

Overview of processing scripts

SuRE data is processed using 4 bash scripts. The scripts implement the entire data processing pipeline for both iPCR and plDNA/cDNA data from plasmid library samples and RNA expression samples. The 4 scripts roughly perform the following functions:

- **cDNA-plDNA-count-BC.bash:** parse cDNA fastq files into count tables
- **iPCR-map-BC.bash:** parse iPCR fastq to bed-like files per sample
- **iPCR-merge-bedpe-Filter-BC-multi-pos.bash:** merge multiple iPCR bed-like files into 1
- **merge-iPCR-cDNA-plDNA.bash:** merge cDNA count data and iPCR bed-like file

Required software

The scripts use various software tools, some standard on most linux environments but most should be installed prior to using these scripts. A complete list is shown in the following table:

name	version
gawk	4.0.1
cutadapt	1.9.1
gzip	-
parallel (GNU)	-
bowtie2	2.1.0
samtools	1.2
python	2.7.6

cDNA-plDNA-count-BC.bash

DESCRIPTION:

This is a bash script (mostly gawk and cutadapt) to process raw fastq files containing data from cDNA/plDNA samples. The barcodes are extracted from the reads and counted. For extracting the barcodes the adapter sequence is aligned with the read (using cutadapt) and the preceding part of the read is defined as the barcode.

Barcodes with length != 20, or which contain N's are discarded. The filtered barcodes and counts are written to stdout, sorted (alphabetically) on barcode.

USAGE/OPTIONS:

cDNA-plDNA-count-BC.bash [options] fastq-filenames

required:

-o: output directory

-a: adapter sequence

optional:

-l: log-filename [stdout]

-n: number of cores used in parallel processes [10]

INPUT:

cDNA/plDNA fastq files

OUTPUT:

tabular txt files (one for each input file) with count and barcode-sequence

iPCR-map-BC.bash

DESCRIPTION:

A bash script (awk, bowtie2, samtools and cutadapt) to process raw fastq files containing data from iPCR samples. The barcodes and gDNA are extracted from the reads and the gDNA is aligned to the reference genome. Barcodes with length != 20, or which contain N's are discarded, as are reads with a MAPQ less than 20. The aligned and filtered paired reads are written to stdout in bedpe-like format including the barcode and sorted (alphabetically) on barcode.

USAGE/OPTIONS:

iPCR-map-BC.bash [options] fastq-filenames

required:

-o: output directory

optional:

-f: forward adapter sequence [CCTAGCTAACTATAACGGTCCTAAGGTAGCGAACCAGTGAT]

-r: reverse adapter sequence [CCAGTCGT]

-d: digestion site [CATG]

-l: log-filename [stdout]

-n: number of cores used in parallel processes [10]

INPUT:

iPCR fastq files

OUTPUT:

tabular txt files in bedpe-like format (one per input fastq)

iPCR-merge-bedpe-Filter-BC-multi-pos.bash

DESCRIPTION:

Bash script (mostly awk) to merge a (set of) raw iPCR bedpe file and remove BC-position pairs if the BC is associated with multiple positions. The BC-position pair with highest iPCR count is retained and remaining pairs are removed.

USAGE/OPTIONS:

iPCR-merge-bedpe-Filter-BC-multi-pos.bash [options] bedpe-like-filenames

required:

```

    -o name of output file
optional:
    -l a logfilename [stdout]
    -n number of cores to use [10]
INPUT:
    raw iPCR bedpe file(s)
OUTPUT:
    tabular txt file with position, strand, alignment info, gDNA sequences, barcode.
    This output file is filtered for barcodes which occur at multiple
    positions. But positions which associate with multiple barcodes are still
    included.

```

merge-iPCR-cDNA-plDNA.bash

DESCRIPTION:

Bash script (mostly awk) to merge data of iPCR, cDNA, and plDNA data from a SuRE experiment into a single tabular text file. All input files are sorted on barcode, the iPCR input is still redundant, i.e. multiple barcodes may be associated with a single position (SuRE-fragment). But the iPCR data is cleaned from barcodes associated with multiple positions. The output is a single tabular txt file with only positional information from the iPCR bedpe file and columns for each cDNA/plDNA file with corresponding counts. The output file is ordered along genomic position.

USAGE/OPTIONS:

```

merge-iPCR-cDNA-plDNA.bash [options]
required:
    -s: sample meta file which has columns with fastq file names for all
        samples (iPCR, cDNA, and plDNA) and a column with (short) sample names.
    -i: name of iPCR bedpe file
    -o name of output file
optional:
    -c: directory containing cDNA count-table file(s) [cDNA/count-tables]
    -p: directory containing plDNA count-table file(s) [plDNA/count-tables]
    -l: log-filename [stdout]
    -n: number of cores used in parallel processes [10]

```

INPUT:

```

iPCR bedpe file; redundant, sorted on barcode
cDNA/plDNA count-tables, sorted on barcode

```

OUTPUT:

```

tabular txt file with position, strand, sample-counts, ordered on position

```

Notes

Filename pattern

Some scripts take fastq filenames as input. Different samples are named using these fastq filenames. The filenames are assumed to contain certain string patterns. If your filenames are not compatible you can either adjust the scripts or your filenames. This dependency on string patterns is in the scripts:

- **cDNA-plDNA-count-BC.bash:** lines 144-149
- **iPCR-map-BC.bash:** lines 179-187

Available on Github

The scripts are available on github: `git@github.com:lpagie/SuRE-dataProcessing-pipeline.git`, in branch *public*.