Common Marmoset Gut Microbiome Profiles in Health and Intestinal Disease

Alex Sheh

September 21, 2020

This file takes FASTQ files and outputs feature counts

1 Obtain the FASTQ files deposited in SRA

The fastq files can be downloaded from SRA #SRP278735 (Bioproject # PRJNA659238)

SRR12514529/SAMN15903944 - Non-stricture control (IBD) - Duodenum - tissue cross-section - Sample #3

SRR12514530/SAMN15903943 - Non-stricture control (IBD) - Duodenum - tissue cross-section - Sample #7

SRR12514531/SAMN15903942 - Non-stricture control (IBD) - Duodenum - tissue cross-section - Sample #9

SRR12514532/SAMN15903941 - Duodenal stricture/ulcer - Duodenum - tissue cross-section - Sample #5

SRR12514533/SAMN15903940 - Duodenal stricture/ulcer - Duodenum - tissue cross-section - Sample #14

SRR12514534/SAMN15903939 - Duodenal stricture/ulcer - Duodenum - tissue cross-section - Sample #21

SRR12514535/SAMN15903938 - IBD - Jejunum - tissue cross-section - Sample #2

SRR12514536/SAMN15903937 - IBD - Jejunum - tissue cross-section - Sample #4

SRR12514537/SAMN15903936 - IBD - Jejunum - tissue cross-section - Sample #6

SRR12514538/SAMN15903935 - Non-IBD control (stricture) - Jejunum - tissue cross-section - Sample #8

SRR12514539/SAMN15903934 - Non-IBD control (stricture) - Jejunum - tissue cross-section - Sample #10

SRR12514540/SAMN15903933 - Non-IBD control (stricture) - Jejunum - tissue cross-section - Sample #15

We rename them and place them in /Raw_Data folder and gzipped For example, after running the SRA toolkit 2.10.8 and obtaining split FASTQ files from SRR12514540. We use "fastq-dump SRR12514540 –split-files we obtained SRR12514540_1.fastq and SRR12514540_2.fastq. These files would be renamed 15_R1.fastq and 15_R2.fastq, and gzipped to fastq.gz files. We placed in /Raw_Data based on the Sample # above.

2 Download references for Callithrix jacchus

We used the March 2009 (WUGSC 3.2/calJac3) assembly of the marmoset genome (calJac3) http://hgdownload.soe.ucsc.edu/goldenPath/calJac3/bigZips/ We downloaded "calJac3.fa.gz" and "calJac3.ncbiRefSeq.gtf.gz" into folder /ref and unzipped the .gz files

