Bridging the Bench-Machine Learning Gap

Dr. Emily A. Beck

Dr. Jake Searcy

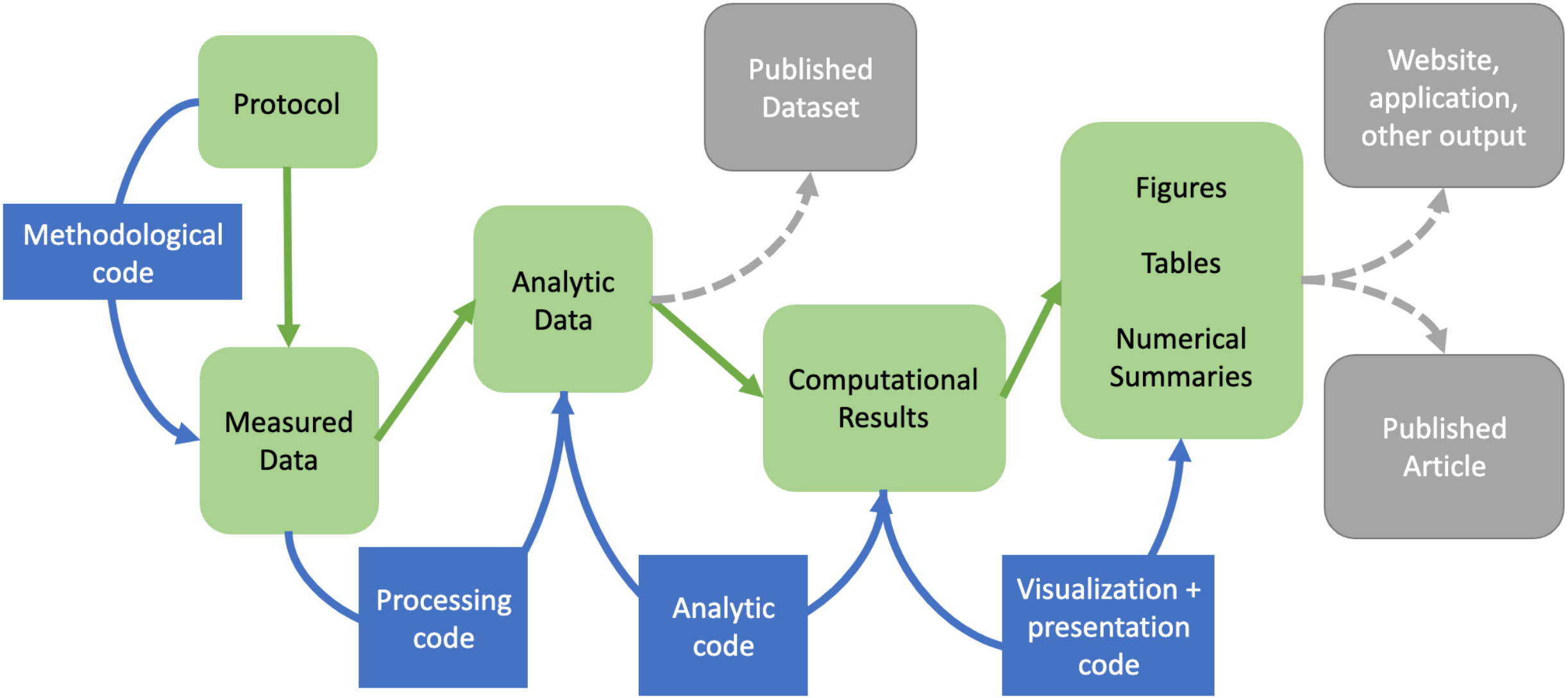
Images from Elin Videvall-Evomics

Learning Objectives

- Become familiar with the R environment
- Learn to upload and use existing packages
- Identify overlaps between linux and R
- Learn to ingest and manipulate files
- Learn some basic plotting skills

