JOHNS HOPKINS
SCHOOL of MEDICINE

Quantitative Molecular Neurogenomics 2023

ME.440.825
Loyal Goff - Course Director
Genevieve Stein-O'Brien - Instructor
Richard Sriworarat - TA
Kyla Woyshner - TA
Jeanette Johnson - TA

COURSE OVERVIEW

Today's Learning Objectives

- Review Course, Overview, Syllabus, Grading, and Expectations
- Confirm/fix setup of computational environment(s) for the course
- Understand expectations for participation
- Understand motivations for modern genomics analysis tools in Neuroscience research
- Exploration of VSCode, bash terminal, & basic python navigation

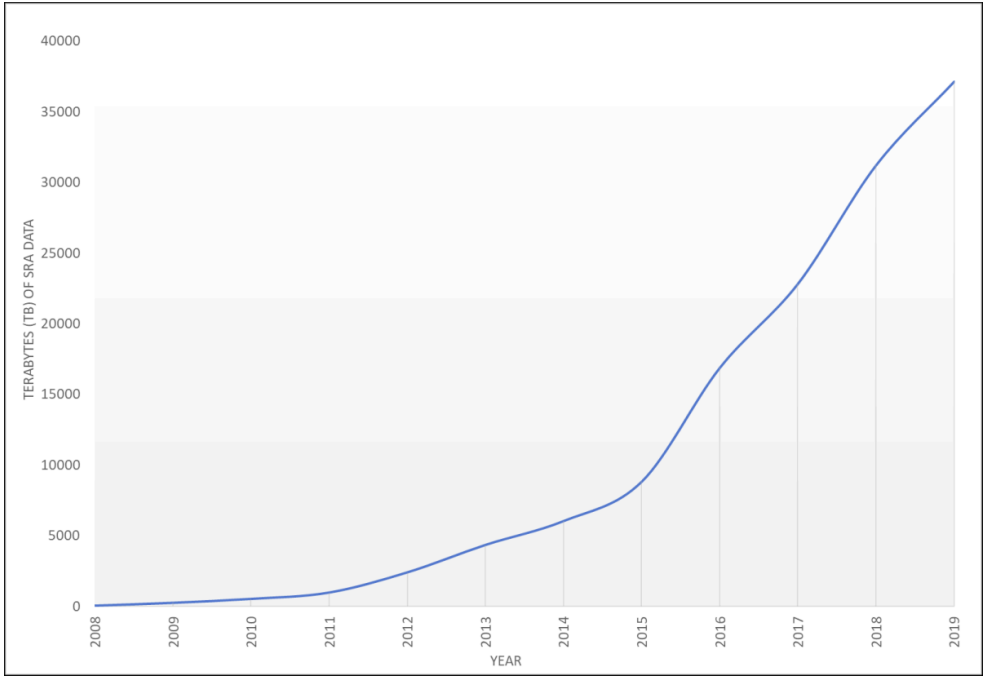
Overview

- Quantitative and computational methods are increasingly required for modern biological research.
- Neuroscience is no exception
- Through this course we want to empower you with the confidence and skills to apply these methods to high-throughput biological data.
- Where possible, we will highlight neuroscience-specific applications, resources, and methods

Why are we here?

- Modern biology is data intensive
- Biology is struggling to deal with the volume and complexity of sequencing data
  - Biologists need more training in data analysis
- Large-scale data is increasingly easier to generate and acquire
  - e.g., High-throughput sequencing & imaging
- But the complexity of experiments has grown as well
- Most modern biology ***data*** requires quantitative methods
  - Often methods are complex as well.

Growth of NIH SRA Database



Source: NCBI NLM Insights

Why can’t I just outsource my data analysis?

