Things I'm Interested In

(that relate to ongoing DIB projects)

Rayna M. Harris

Slides:
github.com/raynamharris/IntegrativeProjectWT2015/docs

DIB Lab Meeting
April 26, 2018

Some talking points

- RMarkdown for presentations

Some talking points

- RMarkdown for presentations
- Update on the Software Carpentry Spanish lessons

Some talking points

- RMarkdown for presentations
- Update on the Software Carpentry Spanish lessons
- My data commons-like thesis chapter

Some talking points

- RMarkdown for presentations
- Update on the Software Carpentry Spanish lessons
- My data commons-like thesis chapter
- How do I `make` this?

Some talking points

- RMarkdown for presentations
- Update on the Software Carpentry Spanish lessons
- My data commons-like thesis chapter
- How do I `make` this?
- Reproducing other people's data analysis

Software Carpentry Spanish Lessons![1]

software carpentry — Teaching basic lab skills for research computing

Our lessons in Spanish

Lesson	Site	Repository	Reference	Instructor Guide	Maintainer(s)
La Terminal de Unix	▣	▣	👁	⊕	Ivan Gonzalez, Clara Llebot, Verónica Jiménez, Silvana Pereyra, Heladia Salgado
Control de versiones con Git	▣	▣	👁	⊕	Ivan Gonzalez, Rayna Harris, Clara Llebot
R para Análisis Científicos Reproducibles	▣	▣	👁	⊕	Rayna Harris, Verónica Jiménez, Silvana Pereyra, Heladia Salgado

The **lesson template** and the **workshop template** are available in Spanish. If you are interested in getting involved with our Spanish lessons **contact us**.

La **la plantilla de lección** y la **plantilla de taller** están disponibles en español. Si está interesado en participar en nuestra lección de español **contáctenos**.

---

[1] https://software-carpentry.org/lessons/

Software Carpentry Spanish Workshops!
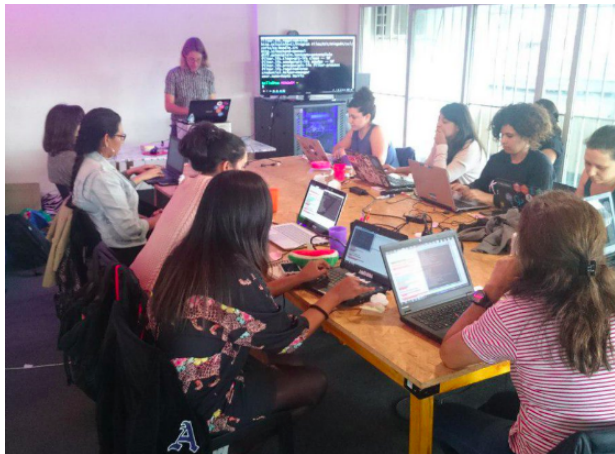
Configurar #git para poder empezar a usarlo, los primeros pasos del taller de @raynamharris pic.twitter.com/hWV9fqsBWA

— R'lyeh Hacklab (@rlyehlab) April 7, 2018

To translate or not to translate[2], that is the question

Términos técnicos, no traducir sólo marcar con negrita

Por favor mantener orden alfabético

- A
  - el **array**
- B
  - sin artículo **bash**
  - el **bit score**
  - los **bytes**
  - el **branch**
- C
  - el **checkout**
  - sin artículo **commit**
  - el **cluster**
  - el tipo de datos **complex** en R
  - el tipo de datos **character** en R
- D
  - el **dataset**
  - el **dataframe**
  - el **Desktop**
  - tipo de datos **double**

---

[2]https://github.com/Carpentries-ES/board/blob/master/Convenciones_Traduccion.md

Quote from Data Commons (TOPMed) that I can relate too

Attributes of both subjects (individuals) and biosamples (e.g. DNA, RNA, serum, urine) are provided in tabular data sets. It is important to distinguish between these two types of identifiers. Some attributes are measured at the subject level (e.g. height, blood pressure of a person), while others are measured at the sample level (e.g. metabolites, genotypes, gene expression levels on serum, DNA, or RNA samples, respectively). Multiple samples may derive from the same individual.

