# iCLIPro Documentation

***Release 0.1.1***

**Tomaz Curk**

December 18, 2014

# CONTENTS

# CHAPTER ONE

# OVERVIEW AND HOW TO CITE

iCLIPro is a Python package that can be used to control for systematic misassignments in iCLIP data.

If you use iCLIPro in your research, please cite this paper (submitted for review):

> Christian Hauer, Tomaz Curk, Simon Anders, Thomas Schwarzl, Anne-Marie Alleaume, Jana Sieber, Ina Hollerer, Madhuri Bhuvanagiri, Jernej Ule, Wolfgang Huber, Matthias W. Hentze and Andreas E. Kulozik
>
> *Improved binding site assignment by high-resolution mapping of RNA-protein interactions using iCLIP*

CHAPTER

TWO

# ICLIPRO USAGE

Usage: iCLIPro [options] in.bam

**Options:**

**-o FOLDER** output folder (default is cwd - current working directory)

**-b INT** genomic bin size (100..1000, default: 300)

**-r INT** number of reads required in bin (20..500, default: 50)

**-s INT** flanking distances when calculating start site overlap ratio (3..15, default: 5)

**-q INT** use only reads with minimum mapping quality (mapq) (0..100, default: 10)

**-g LIST** read len groups (e.g.: “A:16-39,A1:16-25,A2:26-32,A3:33-39,L:20,B:42”)

**-p LIST** generate read overlap maps based on these

comparisons (e.g.: “A1-A3,A2-A3,A1-B,A2-B,A3-B,L-B,A-B”)

**-f INT** flanking region for read overlap maps (default: 50)

**-h** longer help

For given input BAM file [in.bam] the script will generate a number of output files that can be used to check for and diagnose systematic misassignments in iCLIP data.

Main result is stored in file [in_report.txt], for given [in.bam] BAM file.

Read (query template) names in BAM files should include a record of form expressed with this regular expression: `:rbc[ATCGN]+:`. The ending colon can be omitted if random barcode record is placed at the end of the name. Some valid examples:

```
D3FCO8P1:206:C2M53ACXX:8:1207:17086:80291:1:N:0:rbcTGTAC: 272     1       11861 ...
D3FCO8P1:206:C2M53ACXX:8:1101:6625:73240:1:N:0:rbcCCGCC   16      1       11976 ...
D3FCO8P1:206:C2M53ACXX:8:1203:17298:81179:rbcCCGCC:1:N:0  16      1       11976 ...
```

Random barcodes can be specified at the end of the name but must be preceded by colon, for example:

```
D3FCO8P1:206:C2M53ACXX:8:1207:17086:80291:1:N:0:TGTAC     272     1       11861 ...
D3FCO8P1:206:C2M53ACXX:8:1101:6625:73240:1:N:0   :CCGCC   16      1       11976 ...
D3FCO8P1:206:C2M53ACXX:8:1203:17298:81179:1:N:0 :CCGCC    16      1       11976 ...
```

If no random barcode information is available, then iCLIPro will most likely be able to work with the original read names. In such case, please check that the read names do not include any text that conforms to the rules for specifying random barcode as it may mislead iCLIPro.

The generated report file includes a list of random barcodes identified by iCLIPro. You should check it first and make sure that proper random barcode information is being used.

**Parameters -b and -r:** Specifies the bin size to use when segmenting the genome. Only bins with enough reads (parameter -r) are then considered in the read overlap testing.

**Parameter -s:** Flanking region when calculating the mean and the median start site overlap ratios.

**Parameter -q:** Consider only reads that pass the minimum mapping quality, ignore the rest.

**Parameter -g:** iCLIPro needs to group reads based on their length. Any number of (overlapping) groups can be specified. For each group, an interval (INT1-INT2) or single value (INT) of the read lengths in the group can be specified.

**Parameter -p (read overlap maps):** Specifies which groups of reads to compare. When performing a comparison (G1-G2), cross-linked sites identified based on group G2 are used as reference (zero position). The relative positioning of sites identified in G1 is the computed and shown in read overlap maps.

**Parameter -f:** Width of the flanking region relative to reference point shown in read overlap maps.

## 2.1 Method

A typical (i)CLIP experiment may result in the detection of RNA fragments of different lengths. Under the assumptions of conventional iCLIP, the start sites of iCLIP fragments should coincide at the cross-linking position in a fragment length-independent fashion.

This interpretation may not hold for some iCLIP libraries (e.g., substantial read-through, binding to long RNA stretches etc). For details, see associated paper by Hauer and coauthors. In summary, we identified a previously unrecognized effect of iCLIP fragment length on the position of fragment start sites and thus assigned binding sites for some RBPs.

iCLIPro is a robust analysis approach that examines this effect and thus can improve the assignment of binding sites from iCLIP data.

iCLIPro's main function is to visualize coinciding and non-coinciding fragment start sites in order to examine the best way how to analyze iCLIP data.

With iCLIPro you can test test and compare the overlap of different reference points in the iCLIP fragments:

- one nucleotide before first mapped nucleotide (conventional assumption)
- center of the read
- end of the read

iCLIPro identifies regions (bins in genome, parameter -b) with a sufficient number of reads (parameter -r) for an read overlap test. Reads from each selected bin are processed separately. Reads get grouped based on their length (parameter -g) and sites from different groups are compared.

The main output of iCLIPro are read overlap heatmaps that identify the best mode of analysis.

## 2.2 Read overlap heatmaps

Read overlap maps are generated by comparing fragment start, center and end sites in the *test* and *reference* groups.