Research workflow

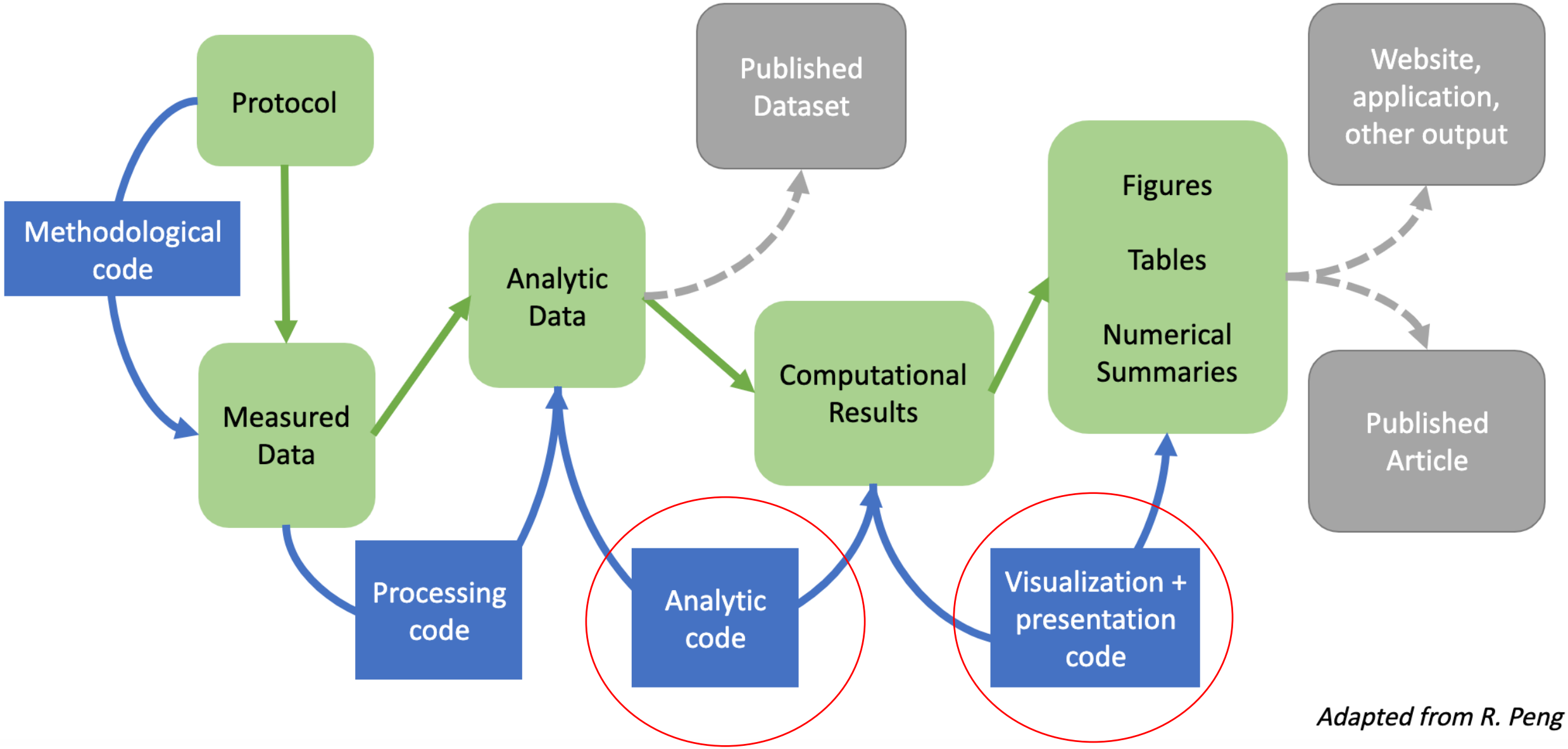
R can be used for many stages of the workflow!



Adapted from R. Peng

Research workflow

R can be used for many stages of the workflow!



Adapted from R. Peng

What is R and why is it so popular with scientists?

- Open source Free!
- Great for data management and manipulation
- Well established packages for statistics and data visualization
- Excellent online resources
- High-level interpreted language (Easy to read)

esources

The R Project: http://www.r-project.org

Quick R Homepage: http://www.statmethods.net

Bioconductor: http://www.bioconductor.org

Longer Intro to R: http://cran.r-project.org/doc/manuals/R-intro.html

Google is your best friend! Want to do something new in R? I bet someone has done it and made their code available.

