



# Bridging the Bench-Machine Learning Gap

Dr. Emily A. Beck

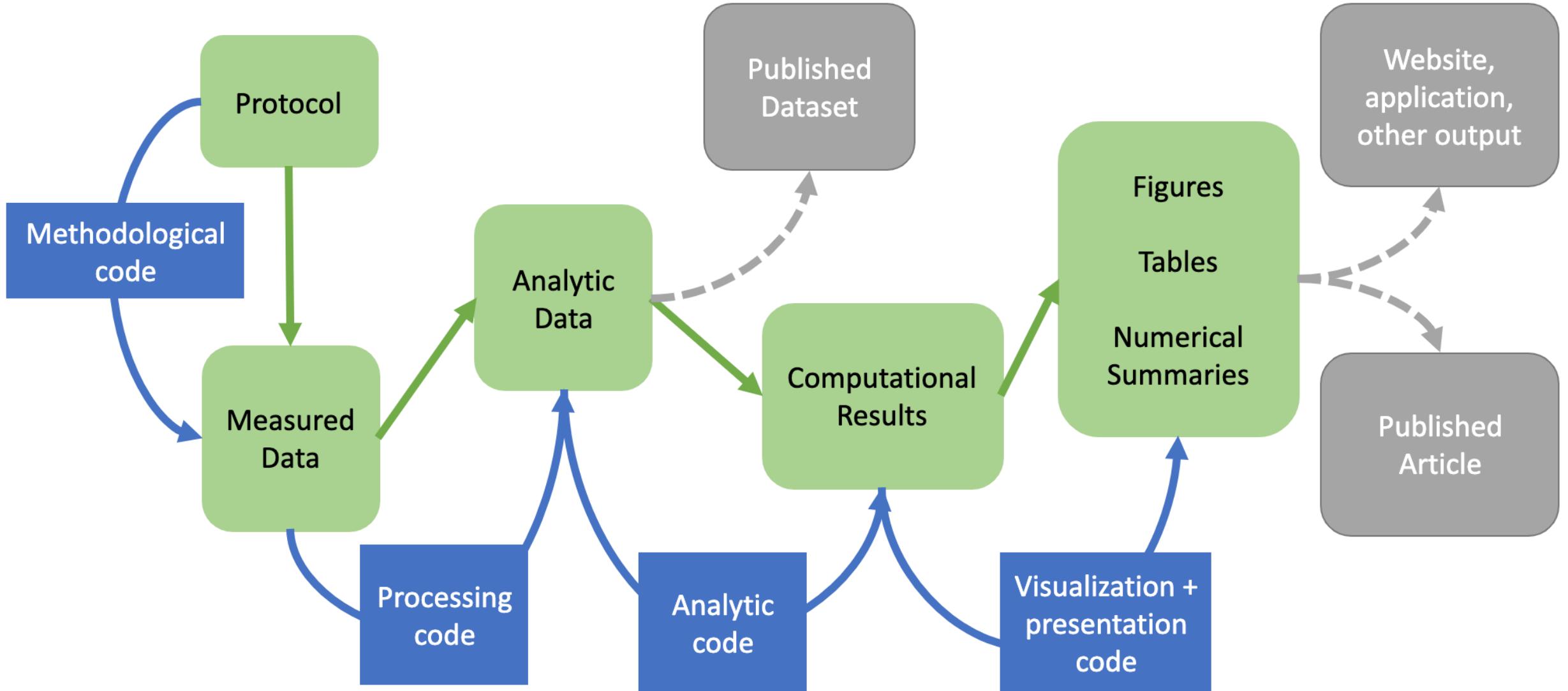
Dr. Jake Searcy

