

Instructions for SURPI Test Run

This document describes how to download a test dataset from the NIH Sequence Read Archive (SRA), analyze the dataset using SURPI, and examine the output files from the pipeline. The test dataset presented here (SRR1106548) is derived from plasma samples spiked with known titers of HIV (10^4 , 10^3 , 10^2 copies/ml) as described in Naccache, et al. under Supplemental Methods. All of the datasets including this one have been deposited under BioProject PRJNA234451 and are currently accessible under the following confidential reviewer / collaborator link:

<ftp://ftp.ncbi.nih.gov/pub/TraceDB/misc/tmp/review/0935b17846ec72839401d500bd0defe6cbc93ef6a1f09b6d0c9ca15d283cbdb9>

The full SRA manifest is as follows. Note that human reads have been removed from all datasets of human origin to maintain privacy of genomic data. The test dataset SRR1106548.sra is highlighted in gray.

Data File	Experiment Title	Base Count
SRR1106116.sra	analysis of serum from individual with hemorrhagic fever for SURPI performance validation	7,368,170,700
SRR1106117.sra	analysis of pediatric diarrheal stool for SURPI performance validation	10,793,664,500
SRR1106119.sra	analysis of serum from individual with hemorrhagic fever for SURPI performance validation	329,022,415
SRR1106121.sra	analysis of sera from individuals with acute hemorrhagic fever from Democratic Republic of the Congo, Africa	49,232,500
SRR1106123.sra	analysis of sera from patients with hepatitis from Transfusion-Viruses Study (TTVS) Transmitted	2,033,681,000
SRR1106126.sra	analysis of sera from individuals with acute febrile illness from Tanzania, Africa	153,793,500
SRR1106129.sra	analysis of cerebrospinal fluid from an individual with Varicella-Zoster Virus (VZV) encephalitis	4,875,721,184
SRR1106547.sra	analysis of lung swabs and tissues from monkeys with fulminant pneumonia infected by titi monkey adenovirus (TMAv)	830,364,902
SRR1106548.sra	analysis of control plasma samples spiked with known titers of HIV (10^4 , 10^3 , 10^2 copies/ml)	600,943,609
SRR1106549.sra	analysis of serum from a returning traveler with a febrile illness	14,775,845
SRR1106550.sra	Comparative ROC curve analysis using reads derived from a study of pediatric diarrhea (stool) in Mexico	500,283,800
SRR1106552.sra	Comparative ROC curve analysis using reads derived from a serum sample from a patient with hantavirus pulmonary syndrome	488,156,597
SRR1106553.sra	Comparative ROC curve analysis using reads derived from nasal swabs from patients infected with influenza A(H1N1)pdm09	38,127,857

The test dataset SRR1106548.sra (above table in grey) includes paired-end data as well as orphan single-end read 1 (R1) and read 2 (R2) data from 3 separately barcoded samples corresponding to spiked HIV titers of 10^4 , 10^3 , 10^2 copies/ml. To restore the original FASTQ files (with the human reads removed), execute the following steps:

- 1) Install the SRA Toolkit (required for fastq-extractBarcodedSRA.sh). Installation instructions can be found at the following link:

http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std

- 2) Extract barcoded reads from SRR1106548.sra.

```
$ fastq-extractBarcodedSRA.sh SRR1106548.sra
```

- 3) Combine extracted FASTQ files into one input file SRR1106548.fastq.

```
$ cat bc*.fastq > SRR1106548.fastq
```

- 4) Run SURPI using SRR1106548.fastq as the input file. SURPI and its dependencies, including the relevant databases, must be already installed.

```
$ SURPI_v22_14.sh -z SRR1106548.fastq  
$ ./go_SRR1106548 &
```

- 5) The following directories will be generated at the end of the pipeline run:

- COVERAGEMAPS_SRR1106548
- DATASETS_SRR1106548
- deNovoASSEMBLY_SRR1106548
- OUTPUT_SRR1106548
- TRASH_SRR1106548

The contents of OUTPUT_SRR1106548 are provided and are described below.

Example SURPI Output (Contents of OUTPUT_SRR1106548)

This is an example of the SURPI output after analyzing example dataset SRR1106548 in comprehensive mode.

Results of Alignments against NCBI Genbank NT (using SNAP)

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

Results of alignment of example dataset preprocessed 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 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.

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

These files have been provided in the OUTPUT_SRR1106548 directory for perusal, but would normally be found in DATASETS_SRR1106548 and deNovoASSEMBLY_SRR1106548.

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

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

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

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

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