Packages for biologists:

Develop, support, and disseminate free open source R software that facilitates reproducible data

Supported by Rstudio, a great resource for apps or for building and hosting your own apps Often in GUI form and are therefore popular but sometimes less customizable.

A network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R

Packages for biologists:

2,140 packages currently available

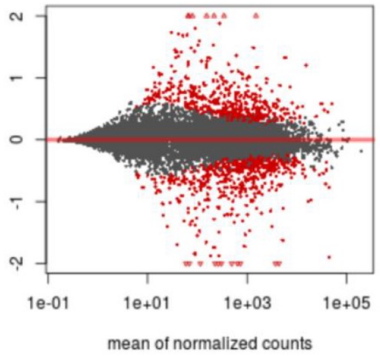
https://bioconductor.org/packages/release/bioc/

- Variant detection: sequence analysis, PolyPhen database
- Annotation: pathway analysis, BioMart, GO, KEGG, NCBI and many others
- High-throughput assays: flow cytometry, mass spec, proteomics
- Transcription factor binding, differential gene expression analysis (DEseq2)
- Methylation and chromatin accessibility screening (ATAC-seq)

Packages for genomics:

Differential analysis of count data – the DESeq2 package

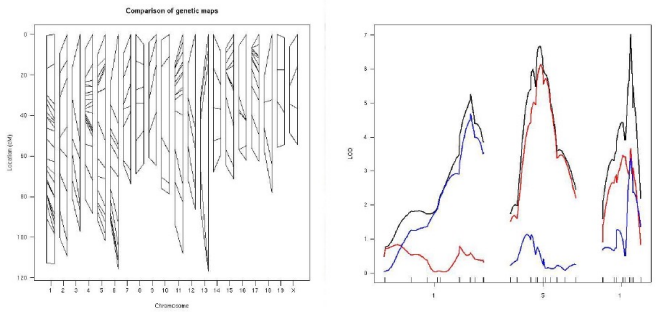
Michael I. Love[1], Simon Anders[2], and Wolfgang Huber[3]



R/qtl: A QTL mapping environment

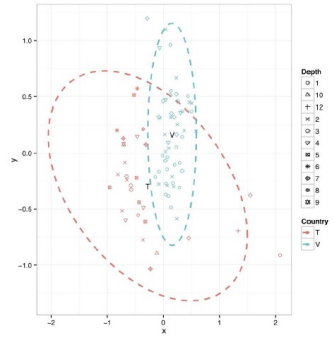
Software for mapping quantitative trait loci in experimental crosses

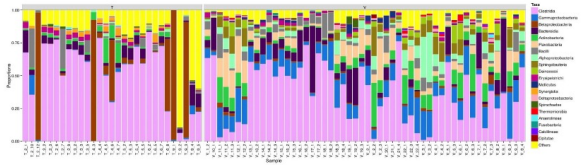
www.rqtl.org



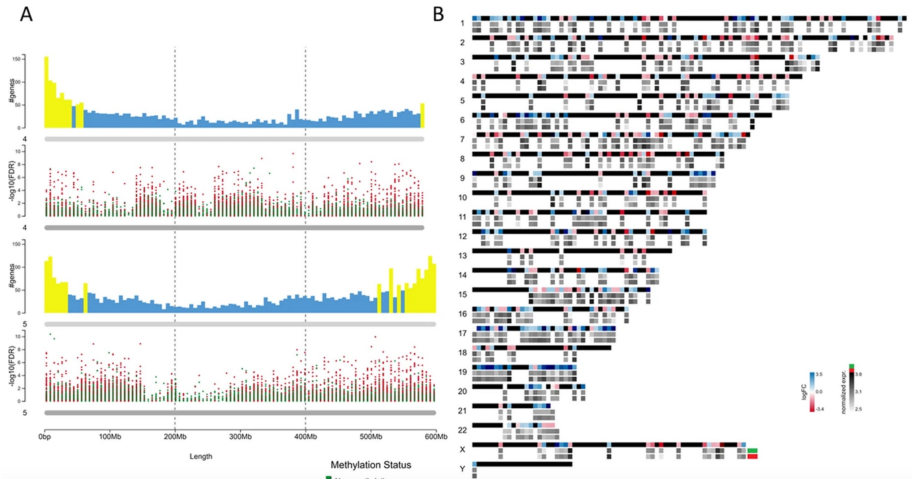
Multivariate Analysis of Ecological Communities in R: vegan tutorial