# 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

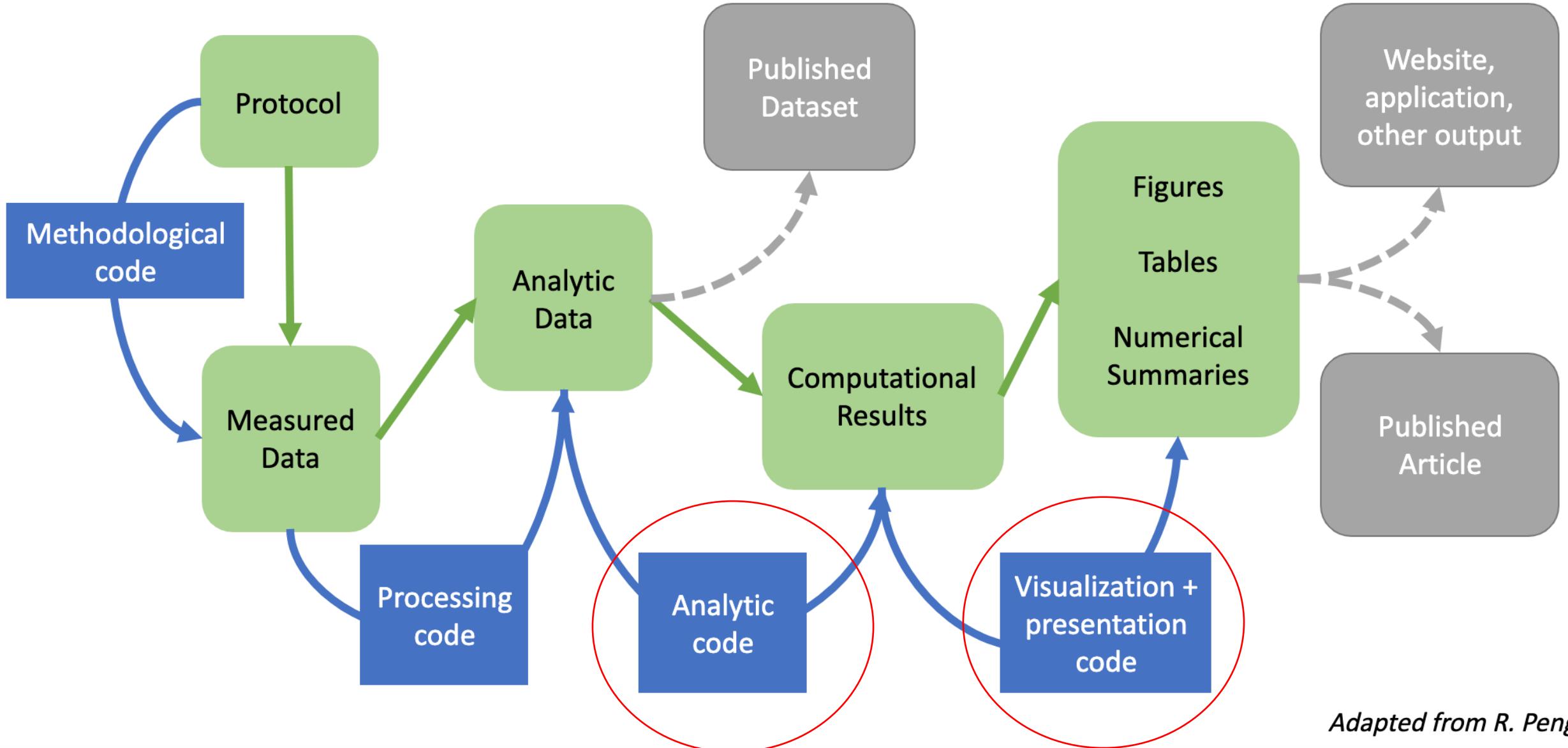
R can be used for many stages of the workflow!



