Matrices, aligned BAM and QC files, and perform basic downstream analysis. We provide a workspace on Terra.bio that can be used for running the software using cloud resources. To run the BeeNetPLUS workflow on Terra, you will need to set up an account.

Running the BeeNetPLUS workflow on Terra.bio requires a single input file with a list of sample names, corresponding FASTQ file names, and metadata information. The input file is uploaded to the cloud so the file locations and sample metadata can be parsed and processed by the workflow using a simple graphical user interface. Analyses of multiple samples can be initiated at once and run in parallel with a single input file and workflow on Terra.

BeeNetPLUS automated analysis follows three main steps:

- QC and pre-processing of raw FASTQ files
- BeeNet™ Analysis: Alignment and annotation of the reads – outputs a BAM file, QC outputs, and Count Matrices (CM)
- Downstream Analysis (Seurat-based)

Figure 2: shows the BeeNetPLUS analysis workflow

///// Setting Up Terra

Set-up a Google Cloud™ Account

In order to create a Terra account, you will need a Google Cloud™ account. An account can be created using a Gmail™account, an institutional or G Suite™ email, or an account associated with a non-Google address.
https://accounts.google.com/signup

Register with Terra.bio

You will need to create a Terra Account from the below link. Terra registration requires a Google Cloud™ account and access associated with an email address. Please check with your organization's policy regarding GCP and Terra security and network settings.
https://support.terra.bio/hc/en-us/articles/360034677651-Account-setup-and-exploring-Terra

Set up Billing

Once you have completed Google Cloud™ and Terra registrations, you will need to set-up a Google Cloud™ billing account and enable Terra access. If you are using an organizational account for Google Cloud™ or Gmail™, you may need to request an administrator to set-up a billing account, or give you access to the existing billing account.

Finally, you will have to create a billing project on Terra. For more details on how to set up a billing account refer to these instructions:
https://support.terra.bio/hc/en-us/articles/360026182251

Grant Terra.bio Access to Data

For analysis on Terra, the data needs to be either on an external Google Cloud™ (GCP) bucket, or uploaded directly to Terra. In both cases the location of the individual files needs to be specified in the input file.

For external cloud buckets the files can either be directly uploaded to a specified folder using the GUI, or from a terminal using Google Cloud™ SDK command line tools.

For more information on GCP storage buckets use the below link:
https://cloud.google.com/storage/docs/creating-buckets

If you are using external GCP buckets to store data, you need to add Terra Proxy Group information to the appropriate GCP bucket. Find the below instructions for how to do this:
https://support.terra.bio/hc/en-us/articles/360045971452-Accessing-data-from-an-external-bucket

Request Access to Workspace and Docker Images

Terra uses docker images to set-up the runtime environment. Honeycomb™ hosts a suite of Docker images on the GCP container registry for use with Terra. Users who are running Terra need to request access to the Docker images by emailing support@honeycomb.bio with your name, institution, the Proxy Group ID (refer to the "Grant Terra.bio Access to Data" section of this document) and your Terra.bio account email address.

You can find your proxy group ID by logging into Terra and clicking the three lines at the top left of the screen. Click on your name, and then click "profile". Your Proxy ID is listed under where it says "Proxy Group". Your proxy group ID should be an email address with the domain firecloud.org.

///// Available References

BeeNet software runs using unique reference index files. These index files are created with an automated script within BeeNet. Available references will be continually updated. Currently available references are listed on the Honeycomb Biotechnologies Support Page. If any other reference is required, please contact support@honeycomb.bio to have us create and host your custom reference genome.