Jari Oksanen

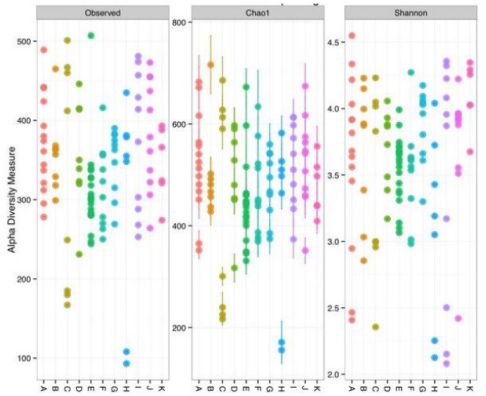


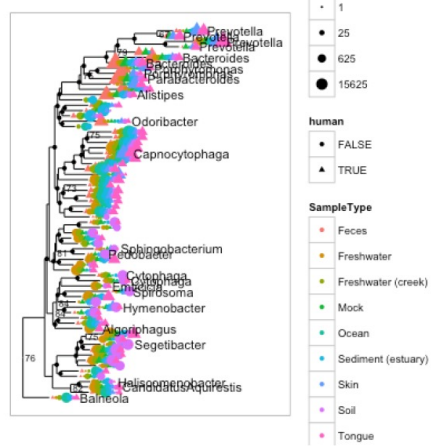


chromomap



phyloseq





http://joey711.github.io/phyloseq/

Packages for biologists:

https://cran.r-project.org/web/packages/available_packages_by_name.html

Genomics and statistics, but also a lot of field specific tools in immunology, environmental biology, ecology etc.

- Ordination
- Cluster Analysis
- Population Dynamics
- Spatial Data Analysis
- Ancestral State recognition
- Phylogenetic Inference
- Association studies
- Quantitative Genetics

Packages for biologists:

iDEP.951 | Load Data | Pre-Process | Heatmap | k-Means | PCA | DEG1 | DEG2 | Pathway | Genome | Bicluster | Network | R

Click here to load demo data

and just click the tabs for some magic!

Reset

1. Optional: Select or search for your species.

Best match | Info

2. Choose data type

- (•) Read counts data (recommended)
- ( ) Normalized expression values (RNA-seq FPKM, microarray, etc.)
- ( ) Fold-changes and corrected P values from CuffDiff or any other program

3. Upload expression data (CSV or text)

Browse... | No file selected

Ready to load data files.

Trusted charities providing Aid in Ukraine selected by CharityWatch, charities and individuals verified by GoFundMe.

April 25, 2022: Gene ID conversion is much faster now, even when species has to be guessed. So is the DEG2 tab.

April 24, 2022: Add a tab for visualizing the fold-change of all genes in all KEGG diagrams across all comparisons!

Feb. 11, 2022: Like iDEP but your genome is not covered? Customized iDEP is now available. Its database includes several custom genomes requested by users. To request to add new species/genome, fill in this Form.

Email Jenny for questions. Dr. Ge is notorisly slow in responding to emails.