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)





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/>

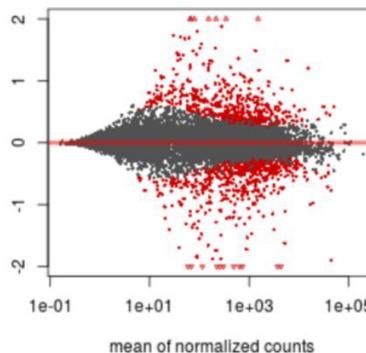
- Variant detection: sequence analysis, PolyPhen database
- Annotation: pathway analysis, BioMart, GO, KEGG, NCBI and many others
- High-throughput assays: flow cytometry, mass spec, proteomics
- Transcription factor binding, differential gene expression analysis (DEseq2)
- Methylation and chromatin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



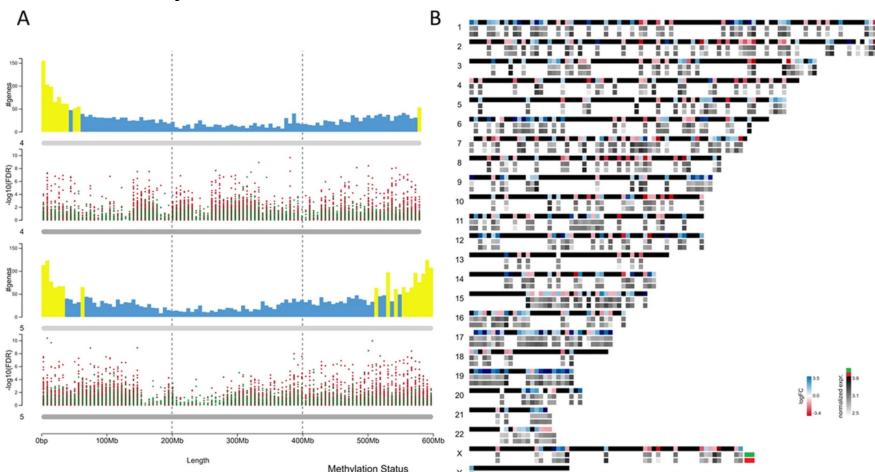
# Packages for genomics:

## Differential analysis of count data – the DESeq2 package

Michael I. Love<sup>1</sup>, Simon Anders<sup>2</sup>, and Wolfgang Huber<sup>3</sup>

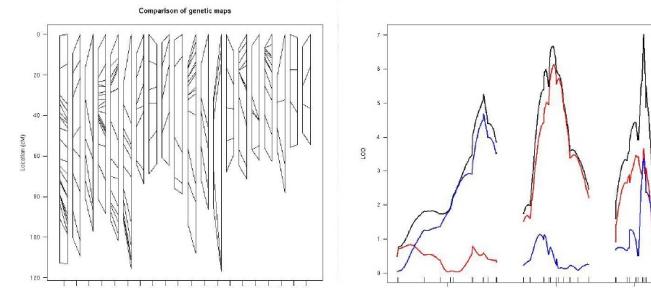


## chromomap



## R/qtl: A QTL mapping environment

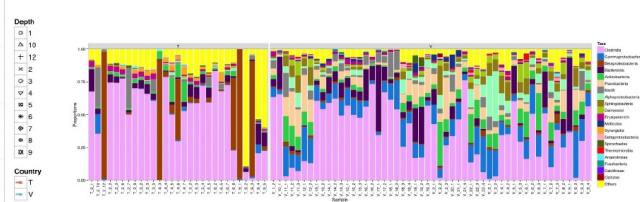
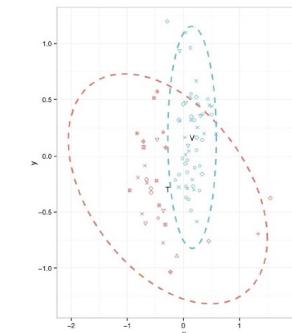
Software for mapping quantitative trait loci in experimental crosses