Human Genome GRCh38 – Input File Tag "hg38"

The reference files were created using the below fasta and gtf :

fasta: http://ftp.ensembl.org/pub/release-104/fasta/homo_sapiens/dna/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz

gtf: http://ftp.ensembl.org/pub/release-104/gtf/homo_sapiens/Homo_sapiens.GRCh38.104.gtf.gz

Source: Ensembl

Human Genome GRCh37 – Input File Tag "hg37"

The reference files were created using the below fasta and gtf :

fasta: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/GRCh37.p13.genome.fa.gz

gtf: http://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_19/gencode.v19.annotation.gtf.gz

Source: Gencode

Mouse Genome mm10 - Input File Tag "mm10"

The reference files were created using the below fasta and gtf :

fasta: http://ftp.ensembl.org/pub/release-104/gtf/mus_musculus/dna_index/Mus_musculus.GRCm39.dna.toplevel.fa.gz

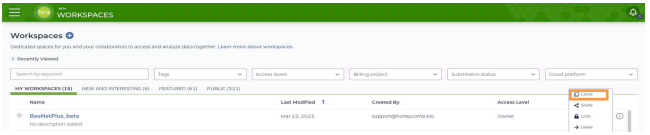
gtf: http://ftp.ensembl.org/pub/release-104/gtf/mus_musculus/Mus_musculus.GRCm39.104.gtf.gz

Source: Ensembl

//// Accessing and Cloning the Workspace

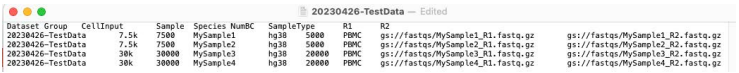
Once you set-up your account and request access with your Terra user email you will receive a share request e-mail from Terra for the BeeNetPLUS Workspace. When you receive the share invitation you should see the "BeeNetPLUS" in your dashboard. You will have reader access to this workspace.

1. Clone the workspace using the 3 dots on the top right corner of the dashboard page. It will prompt you to name and choose billing and once you clone the workspace, this will create a copy of the workspace in your account. Therefore, you will be the owner, which will allow you to run jobs with this workflow.



//// Create and Upload Input File

To run the analysis pipeline, you will need to create an input file with the sample name and GCP data locations (gsutil URIs) for each sample that you want to analyze.



The Input File consist of 9 columns:

- **Dataset:** This is the name of your Experiment. All data from this workflow will be put into a folder with this name. This needs to be the same for all samples within a sample sheet. Independent experiments will require independent runs of BeeNetPLUS.

- **Group:** This is how your samples are grouped together. scRNAseq analysis will be performed between groups. For example, if your experiment compares blood from male and female donors, the Group should be male or female. If your experiment compares treated versus untreated samples, the Group should reflect that metadata.

- **CellInput**: The number of cells loaded into the HIVE.

- **Sample:** The name of your sample. Please do not include any special characters in your sample names.

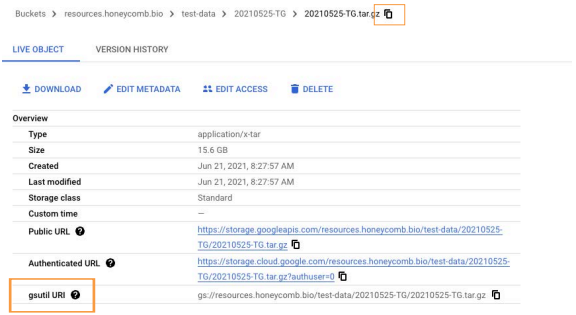
- **Species:** The reference your sample will be aligned to. There are currently 3 reference tags in use: hg37, hg38, and mm10. Custom references can be generated and added to this list (Refer to the References section for more details).

- **NumBC:** The number of cell barcodes BeeNet will output. We recommend 60% of the Cell Input as a good starting point for this metric. This number must be an integer

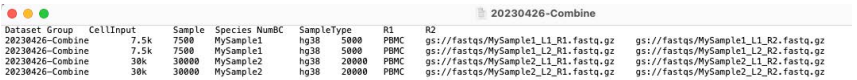
- **SampleType**: The type of sample. To take advantage of Honeycomb automatic-annotation data, enter "FBL", "BM", or "PBMC" if your sample type is filtered blood, bone marrow, or PBMCs.

- **R1**: The Read1 FASTQ file gsutil URI. Gsutil URIs for files on GCP can be found and copied by navigating to the GCP bucket for the individual file and clicking the page icon next to the path. Be sure to insert "gs://" in front of the pasted URL. Alternatively, you can get the full gs://...address by clicking a link further down in the page.

