Day 6 : Introduction to R, RStudio and featureCounts

Taylor Jones, Jesse Kurland and Rutendo Sigauke

Day 6 Overview

- Brief recap/introduction to R
- Running R in Rstudio and in the terminal
- Installing packages
- Counting reads in R

Goal of the day

1. Running `R` in `RStudio`
    a. Install packages
    b. Load files in R
    c. Performing basic statistics
    d. Plotting figures
2. Running `R` in the terminal
3. Counting reads with `featureCounts` in `R`

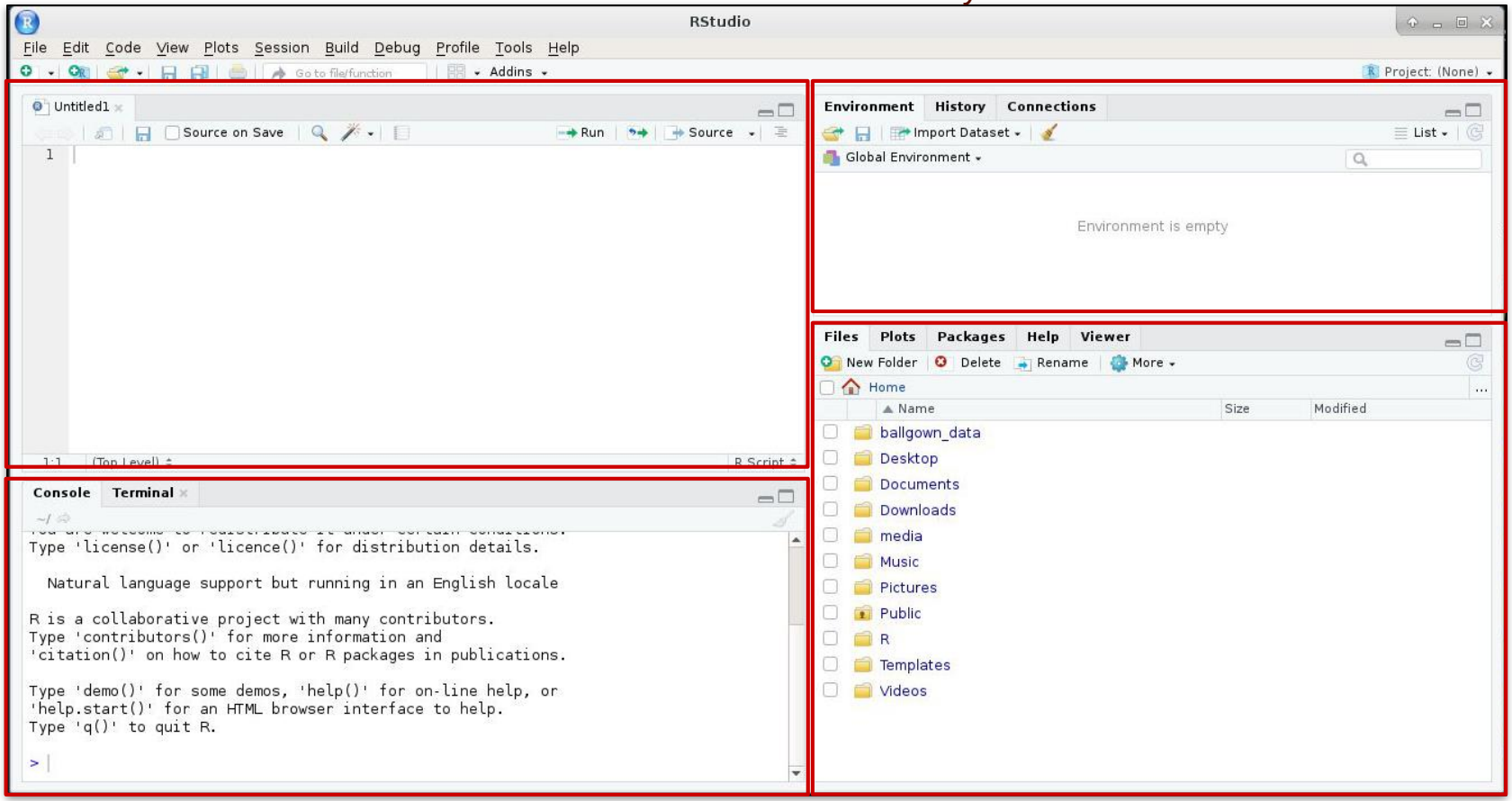
Overview of R

- R is a free statistical computing and graphing software
- Can be installed from their website https://www.r-project.org/
- R can be run in a few environments:
  - RStudio
  - Jupyter

Summary of RStudio

R scripts, R markdown, R notebooks

Summary of all the data loaded in Rstudio



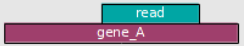
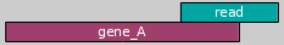
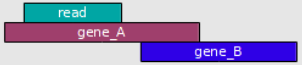
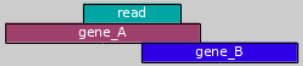
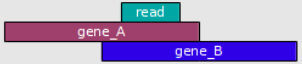
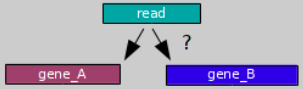
R console, Terminal

Directories, Plots, Packages...