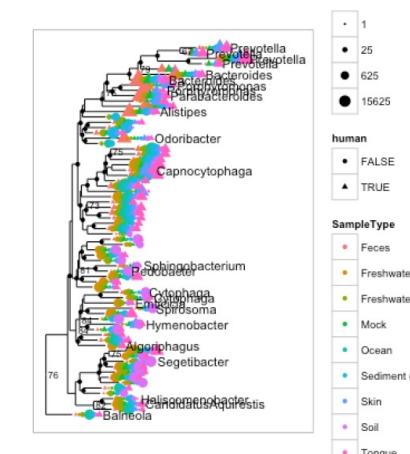
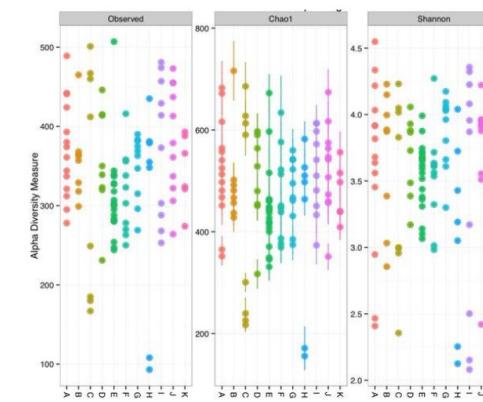
[www.rqtl.org](http://www.rqtl.org)

Multivariate Analysis of Ecological Communities in R: vegan tutorial

Jari Oksanen



## phyloseq



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



# Packages for biologists:



[https://cran.r-project.org/web/packages/available\\_packages\\_by\\_name.html](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 notoriously 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.

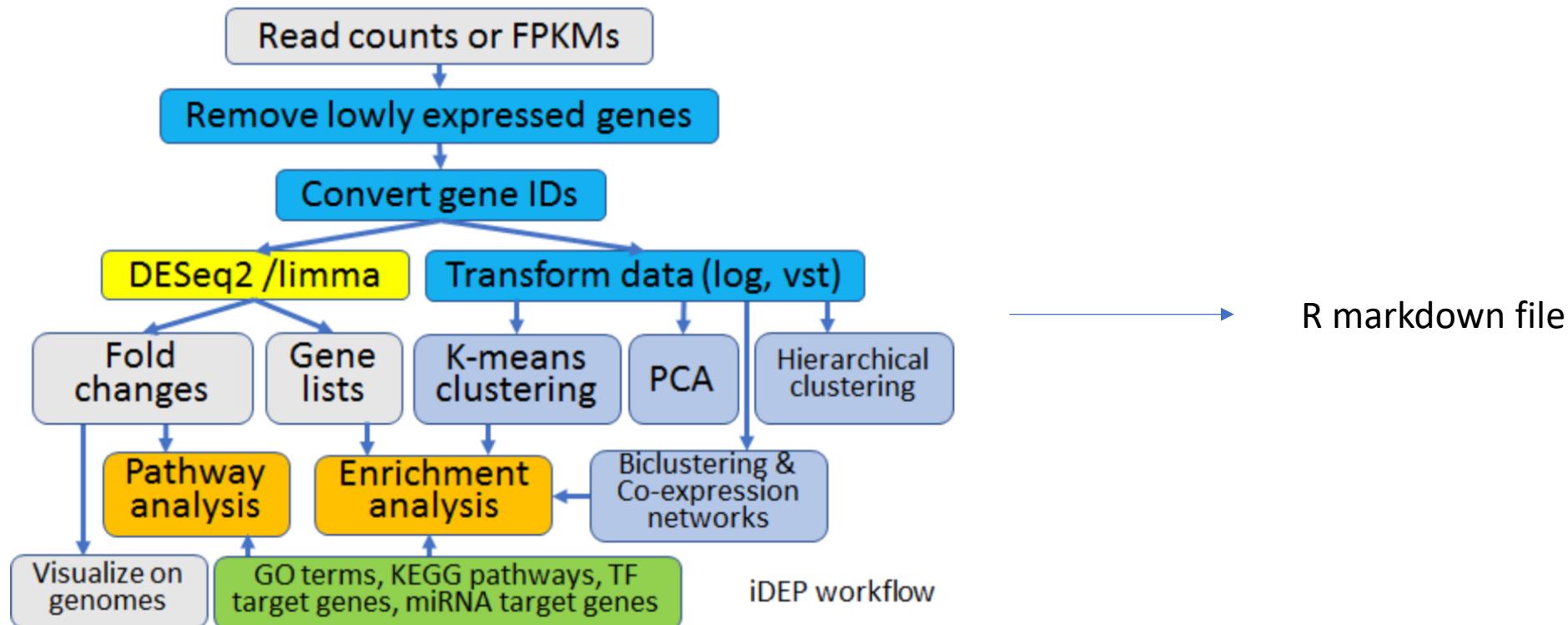
## iDEP: Integrated Differential Expression and Pathway analysis



