Short Read Workshop Day 7
Counting Reads and Differential Expression

Rutendo Sigauke
2023

Day 7 Overview

- M&Ms
- featureCounts
- DESeq2

Goal of the Day

Find genes that are different between samples

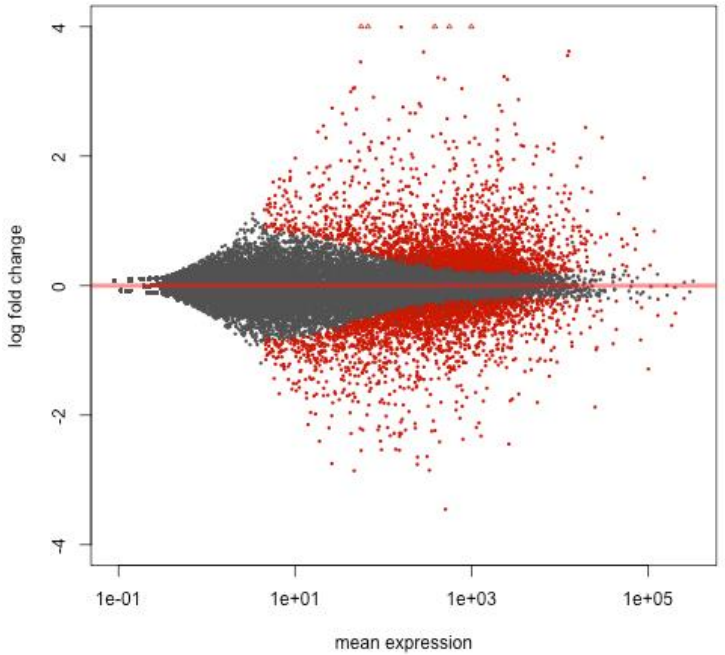
Map reads to reference genome

↓

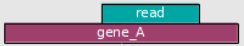
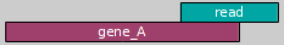
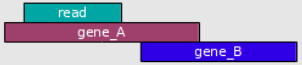
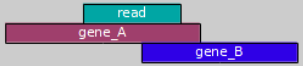
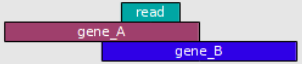
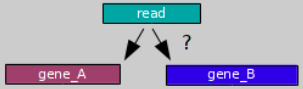
Count reads

