

Rational Function Complexity Manual

Timothy Daley

Andrew Smith

May 9, 2012

Chapter 1

Quick Start

The Rational Function Complexity package is aimed at predicting the yield and number of distinct reads from a genomic library from an initial sequencing experiment. The estimates can then be used to examine the utility of further sequencing, optimize the sequencing depth, or to screen multiple libraries to avoid low complexity samples.

1.1 Installation

1.1.1 Download

Rational Function Complexity is available at <http://smithlab.cmb.usc.edu/software/RationalFunction>

1.1.2 System Requirements

Rational Function Complexity runs on Unix-type system with GNU Scientific Library (GSL), available at <http://www.gnu.org/software/gsl/> and GNU Compilation Collection (GCC) (if you would like to compile it yourself), available at <http://gcc.gnu.org/>. It has been tested on Linux and Mac OS-X.

1.1.3 Installation

Download the source code and decompress it with

```
$ tar xvfz library_complexity.tar.gz
```

Enter the `lc_extrap` directory and run

```
$ make all
```

If the desired input files are in .bam format, bamtools is required, available at <https://github.com/pezmaster31/bamtools>. To compile, the bamtools directory must be specified, i.e.

```
$ make all BAMTOOLS_ROOT=~ /bamtools/
```

If compiled successfully, the executable files are available in `lc_extrap/`.

1.2 Using Rational Function Complexity

1.2.1 Basic usage:

To generate the complexity plot of a genomic library from a sorted read file in .bed or .bam format, use the program *c_curve*. Use -o to specify the output name.

```
$ complexity_plot -o output.txt input.bed
```

To estimate the future yield and bounds on the number of distinct reads of a genomic library using an initial experiment in .bed or .bam format, use the program *lc_extrap*. Use -o to specify the output of the yield estimates, -L to specify the output of the bounds, and -S to specify the output file for the saturation estimates. -v will print more information and will print the library size bounds, if -L is omitted. The options -e and -s set the maximum number of total reads from which yield estimates are desired and the number of reads between estimates, respectively. For confidence intervals of the estimates, -b controls the number of resamples to take and -a controls the confidence level. The last parameter is a .bed or .bam file sorted by chromosome and genomic position.

```
$ library_complexity -o yield.txt -L size_bounds.txt -S saturation_out.txt input.tx
```

1.3 File Format

Input files are mapped read files sorted by chromosome and position in either .bed or .bam format. If files are in .bam format, bamtools is required prior to installation, as detailed in the Installation section ??.

Chapter 2

Detailed usage

2.1 c_curve

c_curve is used to compute the expected complexity curve of a mapped read file by subsampling without replacement and counting the distinct reads. Output is a text file with two columns. The first gives the total number of reads and the second the corresponding number of distinct reads.

-o, -output Name of output file, default prints to screen

-v -verbose Prints more information

-B, -bam Input file is in .bam format. The program must be compiled with bamtools, see Installation ??.

2.2 lc_extrap

lc_extrap is used to compute the expected yield for theoretical larger experiments and bounds on the number of distinct reads in the library and the associated confidence intervals, computed by bootstrapping. Output is a text file with four columns. The first is the total number of reads, second gives the corresponding average expected number of distinct reads, and the third and fourth give the lower and upper limits of the confidence interval. The library size output file will contain the estimated upper and lower bounds on the library size along with the estimated confidence intervals. The saturation output file will consist of 4 columns similar to the yield output, with saturation taking the place of yield.

-o, -output Name of yield output file, defaults prints to screen.

-L, -LIBRARY_SIZE Name of library size bounds output file. If omitted, the option -v will print the bounds.

-S, -SATURATION Name of saturation estimates output file.

-e, -extrapolation_length The maximum number of total reads to compute yield estimates for. The default is 10 billion reads.

-s, -step The step size between yield estimates. Default is 1 million reads.

- b, -bootstraps** The number of bootstraps used to compute the confidence intervals, default is 100. To do a single estimate, set to 0 or 1. This will allow for output of estimated coefficients if VERBOSE is also selected.
- c, -c_level** Confidence of confidence level. Default is 0.95.
- v, -verbose** Print more information.
- B, -bam** Must be specified if the input file is in .bam format. See Installation ?? on how to compile.