- Biological data analysis is not a ‘generalizable’ problem.
- To ***properly*** analyze data, it’s not enough to know how to use the tools. You also need to:
  - Understand the biological question & hypothesis
  - Understand the vagaries of the experimental assay & design
  - Have intuition about the correctness of the results
  - Recognize biologically relevant signals/patterns in the data
- ∴ Even in collaborations, you should understand and have a voice in how your data are being analyzed
  - Data analysis skills ***must*** be acquired by biologists

- “*…if you're a biologist pursuing a hypothesis-driven biological problem, and you're using a sequencing-based assay to ask part of your question, generically expecting a bioinformatician in your sequencing core to analyze your data is like handing all your gels over to some guy in the basement who uses a ruler and a lightbox really well.*” - Sean Eddy (http://cryptogenomicon.org/)

Course Learning Objectives

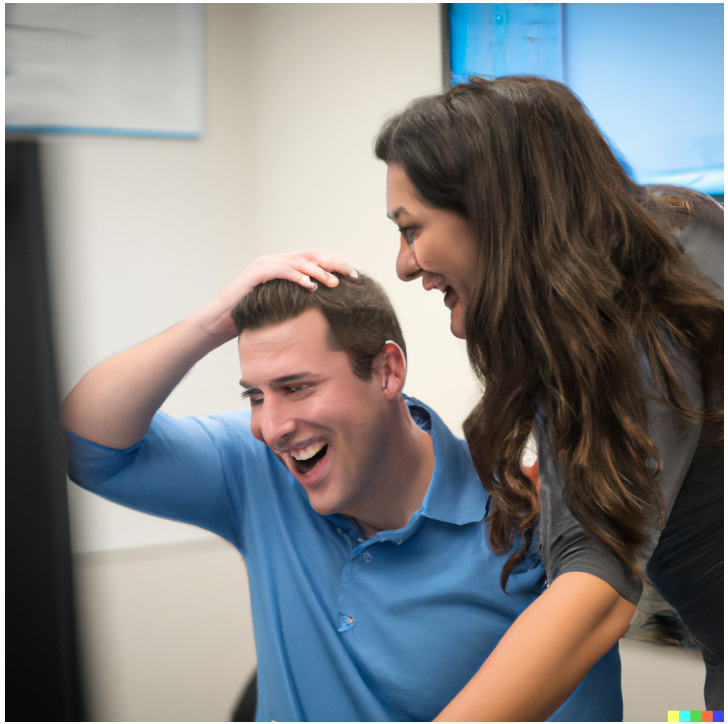
- Develop comfort working within a *NIX environment and with the command line to build and execute bioinformatics workflows
- Develop confidence in using python and other relevant frameworks and languages for bioinformatic analysis and visualization
- Develop knowledge of standard bioinformatic file formats and data structures and how they are connected in an analysis project
- Develop fluency in standard algorithms, statistical tests, and visualizations used in modern computational biology as applied to a diversity of applications.
- Develop good habits for reproducible research

Considerations

- This is the ~~first~~ second year for this course
  - We are ***still*** learning how to teach this
  - Your feedback (in real time) will be appreciated
- The ***primary objective*** for this course is for you to learn and be exposed to principles of data analysis and visualization in neuroscience
- We will ***only scratch the surface*** of modern genomics data analysis
- Your experience in this course will ***largely*** be self-guided, with help and resources made available as needed.
  - Lecture attendance alone is insufficient

Considerations

- There is going to be a high variance in skill set for this class
  - For those with less experience, you are in the right room
  - For those with more experience, you are in the right room
  - Collaboration and coordination across the class is *highly* encouraged.
- Please provide feedback often regarding the pacing of the course (too slow? Too fast?)



