PAR-CLIP analysis suite.

Pavel Morozov

RNA Biology Laboratory

HHMI/ Rockefeler University

pmorozov@rockefeller.edu

PARCLIP_suite is a collection of PERL script for analysis of PAR-CLIP experiment output. Scripts processes FASTQ input, map sequence reads to the reference database and assign reads to various reference sequence categories according to the hierarchy. Reference sequence categories are mRNA, tRNA, miRNA, rRNA, genome, mitochondrial genome and other. Both sequence categories and hierarchy could be adjusted by user to suite needs of particular experiment. The rate of T>C transition is an important criterion in PAR-CLIP analysis and is calculated for each reference sequence category. Results are presented in tab-delimited format. Additional scripts allow extracting subset of sequences by category or/and number and type of mismatches to the reference.

PREREQUISITES.

Suite uses CUTADAPT software to remove adapters from reads and BOWTIE1 software to map reads against reference database.

CUTADAPT: http://cutadapt.readthedocs.io/en/stable/index.html

BOWTIE1: http://bowtie-bio.sourceforge.net/index.shtml

QUICK START.

If you wish quickly run the suite and get results without learning about file formats and analysis options and already have formatted databases and proper configuration file you can just run the next command lines:

```
./prepare_fasta.pl input_fastq[.gz] output_prefix output_directory
```

```
./analyze_parclip.pl input_fasta prefix minimal_length maximal_length configuration_file output_directory
```

After completing calculations load into EXCEL or other spreadshit software the file which you specify as ‘*output_prefix*’ from ‘*output_directory*’.

FILES AND FORMATS.

Input.

Input file should be fastq or gzipped fastq.

Collapsed fasta.

During calculation we use collapsed fasta format which is regular multiple fasta format with read count added to the sequence names and separated from the name by ‘|” symbol, i.e.:

```
>seq1|10

ACGACTATCATCAGCGGGACGAGCGA

>seq2|3

GGCTCTAGCATCACGGATATTACG

…
```

Advantages of this format are elimination of redundancy at the level of reads, faster calculations, lover memory usage and readability, especially if sequences are sorted by count. Disadvantages are additional processing when counting statistics and expenses for converting regular fasta files into collapsed format.

Parameters file

File named ‘*parameters*’ contains parameters necessary for running suite like directory containing scripts and parameters for programs. This file should be edited according with each new system suite is installed on.

Configuration file.

This file contains information about databases, reference sequence categories names and hierarchy of annotation.

Example of simple configuration file:

```
--------------------------------------------------
annotation_db /data5/PAR-CLIP_annotation/annodb_v2/v2_1
annotation_db /data5/PAR-CLIP_annotation/annodb_v2/v2_2
annotation_db /data5/PAR-CLIP_annotation/annodb_v2/v2_miRNA
annotation_db /data5/PAR-CLIP_annotation/annodb_v2/v2_2_34
annotation_db /data5/PAR-CLIP_annotation/hg19/hg19
hier 1 3 2 11
1       b       adapter
2       b       mRNA
3       b       miRNA
11      b       genome
---------------------------------------------------
```

Strings started with *'annotation_db'* token specify database for bowtie formatted database.

Program will map reads against all specified databases and then assign categories to the reads using hierarchy. Each reference sequence name should start with category number followed by the '|' symbol or sequence category would not be recognized.

Hierarchy is stored in the string started with the *'hier'* token. Categories are listed right below the hierarchy string. The order of numbers in the hierarchy string is the order in which read will be assigned a category. Say read having a perfect hit to categories 2, 3 and 11 will be assigned to category 3 (miRNA) by hierarchy in the example above because 3 is the leftmost in the hierarchy string. If read maps with mismatches only hits with lowest number of mismatches are considered.

Category lines have three fields: category number, strand preference, category name. Strand preference tells the program what mappings to take into account with respect to the orientation relative to the reference sequence. Strand preference 'p' stands for forward orientation of mapping, i.e. in the same direction as reference sequence; 'm' stands for reverse orientation of the mapping (reverse compliment of reference); 'b' stands for no orientation preference, i.e. both are forward and reverse orientations of the mappings are allowed. By default strand preference is set to 'b'.

Configuration file can be edited to accommodate various databases and alternative hierarchies.

Intermediate files: complete annotation ‘anno’ and ‘ctab’ files.

