Day 6 Worksheet – featureCounts

Author: Rutendo Sigauke

Introduction:

The featureCounts library is part of Subread (written in C) and RSubread (an R wrapper for Subread), and it is a fast tool optimized for counting reads over features (genes, exons, transcripts …). To see the full utility of Subreads/Rsubread, see their documentation below:

Subread: http://subread.sourceforge.net/

RSubread: http://subread.sourceforge.net/SubreadUsersGuide.pdf

Since counting in compute intensive, this is done on the terminal. Usually we can request multiple threads which make the counting running faster. ***We will be completing the counting section on the server****.*

Install Rsubread:

Before running Rsubread, we have to install the library in to R. Installation can be done in the R console (shown below).

- Type **R** in the terminal

```
[rutendos@ip-172-31-18-92 ~]$ R

R version 3.6.0 (2019-04-26) -- "Planting of a Tree"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-redhat-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.
```

- Rsubread can be found on the **BiocManager**, so to install the counting library, we have to first install **BiocManager**. **BiocManager** library can be installed from the R Comprehensive R Archive Network (CRAN).

```
if (!require("BiocManager", quietly = TRUE))

    install.packages("BiocManager")
```

```
> if (!requireNamespace("BiocManager", quietly = TRUE))  ## Install BiocManager
+   install.packages("BiocManager")
Installing package into ‘/usr/lib64/R/library’
(as ‘lib’ is unspecified)
Warning in install.packages("BiocManager") :
  'lib = "/usr/lib64/R/library"' is not writable
Would you like to use a personal library instead? (yes/No/cancel) yes
Would you like to create a personal library
‘~/R/x86_64-redhat-linux-gnu-library/3.6’
to install packages into? (yes/No/cancel) yes
```

- The above command will list CRAN mirrors from where to download the packages. We will uses 72 for USA (KS) since that is closest mirror to Colorado:

```
--- Please select a CRAN mirror for use in this session ---
Secure CRAN mirrors

 1: 0-Cloud [https]
 2: Australia (Canberra) [https]
 3: Australia (Melbourne 1) [https]
 4: Australia (Melbourne 2) [https]
 5: Australia (Perth) [https]
 6: Austria [https]
 7: Belgium (Brussels) [https]
 8: Brazil (PR) [https]
 9: Brazil (RJ) [https]
10: Brazil (SP 1) [https]
11: Brazil (SP 2) [https]
```

```
71: USA (IA) [https]
72: USA (KS) [https]
73: USA (MI) [https]
74: USA (OH) [https]
75: USA (OR) [https]
76: USA (TN) [https]
77: USA (TX 1) [https]
78: Uruguay [https]
79: (other mirrors)

Selection: 72
```

- Now, we can install **Rsubread** from **BiocManager** to our **R** libraries.

  BiocManager::install("Rsubread")

```
> BiocManager::install("Rsubread")
'getOption("repos")' replaces Bioconductor standard repositories, see
'?repositories' for details

replacement repositories:
    CRAN: https://rweb.crmda.ku.edu/cran

Bioconductor version 3.10 (BiocManager 1.30.18), R 3.6.0 (2019-04-26)
Installing package(s) 'BiocVersion', 'Rsubread'
trying URL 'https://bioconductor.org/packages/3.10/bioc/src/contrib/BiocVersion_3.10.1.tar.gz'
Content type 'application/x-gzip' length 984 bytes
==================================================
downloaded 984 bytes
```

NB: This will take a few seconds. If the library is installed successfully, it can be loaded as shown below without any errors.

```
> library("Rsubread")
```

Make working directories:

Similar to previous worksheets, make the necessary working directories for running featureCounts. Repeat the same process, but this time we will make a directory for macs.

1. Use command **pwd** to determine what directory you are in and if necessary, **cd** to the directory that you want to place your new macs directory in.

2. Make a new directory in your **/scracth/Users/<username>/day6** using the **mkdir** command. Use command **ls -lsh** to confirm the folders are present.

```
$ mkdir featureCounts featureCounts/scripts featureCounts/output
```

3. Copy featureCounts scripts from
/scratch/Shares/public/sread2022/scripts/day6/

4. Edit both scripts using **vim <script>**. This will open the scripts in the text editor.

Edit R script:

- Set your working directory

```
## ----setwd, eval=TRUE----------------------

workdir <- '/PATH/TO/WORKING/DIRECTORY'
setwd(workdir)
getwd()
```

NB: The output folders will be generated based on your **`workdir`**

- Make sure the bam folder path is correct

`/scratch/Shares/public/sread2022/data_files/day6/bam`

```
## ----bamdir, eval=TRUE---------------------------------------
bamdir <- '/scratch/Shares/public/sread/data_files/day6/bam'
```

- Check what annotations are being used, and make sure the path is correct.