Course Goals

- Gain first-hand experience with various data analysis problems in euro-genomics
  - We will formally present only a fraction of problem types and workflows
- Become comfortable writing software and implementing available software tools to analyze large-scale biological data
- Develop reusable and reproducible data analyses
- Identify successful learning mechanisms, strategies, and resources to continue expanding your computational skills beyond this course

The Process

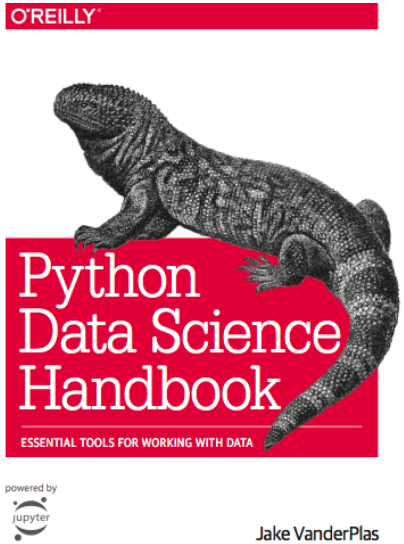
- ***Spend time*** with course materials, vignettes, publicly available tutorials, and published code to dissect individual components/steps in a workflow.
- Make mistakes and try different approaches.
- Find and reverse engineer existing code/analyses to understand how a tool or process is working
- Ask for guidance/support/help and help each other!
- This is just the beginning of a ***lifelong*** learning process
  - New workflows/tools are being developed every day
  - Data and experiments are getting more complex

Scripting is a foundational lab skill

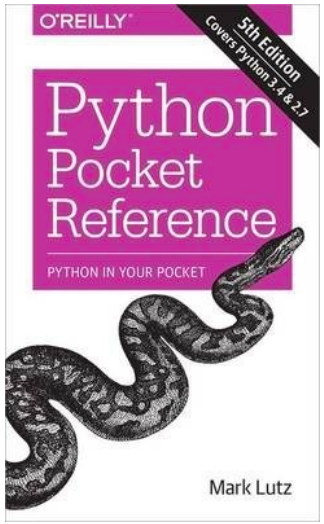
- Essential and easy to learn (like pipetting)
- Don’t need to master a language to script
- Like learning to write a lab protocol
  - you learn by tweaking an existing protocol/script
  - Over time, you will better understand how to write your own from scratch
- Writing scripts and using command line tools may be intimidating but only because we are not yet familiar with these tools.

Recommended text

- Python Data Science Handbook
  - available for free at jakevdp.github.io/PythonDataScienceHandbook/



- Python Pocket Reference
  - Available for free at github.com/shahadot786/Python-Books/blob/master/python-pocket-reference-5th-edition.pdf



Google

- The ***single greatest*** asset to help your coding.
  - Search error messages (you're not the first to have this problem)
  - Find tool vignettes, walkthroughs, tutorials, and examples from which to build off.
  - Search for strategies/code to solve problems or roadblocks

JOHNS HOPKINS
SCHOOL of MEDICINE

GRADING

Grading

- Problem Sets (70%)
  - 10 problem sets, each worth 10 points
  - Lowest scoring pset will be dropped
  - Group/collective work is encouraged
  - Everyone must turn in ***their own*** version with name and date.
  - Due before next Monday class (11:30AM)
- Final Project (30%)
  - Re-analysis of public dataset / study of your choice
    - Reproduction of 2 original findings/figures.
    - 1 new question asked and addressed.
    - Choose a study that is interesting to you and/or relevant to your current research.
  - Outline of study and proposed analysis will be submitted and reviewed by TAs and course instructors (by week 4)
  - Final projects will be presented to class in the style of a brief lab meeting report (10 min)
  - Full analysis (repo, code, notebooks, etc) will be submitted via Canvas

Grading Rubric(s)

- Grading will be directly affected by effort and participation throughout the course.

- Problem set scoring is based on the demonstration of ***concerted*** effort, and not necessarily the correctness of results.
  - I.e. if you worked through the problems effectively but have a bug in your code that leads you to an incorrect answer, we will work with you to correct and not necessarily penalize.

- Final Project grade will be based on:
  - Effort (presentation & notebooks)
  - Ability to achieve analysis benchmarks established in project outline
    - Communication with TAs & Instructors throughout course

Final Project

- Learning the components of a complete *omics workflow is the primary goal for this course

- Start planning now
  - Identify paper/study/dataset early on
  - TAs and Instructors can help review selection

- Consider efforts made by original authors towards reproducibility in your selection
  - Well documented/detailed workflow
  - Code repository available?