Now we proceed to load the libraries

```r
# for ML algorithms
library(Rsubread)
library(edgeR)
```

```
## Loading required package: limma
```

```r
library(gplots)
```

```
##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess
```

```r
library(org.Hs.eg.db)
```

```
## Loading required package: AnnotationDbi

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following object is masked from 'package:limma':
##
##     plotMA

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min
## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: IRanges
```

```
## Loading required package: S4Vectors
```

```
##
## Attaching package: 'S4Vectors'
```

```
## The following object is masked from 'package:gplots':
##
##     space
```

```
## The following object is masked from 'package:base':
##
##     expand.grid
```

```
##
## Attaching package: 'IRanges'
```

```
## The following object is masked from 'package:grDevices':
##
##     windows
```

```
##
```

```r
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18363)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] parallel  stats4    stats     graphics  grDevices utils     datasets
## [8] methods   base
##
```

```
## other attached packages:
##  [1] org.Hs.eg.db_3.10.0  AnnotationDbi_1.48.0 IRanges_2.20.2
##  [4] S4Vectors_0.24.4     Biobase_2.46.0       BiocGenerics_0.32.0
##  [7] gplots_3.0.4         edgeR_3.28.1         limma_3.42.2
## [10] Rsubread_2.0.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.5         knitr_1.29         magrittr_1.5       bit_1.1-15.2
##  [5] lattice_0.20-38    rlang_0.4.7        blob_1.2.1         stringr_1.4.0
##  [9] caTools_1.18.0     tools_3.6.3        grid_3.6.3         xfun_0.16
## [13] KernSmooth_2.23-16 DBI_1.1.0          htmltools_0.5.0    gtools_3.8.2
## [17] bit64_0.9-7        yaml_2.2.1         digest_0.6.25      vctrs_0.3.1
## [21] bitops_1.0-6       memoise_1.1.0      RSQLite_2.2.0      evaluate_0.14
## [25] rmarkdown_2.3      gdata_2.18.0       stringi_1.4.6      compiler_3.6.3
## [29] locfit_1.5-9.4     pkgconfig_2.0.3
```

```r
fastq.files <- list.files(path = "./Raw_Data", pattern = ".fastq.gz$", full.names = TRUE)
buildindex(basename="cj",reference="./ref/calJac3.fa", gappedIndex = TRUE) # index is built

# now we use the "cj" index created from "calJac3.fa" to align all the reads and generate BAM files
align(index="cj",readfile1="./Raw_Data/2_R1.fastq.gz",readfile2="./Raw_Data/2_R2.fastq.gz",
      type="dna", output_file="2.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/3_R1.fastq.gz",readfile2="./Raw_Data/3_R2.fastq.gz",
      type="dna", output_file="3.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/4_R1.fastq.gz",readfile2="./Raw_Data/4_R2.fastq.gz",
      type="dna", output_file="4.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/5_R1.fastq.gz",readfile2="./Raw_Data/5_R2.fastq.gz",
      type="dna", output_file="5.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/6_R1.fastq.gz",readfile2="./Raw_Data/6_R2.fastq.gz",
      type="dna", output_file="6.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/7_R1.fastq.gz",readfile2="./Raw_Data/7_R2.fastq.gz",
      type="dna", output_file="7.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/8_R1.fastq.gz",readfile2="./Raw_Data/8_R2.fastq.gz",
      type="dna", output_file="8.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/9_R1.fastq.gz",readfile2="./Raw_Data/9_R2.fastq.gz",
      type="dna", output_file="9.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/10_R1.fastq.gz",readfile2="./Raw_Data/10_R2.fastq.gz",
      type="dna", output_file="10.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/14_R1.fastq.gz",readfile2="./Raw_Data/14_R2.fastq.gz",
      type="dna", output_file="14.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/15_R1.fastq.gz",readfile2="./Raw_Data/15_R2.fastq.gz",
      type="dna", output_file="15.bam", nthreads=4, minFragLength=50,maxFragLength=600)
align(index="cj",readfile1="./Raw_Data/21_R1.fastq.gz",readfile2="./Raw_Data/21_R2.fastq.gz",
      type="dna", output_file="21.bam", nthreads=4, minFragLength=50,maxFragLength=600)

#flattened GTF file and use "gene_id" using marmoset annotations
cj_annot <- flattenGTF("./ref/calJac3.ncbiRefSeq.gtf", GTF.featureType = "exon",
                       GTF.attrType = "gene_id", method = "merge")

# counts features in the BAM files
fc_2 <- featureCounts("2.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_3 <- featureCounts("3.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_4 <- featureCounts("4.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_5 <- featureCounts("5.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
```

```r
fc_6 <- featureCounts("6.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_7 <- featureCounts("7.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_8 <- featureCounts("8.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_9 <- featureCounts("9.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_10 <- featureCounts("10.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_14 <- featureCounts("14.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_15 <- featureCounts("15.bam", annot.ext = cj_annot, isPairedEnd = TRUE)
fc_21 <- featureCounts("21.bam", annot.ext = cj_annot, isPairedEnd = TRUE)


# Due to the paucity of annotated marmoset databases for pathway analysis, marmoset gene names
# were cross-referenced with human gene names. Exact matches were recorded and human Entrez
# gene IDs were added to a modified GTF file witn additional column "h_id". Unmatched genes
# were marked "NA"
#A second annotation file (cj_annot_h) was created using the human Entrez gene ID
#column "h_id" as the GTF.attrType
load("cj_annot_h.RData")

fc_2h <- featureCounts("2.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_3h <- featureCounts("3.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_4h <- featureCounts("4.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_5h <- featureCounts("5.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_6h <- featureCounts("6.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_7h <- featureCounts("7.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_8h <- featureCounts("8.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_9h <- featureCounts("9.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_10h <- featureCounts("10.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_14h <- featureCounts("14.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_15h <- featureCounts("15.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
fc_21h <- featureCounts("21.bam", annot.ext = cj_annot_h, isPairedEnd = TRUE)
```