Using RNA sequencing to understanding how the hippocampus
remembers things in space and time

RNA sequencing experimental design

Tratment Groups	Sample size		
	DG	CA3	CA1
yoked-consistent	4	3	5
consistent	5	5	4
yoked-conflict	4	3	2
conflict	3	2	4

FYI: Hypothesis-driven versus data-driven behavior analysis tell different but compatible stories

A — Number of Entrances

B — Time to 2nd Entrance

Training Session

C

D — Cognitive Training Group ~ Session

yoked-consistent   consistent   yoked-conflict   conflict
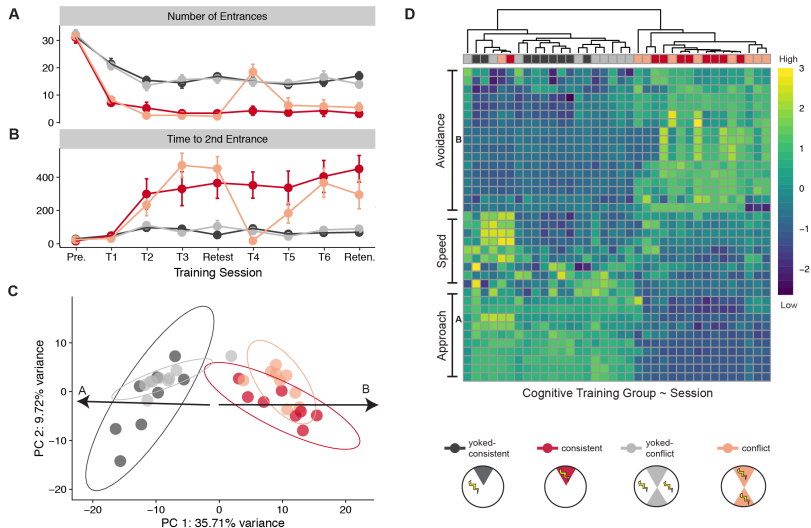
Subfield-specific expression trumps experience-specific expression
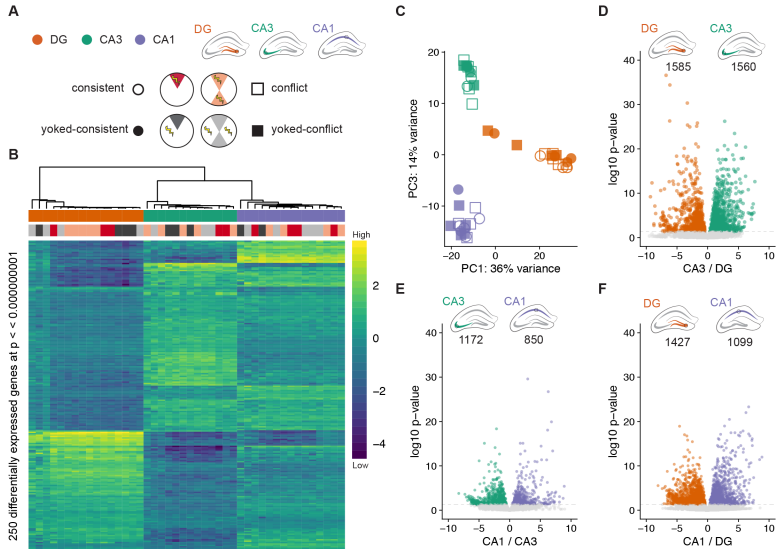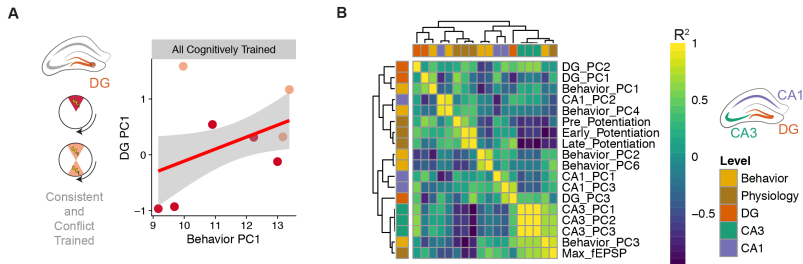
MultiQC[3] for visualizing summary stats!
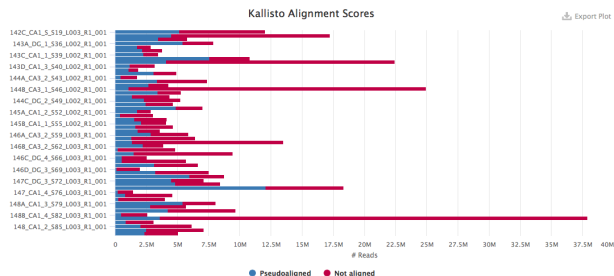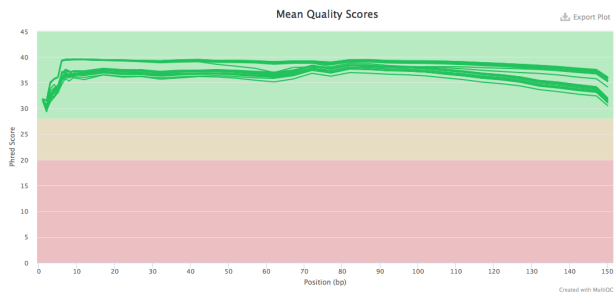
How do I `make this?`[4]

