wind: wORKFLOW FOR PiRNAs AnD BEYONd

Computational workflow for the preprocessing of the GSE68246 dataset regarding Human Breast MCF-7 Cell Line with Cancer Stem Cell Properties

Constantinos Yeles (Konstantinos Geles)

Thu_Oct_29_2020

Contents

The Data set

We will work on a public dataset with GEO accession number: GSE68246, that it has been used in the publications: **Phenotypic and microRNA transcriptomic profiling of the MDA-MB-231 spheroid-enriched CSCs with comparison of MCF-7 microRNA profiling dataset** and **MiRNA Transcriptome Profiling of Spheroid-Enriched Cells with Cancer Stem Cell Properties in Human Breast MCF-7 Cell Line**

Data aqcuisition and preprocessing

i. Downloading the samples

We use a script to download the fastq samples with samtools-kit that it is included in the docker with the name download_SRA.sh

Using the **SRA selector** we download a file with the Accession List and rename the file to **GSE68246_samples.txt**

```
docker run --rm -ti -v $(pwd):/home/my_data  congelos/sncrna_workflow

# run the script to download the SRA
./download_SRA.sh GSE124507_samples.txt 8
```

ii. Preprocessing of the samples

We perform quality control(QC) on the fastq files to get basic information about the samples. We work with the **Fastqc** tool to perform QC.

```
mkdir my_data/qc_first

'fastqc' --threads 6 --outdir=my_data/qc_first/ my_data/downloaded_SRA/GSE_samples/*.fastq.gz

for file in my_data/downloaded_SRA/GSE_samples/*.fastq.gz;
do ./spar_prepare/smrna_adapter_cut.sh $file 6;
done

mkdir my_data/downloaded_SRA/GSE_samples/qc_after

'fastqc' --threads 6 --outdir=my_data/qc_after/ my_data/downloaded_SRA/GSE_samples/*.trimmed.fastq.gz

exit
```

Alignment and Quantification

i. Transcript abundances with Salmon

We will use a public docker image to run salmon

```
# run the docker
docker run --rm -it -v $(pwd):/home/my_data combinelab/salmon

# create the index
salmon index -t ncRNA_transcripts_100bp_RNA_Central_piRNAbank_hg38.fa -i genome_transc_human/ncRNA_Centr

mkdir my_data/smallRNA-breast-cancer/GSE68246/quants/
# run the samples

#!/bin/bash

for fn in my_data/smallRNA-breast-cancer/GSE68246/GSE_samples/*trimmed.fastq.gz;
do samp=`basename ${fn}`;
echo "Processing sample ${samp}";
salmon quant -i my_data/genome_transc_human/ncRNA_Central_piRNAB_hg38_index -l A -r ${fn} --seqBias --g
done

#save as bam files
for file in my_data/smallRNA-breast-cancer/GSE68246/quants/*.sam;
do
regex="${file%%.sam}";
```

```
echo samtools view -O bam -o ${regex}.bam -@ 6 ${file};
done
exit
```

Alignment and quantification of sequenced reads with STAR and Featurecounts

We use the **STAR** aligner and then perform quantification with featureCounts from **Rsubread** package. With the a docker images that contains STAR and **Samtools** we get sorted BAM files and use them for quantification / annotation for smallRNAs.

ii. Alignment with STAR

```
docker run --rm -ti -v "$PWD":/home/my_data congelos/sncrna_workflow

STAR --runMode genomeGenerate --genomeDir my_data/mouse_data/GRCh38 --genomeFastaFiles my_data/mouse_da

mkdir my_data/smallRNA-breast-cancer/GSE68246/star_results

for file in my_data/smallRNA-breast-cancer/GSE68246/GSE_samples/*.trimmed.fastq.gz;
do
samp=`basename ${file}`;
regex="${samp%%.trimmed.fastq.gz}";
echo "Processing sample ${samp} start: $(date)";
STAR --genomeDir my_data/genome_transc_human/human_data/GRCh38_2_7_4a --genomeLoad LoadAndKeep --readFi
echo "end:$(date)";
done
exit
```

Next, we run a docker image which includes varius R packages that will be used futhermore in the downstream analysis following featurecounts for the exploratory data analysis of piRNA data

R docker

```
docker run --rm -v $(pwd):/home/0 -p 8787:8787 -e PASSWORD=12345 -e USER=$UID congelos/rocker_tidyverse
```

From here on we work in R using a browser. we input http://localhost:8787/ on browser and 0 for username and 12345 for password.