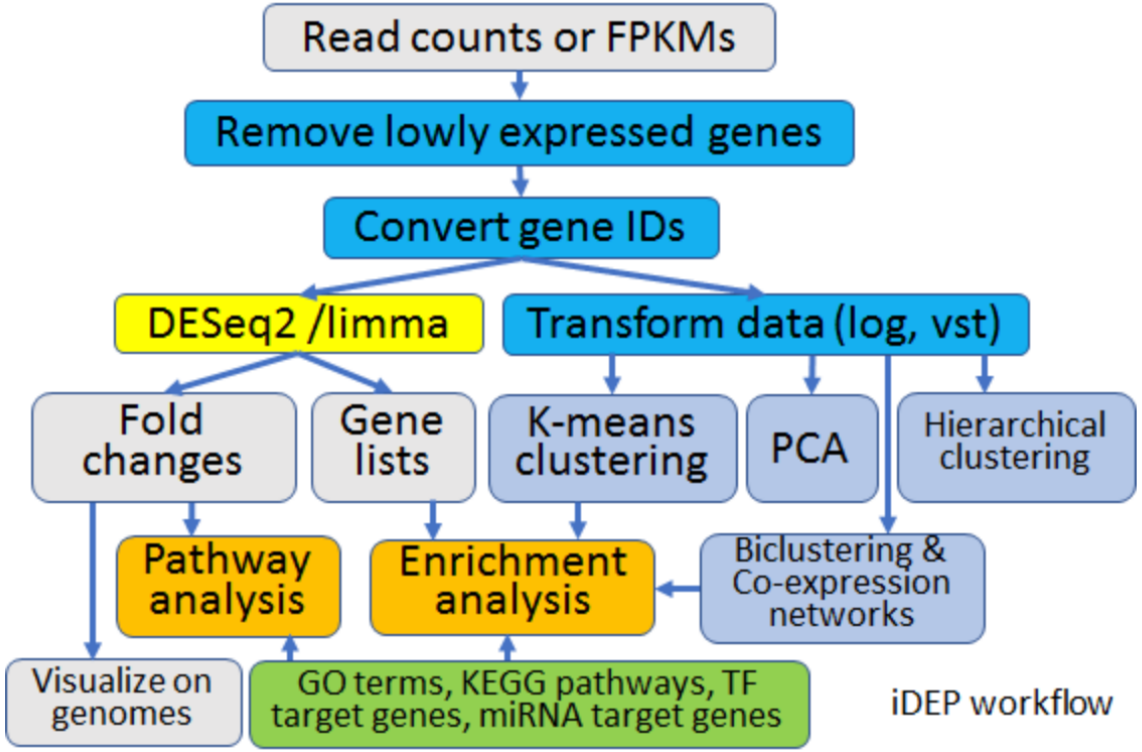
If it is slow, restart from a new browser window (not a new tab). You will be assigned to a new worker computer.

iDEP has not been thoroughly tested. Please let us know if you find any issue/bug.

iDEP: Integrated Differential Expression and Pathway analysis

Packages for biologists:

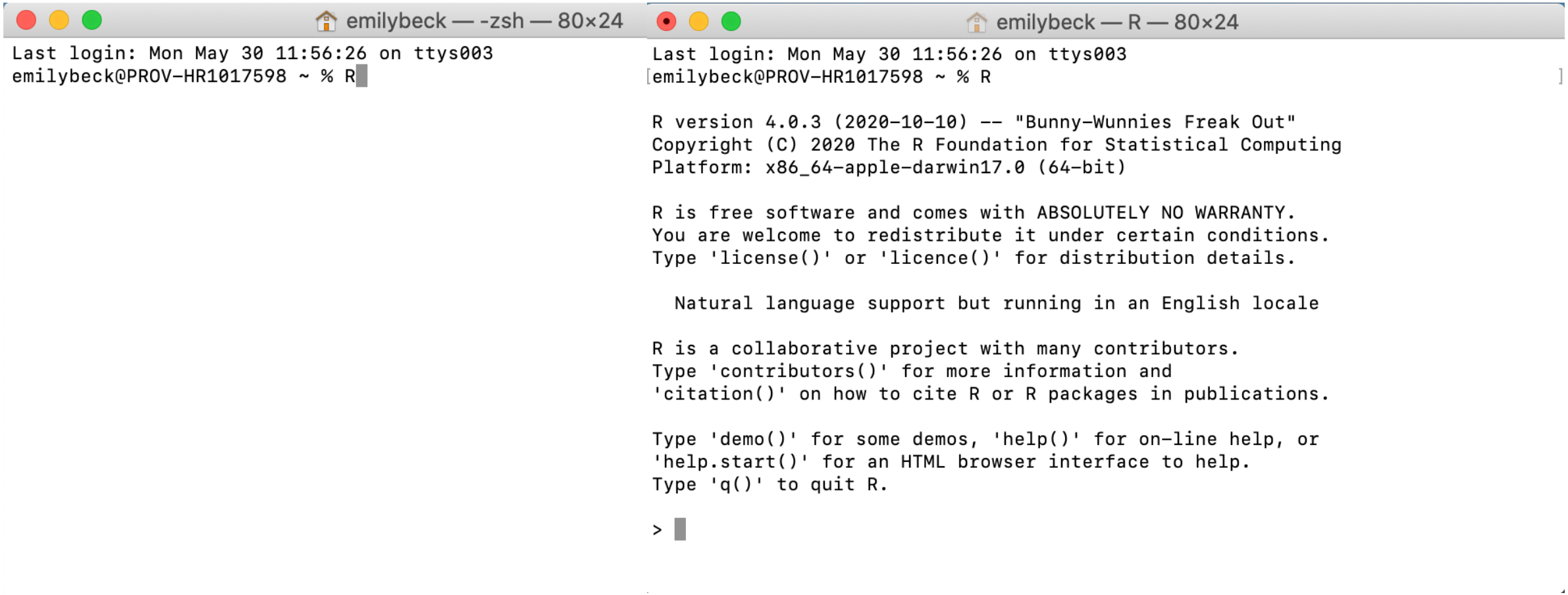
iDEP: Integrated Differential Expression and Pathway analysis