`/scratch/Shares/public/sread2022/data_files/day6/annotations/hg38_ucsc_genes_chr21.gtf`

```
## ----loadGTF, eval=TRUE--------------------------------------------------------------------------

hg38gtf <- "/scratch/Shares/public/sread/data_files/day6/annotations/hg38_ucsc_genes_chr21.gtf"
```

NB: Take a look at the GTF file structure in the commandline (exit vim or R console). Note all the different features represented for each feature. Also, you will see that the file has several columns, with the **first** column is the chromosome ID, the **second** column is the name of the source from which the feature was derived (eg. Ensembl, UCSC or HAVANA). The **third** column is the label for the feature (e.g. exon, CDS, start_codon). This field is used by featureCounts to determine the features to to count reads over. The **fourth** and the **fifth** columns are start and end coordinates respectively. The **sixth** column is the score of the feature, the **seventh** the strand, the **eighth** is phase for CDS features (If phase=0, the codon begin at the first base of CDS nucleotide; if phase=1 the codon begin at the second base of CDS nucleotide; if phase=2 the

codon begin at the third base of CDS nucleotide.). Lastly, the **nineth** column contains additional feature annotations.

```
[rutendos@ip-172-31-18-92 ~]$ head /scratch/Shares/public/sread2022/data_files/day6/annotations/hg38_ucsc_genes_chr21.gtf
chr21   unknown exon    5022493 5022693 .       +       .       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P15779"; transcr
chr21   unknown exon    5022493 5022693 .       +       .       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P7798"; transcri
chr21   unknown exon    5022493 5022693 .       +       .       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P19886"; transcr
chr21   unknown CDS     5022680 5022693 .       +       0       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P7798"; transcri
chr21   unknown CDS     5022680 5022693 .       +       0       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P19886"; transcr
chr21   unknown start_codon     5022680 5022682 .       +       .       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P7798";
chr21   unknown start_codon     5022680 5022682 .       +       .       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P19886";
chr21   unknown CDS     5025009 5025049 .       +       1       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P7798"; transcri
chr21   unknown CDS     5025009 5025049 .       +       1       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P19886"; transcr
chr21   unknown exon    5025009 5025049 .       +       .       gene_id "ICOSLG"; gene_name "ICOSLG"; p_id "P15779"; transcr
```

Edit sbatch script:

- Edit then sbatch script including the **`SBATCH headers`** and path to the **`d6_featureCounts.R`** script.

```
#!/bin/bash

#SBATCH --job-name=<NAME OF JOB>                                # Job name
#SBATCH --mail-type=ALL                                 # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=<YOUR E-MAIL ADDRESS>                       # Where to send mail
#SBATCH --nodes=1                                               # Number of cores job will run on
#SBATCH --ntasks=4                                              # Number of CPU (processers, tasks)
#SBATCH --time=1:00:00                                          # Time limit hrs:min:sec
#SBATCH --partition compute                                     # Job queue
#SBATCH --mem=4gb                                               # Memory limit
#SBATCH --output=/YOUR/EOFILES/PATH/%x_%j.out
#SBATCH --error=/YOUR/EOFILES/PATH/%x_%j.err

################## SET VARIABLES #####################################

FEATURECOUNTS=/PATH/TO/YOUR/d6_featureCounts.R

######################################################################
################## PRINT JOB INFO ####################################

printf "Sample ID: $ROOTNAME"
printf "\nDirectory: $PROJECT"
printf "\nRun on: $(hostname)"
printf "\nRun from: $(pwd)"
printf "\nScript: $0\n"
date

printf "\nYou've requested $SLURM_CPUS_ON_NODE core(s).\n"

######################################################################

Rscript $FEATURECOUNTS
```

- Run the sbatch script!