↓

Perform differential gene expression analysis

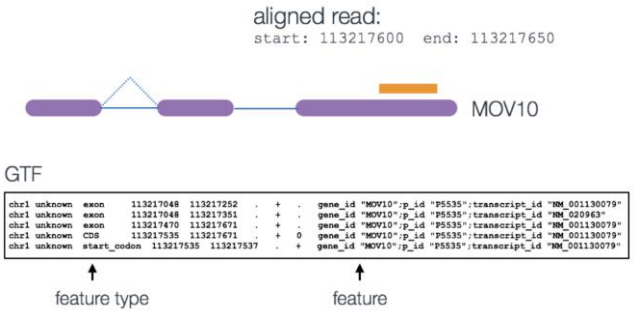


`featureCounts` counts reads over features in R

	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

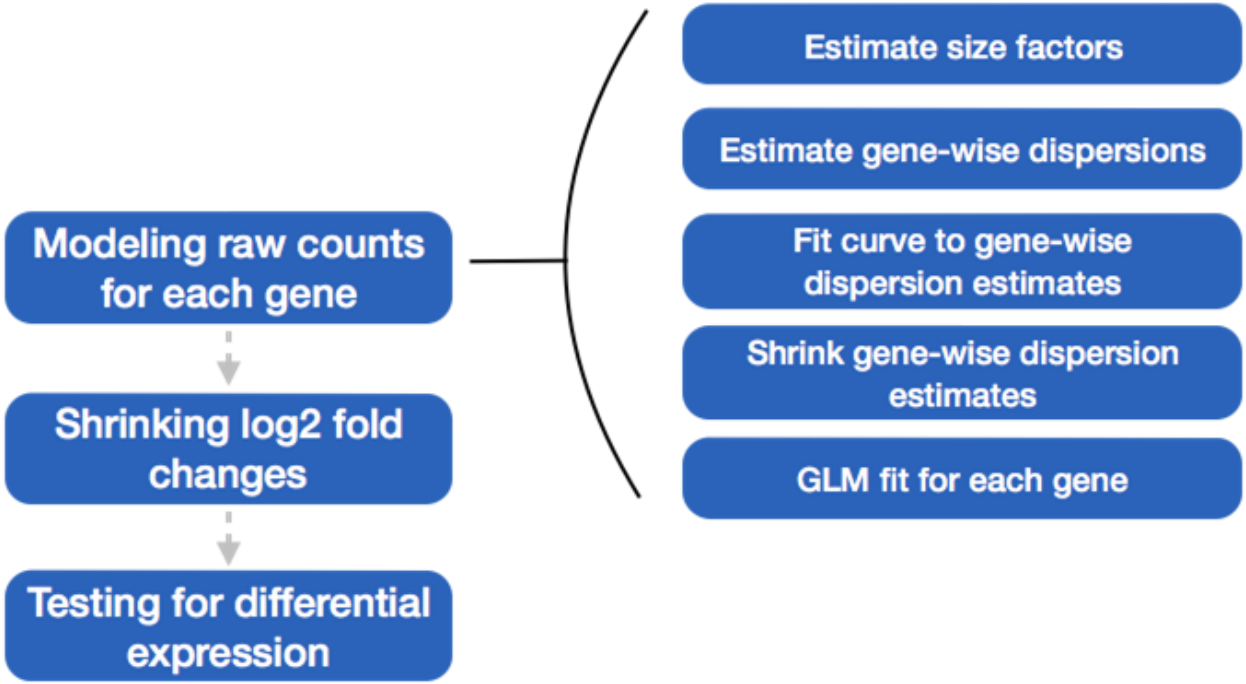
The are other tools for counting reads but featureCounts is more efficient

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.

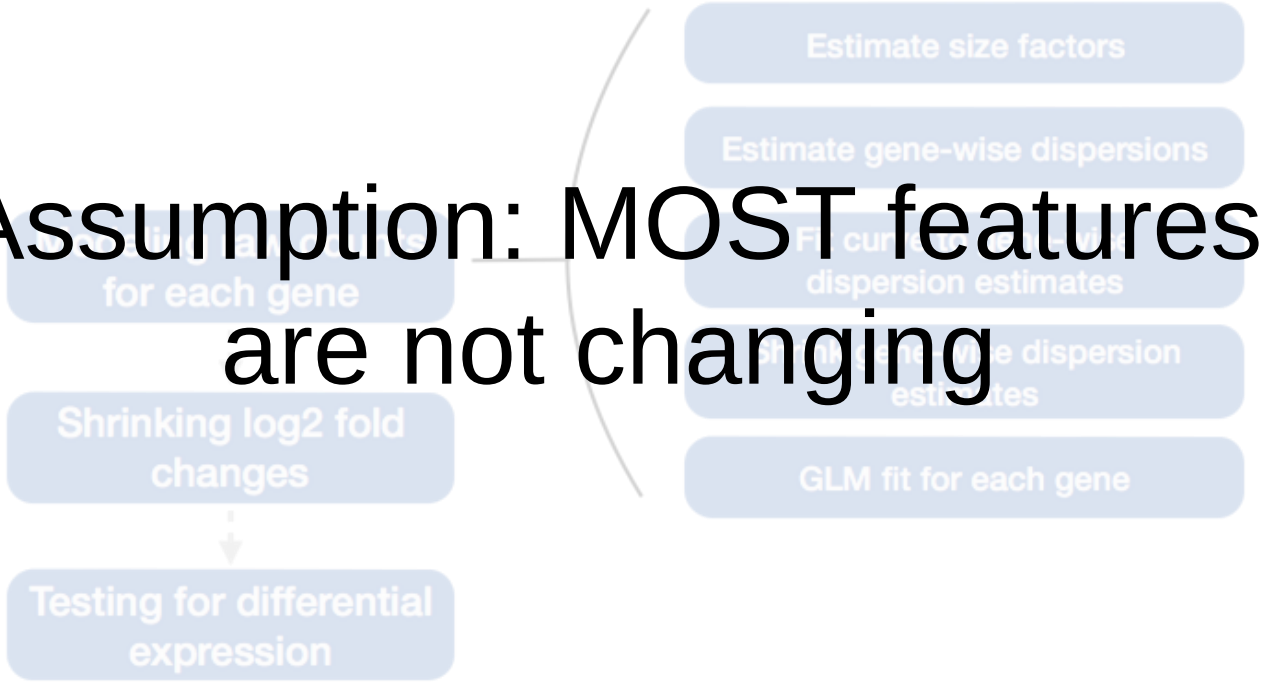
Liao et al. Bioinformatics 2014 doi:10.1093/bioinformatics/btt656