The output of intermediate annotation step is file *‘prefix.anno’* or multiple files *‘prefix_NN.anno’* if split version of suite is used.

This file contains all possible category matches information for each read and might be needed for extracting alternative annotations using different hierarchy.

File with extension ‘ctab’ contains the category assignment information for particular hierarchy and are useful for sequence retrieval and additional statistic analysis.

Final annotation tables.

Final annotation tables are named just as ‘output_prefix’ and contains several tables with read annotation summary. Best way to use them is to load this file into EXCEL and plot all necessary graphs.

INITIAL DATA PREPARATION.

This step converts input file from fastq to fasta, trims adapters as needed and convert fasta file into collapsed fasta. Script should be run from the PARCLIP_suite directory, if input file is located in other directory proper path should be included in file name.

Script prepare_fasta.pl

Usage:

./prepare_fasta.pl input_fastq[.gz] output_prefix output_directory

Where *input_fastq* is sequencing output in fastq format and might be gzipped. If file was archived by other than *gzip* utility it have to be unpacked manually first.

Parameter *output_prefix* is a base name for all output files, i.e. if prefix is “*my_output*” collapsed fasta file will be named “*my_output.fau*” and will be placed in in *‘output_directory’*.

Parameter ‘*output_directory*’ specifies where the results and intermediate files will be stored.

To complete analysis one have to run *analyse_PAR-CLIP.pl* script.

RUNNING ANNOTATION.

Master file for running annotation is *analyse_PARCLIP.pl* with collapsed fasta file obtained from data preparation step and configuration file. Script should be run from the PARCLIP_suite directory, if input file is located in other directory proper path should be included in file name.

Command line is:

```
./analyze_parclip.pl input_fasta prefix minimal_length maximal_length configuration_file output_directory
```

Full path have to be given for collapsed fasta and configuration file.

Annotation will produce *'output_prefix'_summary.txt* file with count of reads by annotation categories and mismatch information.

Another useful file is '*output_prefix.anno*', which contains information about read mappings to various categories before applying hierarchy.

This file can be reused with different hierarchies without running bowtie by *an_anno.pl* script with different configuration files.

RETRIEVING READS BY CATEGORY OR/AND MISMATCHES.

Script *get_sequences.pl*

Usage:

```
./get_sequences -fasta FASTA_FILE -ctab CTAB_FILE -category CATEGORY_NAME -config CONFIG_FILE -minlen NNN -maxlen NNN -mism FIRST_MUT -mism SECOND_MUT
```

Parameters:

-fasta original collapsed fasta file with all sequences;

-ctab ctab file from PAR-CLIP annotation pipeline for this sample;

-category category to select, might be number of category or name

refer to the configuration file used for this run

-config configuration file with path

Optional parameters:

-minlen minimal length of the read, natural number

-maxlen maximal length of the read, natural number

-mism substitutions, up to two can be specifyed third one and after will be ignored; perfect matches only if no mismatches specifyed; zero substitution if NONE, anything if ANY; mismatches can be written as AT or A->T or A>T.

CUSTOM CONFIGURATION AND FORMATTING CUSTOM DATABASES.

Databases are formatted by 'bowtie-build' command from fasta files. Each sequence in fasta file should contain the number of category, i.e.

```
>33|tRNA precursor 1
```

ACTACATCAGCGACGAGCGACATCTATCATCAGCAGCAGCGACGAGGCGCGAGCGAGC

>33|tRNA precursor 2

ATACACGAGCAGCATCATCATCATCATCAGCGACGAGCAGCAGCGGACGAGCGCGACG

etc.

Multiple types can be put together in one fasta file, but it is adviceable to keep similar sequences from different categories separate. I.e. mRNA and mRNA genes better be in separate files, because software makes a limited number of mappings to each database and if say we have bone fide mRNA reads and got 8 hits to mRNA genes we might not see hits to mRNA and thus have a wrong assignment.

Also one have to take care not to use same number for two different categories, which can easily happens when formatting new database from multiple previous attempts.

DEFAULT REFERENCE TRANSCRIPTOME AND SAMPLE DATA.

Default human reference transcriptome is provided as tar/gzip archive *Human_Reference_Transcriptome.tar.gz* and should be placed in Human_Genome_Reference folder under the PARCLIP_suite folder. Configuration file for this reference with default hierarchy is included in this archive.

Sample fastq file is in test.tar.gz archive.