

SURPI

Reference data

July, 2014

SURPI 1.0.16 includes programs that will create the necessary reference data in order to execute the SURPI pipeline. All necessary reference data will be downloaded from external sources (either NCBI, or a server hosted by the Chiu lab). The procedure may take roughly 19-24 hours, depending on your network, processing, and disk I/O speeds. The process will do the following:

- download all raw data (as FASTA files)
- verify (where possible) downloads using md5sum
- decompress downloaded data
- Index downloaded data in one of 3 ways:
 1. SQLite database creation (for taxonomy)
 2. SNAP indexing (using snap index)
 3. RAPSearch indexing (using prerapsearch)

This process is entirely scripted, and is currently being released as a public beta. Below are instructions on how to create SURPI reference data.

- 1) Go to an empty directory on a machine with SURPI 1.0.16 installed. Up to 2TB of free space may be used during the indexing process, though the final indexed databases will take approximately 1.3TB (as of 7/2014)

There is some currently some error-checking built into the procedure (with regard to already downloaded data files), but it is not yet complete. It is best to always start this procedure with an empty directory.

- 2) Type the following to start the indexing process:

```
nohup create_SURPI_data.sh > refdb.log 2> refdb.err &
```

This will start the download/creation of all SURPI necessary data. You can view the refdb.log and refdb.err files to monitor the progress of the reference data creation process. Upon completion, a message resembling the following will appear at the end of the logfile:

```
Sun Jul 6 16:37:53 PDT 2014 create_SURPI_data.sh Completed creation of SURPI reference data.
```

- 3) Once the `create_SURPI_data.sh` program has completed, SURPI reference data should be within the folder entitled `reference`. Move this folder to your desired location, and adjust the parameters in your SURPI config file (in the section entitled *Server related values*) in order to run SURPI using this reference data.

Likely modification will include:

- Replacement of `RAPSearch_NR_db` with path to updated database: (e.g. `rapsearch_nr_07052014_db_v2.12`)
- Adjustment of paths to all databases from `/reference` to wherever you locate the data

Below is the relevant section of the config file:

```
#####  
# Server related values  
#####  
  
# SNAP-indexed database of host genome (for subtraction phase)  
SNAP_subtraction_db="/reference/snap_index_hg19_rRNA_mito_Hsapiens_rna"  
  
# directory for SNAP-indexed databases of NCBI NT (for mapping phase in comprehensive mode)  
# directory must ONLY contain snap indexed databases  
SNAP_COMPREHENSIVE_db_dir="/reference/COMP_SNAP"  
  
# directory for SNAP-indexed databases for mapping phase in FAST mode  
# directory must ONLY contain snap indexed databases  
SNAP_FAST_db_dir="/reference/FAST_SNAP"  
  
#Taxonomy Reference data directory  
#This folder should contain the 3 SQLite files created by the script "create_taxonomy_db.sh"  
#gi_taxid_nucl.db - nucleotide db of gi/taxonid  
#gi_taxid_prot.db - protein db of gi/taxonid  
#names_nodes_scientific.db - db of taxonid/taxonomy  
taxonomy_db_directory="/reference/taxonomy"  
  
#RAPSearch viral database name: indexed protein dataset (all of Viruses)  
#make sure that directory also includes the .info file  
RAPSearch_VIRUS_db="/reference/RAPSearch/rapsearch_viral_aa_130628_db_v2.12"  
  
#RAPSearch nr database name: indexed protein dataset (all of NR)  
#make sure that directory also includes the .info file  
RAPSearch_NR_db="/reference/RAPSearch/rapsearch_nr_130624_db_v2.12"  
  
ribo_snap_bac_euk_directory="/reference/RiboClean_SNAP"
```