**Guidelines for T. pallidum sample receiving, processing, and data storage**

DRAFT, 2024-03-04, by Nicole Lieberman, PhD

**Purpose: Implementation and iterative improvement of this protocol should ensure all sample raw data and physical parameters are readily available.**

All sample logs, raw data, stats, and processed data will be available in S3 directory

**s3://fh-pi-jerome-k-eco/greninger-lab/greninger-lab-file-share/nicole/TP/**

Within this directory, there are several files and folders:

/TP\_Master\_v2.xlsx: ALL samples, all types should be logged here

/manifests\_and\_metadata/

/tprK/

/WGS/

/tpRNAseq/

(other folders are being retroactively organized by Nicole, do not touch)

**Sample Receiving and Login:**

Samples must always be received alongside a manifest (clinical samples may also include metadata). Upon receipt of samples and manifest, tubes should be cross checked prior to storage to ensure:

-all samples present

-tube labels unambiguous

-volumes approximately as expected

Once these checks have been performed, a new folder should be created in the manifests\_and\_metadata directory. The folder should be named in YYYY-MM-DD format by the date samples are received at SLU. Any documents received alongside samples should be stored in this folder.

Samples must also be logged in the TP\_Master\_v2.xlsx master spreadsheet file, including sample type and project, intended analyses (WGS, tprK, tpRNAseq), contact person, date of arrival, etc. Importantly, a library name should be assigned. The library name should:

-begin with a letter, not number

-be less than 20 characters

-contain only letters, numbers, or -. No underscores or periods.

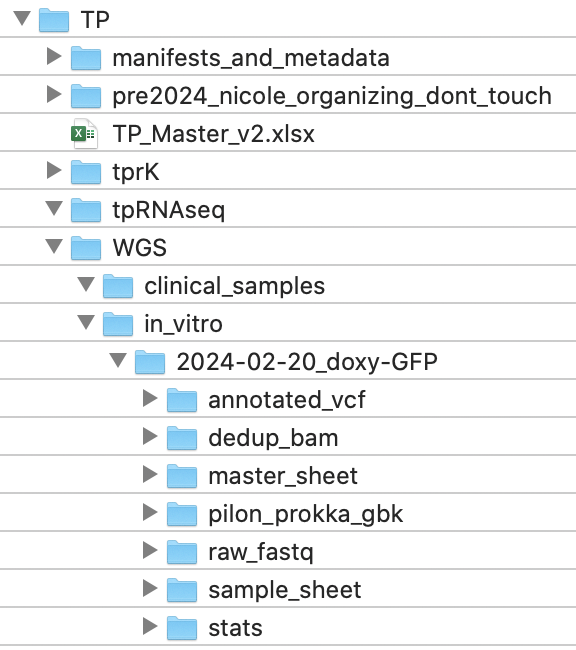
-be completely unambiguous – no duplication with past libraries.

Samples will also receive a unique sample\_id (pre-filled on spreadsheet).

**Data Generation and Storage**

WGS: For samples that need WGS, the relevant lines and header row of the master spreadsheet should be copied to a new file and sent to 1616 alongside the samples. This file should also be placed in the manifest\_and\_metadata folder.

Each analysis/data type has its own directory: WGS, tprK, and tpRNAseq. Within each of these directories, there are folders for in\_vitro and clinical\_samples. All sample QC, raw data, and results will be stored by prep/batch – directory should be named with approximate prep date (or date of arrival at SLU) using YYYY-MM-DD and brief description. An example of directory structure of in vitro samples having WGS is shown below:



-master\_sheet directory should contain all library prep and pooling information

-sample\_sheet directory should contain barcode information

-raw\_fastq should include paired end reads for all samples.

-lanes should be concatenated and names tidied: format should always be:

libname\_R1.fastq.gz, libname\_R2.fastq.gz

-dedup\_bam, pilon\_prokka\_gbk, and annotated\_vcf directories should contain deduplicated bams from initial reference mapping, final prokka-annotated genbank files, and flat files containing SNPs, respectively. No additional files should be included.

-stats directory should contain trimming, mapping, and consensus call information

***Because scripts will be written to manipulate files reproducibly, it is imperative that directory structures and folder names be consistent.***

TprK: For clinical or in\_vitro samples that need tprK sequencing, all samples should be named accordingly: “libname\_tprk”, with the \_tprk added to differentiate between WGS and tprk sequencing. A similar directory structure to WGS will be followed:

/tprK/

-clinical\_samples

-in\_vitro

-YYYY-MM-DD\_sample\_description

-master\_sheet

-raw\_illumina

-trimmed\_merged

-raw\_pacbio

-illumina\_counts

-pacbio\_counts

-raw illumina data should be demultiplexed fastqs

-trimmed\_merged illumina data should have had R1-R2 merged to single end

-raw\_pacbio should contain the raw data from the sequencer (LARGE file).

-illumina\_counts should contain only the final\_data\_seqs.csv files for each sample

-pacbio\_counts should contain the demultiplexed, denoised fastas with primers removed (.noprimers.filtered.RAD.nolines.fix.fasta)

Following analyses, a folder called “analysis” or similar should be added to each directory as appropriate.