rDD-PIPE.README

Simon Zhongyuan Tian

6/7/2022

0. Operation system

```
# This pipeline is running in an Ubuntu 20.04 for rDD library processing
```

1. Create a Directory as working-dir for running rDD library

```
# for example
$ mkdir /mnt/hgfs/script/rDDPP
```

2. Link FASTQ files

```
# jump into the working-dir
$ cd /mnt/hgfs/script/rDDPP

# create a folder to contain FASTQ files
$ mkdir  FQ.demo/
$ cd FQ.demo/

# copy files to this folder
$ cp /path-to-fastq/GM12878_rDD_v-snoRNA1_rep3_BR_1.fq.gz .
$ cp /path-to-fastq/GM12878_rDD_v-snoRNA1_rep3_BR_2.fq.gz .

# or softlink
$ ln -s /path-to-fastq/GM12878_rDD_v-snoRNA1_rep3_BR_1.fq.gz
$ ln -s /path-to-fastq/GM12878_rDD_v-snoRNA1_rep3_BR_2.fq.gz
```

3. Install supported softwares

```
$ sh PRE01_install_soft.sh
```

4. Generate reference genome (hg38-EBV.B958)

```
$ sh PRE02_hg38B_genome.sh
```

5. Modify the configuration file

```
# modify the information in the following config file without execution.

$ vim PRE03_config.sh
```

```
"""
......
echo "NTHREAD=14" > $FC   ## cores you want to use to run this pipeline.
echo "MEM=32g" >$FC  ## RAM you want to use to run this pipeline.
echo "LINKER=LR" >$FC  ## select linker types from:  BR, LR.
echo "fasta="${PIPEDIR}"/ref_genome/hg38B/hg38B.fa" > $FC
                                          ## reference genome fasta file
echo "genome="${PIPEDIR}"/genome_size/hg38B.fa.size" > $FC
                                                ## reference genome size
echo "SPLT=Y" >$FC
              ## split loops and coverage by species (hg38-EBV): Y=yes; N=not.
......
"""
```

6. Run rDD-PIPE pipeline

```
$ sh RDD00_RUN_rDDPP.sh

> please input the forlder of FASTQ
> FASTQ file should like this: rHG011_1.fq.gz rHG011_2.fq.gz

$ FQ.demo

>select config file

PRE03_config.sh

## rDD-PIPE started when you see  a new file named:

> FQ.demo.PRE03_config.sh.20220602-021820.START...
```

7. Check executing status

```
# rDD-PIPE running log is recorded in this log file in realtime

>FQ.demo.PRE03_config.sh.20220602-021820.log

$ less FQ.demo.PRE03_config.sh.20220602-021820.log
"""
sh RDD01_run_script.sh FQ.demo PRE03_config.sh FQ.demo.log 2>&1 &
FQ.demo
PRE03_config.sh
GM12878_rDD_v-snoRNA1_rep3_BR
#!/bin/bash
LIB=GM12878_rDD_v-snoRNA1_rep3_BR
NTHREAD=14
MEM=32g
datadir=/mnt/hgfs/script/rDDPP/GM12878_rDD_v-snoRNA1_rep3_BR
mainprog=/mnt/hgfs/script/rDDPP/rDD-PIPE/util/cpu-dir/cpu
JUICER=/mnt/hgfs/script/rDDPP/rDD-PIPE/util/juicer_tools.1.7.5_linux_x64_jcuda.0.8.jar
LINKER=LR
fasta=/mnt/hgfs/script/rDDPP/rDD-PIPE/ref_genome/hg38B/hg38B.fa
```

```
genome=/mnt/hgfs/script/rDDPP/rDD-PIPE/genome_size/hg38B.fa.size
SPLT=Y
bash ./10.filter_linker.pipe.sh
Linker detection on:  GM12878_rDD_v-snoRNA1_rep3_BR_1.fq.gz and
                      GM12878_rDD_v-snoRNA1_rep3_BR_2.fq.gz
LINKER=LR
bash ./22.map_single_linker_2tags.pipe.sh
Thu 02 Jun 2022 02:23:30 AM PDT STARTED GM12878_rDD_v-snoRNA1_rep3_BR cpu memaln ..
Thu 02 Jun 2022 02:23:30 AM PDT Mapping paired tags ..
......
"""
```