DESeq2 Recap



https://hbctraining.github.io/DGE_workshop_salmon/lessons/04_DGE_DESeq2_analysis.html

DESeq2 Recap



Assumption: MOST features are not changing

https://hbctraining.github.io/DGE_workshop_salmon/lessons/04_DGE_DESeq2_analysis.html

Counting reads with `featureCounts`

- Follow featureCounts worksheet:
  - Open `R` and install `Rsubread`
  - Get `d7_featureCounts.R` and `d7_featureCounts.sbatch` scripts
  - Edit both scripts and execute the `sbatch` script

Challenge Question

- What feature would you used to count reads for RNA-seq?

A. Gene

B. Exon

C. Transcripts

Challenge Question

- What feature would you used to count reads for RNA-seq?

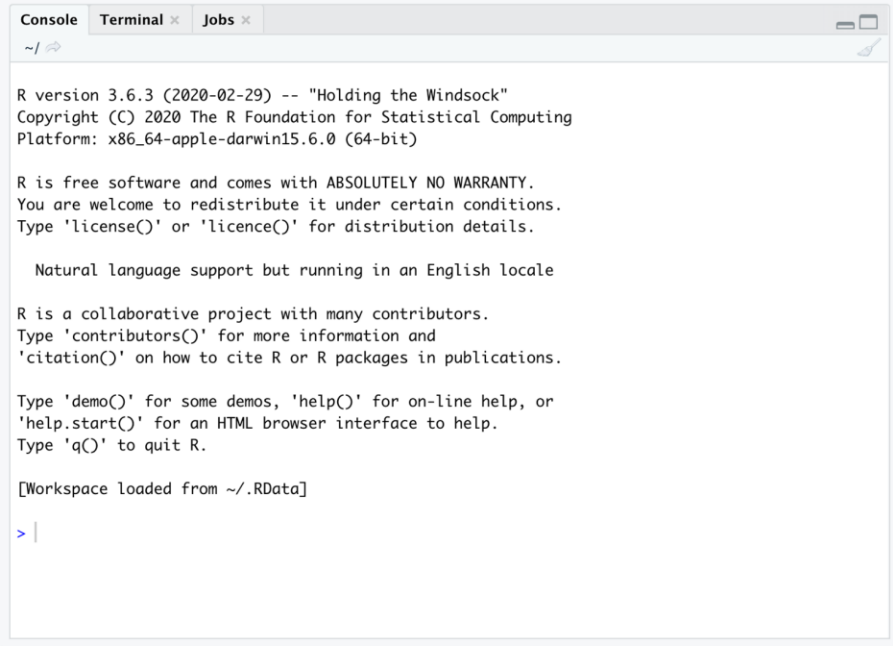
A. Gene

B. Exon ✓

C. Transcript

Run DESeq2…

- Follow DESeq2 worksheet
  - This will be run in an R console in RStudio



Challenge Question

- How would you run DESeq2 on the supercomputer?

Challenge Question

- How would you run DESeq2 on the supercomputer?
  - Install DESeq2 in your R packages directory
  - Make a conditions table that matches your count table
  - Run the R script through an sbatch script

Homework

- Explore DE genes with a heatmap
- Run DESeq2 to explore differential expression with nicotine and caffeine treatments