QRF User Guide

For QRF v0.01

Yaping Liu

Jul. 15. 2015

Contents

1 Introduction

QRF: a random forest model to boost meQTL/eQTL prediction power by using T statistics calculated from common eQTL mapping program (e.g. Matrix EQTL), probes genetic distance-expected genetic distance (recombination rate) and normalized HiC signal. QRF is a public available free software (MIT license) mainly written in Java and wrapped up by perl script. Copyright belongs to Computational Biology Group, CSAIL, MIT .

2 Prerequisites

1. System: Already tested in Linux (CentOS 6) and Mac OSX 10 (10G memory is the minimum requirement).
2. Java: Java(TM) SE Runtime Environment 1.6 (Linux, Mac OSX) or later.
3. Perl: perl scripts in the utilities require Perl v 5.8.8 or later.
4. Specify $PATH: Specify QRF's root directory in the environment

3 Quick Start

3.1 Download QRF program.

Download zipped file from `http://compbio.mit.edu/QRF/`

3.2 Download additional input files for test

All of the input example file could be downloaded from our website: `http://compbio.mit.edu/QRF/`. All *.zip files need to be unzipped firstly. configure.txt will be under software root directory. All the other files will be under ./test_data directory.

1. configure.txt file: specify all of the input files name below and working directory
2. SNP probes location file: SNP.locs.
3. SNP information file: SNP.tab.
4. CpG/Gene_expression probes location file: DNAm.locs.
5. CpG/Gene_expression information file: DNAm.tab.
6. Covariant information file: cov.tab.
7. Recombination rate big wig file: genetic_map_1kGv3_test.hg19.bw.
8. Recombination rate expectation index file: All_length_1M.all_chr.hg19.permutate100.1kGv3.summary.txt (*Download separately*).

9. KR normalized HiC signal file: GM12878_chr1_1kb.KRnorm_raw_test.sparseBed.txt.

10. training file: training_2k.txt

3.3 Run QRF program in terminal.

1. All of the input file should be put into the same directory specified in configure.txt. The following command would output meQTL calling result from matrixEQTL and result from QRF:

```
perl QRF_pipeline.pl test configure.txt
```

4 Interpret output files

4.1 *.afterRandomForest.annotate.txt file

Each row is one pair of meQTL. Here is the explanation of each column:

chr: chromosome name

start: genomic coordinate.(0-based)

end: genomic coordinate.(1-based)

rsid: SNP rsid.

cpg id: CpG's id as in your input DNAm.locs.

matrixEQTL: FDR corrected p value by matrixEQTL

QRF: permutation p value by QRF

5 Input files format

configure.txt file

specify all of the input files name information and working directory. Modify it as you need.

MatrixQtlResult

Optional, when you provided the matrixEQTL result file, QRF will not run matrixEQTL anymore.

SNP_loc

SNP probes location file. As input in matrixEQTL.

gene_loc

CpG/Gene_expression probes location file. As input in matrixEQTL.

SNP_tab

SNP information file. As input in matrixEQTL.

gene_tab

CpG/Gene_expression information file. As input in matrixEQTL.

covar_tab

Covariant information file. As input in matrixEQTL.

recombination_bw

Recombination rate big wig file. Could be downloaded from 1000 genome project, Hapmap, UCSC table browser or our website.

recombination_expect

Recombination rate expectation index file. Could be download from our website.

HiC_KRnorm

KR normalized HiC signal file. Could be converted by "normalize_hic2014_to_sparseBed.pl" script under perl directory (detailed described in Usage Detail section)

training_file

training file for QRF, could be generated by mode 2 in QRF_pipeline.pl or downloaded directly from QRF website

6 Usage Detail

6.1 QRF_pipeline.pl

```
perl QRF_pipeline.pl [option] prefix configure.txt
```

6.1.1 General options

--help

Generate this help message.

--mode < mode_number >

1. Detect meQTL/eQTL; 2. Generate training file; (Default: `'--mode 1'`)

--region < interval >

Specify the region for mode 1 or 2. Should be format like: chr1:1-1000 or chr1 (Default: `'--region chr1'`)

--fdr < FDR >

Specify the false discovery rate for Matrix EQTL and QRF (Default: `'--fdr 0.01'`)

--mem < memory >

Specify the number of gigabytes in the memory to use (Default: `'--mem 15'`)

--qrf_path < QRF_path >

Specify the QRF root direcotry (Default: not specified. use environment variable $QRF)

6.1.2 Options for mode 1, run QRF

--tree_num < tree_number >

Specify the number of tree used for QRF (Default: `'--tree_num 1000'`)

--class_index < class_index >

Specify the column number that are the label column for QRF (Default: `'--class_index 5'`)

--label_class < label_class >

Specify the class label name used for QRF (Default: `'--label_class meqtl'`)

--permutation_times < permutation_times >

Specify the number of permutation used for QRF. 0 or negative value means not enabled (Default: `'--permutation_times 1'`)

--sep < String >

Specify the string used to seperate the column (Default: tab delimit)

--sub_sampling < num_of_samples >

Specify the number of sample used for QRF (Default: not enabled, and use all of the samples)

--hic_resolution < hic_resolution >

Specify the resolution of HiC signal used, default is 1kb (Default: `'--hic_resolution 1'`)

6.1.3 Options for mode 2, generate training file

--positive_probes_sampling < number_of_positive_probes >

Specify the number of positive probes used for QRF training (Default: `'--positive_probes_sampling 10000'`)

--negative_probes_sampling < number_of_negative_probes >

Specify the number of negative probes used for QRF training (Default: `'--negative_probes_sampling 10000'`)

6.2 normalize_hic2014_to_sparseBed.pl

```
perl normalize_hic2014_to_sparseBed.pl  chr1_1kb.KRnorm chr1_1kb.RAWobserved > chr1_1kb.KRnorm.sparseBed.txt
```

6.2.1 Input

chr1_1kb.KRnorm

KRnorm normalized efficient as provided in Rao et al. Cell 2014. (GSE63525)

chr1_1kb.RAWobserved

Raw count as provided in Rao et al. Cell 2014. (GSE63525)

chr1_1kb.KRnorm.sparseBed.txt

output HiC signal for QRF

7 Build on source code

Source code is available on Github website:

https://github.com/dnaase/QRF

All of the required libraries are available in `https://github.com/dnaase/QRF/lib/`.

8 Contact for help

For any of question on QRF, please join our google group `https://groups.google.com/d/forum/qrf-help/` for help

9 Cite QRF

Please use the following publication to cite QRF: