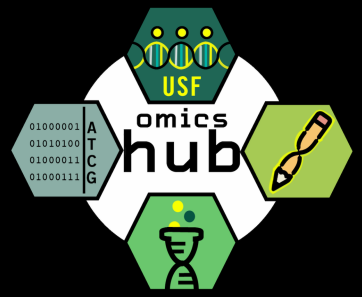
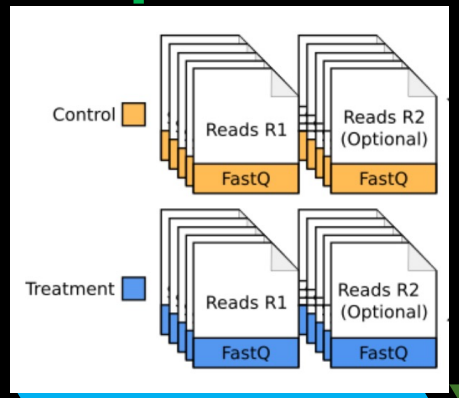
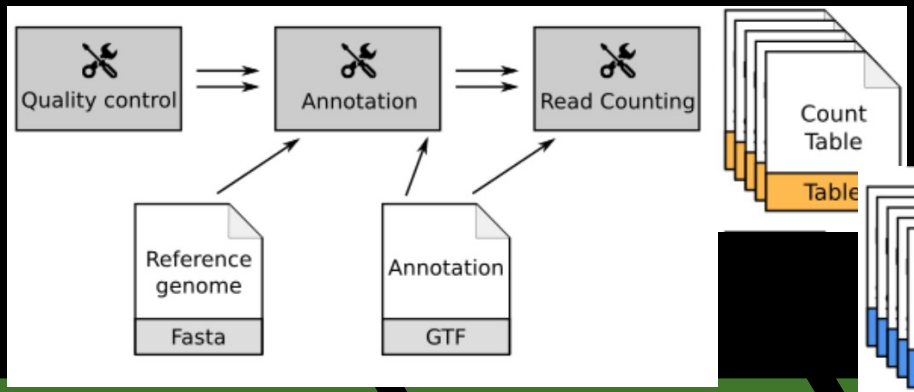
R and RStudio basics for biologists

