Bioinfo-R-matics

Ian Misner, PhD

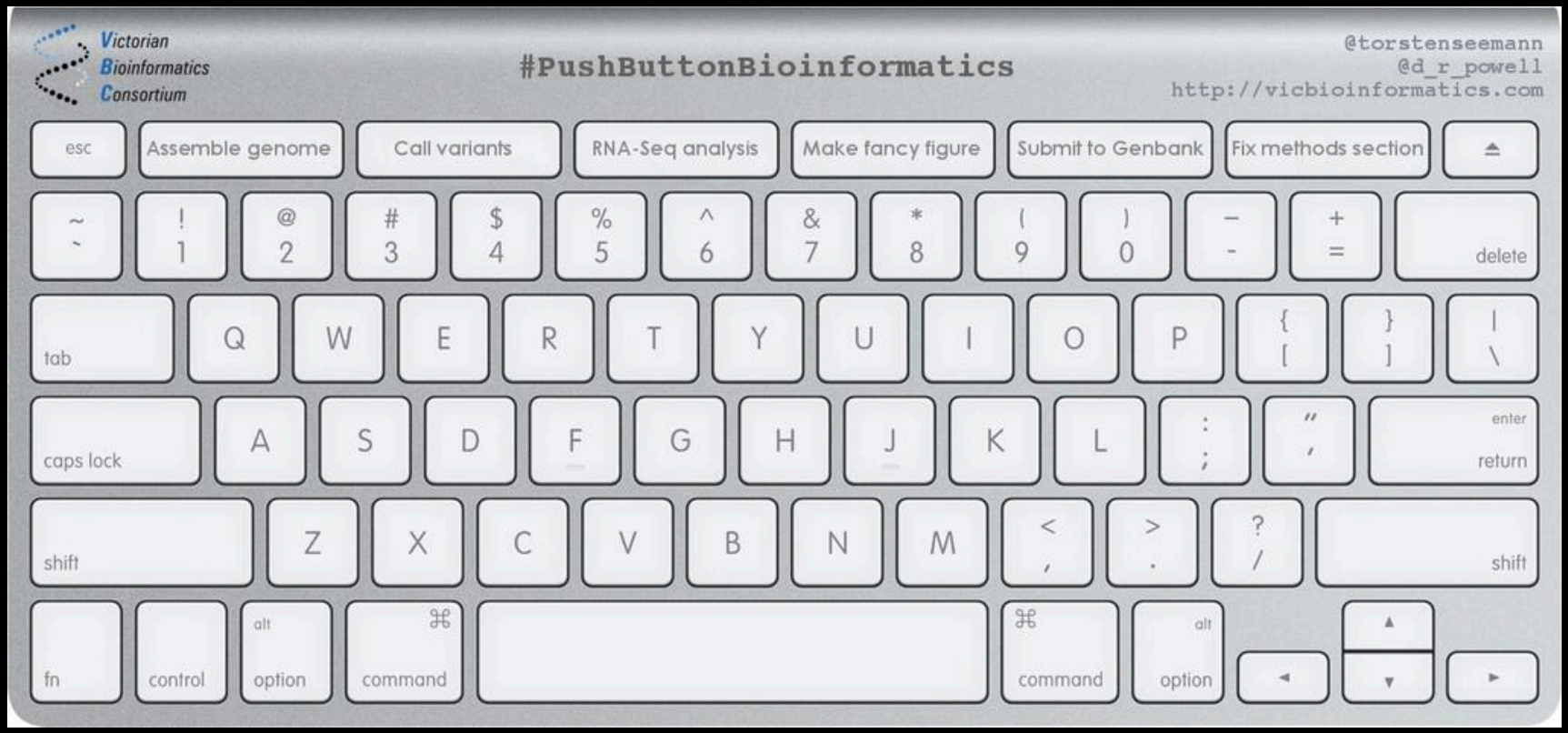
UMD Bioinformatics Core

- Mission: To provide users with the bioinformatic services, support, and education necessary to advance their research program.
- The Core has partnered with the Division of IT to provide the necessary computational resources needed for these demanding analyses.