Learning R in RStudio

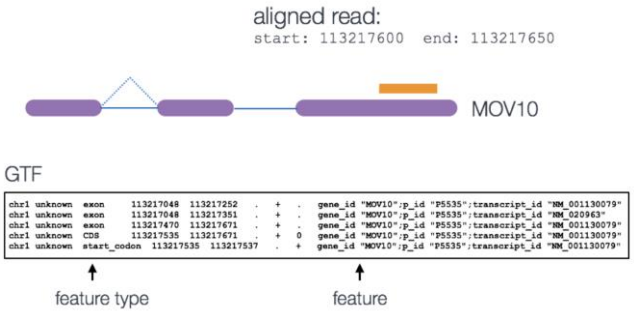
- Let us go over the `Learning_R.R` worksheet:
  - Introduction to R and R Markdown
  - Introduction to the iris dataset
  - Installing and loading libraries
    - tidyverse
  - Generating summary statistic in R
  - Making plots with ggplot2
  - Manipulating data.frames

`featureCounts` counts reads over features in R

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

There are several options in `featureCounts`

```
fc <- featureCounts(files=bam_file_list,
                    annot.ext=gtf,
                    isGTFAnnotationFile=TRUE,
                    GTF.featureType="exon",
                    GTF.attrType="gene_id",
                    useMetaFeatures=TRUE,
                    allowMultiOverlap=TRUE,
                    largestOverlap=TRUE,
                    countMultiMappingReads=TRUE,
                    isPairedEnd=TRUE,
                    strandSpecific=1,
                    nthreads=N)
```



https://hbctraining.github.io/Intro-to-rnaseq-hpc-O2/lessons/05_counting_reads.html

There are different ways to interact with R

R console

```
(base) cu-biot-14-10:~ rutendo$ R

R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin15.6.0 (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.

[Previously saved workspace restored]

>
```

Enter **R code** here

More interactive

Submit an R script as a job

```
#!/bin/bash

#SBATCH --job-name=feature_counts                          # Job name
#SBATCH --mail-type=ALL                    # Mail events (NONE, BEGIN, END, FAIL, ALL)
#SBATCH --mail-user=email@colorado.edu     # 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=/scratch/Users/rutendos/e_and_o/%x_%j.out
#SBATCH --error=/scratch/Users/rutendos/e_and_o/%x_%j.err

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

FEATURECOUNTS=/scratch/Users/rutendos/day6/featureCounts/scripts/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 **R script** here

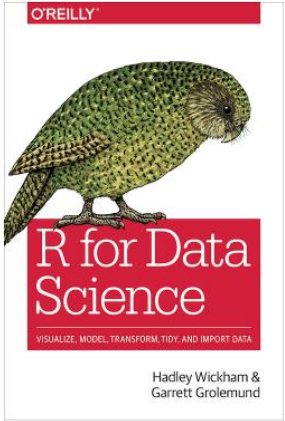
For more compute intensive scripts

Counting reads with `featureCounts`

- Follow `featureCounts` worksheet:
  - Open `R` and install `Rsubread`
  - Copy `d6_featureCounts.R` and `d6_featureCounts.sbatch` scripts from `/scratch/Shares/public/sread2022/scripts/day6` to your `/scratch/Users/<username>/`
  - Edit both scripts and execute the `sbatch` script

More resources for R

- ggplot2 website https://ggplot2.tidyverse.org/
- R-bloggers https://www.r-bloggers.com/
- Quick-R https://www.statmethods.net/
- R for Data Science (by Hadley Wickham & Garrett Grolemund) http://r4ds.had.co.nz/



Other tools for counting reads

Method	Number of reads	Number of fragments	Time (min)	Memory (MB)
featureCounts	4 385 354	4 796 948	1.0	16
SummarizeOverlaps (whole genome at once)	4 385 354	3 942 439	12.1	3400
SummarizeOverlaps (by chromosome)	4 385 354	3 942 439	41.7	661
htseq-count	4 385 207	4 769 913	22.7	101

`featureCounts` is faster and more efficient.

Liao et al. Bioinformatics 2014 doi:10.1093/bioinformatics/btt656

Homework

- Complete the `Learning_R_Additional_Practice.R`

This homework will go over most of the topics covered today, but on a different dataset. There will be more advanced questions that build on what was in the inclass session.

- Install `DESeq2`

This library takes in counts as input and performs differential gene expression analyses on the input features. You will be using this library in Day7. Install this on your local machine too.