Final Project

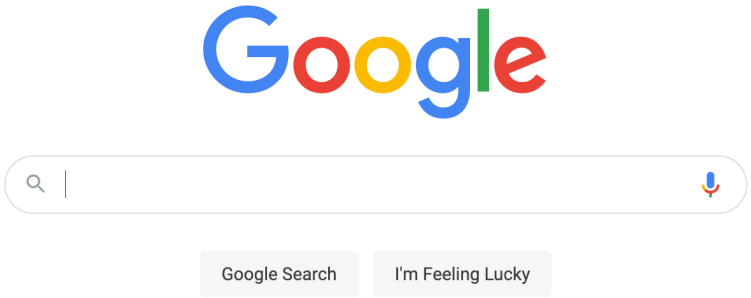
- Work ***with*** each other
  - Group work is encouraged in class and in research.
- Each of you will have a different path forward in your analyses
- Most of you will use tools/components in your workflow that we will not ***explicitly*** discuss in the course materials.
  - Use the learning/help strategies from this course to learn to connect tools for a proper workflow
- Check in ***often*** with instructors, TAs, classmates, and lab mates to help evaluate progress and troubleshoot.

Final Project - Deliverables

- Class presentation of study re-analysis
  - Study overview
  - Experimental Questions
  - Proposed re-analysis workflow
  - Comparison to originally published results
  - Presentation of novel finding/analysis
- Reproducible code base
  - Fully executed re-analysis
  - Repository (preferred) and/or compressed archive of all project-related analysis

Final Project - Resources

- Primary paper
- Project outline - submitted for review
- Vignettes and tutorials
- Instructors/TAs
- Classmates
  - JHUNeurogenomics2023 Slack workgroup:
  - https://join.slack.com/t/jhuneurogenom-jso4949/shared_invite/zt-20ga7u16d-DXsIXCuFMHrkOsrRDLrDqw
- Package/tool/resource documentation
-



JOHNS HOPKINS
SCHOOL of MEDICINE

COMPUTE ENVIRONMENT(S)

Your compute needs will evolve during the course

- Local installation(s) at first
- As projects progress, you may need more:
  - Memory
  - Disk Space
  - CPUs
- Transition to a high-performance environment as needed
  - Rockfish (ARCH)
  - We will provide instructional material for accessing/operating w/in rockfish HPC

Rockfish

- Advanced Research Computing at Hopkins (ARCH)
  - State of the art high-performance computing resource
  - Allocation provided for this course

- Information:
  - Account Info & Resource Allocation: https://coldfront.rockfish.jhu.edu/
  - User Guide: https://www.arch.jhu.edu/access/user-guide/
  - FAQ: https://www.arch.jhu.edu/access/faq/

Check-ins

- Win/Mac/Linux?
- Signed up for Rockfish access?
- Install R/RStudio locally?
- Complete/knit Module 1 Notebooks?
- Open and work on Module 1 pset?
- Experience with HPC?
- Experience with git/version control?
-