5. You can find the summary of counts in the output counts folder. There are five different files:

- `featureCounts_gene_rnaseq.txt` : GeneID, Length, Counts

- `coverage.csv` : Counts

- `.stat.csv` : Coverage Statistics

- `.annotation.csv` : GeneID, Chromosome, Start, End, Strand, Length

- `.targets.csv` : file names for input `bam` files

```
[rutendos@ip-172-31-18-92 counts]$ pwd
/scratch/Users/rutendos/day6/featureCounts/output/counts
[rutendos@ip-172-31-18-92 counts]$ ls -lah
total 196K
drwxrwxr-x 2 rutendos rutendos 6.0K Jul 18 08:56 .
drwxrwxr-x 3 rutendos rutendos 6.0K Jul 18 08:56 ..
-rw-rw-r-- 1 rutendos rutendos 162K Jul 18 08:56 chr21_Ethan_Eric.annotation.csv
-rw-rw-r-- 1 rutendos rutendos 5.6K Jul 18 08:56 chr21_Ethan_Eric.coverage.csv
-rw-rw-r-- 1 rutendos rutendos 6.5K Jul 18 08:56 chr21_Ethan_Eric_featureCounts_gene_rnaseq.txt
-rw-rw-r-- 1 rutendos rutendos  539 Jul 18 08:56 chr21_Ethan_Eric.stat.csv
-rw-rw-r-- 1 rutendos rutendos   72 Jul 18 08:56 chr21_Ethan_Eric.targets.csv
```

6. Open and explore each of the files in the terminal (with `head` or `less`). You can also move them to your local computer using `scp` or `rsync`, and load the files in RStudio.

- `chr21_Ethan_Eric.annotation.csv`

```
[rutendos@ip-172-31-18-92 counts]$ head chr21_Ethan_Eric.annotation.csv
"","GeneID","Chr","Start","End","Strand","Length"
"1","ICOSLG","chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21
;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21"
5;5032053;5032053;5032053;5033408;5033408;5033408;5034582;5034582;5040585;44222995;44226827;44226827;44
;44235272;44235272;44236867;44236867;44236867;44238448;44238448;44238448;44238448;44240804;44240804;442
;5028225;5028225;5032217;5032217;5032217;5033443;5033443;5033443;5036782;5036782;5040668;44223078;44229
235562;44235562;44235562;44235562;44237065;44237217;44237217;44238488;44238488;44238488;44238488;442410
-;-;-;-;-;-;-;-;-;-;-;-;-;-;-",6757
"2","C21orf33","chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr21;chr
118691;5121711;5124756;5124756;5125879;5125879;5127898;5127898;5128214;5128214;44133612;44133612;441341
117231;5117231;5118847;5118847;5121803;5124875;5124875;5125992;5125992;5127950;5127950;5128440;5128440;
61;44145723;44145723","-;-;-;-;-;-;-;-;-;-;-;-;-;+;+;+;+;+;+;+;+;+;+;+;+;+",3334
```

- `chr21_Ethan_Eric.coverage.csv`

```
[rutendos@ip-172-31-18-92 counts]$ head chr21_Ethan_Eric.coverage.csv
"","chr21Eric.repA.sorted.bam","chr21Ethan.repA.sorted.bam"
"ICOSLG",1253,2266
"C21orf33",1333,2833
"PWP2",490,1438
"LINC00313",0,0
"LINC00319",0,0
"SIK1",426,1083
"CBS",114,135
"U2AF1",1817,3023
"CRYAA",0,2
```

- chr21_Ethan_Eric_featureCounts_gene_rnaseq.txt

```
[rutendos@ip-172-31-18-92 counts]$ head chr21_Ethan_Eric_featureCounts_gene_rnaseq.txt
GeneID  Length  chr21Eric.repA.sorted.bam       chr21Ethan.repA.sorted.bam
ICOSLG  6757    1253    2266
C21orf33        3334    1333    2833
PWP2    6520    490     1438
LINC00313       1158    0       0
LINC00319       6004    0       0
SIK1    9404    426     1083
CBS     5456    114     135
U2AF1   2040    1817    3023
CRYAA   2288    0       2
```

- chr21_Ethan_Eric.stat.csv

```
[rutendos@ip-172-31-18-92 counts]$ head chr21_Ethan_Eric.stat.csv
"","Status","chr21Eric.repA.sorted.bam","chr21Ethan.repA.sorted.bam"
"1","Assigned",104000,202339
"2","Unassigned_Unmapped",679,578
"3","Unassigned_Read_Type",0,0
"4","Unassigned_Singleton",0,0
"5","Unassigned_MappingQuality",0,0
"6","Unassigned_Chimera",0,0
"7","Unassigned_FragmentLength",0,0
"8","Unassigned_Duplicate",0,0
"9","Unassigned_MultiMapping",0,0
```

- chr21_Ethan_Eric.targets.csv

```
[rutendos@ip-172-31-18-92 counts]$ cat chr21_Ethan_Eric.targets.csv
"","x"
"1","chr21Eric.repA.sorted.bam"
"2","chr21Ethan.repA.sorted.bam"
```