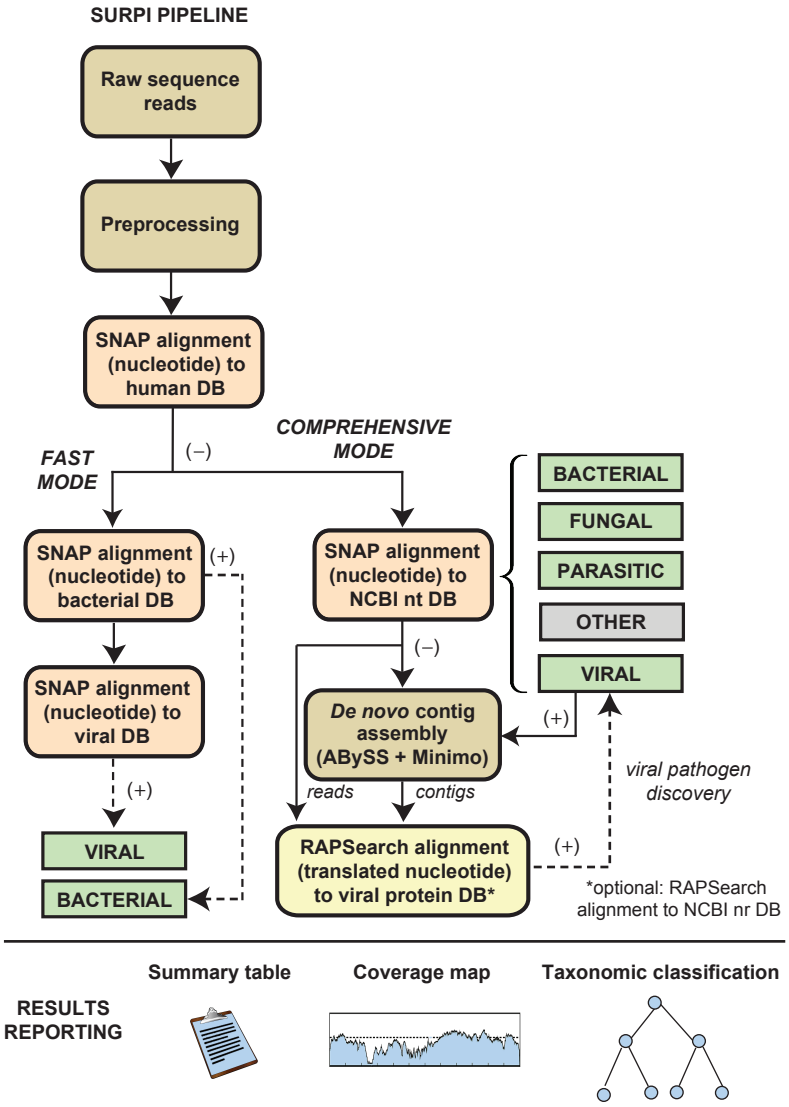
SURPI

Output Interpretation

June, 2014



A schematic overview of the SURPI pipeline. 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.

Using SRR1106548.fastq as a test SURPI input. SURPI and its dependencies, including the relevant databases, must be already installed.

```
$ SURPI.sh -z SRR1106548.fastq
$ ./go_SRR1106548 &
```

SURPI in comprehensive mode generates the following folders:

```
DATASETS_SRR1106548
deNovoASSEMBLY_SRR1106548
LOG_SRR1106548
OUTPUT_SRR1106548
TRASH_SRR1106548
```

Mapping to NCBI Genbank NT

All reads mapping to NCBI Genbank NT (NCBI non-redundant nucleotide collection)

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

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

Files ending in ".annotated" are in SAM format, with taxonomic information added to the last 4 columns.

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

Eukaryotes

Reads mapping to NCBI Genbank NT corresponding to primate sequences:

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

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

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

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

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

Reads mapping to NCBI Genbank NT 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 Genbank NT corresponding to viral sequences:

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

Viruses

Reads mapping to NCBI Genbank NT corresponding to viral sequences:

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

Count tables are parsed from above annotated 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`

Datasets of unmatched reads and *de novo* assembled reads

Reads are preprocessed and computationally subtracted against the human genome, then computationally subtracted against all of GenBank to NT:

`DATASETS_SRR1106548/SRR1106548.NT.snap.unmatched.fulllength.fastq`

Contigs generated by *de novo* assembly of reads in `SRR1106548.NT.snap.matched.fl.Viruses.annotated` and `SRR1106548.NT.snap.unmatched.fulllength.fastq`:

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

Mapping to proteins at lower stringency parameters

Files ending in ".annotated" are the tabular output of RAPSearch and follow –m 8 BLAST format. Taxonomic information has been added to the last 4 columns

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

Reads mapping to viral proteins

Reads are preprocessed and computationally subtracted against the human genome, then computationally subtracted against all of Genbank to NT followed by translated nucleotide alignment against a viral protein database at low-stringency parameters using RAPSearch to identify divergent viral reads:

`SRR1106548.Viral.RAPsearch.e1.annotated`

Count tables parsed from above ".annotated" file

`SRR1106548.Viral.RAPsearch.e1.annotated.family.counttable`
`SRR1106548.Viral.RAPsearch.e1.annotated.genus.counttable`
`SRR1106548.Viral.RAPsearch.e1.annotated.gi.counttable`
`SRR1106548.Viral.RAPsearch.e1.annotated.species.counttable`

Reads mapped to viral proteins, cleaned up by subsequent alignment to all of NR (NCBI non-redundant protein collection)

Reads `SRR1106548.Viral.RAPsearch.e1.annotated` were re-aligned by translated nucleotide alignment to NR proteins using RAPSearch:

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

Count tables parsed from above ".annotated" 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
```

Contigs mapped to viral proteins

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

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

Counttable by family parsed from above .annotated 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
```

Coverage plots

For each barcode, the best coverage map for each viral genus identified in the dataset is shown. Reads contributing to the coverage map are derived from genus-level (or lower level) assignments following SNAP alignment to all of NCBI Genbank NT and RAPSearch translated nucleotide alignment to viral proteins:

```
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 log for the SURPI pipeline

```
SURPI.SRR1106548.log
```

Quality of the input dataset generated using fastQValidator

```
quality.SRR1106548.log
```

Number of reads tallied for all barcodes together and each barcode separately, for the following pipeline steps:input reads, preprocessed reads, human depleted reads, reads aligning to Genbank NT, viral portion of reads mapping to Genbank NT, reads mapping to viral proteins:

```
readcounts.SRR1106548.log
```