Tag-seq scripts: a manual

November 14, 2014

These are the printouts from scripts when run without arguments, in order of their appearance in the walkthrough.

---

ngs_concat.pl :

```
concatenates files by matching pattern in their names

arg1: common pattern for files
arg2: perl-like pattern in the filename to recognize,
        use brackets to specify the unique part, as in
      "FilenameTextImmediatelyBeforeSampleID(.+)FilenameTextImmediatelyAfterSampleID"

Example (to concatenate files names like Sample_Pop1_L1.fastq, Sample_Pop1_L2.fastq):
ngs_concat.pl 'Sample' 'Sample_(.+)_L'
```

---

iRNAseq_trim_launch.pl :

```
Prints out a series of commands to trim Illumina RNA-seq reads, one for each reads file
(rnaseq_clipper.pl | fasx_clipper | fastx_quality_filter )

prints to STDOUT

Arguments:
1: glob to fastq files
2: optional, the position of name-deriving string in the file name
      if separated by underscores,
      such as: input file Sample_RNA_2DVH_L002_R1.cat.fastq
      specifying arg2 as '3' would create output file with a name '2DVH.fastq'

Example:
iRNAseq_trim_launch.pl '\.fq$' > clean

NOTE: use this script if your qualities in the fastq files are 33-based;
if not, use iRNAseq_trim_launch_bgi.pl instead
```

---

rnaseq_clipper.pl :

```
Clips 5'-leader off Illumina fastq reads in RNA-seq

Removes duplicated reads sharing the same degenerate header and
the first 20 bases of the sequence (reads containing N bases in this
region are discarded, too)

prints to STDOUT

arguments:
1 : fastq file name
```

2 : string to define the leading sequence, default '[ATGC]?[ATGC][AC][AT]GGG+'
'keep' : optional flag to say whether the sequences without leader should be kept.
                    By default, they are discarded.

Example: rnaseq_clipper.pl D6.fq

---

iRNAseq_bowtie2map.pl :

Prints out a list of bowtie2 calls to map Illumina RNA-seq reads, one for each reads file.
Make sure to run bowtie2-build on your transcriptome before running this.

prints to STDOUT

Arguments:
1: glob to fastq files
2: database to map to, such as '~/db/transcriptome.fasta'
3: optional, the position of name-deriving string in the file name
            if it is by underscores or dots

Example:
iRNAseq_bowtie2map.pl \"trim$\" ~/db/transcriptome.fasta > maps

NOTE: if you plan to use gmapper (SHRiMP) for mampping, use iRNAseq_shrimpmap_SAM.pl instead.

---

isogroup_namer.pl:

assigns sequences to \"isogroups\" based on cd-hit-est results,
prints a tab-delimited table of sequences - cluster designations

prints to STDOUT

usage:
isogroup_namer.pl [fasta file] [cd-hit-est result, .clstr file]

example:
isogroup_namer.pl transcritptome.fasta transcriptome.fasta.clstr >transcriptome_seq2iso.tab

NOTE:
run cd-hit-est before this; for example, to look for 99% or better
matches between contigs taking 30% of their lengths
cd-hit-est -i transcriptome.fasta -o transcriptome.fasta -c 0.99 -G 0 -aL 0.3 -aS 0.3

---

samcount_launch_bt2.pl :

Prints out list of commands to derive counts from SAM files
generated by bowtie2

prints to STOUT

Arguments:
1: glob to sam files
2: path-filename of clusters2isogroups table

Example:

```
samcount_launch_bt2.pl '\.sam' /path/to/reference/transcriptome_seq2iso.tab > sc
```

NOTE: for SAM files made by gmapper (SHRiMP), use samcount_launch.pl

---

samcount.pl , v.0.2 (November 2014):

counts reads mapping to isogrops in SAM files

Arguments:

arg1: SAM file (by cluster, contig, or isotig)
arg2: a table in the form 'reference_seq<tab>gene_ID', giving the correspondence of reference sequences to genes. With 454-deived transcriptome, the gene_ID would be isogroup; with Trinity-derived transcriptiome, it would be component; you can also use cd-hit-est and isogroup_namer.pl to create

dup.reads=keep|toss : whether to remove exact sequence-duplicate reads mapping to the same position in the reference. Default keep (duplicates are supposed to be tossed at the trimming stage).

aligner=gmapper|bowtie2 : aligner that made the SAM file. Default bowtie2.
bowtie2 is assumed to be used in -k mode.

mult.iso=random|toss : (for aligner=gmapper) if a read maps to multiple isogroups, it is disregarded by default. Set this option to 'random' if you want to randomly pick an isogroup to assign a count to.

Example:

```
samcount.pl A5.fq.trim.sam /path/to/reference/transcriptome_seq2iso.tab \
>A5.fq.trim.sam.counts
```

expression_compiler.pl :

assembles RNAseq counts data into a single table

Arguments:
arg1: [pattern to counts files]

Example:

```
expression_compiler.pl *.counts > allcounts.txt
```