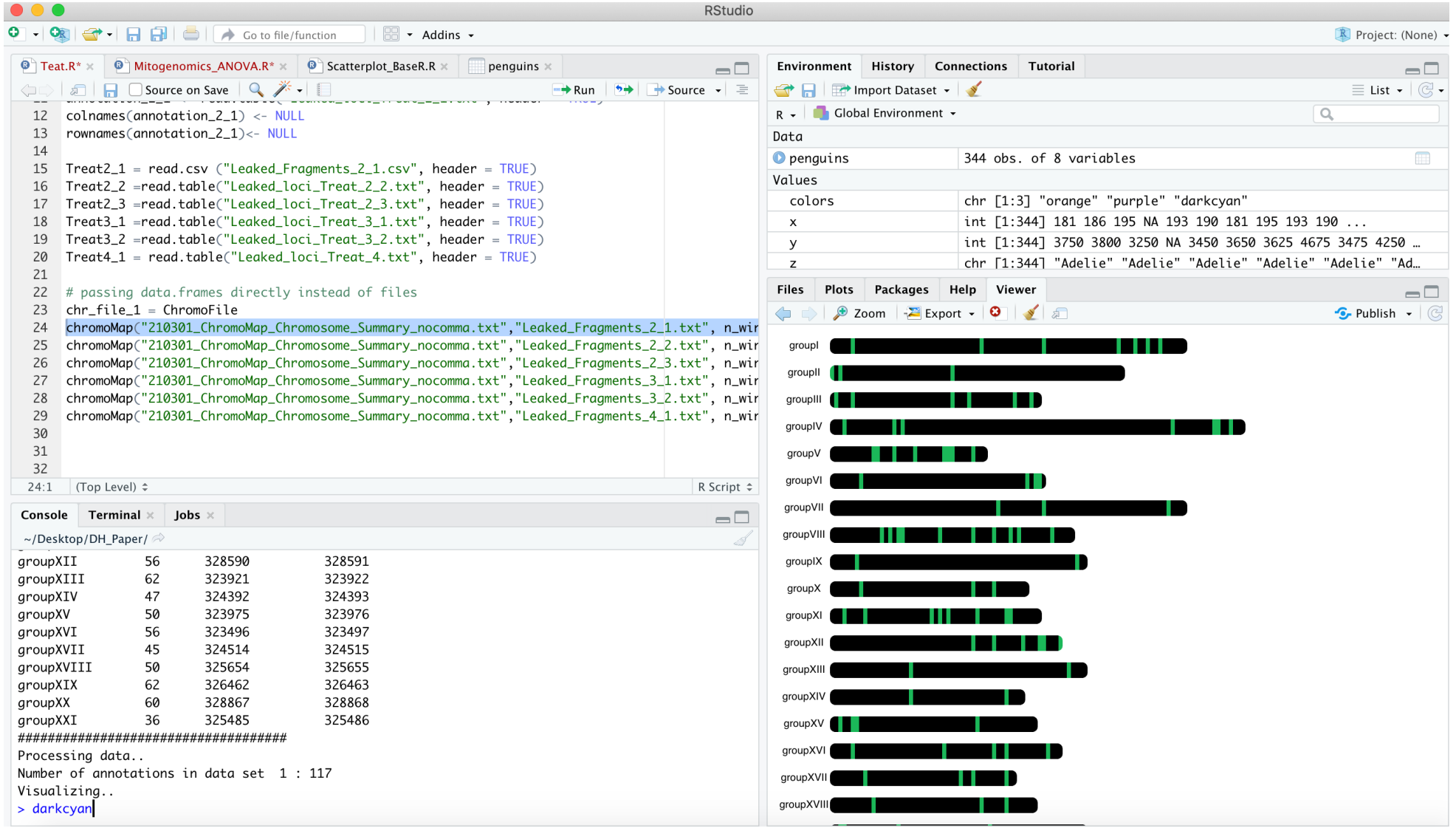
iDEP workflow

→ R markdown file

R can be run many ways
Option 1: In the terminal

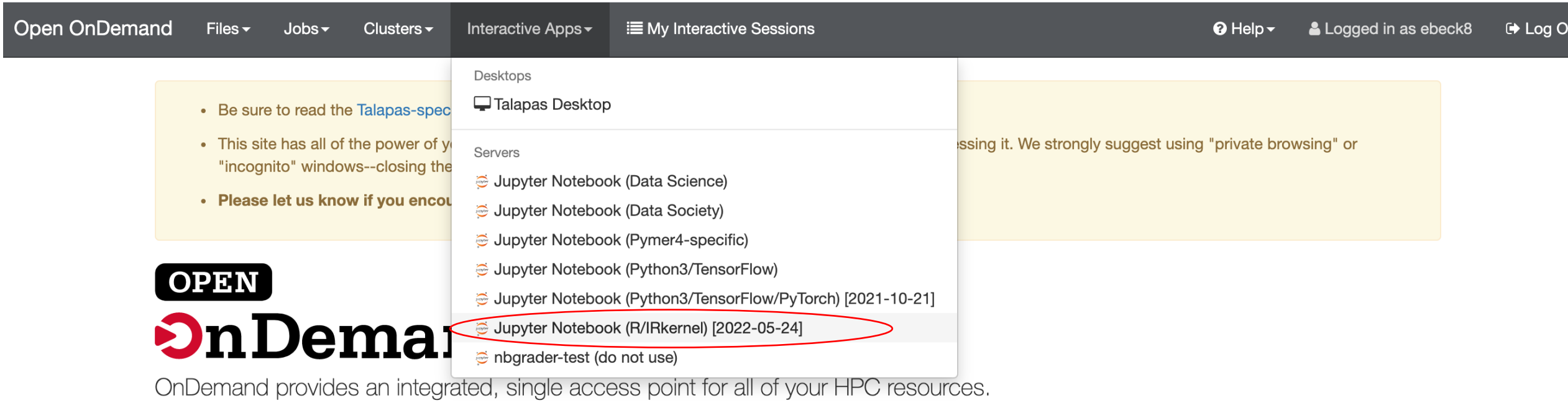


R can be run many ways
Option 2: In R Studio



R can be run many ways:
Option 3: Open OnDemand!

https://talapas-ln1.uoregon.edu/ Login with your UO credentials



Basic R mechanics

Same tips and tricks apply from Linux!

Tab complete
Up-arrow for the last command

To run current line of code:
Ctrl +Enter (PC)
Cmd +Enter (Mac)