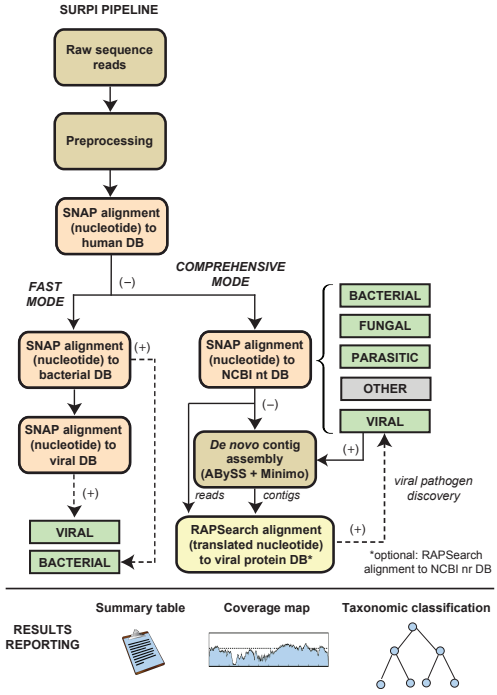
SURPI

Output Interpretation

June, 2014

A schematic overview of the SURPI pipeline (adopted from SURPI paper). Raw NGS reads are preprocessed by removal of adapter, low-quality, and low-complexity sequences, followed by computational subtraction of human reads using SNAP. In fast mode, viruses and bacteria are identified by SNAP alignment to viral and bacterial nucleotide databases. In comprehensive mode, reads are aligned using SNAP to all nucleotide sequences in the NCBI nt collection, enabling identification of bacteria, fungi, parasites, and viruses. Unclassified reads and contigs generated from *de novo* assembly are then aligned to a viral protein database using RAPSearch for pathogen discovery of divergent viruses. SURPI output includes a list of all classified reads with taxonomic assignments, a summary table of read counts, and both viral and bacterial genomic coverage maps.



This document will explain the derivation of the output files. The output discussed is generated by following the *How_to_run_SURPI.pdf* document.

SURPI in comprehensive mode generates the following folders:

```
DATASETS_SRR1106548
deNovoASSEMBLY_SRR1106548
LOG_SRR1106548
OUTPUT_SRR1106548
TRASH_SRR1106548
```

Most of what will be described below, is contained in the folder `OUTPUT_SRR1106548`.

• Files ending in ".annotated" are in SAM format (for SNAP alignments), or -m8 format (for RAPSearch alignments), with taxonomic information added to the end of each row.

• Files ending in ".counttable" are tab-delimited summary tables whereby rows represent taxonomic annotations at various levels (family, genus, species, gi, lineage), columns represent individual barcodes found in the dataset, and cells contain the number of reads.

SNAP alignment (nucleotide) to NCBI nt DB

All reads mapping to NCBI nt DB (NCBI non-redundant nucleotide collection)

Results of alignment of preprocessed dataset and computationally subtracted reads against the human genome at high stringency against NCBI nt DB at high stringency. This file is sorted by the edit distance:

`SRR1106548.NT.snap.matched.fulllength.all.annotated.sorted`

The files below are subsets of this file. They are created by parsing the lineage field.

Eukaryotes
Reads mapping to NCBI nt DB corresponding to primate sequences:

```
SRR1106548.NT.snap.matched.fl.Primates.annotated
```

Reads mapping to NCBI nt DB corresponding to non-primate mammal sequences (e.g. avian, rodent):

```
SRR1106548.NT.snap.matched.fl.nonPrimMammal.annotated
```

Reads mapping to NCBI nt DB corresponding to non-mammalian chordate sequences (e.g. reptiles, fish):

```
SRR1106548.NT.snap.matched.fl.nonMammalChordat.annotated
```

Reads mapping to NCBI nt DB corresponding to non-chordate eukaryotes (e.g. all other eukaryotes, protozoa, nematodes, coral):

```
SRR1106548.NT.snap.matched.fl.nonChordatEuk.annotated
```

Bacteria
Reads mapping to NCBI nt DB corresponding to viral sequences:

```
SRR1106548.NT.snap.matched.fl.Bacteria.annotated
```

Viruses
Reads mapping to NCBI nt DB corresponding to viral sequences:

```
SRR1106548.NT.snap.matched.fl.Viruses.annotated
```

Count tables are parsed from the above file:

```
SRR1106548.NT.snap.matched.fl.Viruses.annotated.family.counttable
SRR1106548.NT.snap.matched.fl.Viruses.annotated.genus.counttable
SRR1106548.NT.snap.matched.fl.Viruses.annotated.gi.counttable
SRR1106548.NT.snap.matched.fl.Viruses.annotated.species.counttable
```

If desired, you can create additional counttables from any .annotated file by using the table_generator.sh program (installed as part of SURPI).

