SURPI

Input file requirements

June, 2014

SURPI input:

SURPI accepts FASTA and FASTQ files as inputs.

As SURPI scales well and recognizes multiplexed data, it is recommended that you concatenate the entirety of a sequencing run (all indexes + Read 1 + Read 2) into one input fastq file for input.

SURPI uses index/barcode information to *de novo* assemble reads separately by index, and reports readcounts for each taxonomic unit separately for each index.

SURPI recognizes multiplexed reads by recognizing the following header format:

```
@TEST:82:H81VWADXX:1:1101:4073:2064#ATGCAG/2
@TEST:82:H81VWADXX:1:1101:4073:2064#68/1
```

(where ATGCAG or 68 represent the index/barcode, and is delimited at the end of the header by #, with /1 and /2 representing the read direction)

Or:

```
@TEST:82:H81VWADXX:1:1101:17666:2067 2:N:0:ATGCAG
```

Which SURPI will convert to:

```
@TEST:82:H81VWADXX:1:1101:17666:2067#ATGCAG/2
```

For SURPI to provide proper readcount statistics, all read headers in a single SURPI input dataset should share a common 3 letter string (eg: M00, HWI, HIS, SCS, SRR for example). SURPI currently selects the string from the first and last reads only.

SURPI currently does not have paired-end functionality – it currently treats read 1 and read 2 as independent reads.

Run SURPI using (for example) SRR1106548.fastq as the input file as follows (SURPI and its dependencies, including the relevant databases, must be already installed.)

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