TELL-Seq™ BaseSpace Phasing Application User Guide

For Research Use Only. Not for use in diagnostic procedures.

Document # 100041 v1.0

December 2022

This document describes procedures for uploading the sequencing data to BaseSpace and running the Universal Sequencing Technology's Tell-Read and Tell-Sort applications. Note that using BaseSpace CLI requires familiarity with working in a command line environment, and builds are available for Linux, Windows, and Mac OS X.

Revision History

Document Number	Date	Description of Change
Doc #100041 v1.0	December 2022	Initial Release

1. Setting up BaseSpace Sequencing Hub (BSSH) CLI

There are a few situations where you need to use BSSH Command Line Interaction (CLI) to prepare your run data for analysis on BaseSpace, for example, uploading your run data (if they are not already on BaseSpace), genome reference files, VCF references, etc.

Follow a few steps outlined below to install BaseSpace Sequencing Hub CLI and connect to BaseSpace.

- Download the latest BaseSpace Sequencing Hub Command Line Interface (CLI).

  Detailed instructions on how to download the latest CLI for different operating systems can be found here:

  https://developer.basespace.illumina.com/docs/content/documentation/cli/cli-overview

- Connect to BaseSpace using the following command:

  % bs auth

  This will provide a URL. Enter this URL into a Web Browser and log into BaseSpace to authenticate this account and link the BS CLI to the BaseSpace account. After authentication, go to existing Projects and Runs on the account can viewed at the command line with the following commands:

  % bs project list

  % bs run list

2. Uploading raw directory of a sequencing run to BaseSpace (optional)

Sequencing data is often uploaded to BaseSpace automatically after the run is completed. To manually upload the raw directory of a run, choose a name to call the run, identify the instrumentation, and upload the data using the following command:

% bs upload run -n <NAME_OF_RUN> -t <MACHINE> <RAW_RUN_DIR>

3. Creating a new Project on BaseSpace (optional)

New Projects can be created on the BaseSpace website directly.

If you prefer to use the CLI, use the following command:
% bs project create -n <NAME_OF_PROJECT>

The new project will be assigned an ID, which is needed when uploading reference sequences.

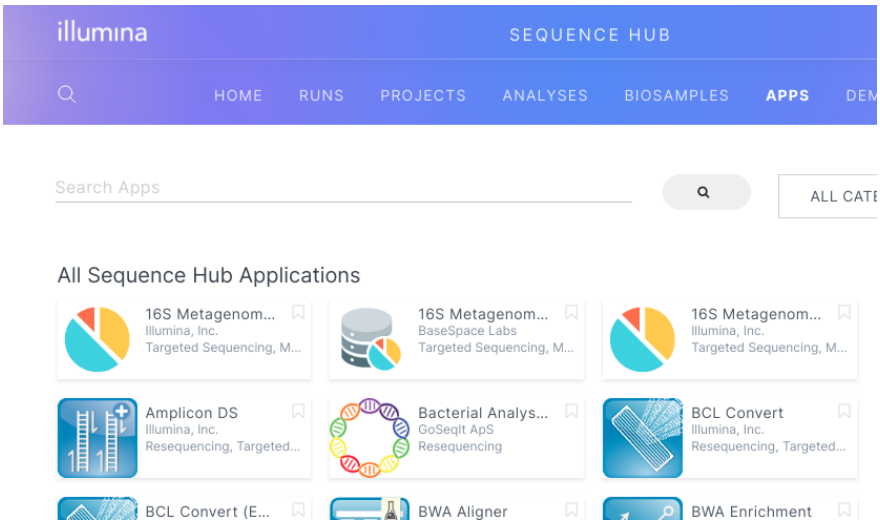
4. Uploading a reference fasta file to BaseSpace (recommended for Tell-Read and Tell-Link)

Upload a reference fasta file to a Project in BaseSpace using the following command:

% bs upload dataset -p <ID_OF_PROJECT> -t common.files <FASTA>

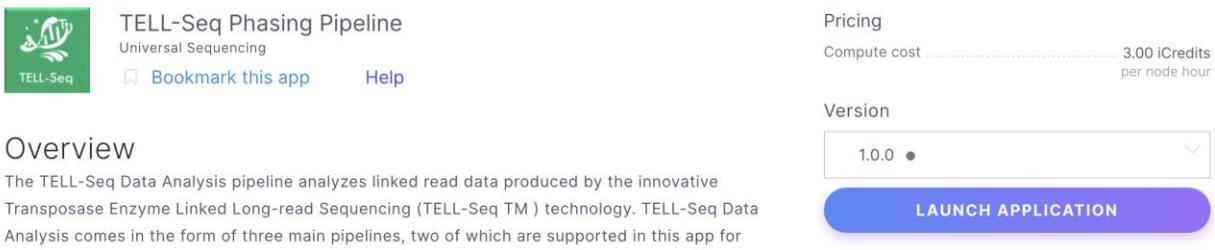
5. Starting Tell-Seq Phasing App from an uploaded run

a. Select the TELL-Seq Phasing App from the list of Apps on BaseSpace.

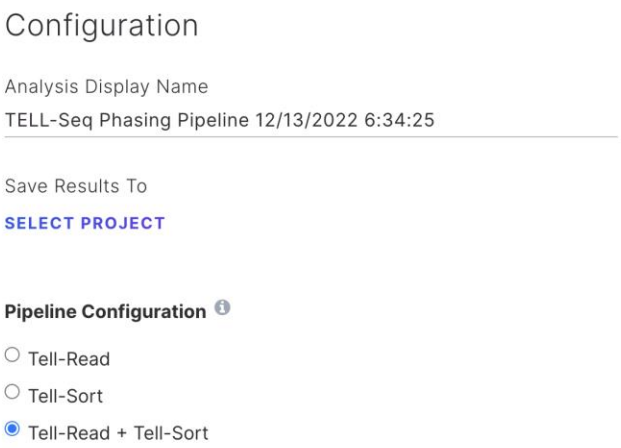


b. Locate the "TELL-Seq Data Analysis Pipeline – Phasing",

c. Click on the TELL-Seq Phasing Pipeline App



d. Click on Launch the Application. You will be directed to the following page:



e. In the Configuration section, 'Analysis Display Name' field will have the name shown in the list of all analyses associated with the user account. By default, the 'Analysis Display Name' will be TELL-Seq Phasing Pipeline followed by date and time. If necessary, users can change the file name by simply typing in 'Analysis Display Name' field.

f. In the Pipeline Configuration section, users can either run an individual pipeline or in combination with other pipelines depending upon the type of analysis required.

Phasing workflows comprise two parts: Tell-Read, followed by Tell-Sort. They can be run with three pipeline configuration options.

1. Run Tell-Read only

Tell-Read processes raw BCL image data from a sequencing Run directory uploaded to Basespace (Refer Uploading to Basespace user guide). It converts BCL data into FASTQ data and goes through several QC steps to generate clean TELL-Seq linked-read FASTQ BioSamples for downstream processes.

This is a required step. Users use Tell-Read only configuration when they want to process multiple samples at the same time. However, due to the runtime limitations that the BaseSpace platform poses, the process will fail if the sample sizes are exceptionally large (more than 1.5 billion pair-end reads). To avoid any kinds of failure, users should just pick a single sample to process. To work around this limit, users can start multiple processes simultaneously, each running a separate sample.

2. Run Tell-Sort only

   This step is to generate the phasing results. Users should select this pipeline configuration only if they have previously generated BioSample FASTQ files on BaseSpace and would like to run the phasing analysis using different reference files or parameters.

   Only one sample can be analyzed in Tell-Sort. For multiple samples, users can start multiple runs concurrently.

3. Run Tell-Read and Tell-Sort combined

   Users can run this pipeline configuration if they have not previously processed sequencing run data and prefer to generate phasing results. Note in this configuration, BioSample FASTQ files will be generated by Tell-Read. They can be used for future Tell-Sort runs as well. This configuration allows only one sample to be selected for the workflow. If multiple samples need to be processed simultaneously, users can run multiple sample analyses concurrently.

To specify reference sequence(s) to use, click on SELECT DATASET FILE(S) and select fasta file(s) that have been already uploaded to BaseSpace.