Workflow

- Part 1: Behavior
  - **01a_behavior_analysis:** behavior data wrangling, statistics, and data visualization
- Part 2.1: UNIX for RNA-seq
  - UNIX workflow Initial processing of RNA sequencing was performed on the Stampede cluster at the Texas Advanced Computing Facility. My workflow for this is described in the 'UNIXworkflow' directory.
  - **00_rawdata:** Download the data to scratch on Stampede with `00_gsaf_download.sh`.
  - **01_fastqc:** Evaluate the quality of the reads using the program FastQC.
  - **02_filtrimreads:** Filter low quality reads and trim adapters using the program cutadapt.
  - **03_fastqc:** Evaluate the quality of the processed reads
  - **04_kallisto:** Quantify transcript-level expression using Kallisto
- Part 2.2: R for RNA-seq
  - **02a_rnaseq_makedfs:** converting the Kallisto transcript counts to gene counts and wrangling the categorical data about the samples
  - **02b_rnaseqALL:** analyzing all the RNA-seq data together
  - **02c_rnaseqSubfield:** analyzing the data for each hippocampal subfield separately
  - **02d_rnaseqAvoidance:** combining the two yoked group and the two training groups before analyzing and then analyzing each subfield separately
  - **02e_GO_MWU** an analysis of gene ontology. Note: this directory contains scripts and data
- Part 3: Electrophysiology
  - **03_ephys:** analysis of electrophysiology data
- Part 4: Integrative analysis
  - **04_integration:** correlations across levels, mostly using PCA data

---

[4]https://github.com/raynamharris/IntegrativeProjectWT2015

Repository organization and workflow

1. UNIXworkflow - runs on The Texas Advanced Computing Center's (TACC) cluster 'Stampede'

2. data - mostly raw data and meta data, some results included

3. docs - my talk for the R-Ladies Buenos Aires group

4. scripts

   - 00_cpfiles.sh - a bash script for copying files into and out of this repo
   - 01_behavior.md - output of a knittable analysis script for wrangling the avoidance behavior data. The output is use for the `Fig*.Rmd` scripts that accompany each figure
     - **Figures** 1, 2, 3, 4, 5, 6, 7 - Because I had so many behavior figures and accompanying statistics, I created a Rmd script containing the code to make the figure and calculate the statistics. The data were generated in 01_behavior.Rmd.
   - 02_RNAseq.md - output of an knittable analysis of differential gene expression of RNA-seq data with DESeq2, GO-WMU, and WGCNA
   - 03_wgcna.R - R script for weighted gene co-expression network analysis
   - 04_ephys.Rmd - analysis of hippocampal electrophysiology to measure synaptic properties -- note: must be run from the R command line as knitting doesn't currently work well
   - 05_Ceolin.md - a **replication** of a recently published paper and a **comparison** to my own data
   - 06_GO_MWU - directory with data, code, and results for Gene Ontology analysis

5. results

   - In the R script accompanying this file, 08_results.md, I write the results section for my manuscript.
   - This document is an executable way for me to keep the results section under version control and with newest version of the figure.
   - I alternate between exporting the result to a `md_document` and a `word_document` for writing my thesis

---

[5]https://github.com/raynamharris/FMR1CA1rnaseq

Figures from Ceolin et al. (2017) Cell Type-Specific mRNA Dysregulation in Hippocampal CA1 Pyramidal Neurons of the Fragile X Syndrome Mouse Model. Front. Mol. Neurosci. 10:340. doi: 10.3389/fnmol.2017.00340

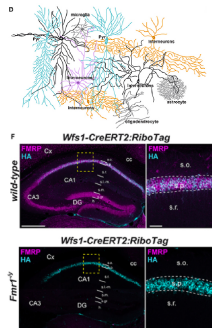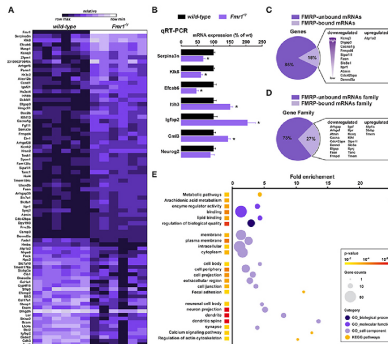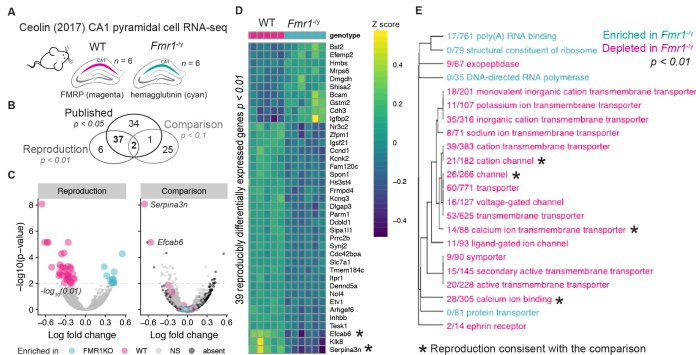# Reproducing[7] other people's data