- **R2**: The Read2 FASTQ file gsutil URI. Gsutil URIs for files on GCP can be found and copied by navigating to the GCP bucket for the individual file and clicking the page icon next to the path. Be sure to insert "gs://" in front of the pasted URL. Alternatively, you can get the full gs://address by clicking the gsutil URI further down the page.

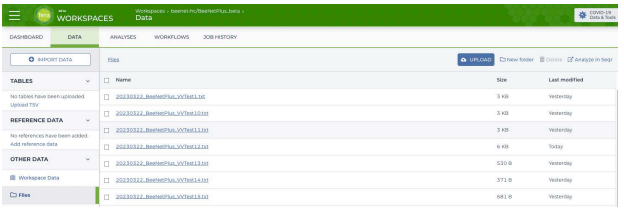


Note: for Samples with more than 1 pair of FASTQ files, please put each pair of fastqs on a separate line of the input file. BeeNetPLUS will combine FASTQ pairs with the same sample name, so please ensure the sample names match for each pair of FASTQs (see below for example).



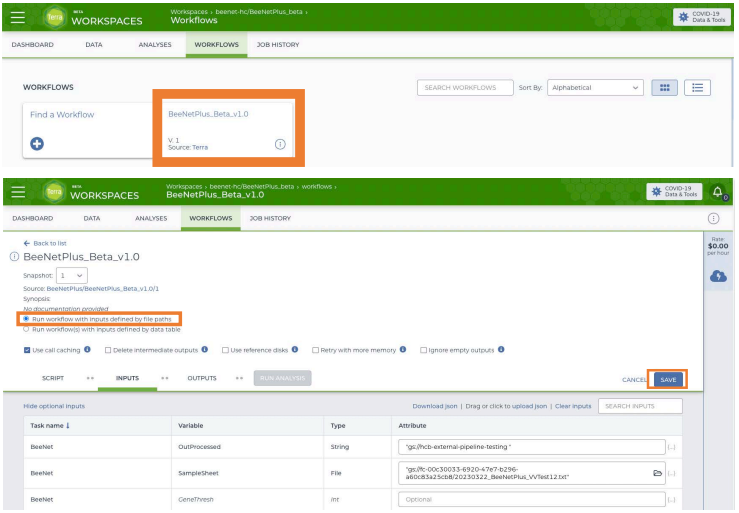
Input File format: All fields must be tab separated (.txt file)

The Input file can be uploaded to any GCP storage bucket that Terra.bio is authorized to access and the URI for the file can be used as an input. Alternatively, it can be uploaded directly to Terra. To upload your input file to Terra, go to DATA>Files. Click on the blue "Upload" folder. Choose your input file and click update. When you use this format, you can click on the folder icon in the Inputs and choose your input file.



///// Running Analysis

Once you have the workspace cloned, go to your dashboard/workflows tab and choose the version 1 workflow. This will take you to an input page. Select **"Run workflow with inputs defined by file paths"** option and fill out the Inputs based on the information on the next section.

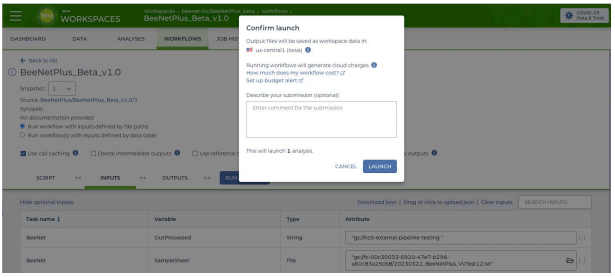


Workflow Inputs tab

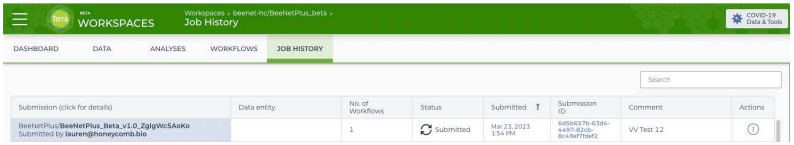
The BeeNetPlus workflow has 2 required and 4 optional inputs. String and File Types must be in quotations.

