GSER

A pipeline for genome size estimation and quality assessment of sequenced genome libraries

INSTALLATION INSTRUCTIONS

---

1. Dependencies

GSER is implemented in Bash, and requires no installation. However, other software that are part of the pipeline are required. There are 2 required software (R and ntCard), and 3 R libraries required.

SOFTWARE

- **R 3.6.3**

-Source page: https://www.r-project.org/

```
$ wget https://cran.r-project.org/src/base/R-3/R-3.6.3.tar.gz
$ tar -xvzf tar -xvzf R-3.6.3.tar.gz
$ cd R-3.6.3/
$ ./configure
$ make
$ sudo make install
```

Check if installation was correct using

```
$ R --version
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under the terms of the
GNU General Public License versions 2 or 3.
For more information about these matters see
https://www.gnu.org/licenses/.
```

- **ntCard 1.1.0**

-Source page: https://github.com/bcgsc/ntCard

```
$ wget https://github.com/bcgsc/ntCard/archive/v1.1.0.tar.gz
$ tar -xvzf v1.1.0.tar.gz
$ cd ntCard-1.1.0/
$ ./autogen.sh
$ ./configure
$ make
$ sudo make install
```

Check if installation was correct using

```
$ ntcard Version 1.1.0
Written by Hamid Mohamadi.
Copyright 2018 Canada's Michael Smith Genome Science Centre
```

R LIBRARIES

Once R is installed, enter the R command line by typing "R" in a terminal:

```
$ R

R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
```

- **ShortRead**

Source page: https://www.bioconductor.org/packages/release/bioc/html/ShortRead.html

```
> if (!requireNamespace("BiocManager", quietly = TRUE))
      install.packages("BiocManager")
> BiocManager::install("ShortRead")
```

- **ggplot2**

Source page: https://cran.r-project.org/web/packages/ggplot2/

```
> install.packages("ggplot2")
```

- **reshape2**

Source page: https://cran.r-project.org/web/packages/reshape2/

```
> install.packages("reshape2")
```

- **Plotly (OPTIONAL)**

Source page: https://plotly.com/r/getting-started/

```
> install.packages("plotly")
```

2. Setting up GSER

Download the GSER tarball, and uncompress it:

```
$ wget https://github.com/mobilomics/GSER/blob/master/gser_v1.0.6.tar.gz
```

$ tar -xvzf gser_v1.0.6.tar.gz

Grant execution permissions to the pipeline script:

```
$ chmod u+x gser_v1.0.6/gser_v1.0.6.sh
```

For simplicity of use, add the GSER full path to your PATH environment variable. First get the full path:

```
$ readlink -e gser_v1.0.6/gser_v1.0.6.sh
```

Then copy the output of the previous command, and add it to the PATH variable:

```
$ export PATH=$PATH:/path/to/GSER
```

SAMPLE USAGE

Once GSER is in your PATH variable, you can execute it as

```
$ gser_v1.0.6.sh CONFIG THREADS OUTPUTDIR

CONFIG :  Configuration file in the format described below
THREADS : Number of threads to use
OUTPUTDIR: Name of the directory in which the results will be stored
```

Important consideration

The CONFIG file is **tab-separated, 2 column, plain text file** in the following format

```
K kmersize1,kmersize2,kmersize3,..,kmersizeN
Group1  fullpath to fastqfile
Group1  fullpath to another fastqfile
Group2  fullpath to fastqfile
```

Line 1 must have K in the first column, and a comma-separated list of integer numbers to be used as k-mer sizes Line 2 and onwards must have a Group specification and then the full path (obtained with readlink -e) to the FASTQ file corresponding to that group

Example CONFIG file for *Xenopus laevis* genome data available at SRA Study ID SRP071264 (https://trace.ncbi.nlm.nih.gov/Traces/study1/?acc=SRP071264):

```
K       10,11,12,13,14,15,16,17,18,19,20,30,40,50,60,70,80,90,100,110,120,130,140,150
PE_225  /home/user/SRP071264/SRR3210959_1.fastq
PE_225  /home/user/SRP071264/SRR3210959_2.fastq
PE_450  /home/user/SRP071264/SRR3210971_1.fastq
PE_450  /home/user/SRP071264/SRR3210971_2.fastq
PE_900  /home/user/SRP071264/SRR3210972_1.fastq
PE_900  /home/user/SRP071264/SRR3210972_2.fastq
MP_1500 /home/user/SRP071264/SRR3210973_1.fastq
```

```
MP_1500 /home/user/SRP071264/SRR3210973_2.fastq
MP_4000 /home/user/SRP071264/SRR3210974_1.fastq
MP_4000 /home/user/SRP071264/SRR3210974_2.fastq
```

In this example, Genome Size Estimations will be done for k-mer sizes from 10 to 150, in steps of 10. 5 groups were defined according to the type of library (PE: Paired-End and MP: Mate Pair) and insert size (225, 450, 900 for PE, and 1500, 4000 for MP).

CONTACT

Please send any inquiries about usage and/or bugs to mobilomics@gmail.com