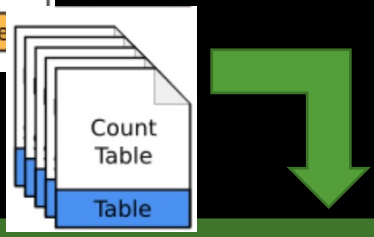
Jenna Oberstaller, PhD



RNAseq skill-set overview

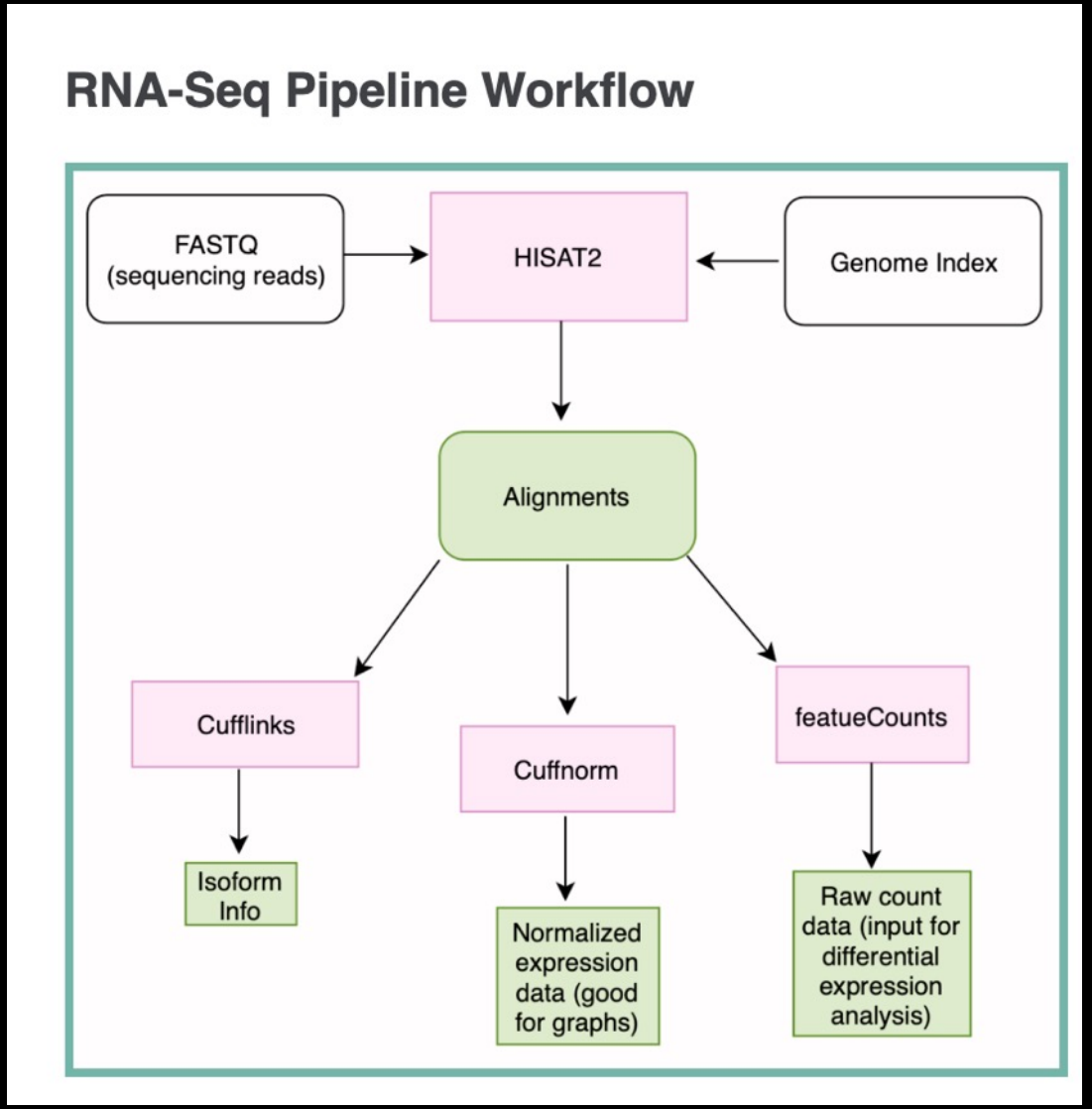






	STEP 0	STEP 1	STEP 2	STEP 3	STEP 4
	Next-Gen sequencing at the bench: Illumina workflow	FastQ files: Quality assessment, control and processing	Mapping your reads to a reference genome	Generating read-counts tables	Differential expression analysis
Tools	BaseSpace	FastQC MultiQC	hisat2 samtools	featureCounts	deSeq2
Platform	Web	UNIX (student cluster)	UNIX (student cluster)	UNIX (student cluster)	Rstudio (local)
Output	fastq.gz	fastq.gz	.bam .sam	.txt (table)	.txt (table)

REVIEW


RNA-Seq Pipeline Workflow
FASTQ (sequencing reads)
HISAT2
Genome Index
Alignments
Cufflinks
Cuffnorm
featueCounts
Isoform Info
Normalized expression data (good for graphs)
Raw count data (input for differential expression analysis)

What is R?

An open-source statistical programming language

- How is it different from what we did yesterday?

We used UNIX on the cluster to process our data down to transcript-counts; now we need stats and plots to analyze it!

- Why do we use it?

Free; particularly great for data-analysis and powerful visualizations; lots of ‘omics tools available

- How do we use it?