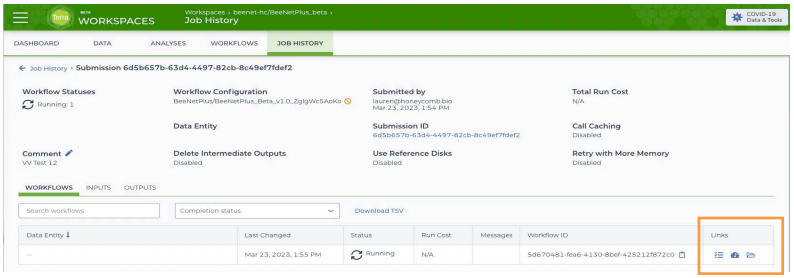
- **OutProcessed:** GCP bucket location for output files. You need to make sure that Terra has access to this bucket by adding the proxy as explained in Grant Terra.bio access to your data section of this document. Output must be in quotations.
- **SampleSheet:** URI Link to the input file with sample names and data location. Input file can either be uploaded to a GCP bucket or directly to Terra workspace data folder. SampleSheet URI must be in quotations. More information about creating the sample sheet can be found in the "Input File" section of this document.
- **GeneThresh:**The minimum number of genes/cell that is required to define a cell. Default is 100. This must be an integer and must not contain quotations.
- **SeuratRAM**: The amount of RAM to allocate to the software to perform downstream analysis. Default is 240G. This must be a string and must contain quotes (e.g. "300G").
- **SeuratStorage**: The amount of disk space to allocate to the software to perform downstream analysis. Default is 400G. This must be written as such: "local-disk 400 SSD" for 400GB.
- **TranThresh:** The minimum number of genes/cell that is required to define a cell. Default is 200. This must be an integer and must not contain quotations.

Once all the inputs are completed in the correct format start the analysis by clicking 'Run Analysis'. A prompt will appear to confirm launch. To keep track of your experiments, we recommend putting the dataset name in the Comment Section. Click 'launch' to start your run. You should now see this analysis queued in your job history.



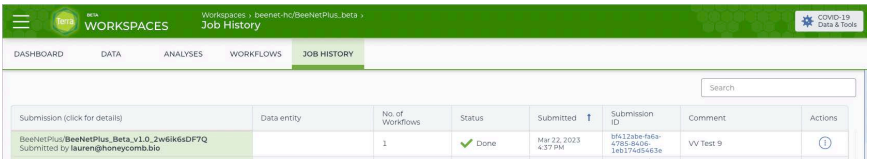
The status will change to "Running" and you can follow its status by clicking on this submission. You can access the data folders in the cloud by clicking the workflow or





///// Accessing Outputs

All your runs will be listed under job history tab on your workspace. If your analysis is completed successfully your run will show as green and will be displayed as "Done".



To access results, navigate to the GCP storage bucket specified in the workflow input for the output files (refer to workflow inputs for details). The files will be moved to that bucket with the following file structure:

out/dataset/

For example if you have specified "gs://honeycomb-output" as your output directory and "Experiment_1" as your dataset name your output files can be found at:

gs://honeycomb-output/Experiment_1/

This output folder will contain files from BeeNet™ alignment and count matrix creation. These file folders are separated by sample name. This output folder will also contain files from Secondary Analysis performed using Seurat. This file folder will be named **Analysis-YYYYMMDD-GeneThres-TranThres**.

For more information on how to use Terra.bio workspaces and storage please refer to their instructions at https://support.terra.bio/hc/en-us/articles/360024743371

///// BeeNet™ Output File Naming

BeeNet™ outputs an aligned BAM file, which is used to create the count matrices. In addition to the count matrix files there are multiple summary files with QC information related to the sequencing data. Files will be named automatically based in the –sample-name flag as below:

SampleName_YYYYMMDD_filename.extension

e.g., Sample1_20201201_CM.Reads.tsv.gz

file extensions:

RCM – Read Count Matrix

TCM – Transcript Count Matrix

CMSummary – Count Matrix Summary

///// BeeNetPLUS scRNAseq Analysis Output File Naming

Output files from the BeeNetPLUS workflow will be automatically named based on the information provided in the sample sheet in the Dataset, Group, and Sample columns. Output files are zipped into one file for download to your local computer, cloud, or cluster.