What is Bioinformatics?

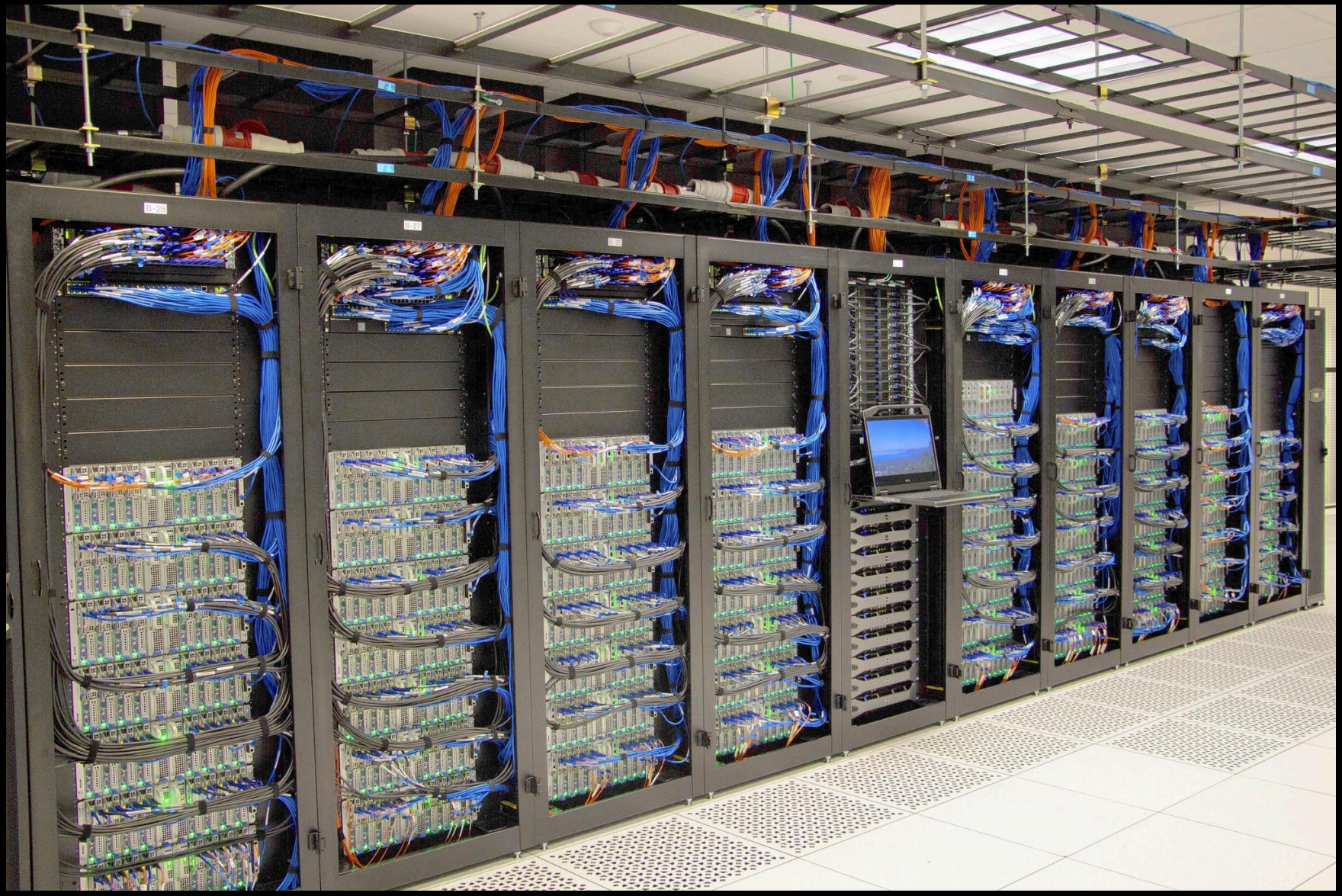
- Interdisciplinary field combining:
  - Computer science
  - Statistics
  - Mathematics
  - And Biology
- Lots of different areas of expertise:
  - Biological programing
  - Software development
  - Hardware development
  - Experimental design
- Difficult for an individual to be an expert in all areas.
- HIGHLY COLLABRATIVE!