with RStudio: an **I**ntegrated **D**evelopment **E**nvironment for R

R: Engine

R: Do not open this

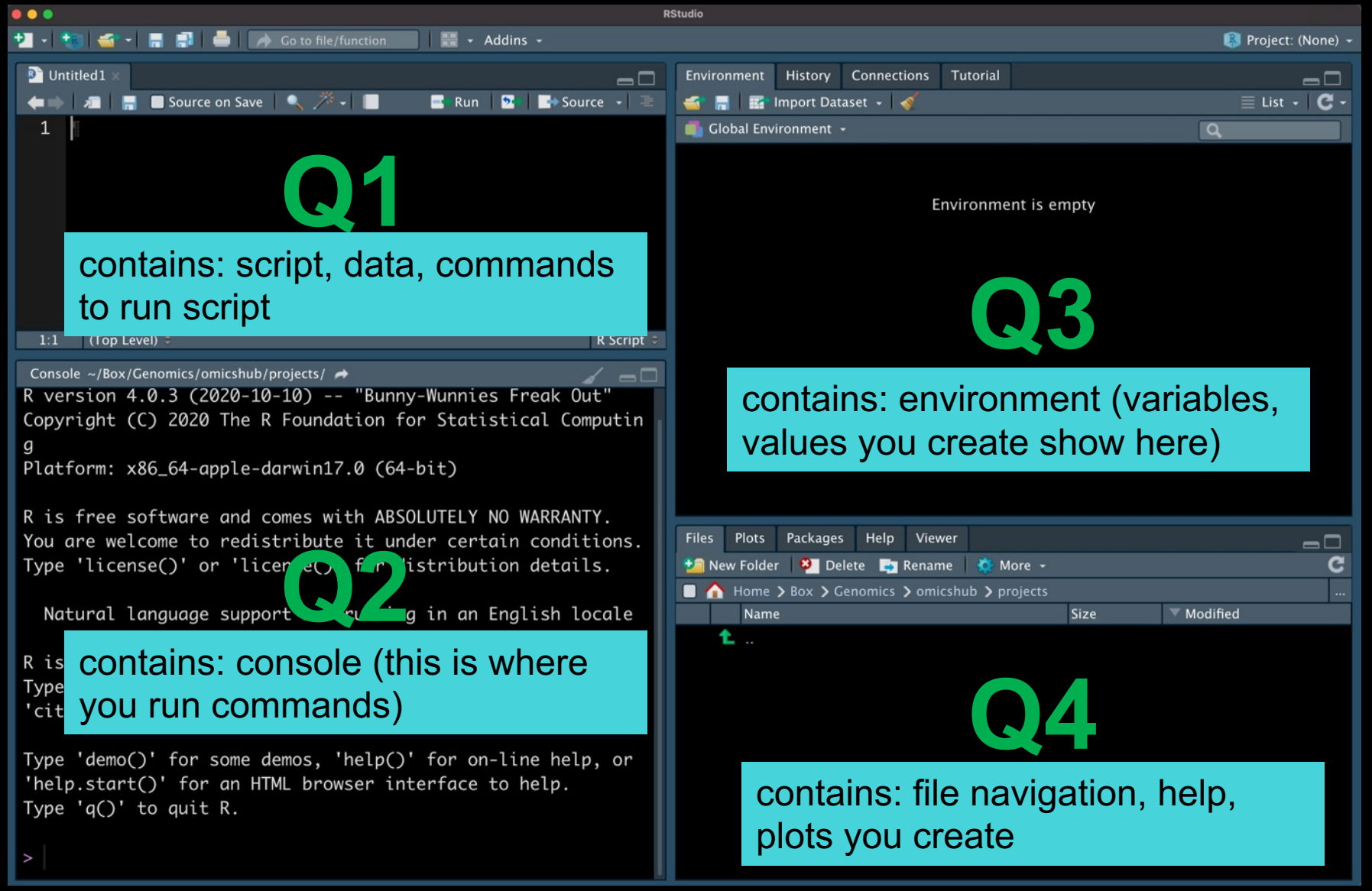
RStudio: Dashboard

RStudio: Open this

R vs. RStudio

Credit: ModernDive

Looking around Rstudio: Basics



Packages

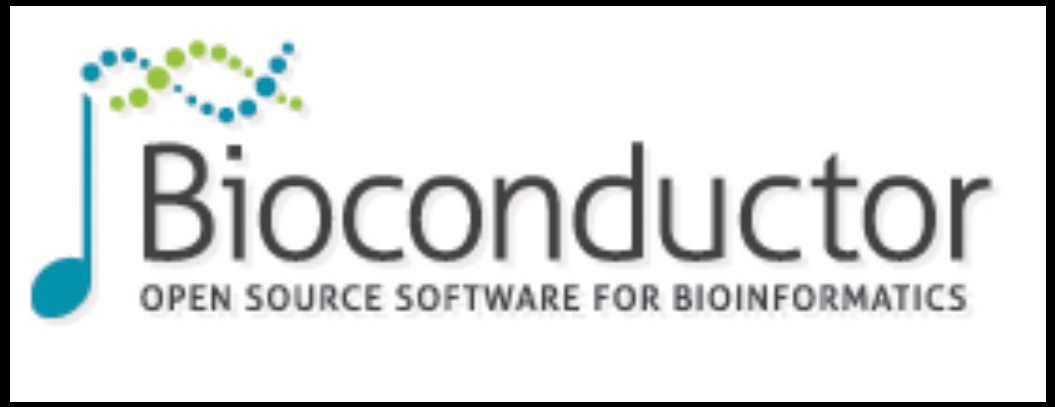
- “Code a generous person has written, tested, bundled and gave away FOR FREE because they are so terribly nice”

- Mostly designed to solve a specific problem
  - pull together functions to solve that problem
    - e.g.: general packages for “data-wrangling”, or visualization, etc.
    - or more specific packages designed already for someone trying to do exactly what you’re trying to do, such as differential gene-expression analysis