RAPSearch alignment (translated nucleotide) to protein DB*

**There are 2 usages of NR within SURPI.*

1. maps unmatched reads directly to all of nr (slower method)

2. first maps unmatched reads to a viral protein database, then maps the hits to all of nr for cleanup (faster method ideal for viral discovery). The output for this method is described below.

Reads mapping to viral proteins

Reads that are not aligned to NCBI nt DB are mapped by translated nucleotide alignment against a viral protein database at low-stringency using RAPSearch. This process identifies divergent viral reads. However, due to the low-stringency parameters required to detect divergent viruses (default e-value = 10) and the reduced database utilized, this output may contain many non-specific hits.

```
SRR1106548.Viral.RAPsearch.e1.annotated
```

Count table parsed from the above file:

```
SRR1106548.Viral.RAPsearch.e1.annotated.species.counttable
```

Clean up by alignment to NR (NCBI non-redundant protein collection)

To aid interpretation by removing potential non-specific hits, `SRR1106548.Viral.RAPsearch.e1.annotated` is cleaned up by realigning to the entirety of the NCBI non-redundant protein database. Reads still mapping to viral protein sequences are here:

```
SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated
```

Count tables parsed from the above file:

```
SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated.family.counttable
SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated.genus.counttable
SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated.gi.counttable
SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated.species.counttable
```

Reads removed from `SRR1106548.Viral.RAPsearch.e1.annotated` in the cleanup process are here:

```
SRR1106548.Viral.RAPSearch.e1.annotated.not.in.NR.annotated
```

Count table parsed from the above file:

```
SRR1106548.Viral.RAPSearch.e1.annotated.not.in.NR.annotated.species.counttable
```

Contigs mapped to NR

Contigs are generated by *de novo* assembly as describe in Supplemental Methods, and are found here.

```
deNovoASSEMBLY_SRR1106548/all.SRR1106548.NT.snap.unmatched_addVir_uniq.fasta.unitigs.cut151.264-mini.fa
```

De novo assembled contigs are mapped by translated nucleotide alignment to NR proteins using RAPSearch:

```
SRR1106548.Contigs.NR.RAPSearch.e0.annotated
```

Count tables parsed from the above file:

```
SRR1106548.Contigs.NR.RAPSearch.e0.annotated.family.counttable
SRR1106548.Contigs.NR.RAPSearch.e0.annotated.genus.counttable
SRR1106548.Contigs.NR.RAPSearch.e0.annotated.gi.counttable
SRR1106548.Contigs.NR.RAPSearch.e0.annotated.species.counttable
```

Note that contigs identified as viral by translated nucleotide alignment are included in

```
SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated
```

Coverage plots

For each barcode, the best coverage map for each viral genus identified in the dataset is generated. Reads contributing to the coverage map are derived from genus-level assignments from `SRR1106548.NT.snap.matched.fl.Viruses.annotated` and `SRR1106548.Contigs.and.NTunmatched.Viral.RAPsearch.e1.NR.e0.Viruses.annotated`

```
bar.CGATGT.SRR1106548.NT.snap.matched.fl.Viruses.annotated.genus.top.pdf
bar.GCCAAT.SRR1106548.NT.snap.matched.fl.Viruses.annotated.genus.top.pdf
bar.TGACCA.SRR1106548.NT.snap.matched.fl.Viruses.annotated.genus.top.pdf
```

Log files

Configuration file containing parameters used to run the pipeline

```
SRR1106548.config
```

Run/error log for the SURPI pipeline

```
SURPI.SRR1106548.log
SURPI.SRR1106548.err
```

Quality of the input dataset generated using fastQValidator

`quality.SRR1106548.log`

Number of reads tallied for entire dataset, each barcode separately, and by PE direction (1/2)

`readcounts.SRR1106548.log`

counted in the following order

input reads
preprocessed reads
human depleted reads
reads aligning to NCBI nt DB
viral portion of reads mapping to NCBI nt DB
bacterial portion of reads mapping to NCBI nt DB
non-Chordate Eukaryotic portion of reads mapping to NCBI nt DB
reads not aligning to NCBI nt DB
reads mapping to viral proteins (cleaned up)