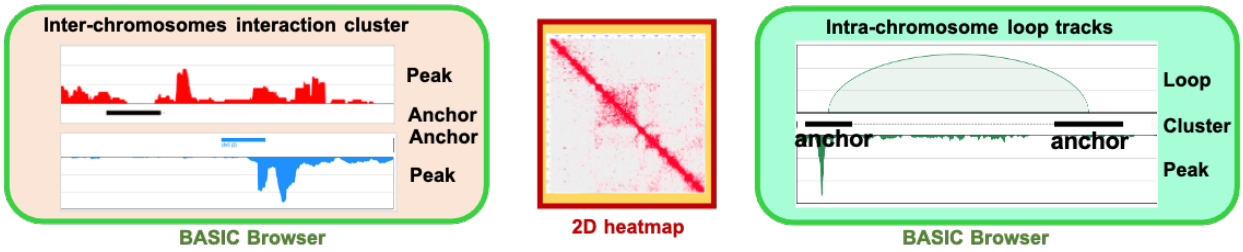
8. Results description

Following files are main results of a rDD library:

```
"""
GM12878_rDD_v-snoRNA1_rep3_BR.cluster
                        ## is the chromatin interaction cluster (loop) file
GM12878_rDD_v-snoRNA1_rep3_BR.for.BROWSER.bedgraph
                        ## is the signal coverage file
GM12878_rDD_v-snoRNA1_rep3_BR.hic
                        ## is the file for  2D heatmap to display in Juicerbox
GM12878_rDD_v-snoRNA1_rep3_BR.no_input_all_peaks.narrowPeak
                        ## is the peak called by macs2

BASIC.DIR/GM12878_rDD_v-snoRNA1_rep3_BR.HG38B.HH.cluster
                        ## is the contacts between Host-Host
BASIC.DIR/GM12878_rDD_v-snoRNA1_rep3_BR.HG38B.EE.cluster
                        ## is the contacts between EBV-EBV
BASIC.DIR/GM12878_rDD_v-snoRNA1_rep3_BR.HG38B.HE-E.anchor
                         ## is the EBV side anchor of the contacts between Host-EBV
BASIC.DIR/GM12878_rDD_v-snoRNA1_rep3_BR.HG38B.HE-H.anchor
                        ## is the Host side anchor of  the contacts between Host-EBV
BASIC.DIR/GM12878_rDD_v-snoRNA1_rep3_BR.HG38B.H.bdg
                        ## is the signal coverage of Host genome
BASIC.DIR/GM12878_rDD_v-snoRNA1_rep3_BR.HG38B.E.bdg
                        ## is the signal coverage of EBV genome
"""
```

9. Data vasualization



10. Quality control Table

GM12878_rDD_v-snoRNA1_rep3_BR.final_stats.tsv

Item	Value
Library_ID	GM12878_rDD_v-snoRNA1_rep3_BR
Reference_genome	hg38B.fa.size
Total_read_pairs	14,096,056
Read_pairs_with_linker	11,655,019
Fraction_read_pairs_with_linker	0.83
One_tag	5,035,442
PET	6,379,723
Uniquely_mapped_PET	4,976,501
Non-redundant_PET	2,135,758
Redundancy	0.57
Non-redundant_tag	9772813
Peak	50,413
Self-ligation_PET	502,461
Inter-ligation_PET	1,633,297
Intra-chr_PET	516,168
Inter-chr_PET	1,117,129
ratio_of_intra/inter_PET	0.46
Singleton	1,524,230
Intra-chr_singleton	462,324
Inter-chr_singleton	1,061,906
PET_cluster	48,302
ratio_of_intra/inter_cluster	0.89
Intra-chr_PET_cluster	22,808
pets_number_2	18,976
pets_number_3	2,508
pets_number_4	750
pets_number_5	321
pets_number_6	130
pets_number_7	63
pets_number_8	29
pets_number_9	14
pets_number_10	7
pets_number>10	10
Inter-chr_PET_cluster	25,494
pets_number_2	22,655
pets_number_3	2,017
pets_number_4	507
pets_number_5	173
pets_number_6	75
pets_number_7	42
pets_number_8	11
pets_number_9	7
pets_number_10	4
pets_number>10	3
Host-Host_Loops	47,957
EBV-EBV_Loops	212
Host-EBV_Loops	133

11. Quality control Plots

GM12878_rDD_v-snoRNA1_rep3_BR.QC.pdf