- You could write all your own functions, and hopefully one day you will
  - BUT TODAY IS NOT THAT DAY

Bioconductor

A repository of R packages for omics data analysis



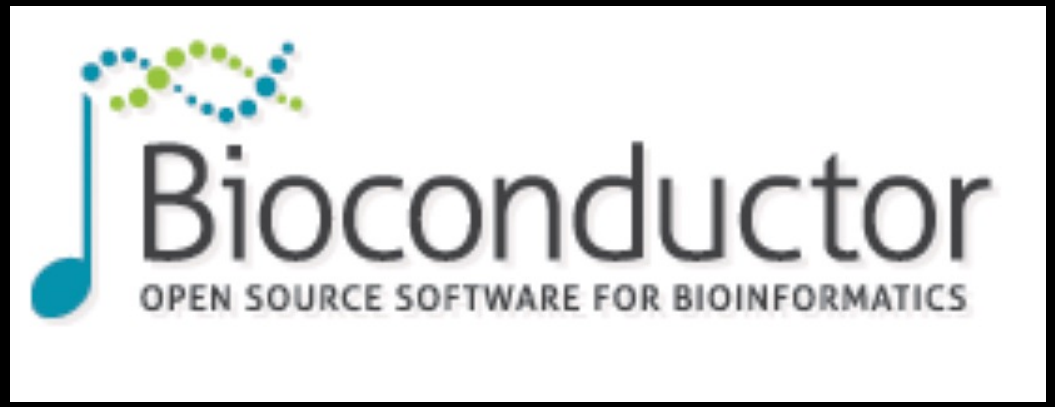
https://www.bioconductor.org/about/

MISSION:

- To promote the statistical analysis and comprehension of current and emerging high-throughput biological assays.

Most *Bioconductor* components are distributed as R packages. The functional scope of Bioconductor packages includes the analysis of DNA microarray, sequence, flow, SNP, and other data.

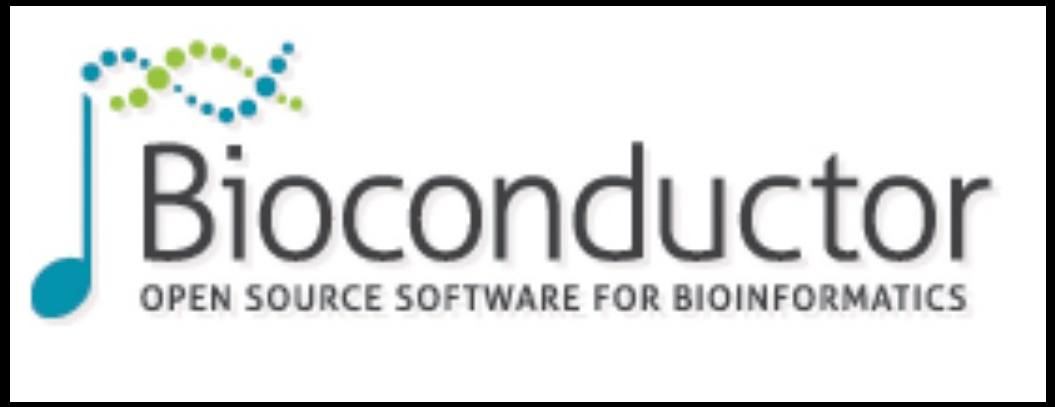
- Bioconductor is committed to open source, collaborative, distributed software development and literate, reproducible research.



https://www.bioconductor.org/about/

PROJECT GOALS (highlights)

- To provide widespread access to a broad range of **powerful statistical and graphical methods** for the analysis of genomic data.
- To further scientific understanding by producing **high-quality documentation** and **reproducible research.**
- To train researchers on computational and statistical methods for the analysis of genomic data.



https://www.bioconductor.org/about/

PROJECT GOALS (highlights)

- To provide widespread access to a broad range of **powerful statistical and graphical methods** for the analysis of genomic data.

- Each Bioconductor package contains one or more vignettes to show you how to use them
    - SOME are useful

Further resources

- More DESeq2:
  - https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html
  - https://bioc.ism.ac.jp/packages/2.14/bioc/vignettes/DESeq2/inst/doc/beginner.pdf

- Further relevant tutorials
  - de-gene analysis
    - https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/ref-based/tutorial.html#functional-enrichment-analysis-of-the-de-genes
  - Functional enrichment
    - https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/goenrichment/tutorial.html