BINF 5003: Data Mining, Modeling, and Biostatistics

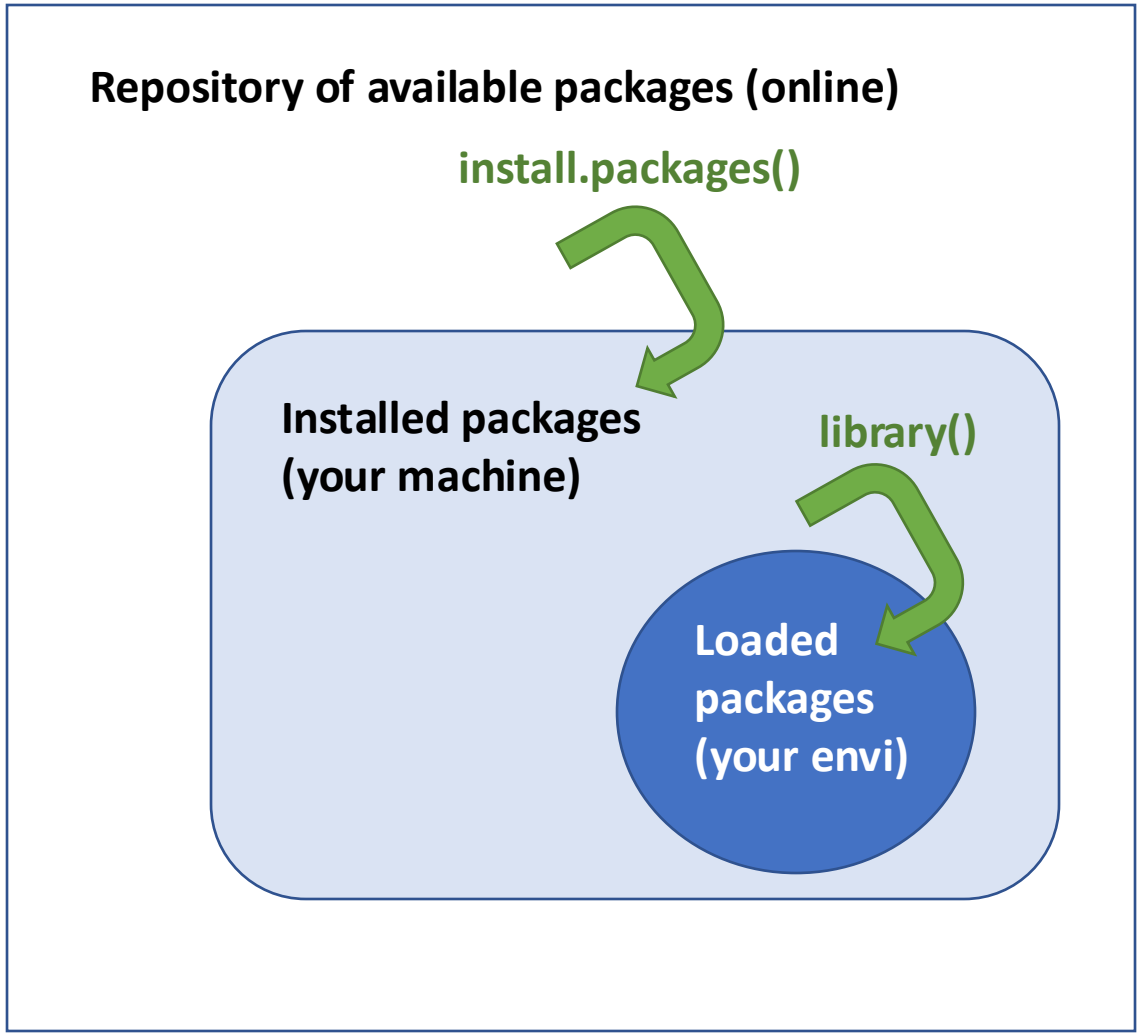
Week 8

Module 5 – Differential Gene Expression

Overview

- New repository for packages! Bioconductor
  - Compare and contrast with CRAN
- New repository for genomic data! GEO
  - Gene Expression Omnibus (GEO) database
- New data structure: DESeq & expressionset
- High dimensionality in data

Refresher: Packages are collections of functions



The Comprehensive R Archive Network

- Maintained by just a handful of people
- Maintains all the versions of R and as of Aug 2023, 19905 packages
- Install with `install.packages()`

Dependencies

- Some packages, such as many of the tidyverse ones, depend on functions developed in other packages

- If a package as dependencies, that means it depends on having these other packages installed
  - Other packages need to be installed before the current one can

Depreciated/Deprecated

- “diminishing in value over time”
- The function is about to be retired
- Warning to find alternatives
  - No longer supported
  - May be removed in a future update

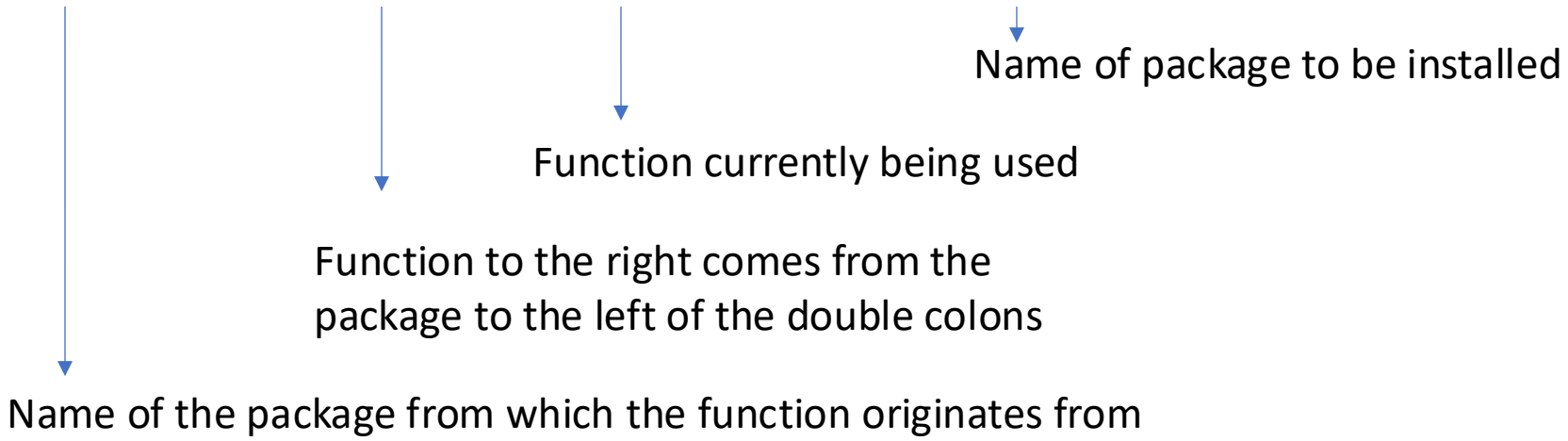
Bioconductor

- Repository specializing in “rigorous and reproducible analysis of data from current and emerging biological assays”
- All software using R statistical programming language
- Annual conference and educational training for support

Package structure

- Need to install Bioconductor/BiocManager first
  - Install the function to install other Bioconductor packages

```
BiocManager::install("oligo")
```



Installing Bioconductor packages

Install *Bioconductor* Packages

To install core packages, type the following in an R command window:

```
if (!require("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
BiocManager::install()
```

Install specific packages, e.g., "GenomicFeatures" and "AnnotationDbi", with

```
BiocManager::install(c("GenomicFeatures", "AnnotationDbi"))
```

Should you always update packages to the most recent version?

Should you always update packages to the most recent version?

- No!

- Functions may change
  - Cause your code to break
  - Cause your code to add differently without your knowledge

- Update after you’re done a project

- Make sure to keep track of the versions for publications

Record keeping

- So helpful when you need to write your Methods section!
- Helpful for facilitating collaborations or troubleshooting issues

```
> sessionInfo()
R version 4.1.1 (2021-08-10)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 19045)

Matrix products: default

locale:
[1] LC_COLLATE=English_Canada.1252
[2] LC_CTYPE=English_Canada.1252
[3] LC_MONETARY=English_Canada.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_Canada.1252

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

other attached packages:
 [1] tidyselect_1.2.0 datasauRus_0.1.6 lubridate_1.9.2
 [4] forcats_1.0.0    stringr_1.5.0    dplyr_1.1.2
 [7] purrr_1.0.1      readr_2.1.4      tidyr_1.3.0
[10] tibble_3.2.1     ggplot2_3.4.2    tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] pillar_1.9.0          compiler_4.1.1
 [3] BiocManager_1.30.21.1 tools_4.1.1
```

Gene Expression Omnibus (GEO) database

- “Public functional genomics data repository supporting MIAME-compliant data submission”
- Mostly microarray and genome sequencing data
- Light built in tools to peek at the data in browser
- Many tools developed to integrate with R

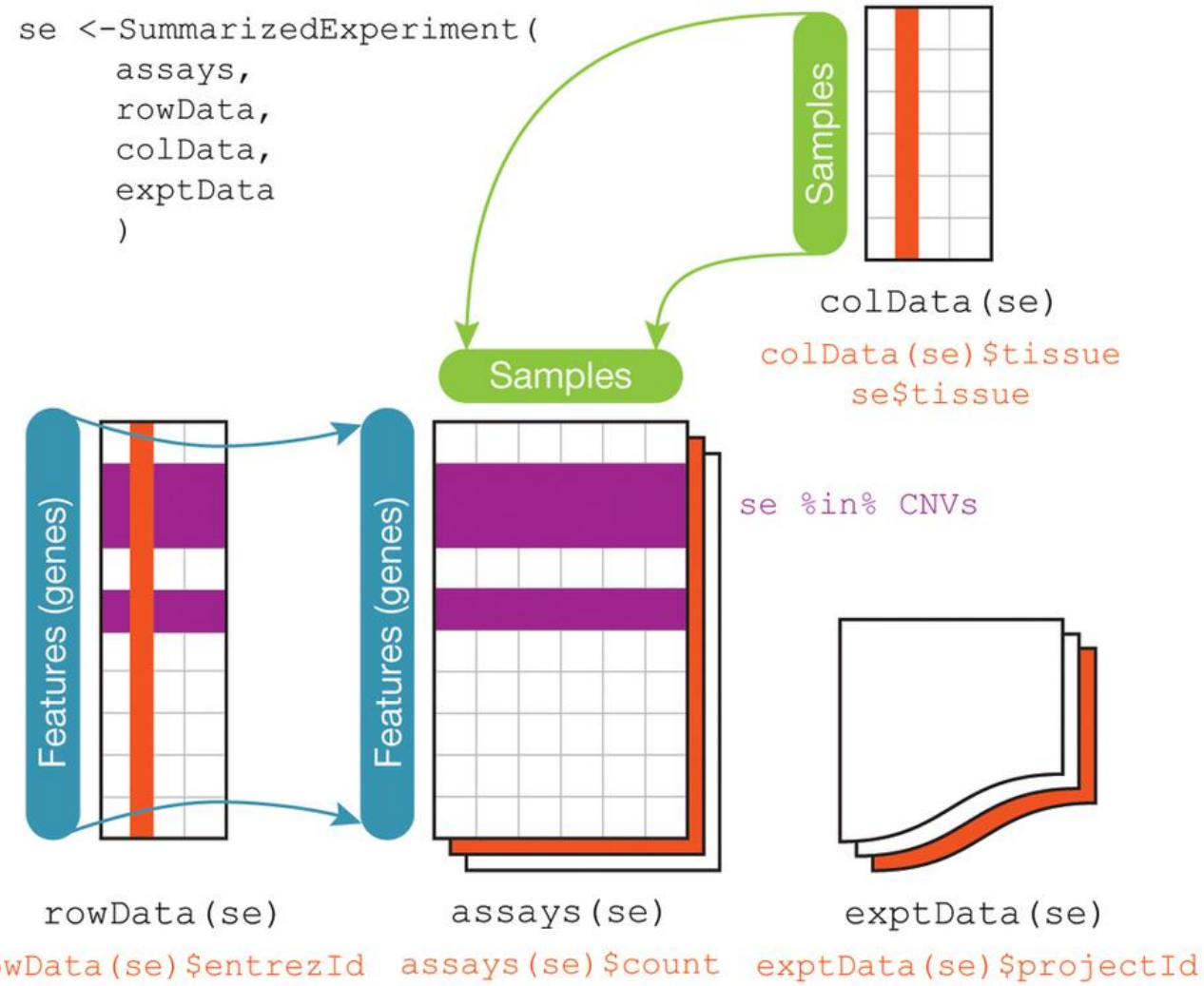
Series GSE183071 Query DataSets for GSE183071

Status	Public on Dec 31, 2021
Title	A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity
Organism	Homo sapiens
Experiment type	Expression profiling by array
Summary	COVID-19 symptoms range from mild to severe illness; the cause for this differential response to infection remains unknown. Unraveling the immune mechanisms acting at different levels of the colonization process might be key to understand these differences. We carried out a multi-tissue (nasal, buccal and blood) gene expression analysis of immune-related genes from patients affected by different COVID-19 severities, and healthy controls.
Overall design	Blood, nasal epithelium, and saliva samples were sampled from patients and controls at the same timepoints. Blood samples (2.5ml) were collected in PAXgene tubes (BD). RNA was isolated using PAXgene blood miRNA extraction kit (Qiagen) following manufacturer recommendations. Saliva and nasal epithelium samples were collected in Oragene CP-190 kit (DNA Genotek). Patients and healthy controls were self-reported as of South-European ancestry; and cases were classified at the time of sample collection by severity illness: severe (ICU admission), moderate (non-ICU but admitted to hospital) and mild (domiciliary lockdown patients with mild symptoms or asymptomatic). Sample numbers are distributed as follows: (i) whole blood from 41 patients and 13 controls, (ii) nasal epithelium from 38 patients and 11 controls, and (iii) saliva from 41 patients and 12 controls. The complete tissue sample set could be collected for 27 out of the 52 patients. Immunological gene expression patterns of different tissues were evaluated using a SPRINT nCounter system (NanoString Technologies) and the Human Immunology v2 Panel.
Contributor(s)	Gómez-Carballa A
Citation(s)	Gómez-Carballa A, Rivero-Calle I, Pardo-Seco J, Gómez-Rial J et al. A multi-tissue study of immune gene expression profiling highlights the key role of the nasal epithelium in COVID-19 severity. *Environ Res* 2022 Jul;210:112890. PMID: 35202626

ExpressionSet object

- Complex, linked data types

- Expression table of features x samples is central
- Linked to
  - featureData
  - phenotypeData
- Deleting a row in the expression set will also delete the corresponding row in the featureData



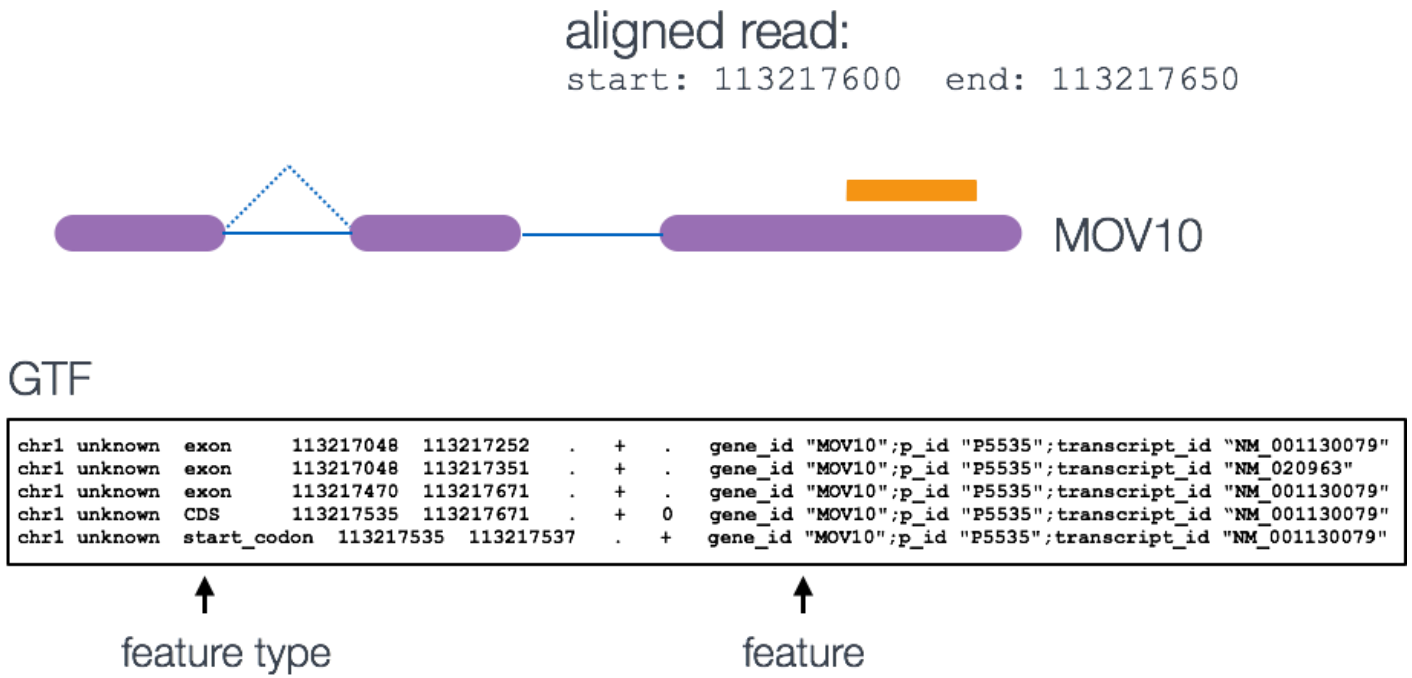
Differential gene expression

- Identify differences in the transcriptome (gene expression) across samples
  - E.g., differences between multiple biological conditions (e.g. drug treated vs. untreated samples)

- Many possible tools, depending on the type of data & your questions
  - DESeq2 (RNA-seq)
  - limma (microarray)
  - edgeR (RNA-seq)

