SURPI

Input file requirements

June, 2014

SURPI accepts FASTA and FASTQ files as inputs, and uses multiplexing information when available.

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

We routinely concatenate all Illumina FASTQ files from one (HiSeq lane / MiSeq run) into a single FASTQ file as input to SURPI.

The barcode recognized is between the hash and the forward slash (if present). Below are some acceptable FASTQ headers:

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

The above FASTQ headers will all be recognized into the 3 separate barcodes **ATGCAG, chiulab,** and **68.**

SURPI will also accept a FASTQ file in the following format (current MiSeq or HiSeq standard format):

```
@TEST:82:H81VWADXX:1:1101:4073:2064 1:N:0:68
@TEST:82:H81VWADXX:1:1101:4073:2064 2:N:0:ATGCAG
@TEST:82:H81VWADXX:1:1101:4073:2064 1:N:0:chiulab
```

Which SURPI will internally convert to:

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

For SURPI to provide proper read count statistics, all read headers in a single SURPI input dataset should share a common 3 letter string (e.g. TES, M00, HWI, HIS, SCS, SRR, etc.). 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.