iv. featureCounts

```
library(Rsubread)
library(tidyverse)
list.BAM <- list.files(path = "GSE68246/star_results",
                       pattern = ".bam$",
                       recursive = TRUE,
                       full.names = T)
```

```
path_gtf <- "../genome_transc_human/ncRNA_transcripts_100bp_RNA_Central_piRNAbank_hg38.gtf"
todate <- format(Sys.time(), "%d_%b_%Y")

fc <- featureCounts(files = list.BAM,
                    annot.ext =  path_gtf,
                    isGTFAnnotationFile = TRUE,
                    GTF.featureType = "exon",
                    GTF.attrType.extra = c("gene_type", "sRNA_id", "seq_RNA"),
                    nthreads = 6,
                    useMetaFeatures = TRUE,
                    allowMultiOverlap = TRUE,
                    minOverlap = 10,
                    largestOverlap = TRUE,
                    fraction = TRUE,
                    strandSpecific = 0,
                    verbose = TRUE,
                    reportReads = "BAM",
                    reportReadsPath = "GSE68246/star_results")
fc %>% write_rds(str_glue("GSE68246/feature_counts_GSE68246_{todate}.rds"))
```

Next we will follow the workflow of data_exploration_salmon_fc ## R Session Info

```
sessionInfo()
R Under development (unstable) (2019-12-06 r77536)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Debian GNU/Linux 10 (buster)

Matrix products: default
BLAS/LAPACK: /usr/lib/x86_64-linux-gnu/libopenblasp-r0.3.5.so

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=C
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] forcats_0.5.0       stringr_1.4.0       dplyr_0.8.4         purrr_0.3.3
 [5] readr_1.3.1         tidyr_1.0.2         tibble_2.1.3        ggplot2_3.2.1
 [9] tidyverse_1.3.0     Rsubread_2.1.2      BiocManager_1.30.10

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.3       cellranger_1.1.0 pillar_1.4.3     compiler_4.0.0
 [5] dbplyr_1.4.2     tools_4.0.0      lubridate_1.7.4  jsonlite_1.6.1
 [9] lifecycle_0.1.0  nlme_3.1-144     gtable_0.3.0     lattice_0.20-40
[13] pkgconfig_2.0.3  rlang_0.4.5      reprex_0.3.0     Matrix_1.2-18
[17] cli_2.0.2        DBI_1.1.0        rstudioapi_0.11  xfun_0.12
[21] haven_2.2.0      knitr_1.28       withr_2.1.2      xml2_1.2.2
```

```
[25] httr_1.4.1       fs_1.3.1         generics_0.0.2   vctrs_0.2.3
[29] hms_0.5.3        grid_4.0.0       tidyselect_1.0.0 glue_1.3.1
[33] R6_2.4.1         fansi_0.4.1      readxl_1.3.1     modelr_0.1.6
[37] magrittr_1.5     backports_1.1.5  scales_1.1.0     rvest_0.3.5
[41] assertthat_0.2.1 colorspace_1.4-1 stringi_1.4.6    lazyeval_0.2.2
[45] munsell_0.5.0    broom_0.5.5      crayon_1.3.4
```

We work on :

```
[root@localhost GSE124507_brain_project]# cat /etc/*-release

CentOS Linux release 7.8.2003 (Core)
NAME="CentOS Linux"
VERSION="7 (Core)"
ID="centos"
ID_LIKE="rhel fedora"
VERSION_ID="7"
PRETTY_NAME="CentOS Linux 7 (Core)"
ANSI_COLOR="0;31"
CPE_NAME="cpe:/o:centos:centos:7"

[root@localhost GSE124507_brain_project]# docker version

Client: Docker Engine - Community
 Version:           19.03.8
 API version:       1.40
 Go version:        go1.12.17
 Git commit:        afacb8b
 Built:             Wed Mar 11 01:27:04 2020
 OS/Arch:           linux/amd64
 Experimental:      false

Server: Docker Engine - Community
 Engine:
  Version:          19.03.8
  API version:      1.40 (minimum version 1.12)
  Go version:       go1.12.17
  Git commit:       afacb8b
  Built:            Wed Mar 11 01:25:42 2020
  OS/Arch:          linux/amd64
  Experimental:     false
 containerd:
  Version:          1.2.13
  GitCommit:        7ad184331fa3e55e52b890ea95e65ba581ae3429
 runc:
  Version:          1.0.0-rc10
  GitCommit:        dc9208a3303feef5b3839f4323d9beb36df0a9dd
 docker-init:
  Version:          0.18.0
  GitCommit:        fec3683
```

```
[root@localhost GSE124507_brain_project]# git version
git version 1.8.3.1

[root@localhost GSE124507_brain_project]# pigz --version
pigz 2.3.4
```