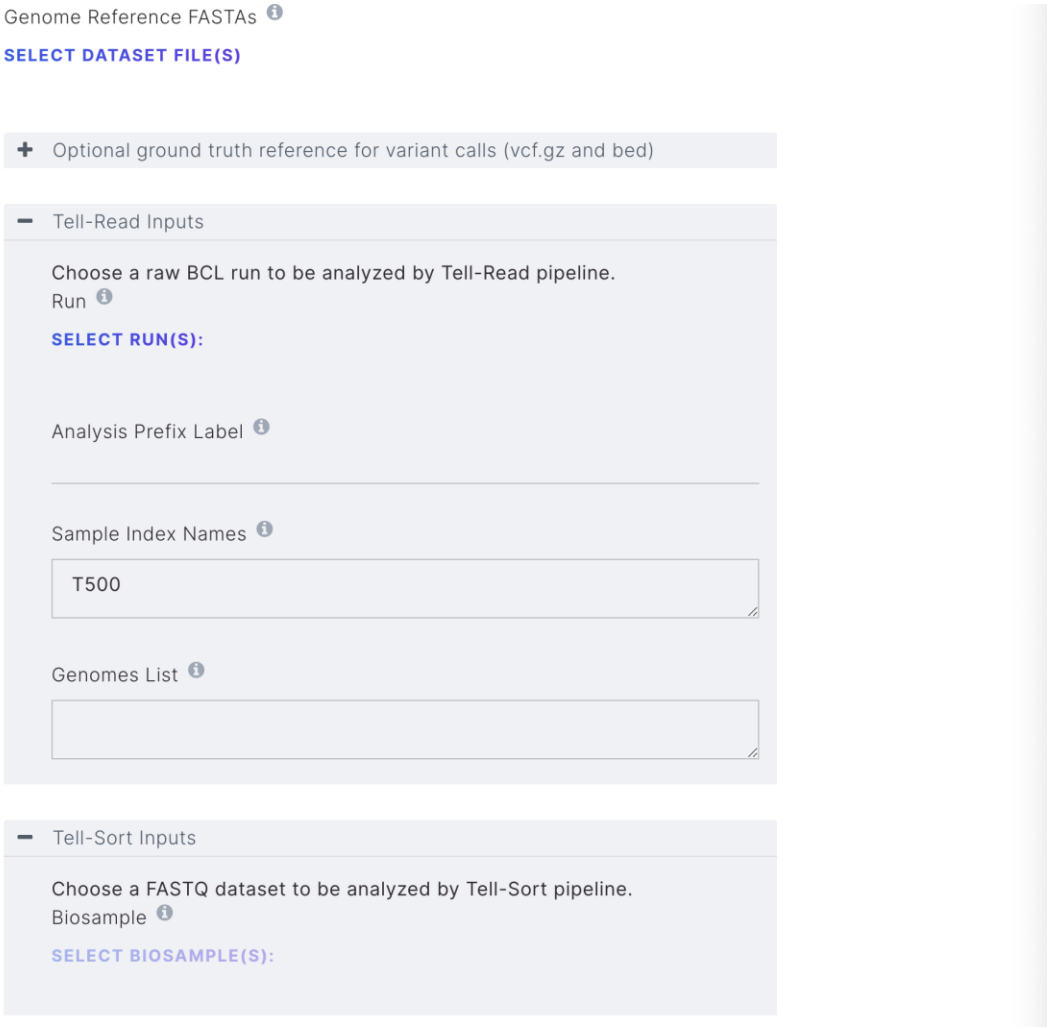
In this step, select and add all fasta references needed to be used for samples to be analyzed. However, for each genome in the samples, its reference file only needs to be selected once.

The Analysis Prefix Label is a user chosen identifier to be used to group and identify the analysis results.

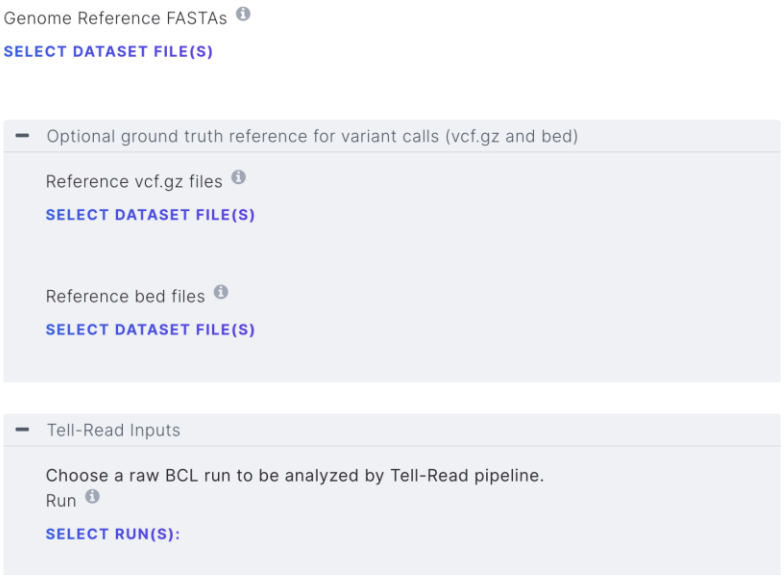
In the Sample Index Names field, user should input a comma delimited TELL-Seq sample IDs, e.g., T501, T502, etc.