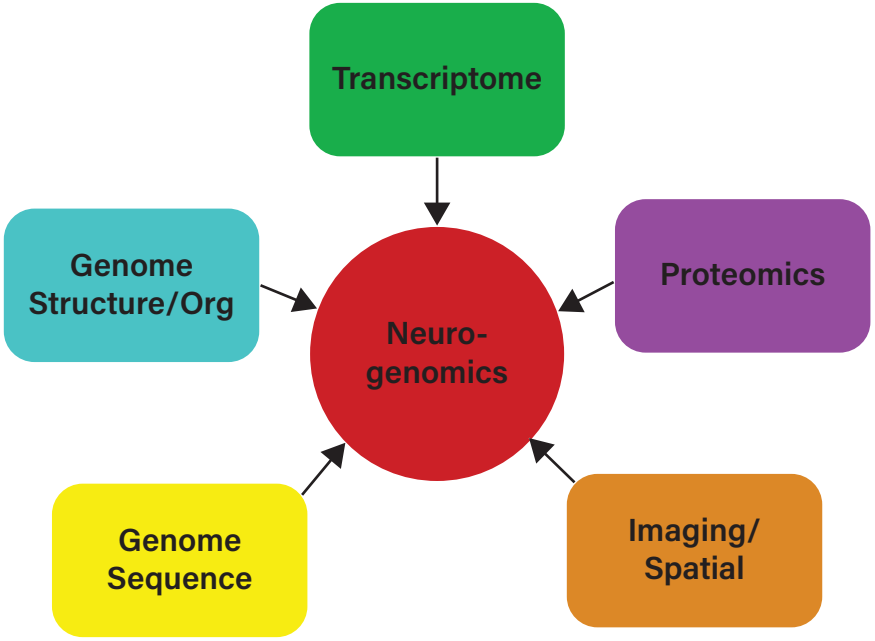
Questions about compute environment/setup?

JOHNS HOPKINS
SCHOOL of MEDICINE

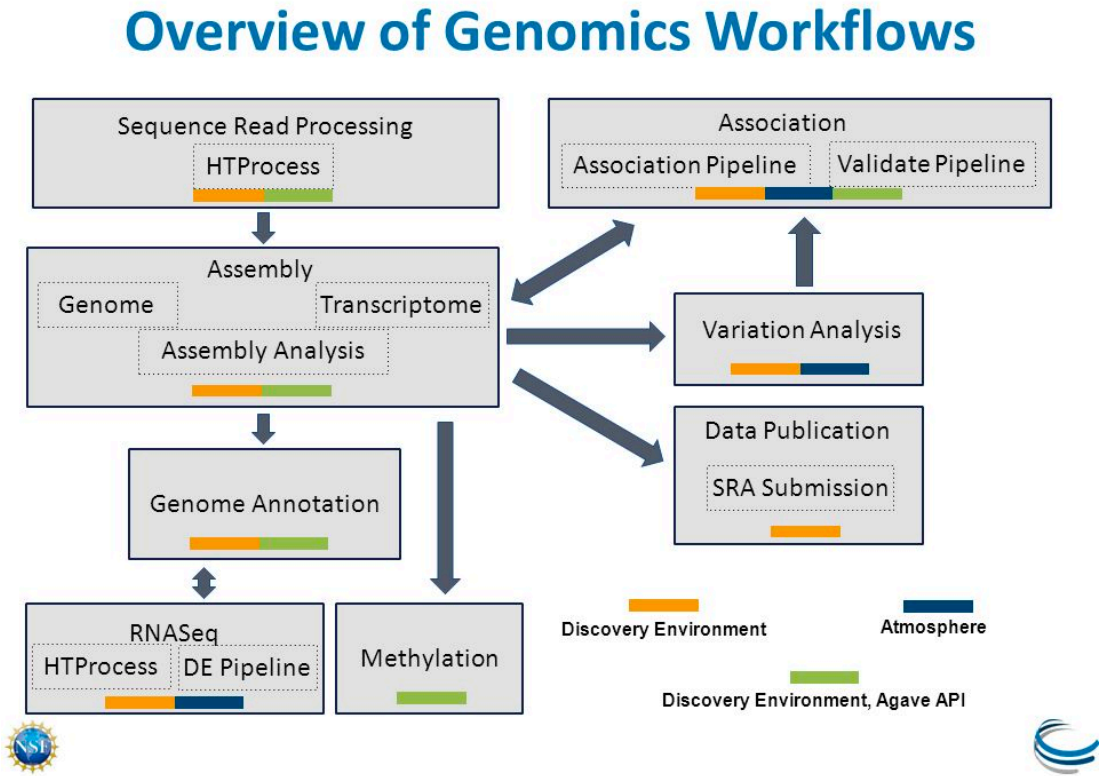
MODERN [NEURO]GENOMICS: A BRIEF OVERVIEW