///// Output Files List

There will be a single unsorted, unindexed BAM file for each sample. The cell barcode for each read is contained in the XC tag for each read.

List of expected output files are as below:

BeeNet Outputs:

Sample1_20201201.bam

Sample1_20201201_RCM.tsv.gz

Sample1_20201201_TCM.tsv.gz

Sample1_20201201_CMSummary.tsv

Sample1_20201201_ ReadsQC.tsv

Sample1_20201201_ SampleQC.tsv

FASTQC outputs for each lane:

L1_R1_fastqc.html

L1_R2_fastqc.html

L1_R1_fastqc.zip

L1_R2_fastqc.zip

List of expected output files are as below:

scRNAseq AnalysisOutputs (contained in a zipped folder):

Experiment1.txt

ClusterMetrics-Experiment1.txt

ClusterMetricsByGroup.txt

FullCluster_Enrichments.txt

FullCluster_Enrichments_CellType.txt

FullDataset-Experiment1.Rdata

HCB-scRNAseq-Experiment1.html

HCB-scRNAseq-Experiment1_files

SamplesMetrics-Experiment1.txt

In the Analysis output folder, there contains an .html report that summarizes the results of the BeeNetPLUS workflow. This document can be viewed by navigation to the location on the GCP and hitting the 'Download' button. This will open up the report in your Internet browser to be viewed. All files and images embedded within the html report can be found in the folder **HCB-scRNAseq-<DATASET>_files**. There are also additional images found in that folder that were not embedded in the report.

NOTE: Do not save the HTML report using File -> Save As on your Internet browser. This will affect the formatting of the report. There is an additional copy of the HTML report embedded in the zipped output file for download to avoid this error.

///// BeeNet™ Output File Descriptions

All Files with barcodes are alphabetized by the barcodes.

***RCM.tsv.gz**- displays the number of reads for each unique cell barcode that maps to a specific gene in the reference genome

- **Gene:** UniProtKB Gene Name or HGNC Symbol
- **'Headers':** barcodes for each cell - unique cell barcode of 12 bases

***TCM.tsv.gz** - displays the number of unique molecule counts for each transcriptome that mapped to a specific gene in the reference genome

- **Gene:** UniProtKB Gene Name or HGNC Symbol
- **'Headers':** unique transcriptome barcode of 12 bases

***CMSummary.tsv** - displays the number of total Genes and number of molecule counts for each transcriptome barcode

- **Barcode:** unique cell barcode of 12 bases
- **nGenes:** total number of genes for each cell barcode
- **nTran:** total number of molecule counts for each cell barcode

***ReadsQC.tsv** - QC metrics per cell barcode

- **TotalReads:** Total reads for an individual cell barcode
- **MappedReads:** Reads from an individual cell barcode that map to the reference genome
- **ExonReads:** Reads from an individual cell barcode that map to exons
- **FilteredReads:** Total Reads that have passed the filtering prior to alignment
- **PolyAReads:** Reads that contain PolyA
- **5PFReads:** Reads that were filtered out due to having adapter sequence present in 5' end
- **3PFReads:** Reads that were filtered out due to having adapter sequence present in 3' end
- **badBaseBC:** Reads that were filtered out due to 2 or more bases in cell barcode with poor phred scores

***SampleQC.tsv** – QC metrics for all the reads in the FASTQ files for the sample

- **TotalReads:** Total reads in the FASTQ files for the sample
- **MappedReads:** Total Reads that map to the reference genome for the sample
- **ExonReads:** Total Reads that map to exons for the sample
- **FilteredReads:** Total Reads that have passed the filtering prior to alignment for the sample
- **PolyAReads:** Total Reads that have passed the filtering
- **5PFReads:** Total Reads that were filtered out due to having adapter sequence present in 5' end
- **3PFReads:** Total Reads that were filtered out due to having adapter sequence present in 3' end
- **badBaseBC:** Total Reads that were filtered out due to 2 or more bases in cell barcode with poor phred scores

///// BeeNetPLUS Analysis Output File Descriptions

<Dataset>.txt- This is a copy of the sample sheet that was imported into the Terra.bio workflow.