In the Genome List field, users should input a comma delimited list of genome reference names. The genome reference name is the portion of the fasta file name without ‘.fasta’, or ‘.fa’ file extension. For example, in the Genome Reference FASTAs, the genome reference file user selected is hg38.fasta, the genome reference name to be used in Genomes List should be hg38.

The number of items in Sample Index Names and Genomes List fields must match and in one-to-one correspondence. Even if two samples in the Sample Index Names use the same genome reference, the corresponding genome reference name must show up twice in the Genome List.



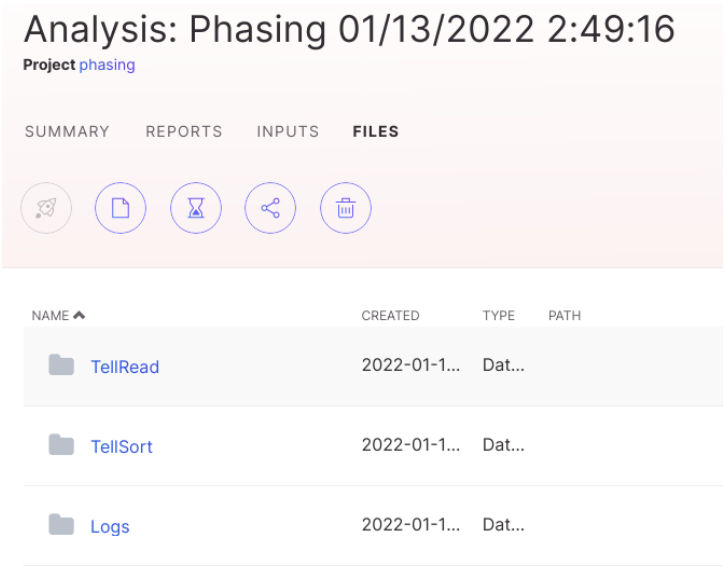
Click on '+' to specify optional ground truth reference for variant calls.



Click on Launch Application at the bottom of the page to start the analysis.

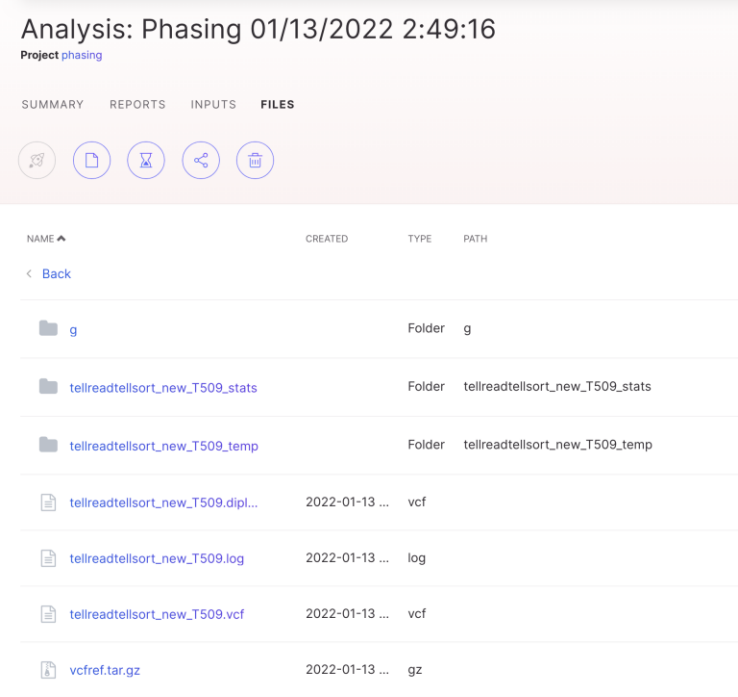
6. Viewing and Downloading Results from BaseSpace