# Packages for biologists:

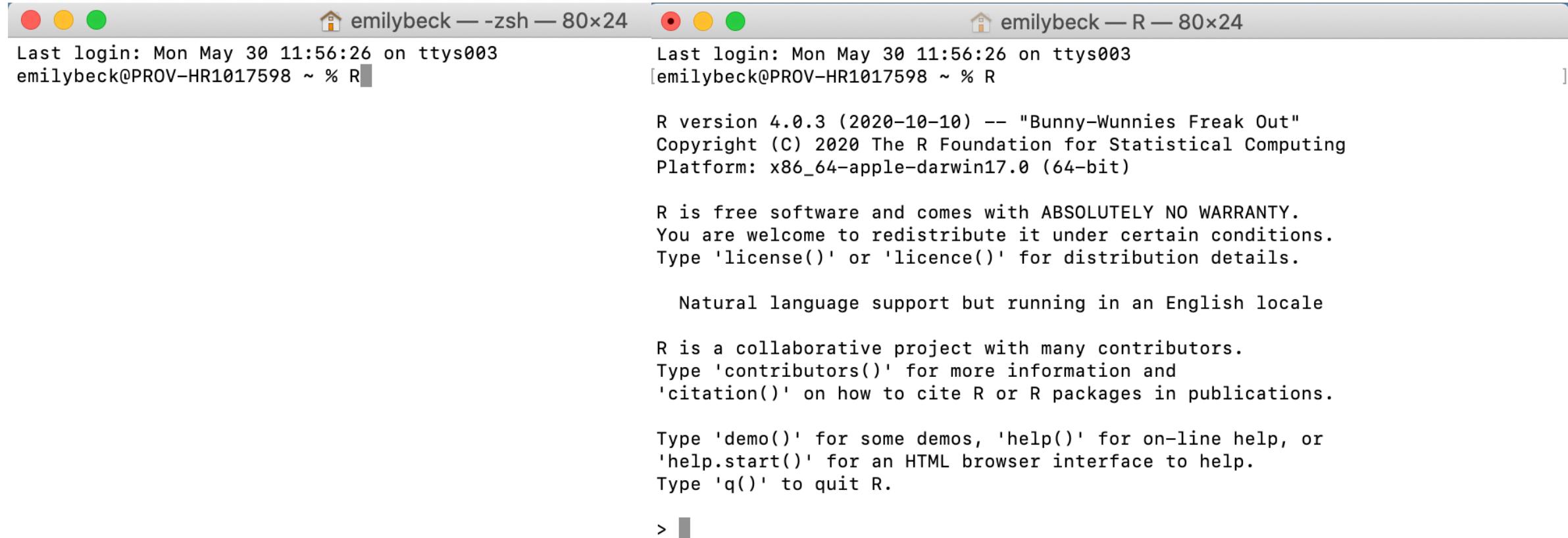


## iDEP: Integrated Differential Expression and Pathway analysis



# R can be run many ways

## Option 1: In the terminal



The image shows two side-by-side terminal windows on a Mac OS X desktop. Both windows have a title bar with a house icon and the text "emilybeck —". The left window is titled "zsh — 80x24" and the right window is titled "R — 80x24". Both windows show the same text output from the R startup process.

```
Last login: Mon May 30 11:56:26 on ttys003
emilybeck@PROV-HR1017598 ~ % R

Last login: Mon May 30 11:56:26 on ttys003
[emilybeck@PROV-HR1017598 ~ % R

R version 4.0.3 (2020-10-10) -- "Bunny-Wunnies Freak Out"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (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.

> ]
```