ClusterMetrics-<Dataset>.txt: Sequencing Metrics associated with each Seurat Cluster

- **TotalReads:** Total reads for an individual cell barcode
- **MappedReads:** Reads from an individual cell barcode that map to the reference genome
- **ExonReads:** Reads from an individual cell barcode that map to exons
- **nCount_RNA:** Median number of transcripts per cell barcode
- **nGenes:** Median number of genes per cell barcode
- **percMito:** The fraction of reads that map to mitochondrial genes
- **nCells:** The number of cell barcodes

ClusterMetricsbyGroup.txt: Sequencing Metrics associated with each Group. See above for description of each row.

FullCluster_Enrichments.txt: Differential Gene Expression analysis by cluster. This analysis was performed using the FindAllMarkers function in Seuratv4.0.5. See the html output report for explanations for each column.

FullCluster_Enrichments_CellType.txt: Differential Gene Expression analysis by cell type. This analysis was performed using the FindAllMarkers function in Seuratv4.0.5. See the html output report for explanations for each column.

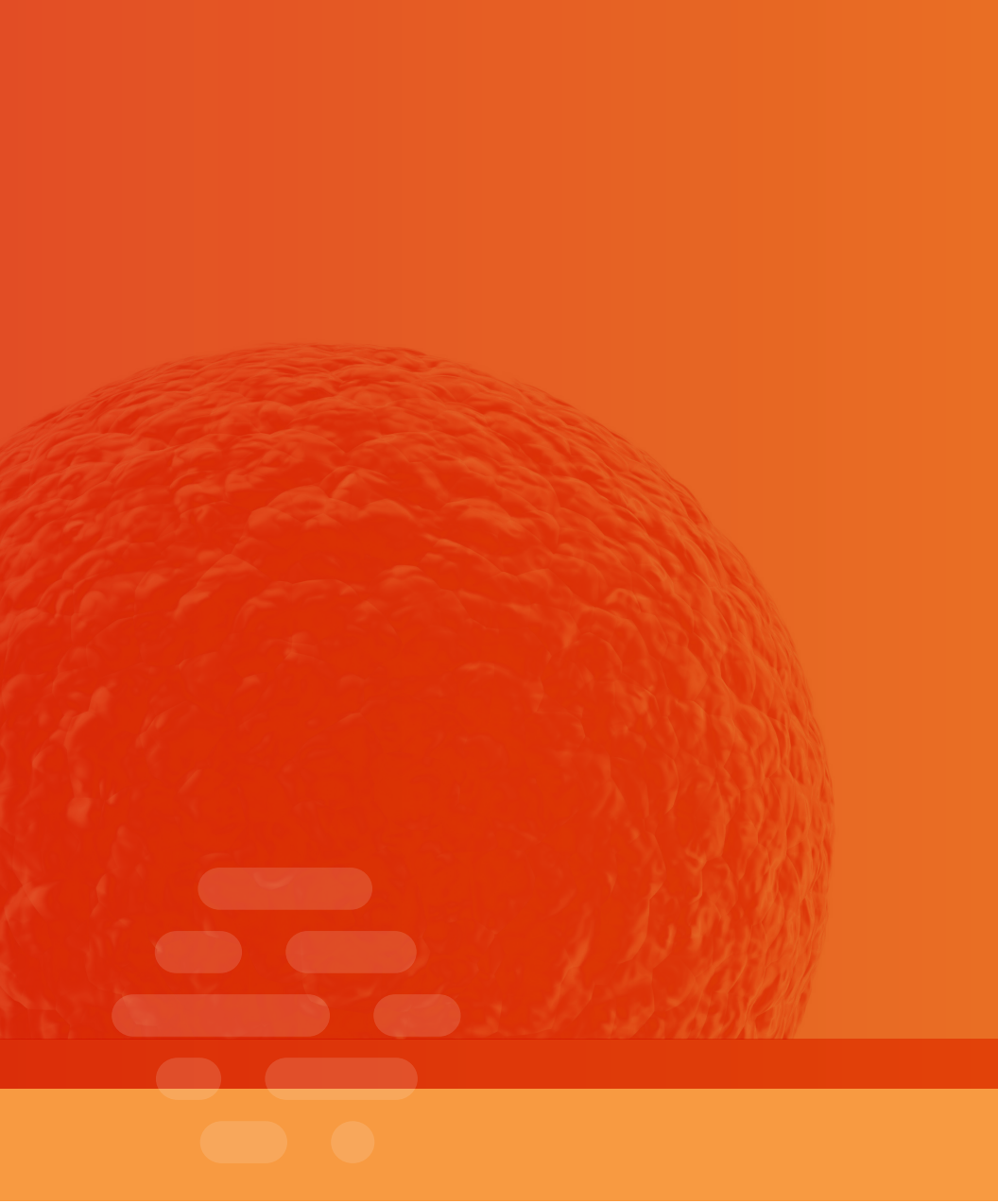
FullDataset-<DATASET>.Rdata: R object containing the Seurat object, metrics, cell type annotation, and metadata information. This R object can be used for further analysis.