Count Data

- Counting reads associated with genes



Count Data

- Counting reads associated with genes

Each column is a sample

Each row is a gene

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

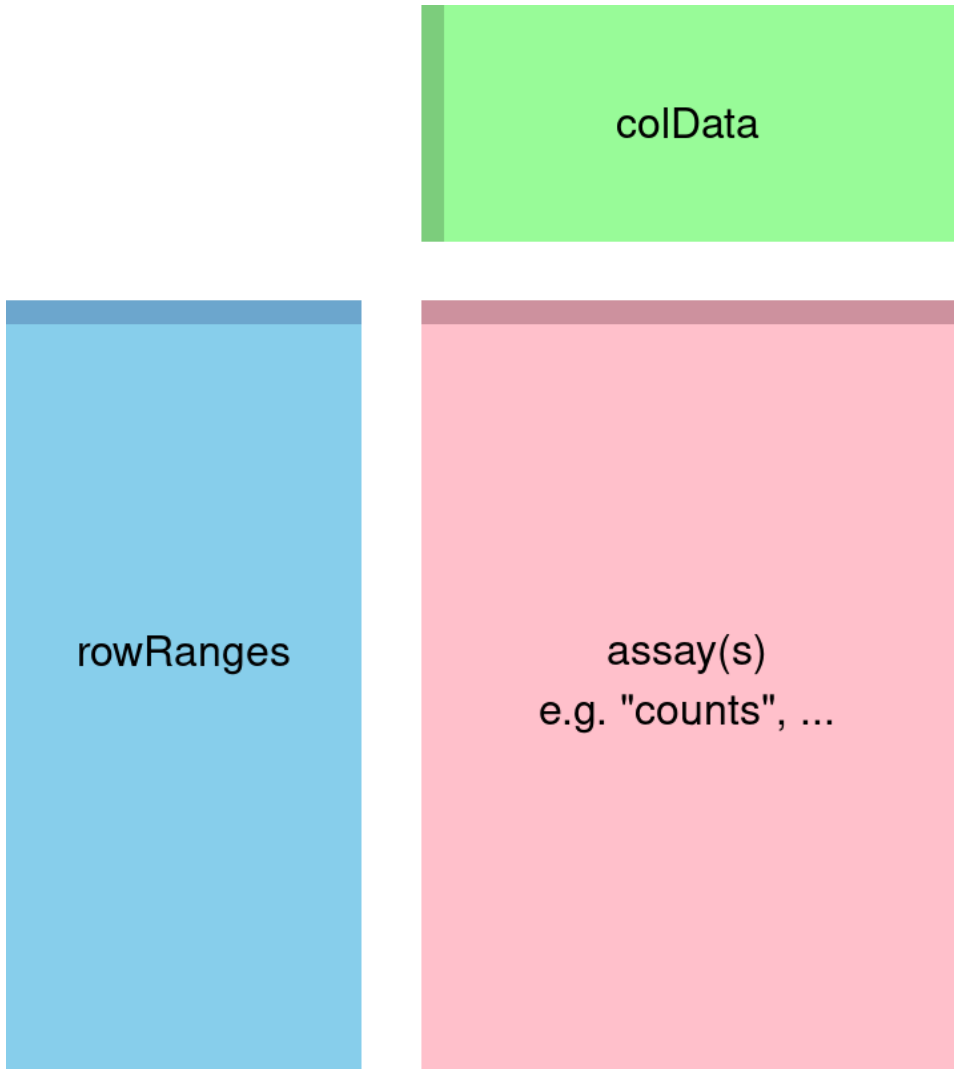
DESeq

- DESeq2 is used to analyze read count data and identify differentially expressed genes

- DESeq2 uses an object called the DESeqDataSet, and requires a formula

DESeqDataset

- DESeq2 must take this specialized dataset
  - Similar to all other gene expression data
- Sometimes, your dataset might be of a different class (e.g., *SummarizedExperiment*)
  - Use function DESeqDataSet() to coerce