# R can be run many ways

## Option 2: In R Studio

RStudio

Project: (None)

Teat.R\*   Mitogenomics\_ANOVA.R\*   Scatterplot\_BaseR.R\*   penguins

File Edit View Insert Cell Help Addins

Go to file/function Run Source

```
12 colnames(annotation_2_1) <- NULL
13 rownames(annotation_2_1)<- NULL
14
15 Treat2_1 = read.csv ("Leaked_Fragments_2_1.csv", header = TRUE)
16 Treat2_2 =read.table("Leaked_loci_Treat_2_2.txt", header = TRUE)
17 Treat2_3 =read.table("Leaked_loci_Treat_2_3.txt", header = TRUE)
18 Treat3_1 =read.table("Leaked_loci_Treat_3_1.txt", header = TRUE)
19 Treat3_2 =read.table("Leaked_loci_Treat_3_2.txt", header = TRUE)
20 Treat4_1 = read.table("Leaked_loci_Treat_4.txt", header = TRUE)
21
22 # passing data.frames directly instead of files
23 chr_file_1 = ChromoFile
24 chromoMap("210301_ChromoMap_Chromosome_Summary_nocomma.txt","Leaked_Fragments_2_1.txt", n_wir
25 chromoMap("210301_ChromoMap_Chromosome_Summary_nocomma.txt","Leaked_Fragments_2_2.txt", n_wir
26 chromoMap("210301_ChromoMap_Chromosome_Summary_nocomma.txt","Leaked_Fragments_2_3.txt", n_wir
27 chromoMap("210301_ChromoMap_Chromosome_Summary_nocomma.txt","Leaked_Fragments_3_1.txt", n_wir
28 chromoMap("210301_ChromoMap_Chromosome_Summary_nocomma.txt","Leaked_Fragments_3_2.txt", n_wir
29 chromoMap("210301_ChromoMap_Chromosome_Summary_nocomma.txt","Leaked_Fragments_4_1.txt", n_wir
30
31
32
```

24:1 (Top Level) R Script

Console Terminal Jobs

~/Desktop/DH\_Paper/

group	value
groupI	black
groupII	green
groupIII	green
groupIV	black
groupV	green
groupVI	green
groupVII	black
groupVIII	black
groupIX	green
groupX	black
groupXI	green
groupXII	black
groupXIII	black
groupXIV	green
groupXV	green
groupXVI	black
groupXVII	black
groupXVIII	black

#####
Processing data..
Number of annotations in data set 1 : 117
Visualizing..
> darkcyan

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

<https://talapas-ln1.uoregon.edu/>

Login with your UO credentials

The screenshot shows the Open OnDemand web interface. At the top, there is a navigation bar with links for "Open OnDemand", "Files", "Jobs", "Clusters", "Interactive Apps", "My Interactive Sessions", "Help", and user authentication information. Below the navigation bar, there is a message box containing instructions:

- Be sure to read the [Talapas-specific notes](#)
- This site has all of the power of your local "incognito" windows--closing them will not log you out.
- Please let us know if you encounter any problems.

On the left side, there is a large "OPEN onDemand" logo. In the center, there is a list of interactive sessions categorized into "Desktops" and "Servers". The "Servers" section includes:

- Talapas Desktop
- Jupyter Notebook (Data Science)
- Jupyter Notebook (Data Society)
- Jupyter Notebook (Pymer4-specific)
- Jupyter Notebook (Python3/TensorFlow)
- Jupyter Notebook (Python3/TensorFlow/PyTorch) [2021-10-21]
- Jupyter Notebook (R/IRkernel) [2022-05-24]** (This item is circled in red.)
- nbgrader-test (do not use)

At the bottom of the page, there is a footer message: "OnDemand provides an integrated, single access point for all of your HPC resources."

Message of the Day

# Basic 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)