The data underlying the high-resolution overlap heatmaps is used to calculate a ratio of overlapping and non-overlapping start sites thus enabling the decision to be made as to whether the start or the center of the fragments should be used as a reference point for most accurately defining the binding site. This overlap start site ratio is reported at the end of the generated report file. When calculating the start site overlap ratio a default flanking distance of 5 nt is used (parameter -s, see paper).

A ratio well above 1 suggests to use the start sites of iCLIP fragments to detect binding sites (e.g., mean overlap start site ratio of 1.31 for U2AF65). A ratio below 1 favors the use of the center position for binding site assignment (e.g., mean overlap start site ratio of 0.88 for eIF4A3, see paper for details).

Sites identified based on a reference group are used to define the reference (zero) position in the map. The regions (-50 to +50, x-axis on plots, parameter -f) relative to the reference positions are then scanned and number of co-occuring sites in test group is recorded.

The x-axis shows the offset of the sites of the test group (shorter reads) relative to the sites of the reference group (usually longer reads). The y-axis shows the fragment length. The color in the heatmap represents the number of fragments that co-occur at a given offset relative to the longer reference fragments.

In case of the fragment start sites, a peak at the start reference position 0 corresponds to coinciding start sites, whereas a distribution downstream of the reference position 0 arises from start sites of smaller fragments that occur at length-dependent offsets from the reference start sites.

CHAPTER

THREE

# INSTALLATION

iCLIPro will be made available from the Python Package Index (PyPI), iCLIPro package on PyPI. For now, please download this source file.

You need Python version 2.6 or later (Python 3 was not tested yet).

Please, install first matplotlib (plotting) and pysam (reading BAM files).

**iCLIPro is a typical Python package. To install it from source follow these steps:**

- download source
- unpack the tarball (`tar -xvzf iCLIPro-0.1.1.tar.gz`)
- go into the unpacked folder (`cd iCLIPro-0.1.1`)
- type to install for current user:

```
    python setup.py install --user
```

A system-wide installation (requires admin rights) can be performed instead:

```
python setup.py build
sudo python setup.py install
```

**To test the installation, in command-line:**

1. change working folder to other than the build folder (*e.g.*, `cd ~`)
2. type `python` to enter Python
3. then type `import iCLIPro` to see if the package can be imported
4. exit Python and in command-line type `iCLIPro`

If you get an error message when importing iCLIPro in Python (step three above), then please make sure that the environment variable `PYTHONPATH` points to the iCLIPro package.

If you get an error message when trying to run the iCLIPro script (step four above), then please make sure that the environment variable `PATH` points to the script (also found in source `scripts/iCLIPro`).

# CHAPTER FOUR

# EXAMPLES

Scripts used to generate some of the figures in paper are available in folder *examples*:

- 00_get_raw_BAMs.sh: downloads the raw BAM files needed for the analysis, files are stored in folder `raw_data` and `data`.
- 01_run_iCLIPro.sh: generates figures 4b, 4c, 6a, 6b, S3a-e, S4a-h, S6a-h that are used in the paper, figures are stored in folder `results\figures`.

To learn more about each step, please check the individual scripts and modify them according to your needs.

You can just download the complete examples folder http://www.biolab.si/iCLIPro/examples/ and explore the outputs of scripts.

The scripts should render figures like Figure 4b in the paper:

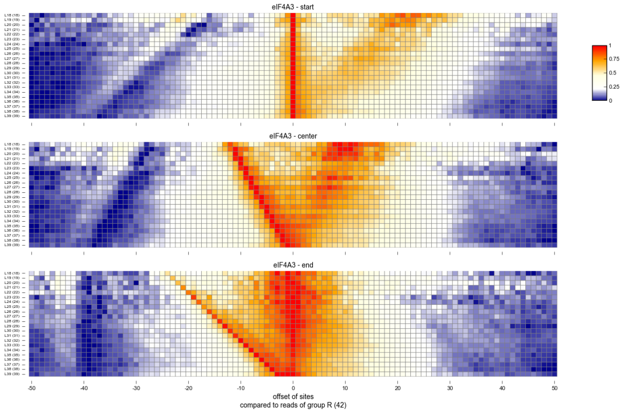

Figure 4.1: Figure 4b in paper.

CHAPTER

FIVE

# VERSION HISTORY

## 5.1 Version 0.1.1

2014-12-16, source v0.1.1

- Among reads mapping to same start site and with same random barcode we take their median center and end positions.
- Added grouping of reads by end position.
- Added heatmap of offset frequencies of each specified read group (read lengths).
- Changed default minimum mapping quality (MAPQ) to 10 (before was 30).
- Added helper script iCLIPro_bam_splitter, which can be used to generate bedGraph files from groups of reads of specified lengths.

## 5.2 Version 0.1.0

2014-07-08, source v0.1.0

First version.

CHAPTER

SIX

# AUTHOR

iCLIPro is developed by Tomaz Curk at University of Ljubljana, Faculty of Computer and Information Science, Bioinformatics Laboratory.

This software is the result of a collaboration with the groups of prof. dr. Andreas E. Kulozik, MD, prof. dr. Matthias W. Hentze, MD, dr. Wolfgang Huber and prof. dr. Jernej Ule.

Special thanks to Christian Hauer who was instrumental during the inception and development of this tool.

You can contact me at tomaz.curk@fri.uni-lj.si.

CHAPTER

SEVEN

# LICENSE

iCLIPro is free software: you can redistribute it and/or modify it under the terms of the GNU General Public License as published by the Free Software Foundation, either version 3 of the License, or (at your option) any later version.

This program is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU General Public License for more details.

The full text of the GNU General Public License, version 3, can be found here: http://www.gnu.org/licenses/gpl-3.0-standalone.html