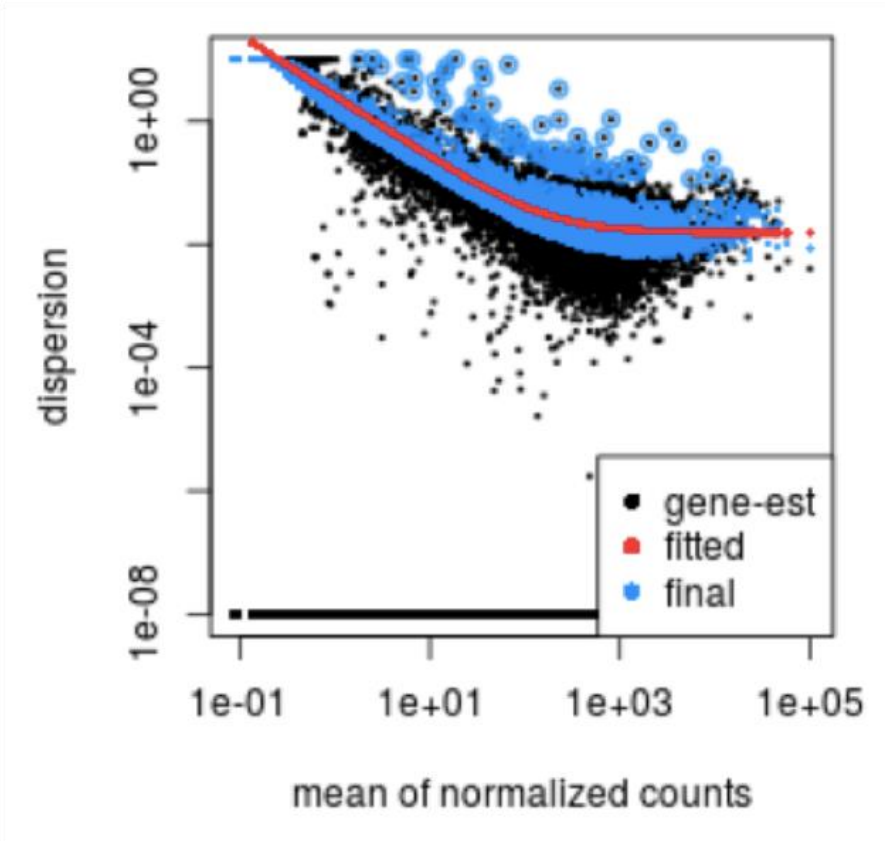
DESeq Formula

- Formula or design: what are we testing for and what are we controlling?

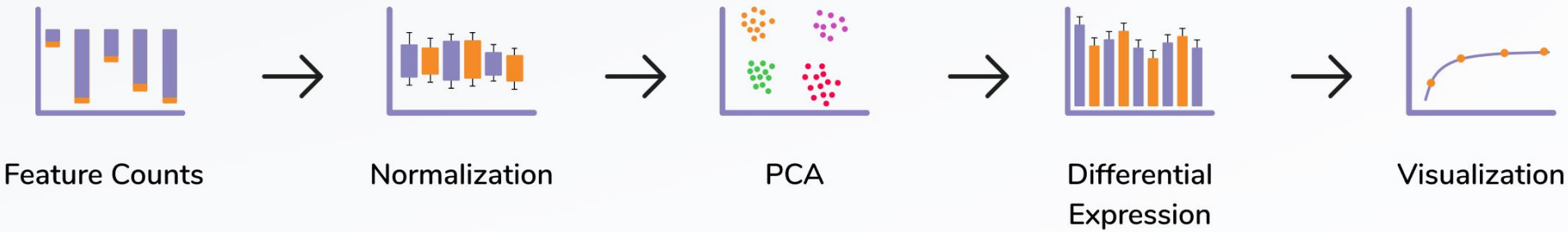
- design: ~ sex + treatment
  - “expression differences between treatment while taking into account the sex of the patient"

Differential gene expression analysis workflow

1. Estimate size factors
   - sizeFactors()
2. Estimate dispersion
3. Fit curve to gene-wise dispersion estimates
4. Shrink gene-wide dispersion to predicted estimates from curve
   - plotDispEsts()
5. Fit model



Bioinformatics Workflow



Wrap-up

- Bioconductor hosts R packages for analyzing biological assays
  - Different way to install, but function the same after you load it in
- Gene Expression Omnibus (GEO) database is home a large collection of published microarray and sequencing data
  - Facilitating open access science
- expressionSet objects are linked tables with the actual data and tables elaborating on the samples and features
- Count data aligns reads with genes to get a measure of gene expression
- DESeq2 requires a DESeqDataSet and a formula to estimate gene expression of RNA-Seq data using a modelling technique
- A bioinformatics workflow typically requires normalization and visualization of differential gene expression