Rational Function Complexity Manual

Timothy Daley Andrew Smith

April 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 availible 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 Complilation 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 and -L to specify the output of the bounds. -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 input.txt
```

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 1.1.3.

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

-l, -lower Lower limit of sampling, default is 1M reads

-u, -upper Upper limit of sampling, default is the number of reads in the input file.

-s, -step Step size between sampling points, default is 1M reads.

-v -verbose Prints more information

-b, -bam Input file is in .bam format. The program must be compiled with bamtools, see Installation 1.1.3.

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 test 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.

-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.

-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.

-a, -alpha Significance level of confidence intervals. Default is 0.05, corresponding to 95% confidence intervals.

-t, -terms Maximum number of terms to use in the rational function approximation Must be at least 8. Default uses the maximum number of terms possible. Less terms corresponds to more conservative estimates.

-v, -verbose Print more information.

-B, -bam Must be specified if the input file is in .bam format. See Installation 1.1.3 on how to compile.

-smooth Smooth histogram before approximation. Default is no smoothing.