SURPI

Generating SURPI input from SRA source
June, 2014

Converting a dataset from the NIH Sequence Read Archive (SRA) to a fastq file for SURPI.

The test dataset used here (SRR1106548) is derived from plasma samples spiked with known titers of HIV (10^4, 10^3, 10^2 copies/ml) as described in Supplemental Methods Naccache et al 2014 and is found here: http://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR1106548

The test dataset SRR1106548.sra 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
```