HCB-scRNAseq-<DATASET>.html: html output report containing the results of BeeNet and scRNA-seq analysis.

SamplesMetrics-<DATASET>.txt: Sample Metrics for each sample in the analysis.

- **AnalysisDate:** Date the analysis was performed.
- **Dataset:** Name of the Experiment
- **SampleID:** Name of the sample
- **Group:** Group identifier for each sample
- **SampleType:** Sample Type for each sample
- **GeneThreshold:** User-specified minimum genes/cell threshold
- **TranscriptThreshold:** User-specified minimum transcripts/cell threshold
- **CellInput**: Number of cells loaded into the HIVE
- **NumBC:** Number of barcodes input into BeeNet
- **TotalReads:** Median Total reads for an individual cell barcode within a sample
- **MappedReads:** Median Reads from an individual cell barcode that map to the reference genome within a sample
- **ExonReads:** Median Reads from an individual cell barcode that map to exons within a sample
- **nCount_RNA:** Median number of transcripts per cell barcode
- **nGenes**: Median number of genes per cell barcode

- **percMito:** The percent of molecules that map to mitochondrial genes
- **nCells:** The number of high quality cell barcodes recovered
- **FractionRecovered:** The number of high quality cells recovered divided by the number of cells loaded into the HIVE.
- **Complexity:** The number of exon reads per cell barcode divided by the number of transcripts recovered for that barcode. This value is the median number of that calculation for each sample. This is a measurement of sequencing saturation
- **SampleTotalReads:** The total number of reads for the entire sample
- **SampleFilteredReads:** The total number of filtered reads for the entire sample
- **SampleMappedReads:** The total number of mapped reads for the entire sample
- **SampleExonReads:** The total number of exon-mapped reads for the sample
- **HQCellsTotalReads:** The total number of reads for all high quality cells that passed gene and transcript thresholds
- **FracReadsHQCells**: The fraction of total reads that map to high quality cells
- **PctExon:** The percentage of reads that map to exons in high quality cells
- **FracPassFilter:** The fraction of reads that passed filtering in high quality cells
- **SeqSat:** Sequencing Saturation: 1- (number of unique valid cell-barcode,transcript combinations) / number of mapped reads)
- **nFastq:** The number of paired fastq reads associated with a sample
- **AlignedRef:** The reference genome that was used for alignment
- **BeeNet:** The version of BeeNet used
- **SecondaryPipeline:** The version of the downstream analysis pipeline used
- **WDL:** Version of BeeNetPLUS used



HONEYCOMB