What it isn't…


Victorian Bioinformatics Consortium
#PushButtonBioinformatics
@torstenseemann
@d_r_powell
http://vicbioinformatics.com
esc
Assemble genome
Call variants
RNA-Seq analysis
Make fancy figure
Submit to Genbank
Fix methods section
delete
tab
caps lock
enter
return
shift
shift
fn
control
alt
option
command
command
alt
option

Bioinformatics Core Services

- Raw data processing
- Genome and transcriptome assembly
- RNAseq analysis
- Variant discovery
- Grant writing support
- Experimental design assistance
- Workflow and pipeline construction
- Custom analyses



UNIVERSITY OF MARYLAND 18 56

BIOINFORMATICS CORE

What is R?

- R is a language and environment for statistical computing and graphics.
- Full-featured programming language.
- R has coherent, and extensive documentation
- The graphical facilities for data analysis can display on screen or be saved as hard copy.
- Think of it as a statistics environment where you can implement statistical techniques.
- Fully extensible, CRAN, bioconductor, etc …

Why R?

- R is FREELY available on any computer system.
  - Documents your data analysis and makes it reproducible.
- R is command line program.
  - You cannot just click menus and get results you have to know what your doing and therefore you learn and remain active in your analysis.
- It's free.
  - Not to say that it was free to create so please cite R and packages appropriately.

Why R?

- Combines many programs into one environment.
  - Excel, SAS, JMP, SigmaPlot, etc…
- Make publication quality figures.
  - Export graphs as PDFs and they are resolution independent.
- It's easy to write down and save your "instructions" for your analysis. i.e. Script
  - The script becomes a permanent, repeatable, annotated, cross-platform, sharable record of your analysis.

BEFORE R: REMEMBER THE SCIENTIFIC PROCESS!

Experimental Design

1. What is the biological question?
   - Do the amino acid polymorphisms at the *Pgm* locus have an effect on glycogen content (GC)?
2. Put the question in the form of biological null and alternative hypothesis.
   - Different amino acid sequences do not affect the biochemical properties of PGM, so GC is not affected by PGM sequence.
   - Different amino acid sequences do affect the biochemical properties of PGM, so GC is affected by PGM sequence.

Adapted from "Handbook of Biological Statistics 3rd Ed.

Experimental Design

3. Put the question in the from of statistical null and alternative hypothesis.
   - Flies with different sequences of the PGM enzyme have the same average GC.
   - Flies with different sequences of PGM have different average GC.
4. Determine which variables are relevant to the question.
   - Variables are glycogen content and PGM sequence.

Adapted from "Handbook of Biological Statistics 3rd Ed.

Experimental Design

5. Determine the type of each variable.
   - GC is a measurement variable.
   - PGM sequence is a nominal variable with four possible values (V-V, V-L, A-V, or A-L)
6. Design an experiment that controls or randomized the variables.
   - Other variables include age and where in vial flies pupated. These were controlled in the experiment. All files were the same age and files were taken at random from the vials.

Adapted from "Handbook of Biological Statistics 3rd Ed.

Experimental Design

7. Based upon the # variables, kind, expected fit to assumptions, and hypothesis chose the best statistical test.
   - The goal here is to compare a single measurement variable against each nominal variable. This is perfect for a one-way anova.
8. If possible do a power analysis to determine the proper sample size.
   - A power analysis would require an estimate of the SD of GC (from the literature) and number for the effect size. In this case researchers used as many files as feasible.

Adapted from "Handbook of Biological Statistics 3rd Ed.

Experimental Design

9. **Do the experiment!**
   - **Experiment was completed and GC measured in each fly.**
10. Examine the data to see if it meet the assumptions of the statistical test you chose.
   - Anova assumes:
     - Measurement variable, GC, is normal distribution…
     - and homoscedastic (variance in GC of PGM sequences are equal).
     - Histograms confirm this to be true of the data.

Adapted from "Handbook of Biological Statistics 3rd Ed.

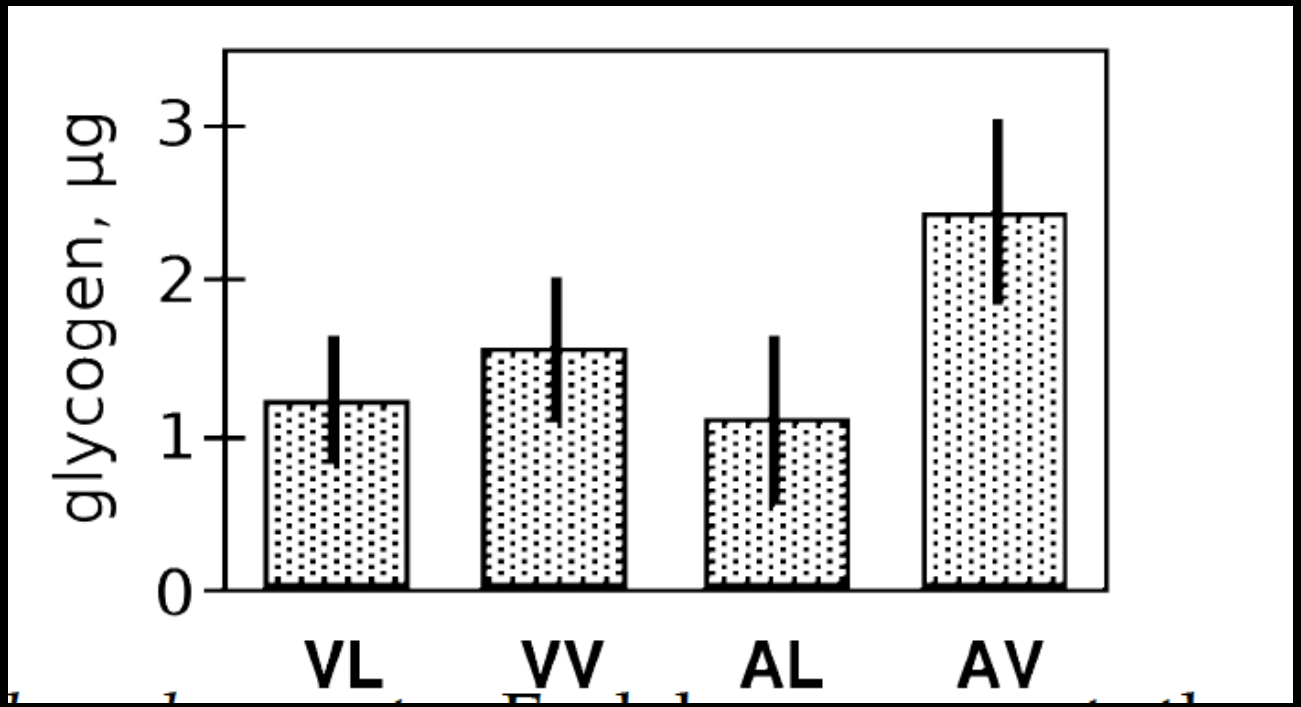
Experimental Design

11. Apply the statistical test you chose, and interpret results.
    - Complete one-way anova (USING R). The interpretation is that files with some PGM sequences have different average GC than flies with other PGM sequences
12. Communicate your results effectively, usually a graph or table.

Adapted from "Handbook of Biological Statistics 3rd Ed.

Experimental Design

12. Graph in R

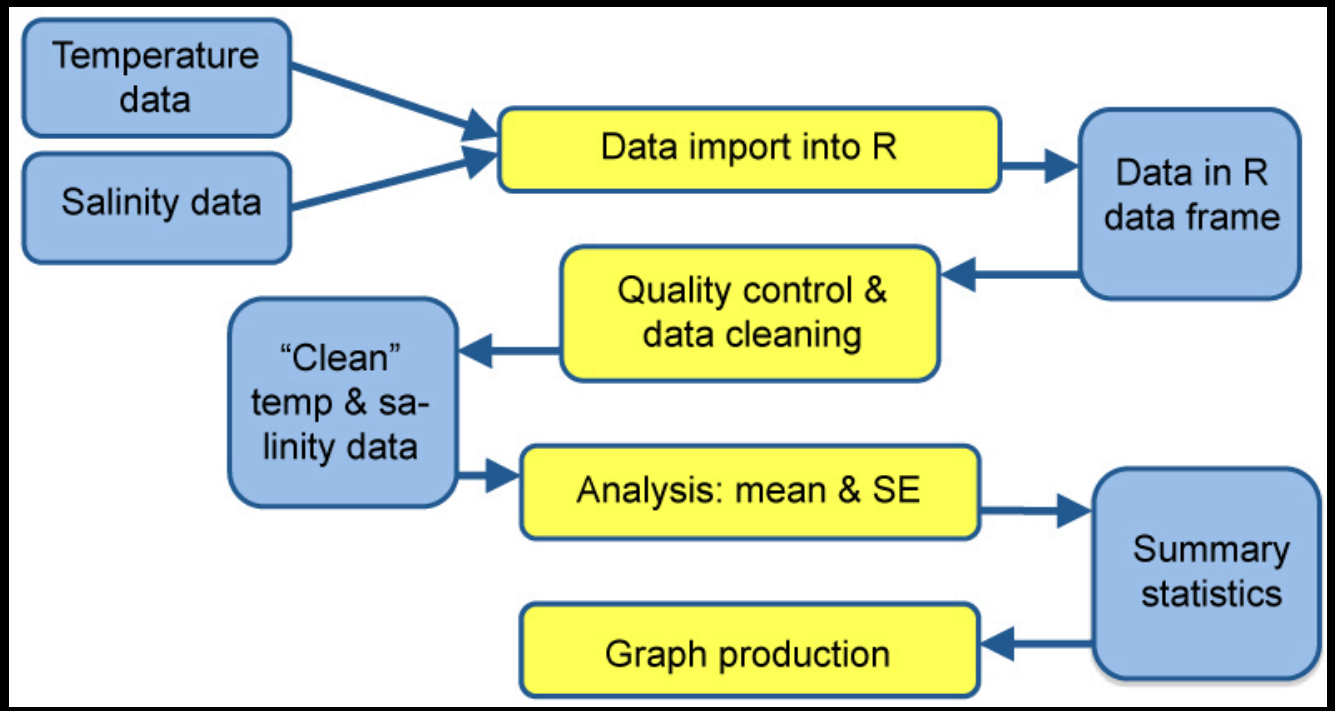


From "Handbook of Biological Statistics 3rd Ed. Study Verrelli and Eanes (2001)

Interface with R

What is a script?

script – a list of commands that are stored in a file and perform a sequence of computational tasks without the need user interaction.



http://carlystrasser.net/2012/06/

Why a script?

Reliability

Efficiency

Repeatability

Automation

"Parallelization"

Pipeline production

Parts to an R script

Text document (.R)

1. Import the libraries
2. Import data
3. establish variables
4. perform tasks on variables
5. output processed data

Output text data

sink()

Nothing in ()'s = print to terminal

sink("myfile", append=FALSE, split=FALSE)

Send to "myfile"

"append" determines if it overwrites or not.

"split" determines if the output is sent to both screen and file or just file.

Doesn't not redirect graphical output!!

Output graphical data

pdf()
png()
jpeg()
bmp()
Use dev.off() at end of script to close redirect to graphical output destination!!!

ex. pdf("myfile.pdf")
    plot(x)
    dev.off()

Errors Errors Everywhere!!

You will get errors!
It is not mad at you!
It is not mocking you!
It is informing you!
Listen to them and you will find answers quickly!
Computer-based problems solving!!

- http://www.r-project.org/
- www.google.com
- http://stackoverflow.com/