Upon completion on the analysis, click on the “FILES” link:

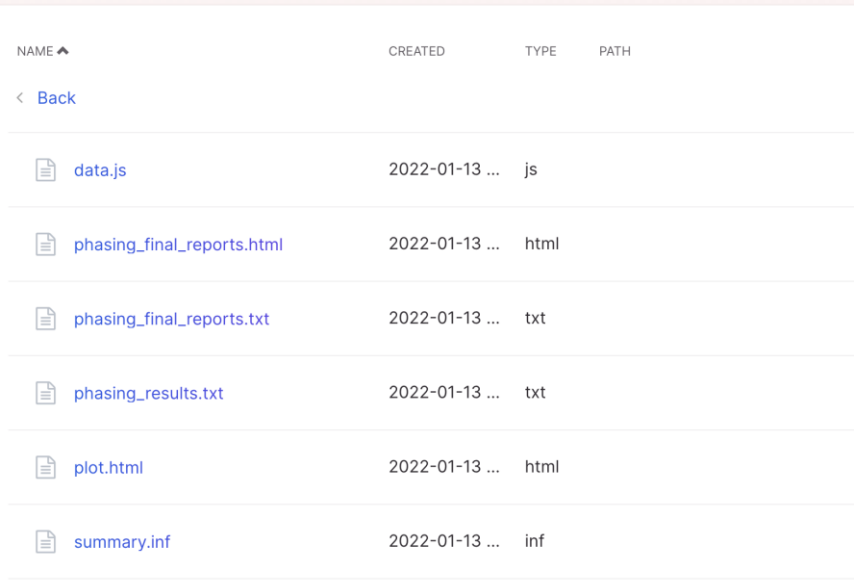


There are two main folders in the output - TellRead and TellSort.

- For Phasing analysis results, select the **TellSort** folder to view the files

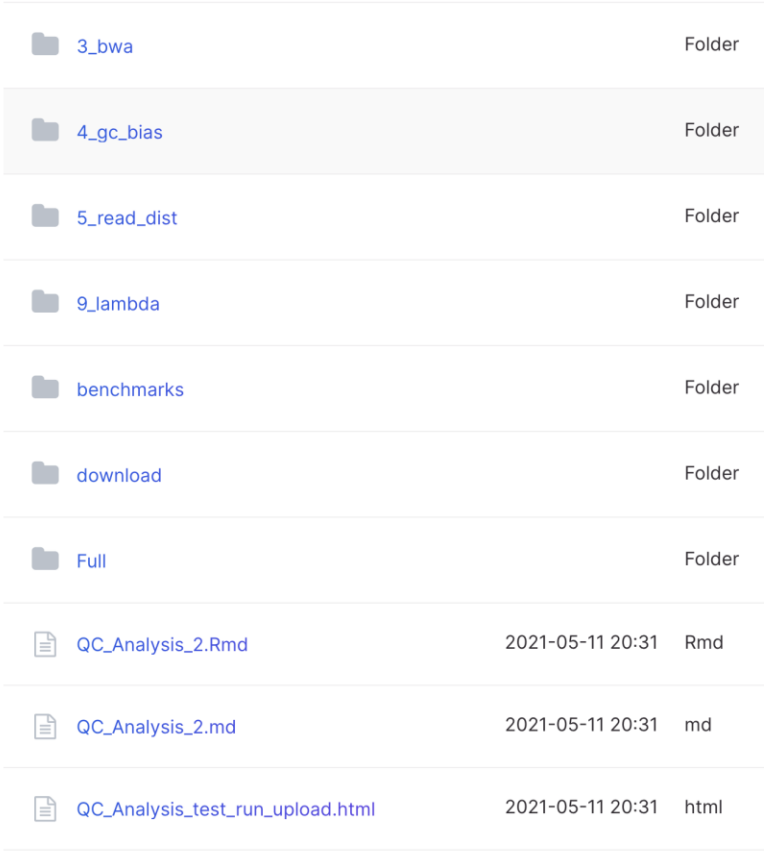


Click on <run>_<sample>_stats folder to view the content.



Click ‘Phasing_final_report.html’ to download phasing report file.

- For accessing the QC data of the sequencing run, select the **TellRead** folder.



To view the QC report, download the QC_Analysis_xxxxxx.html file, where xxxxxx is the Analysis Prefix Label.