Neuro-genomics

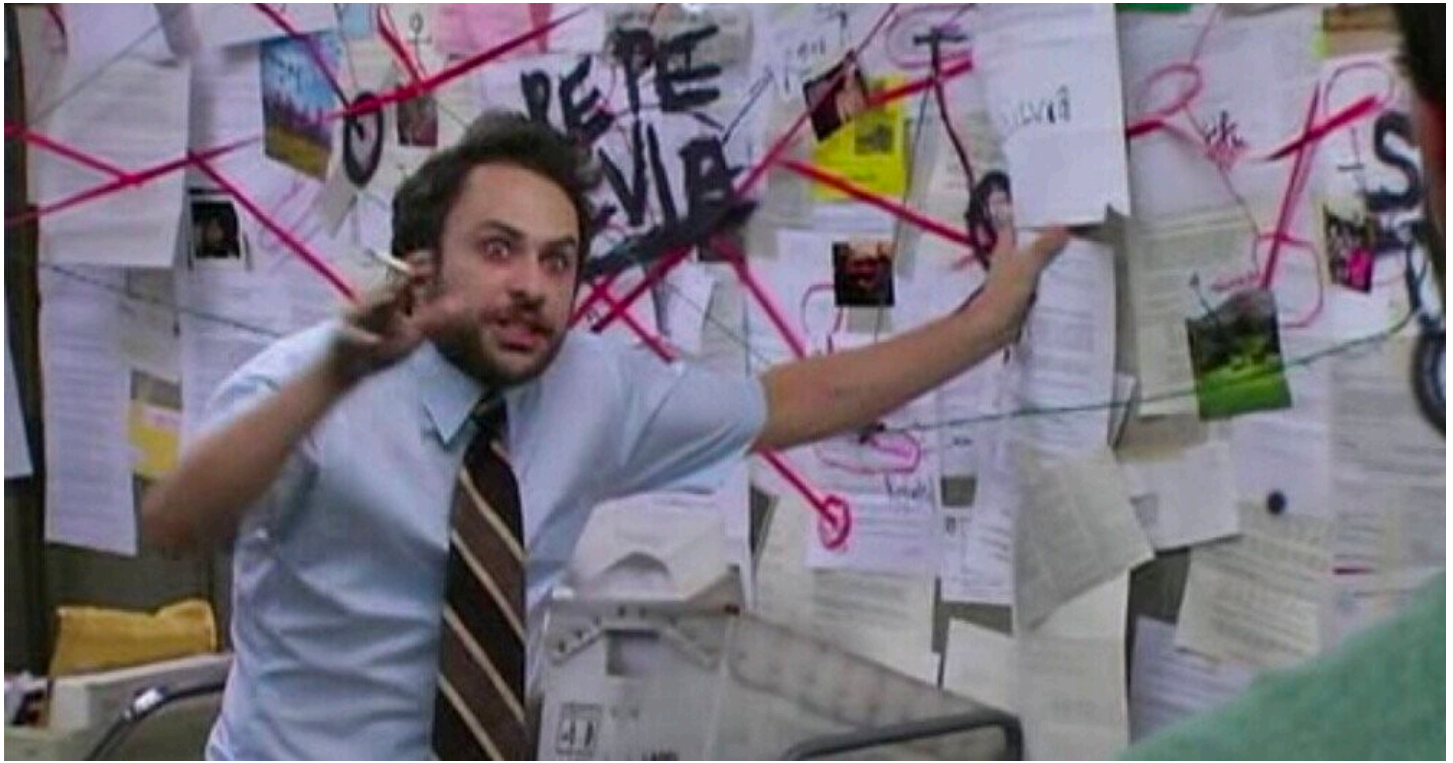
- How does the genome of an organism influence the development and function of the nervous system?
- Genome and derivatives
- This course will focus on high-throughput sequencing-based assays as applied to neuroscience questions



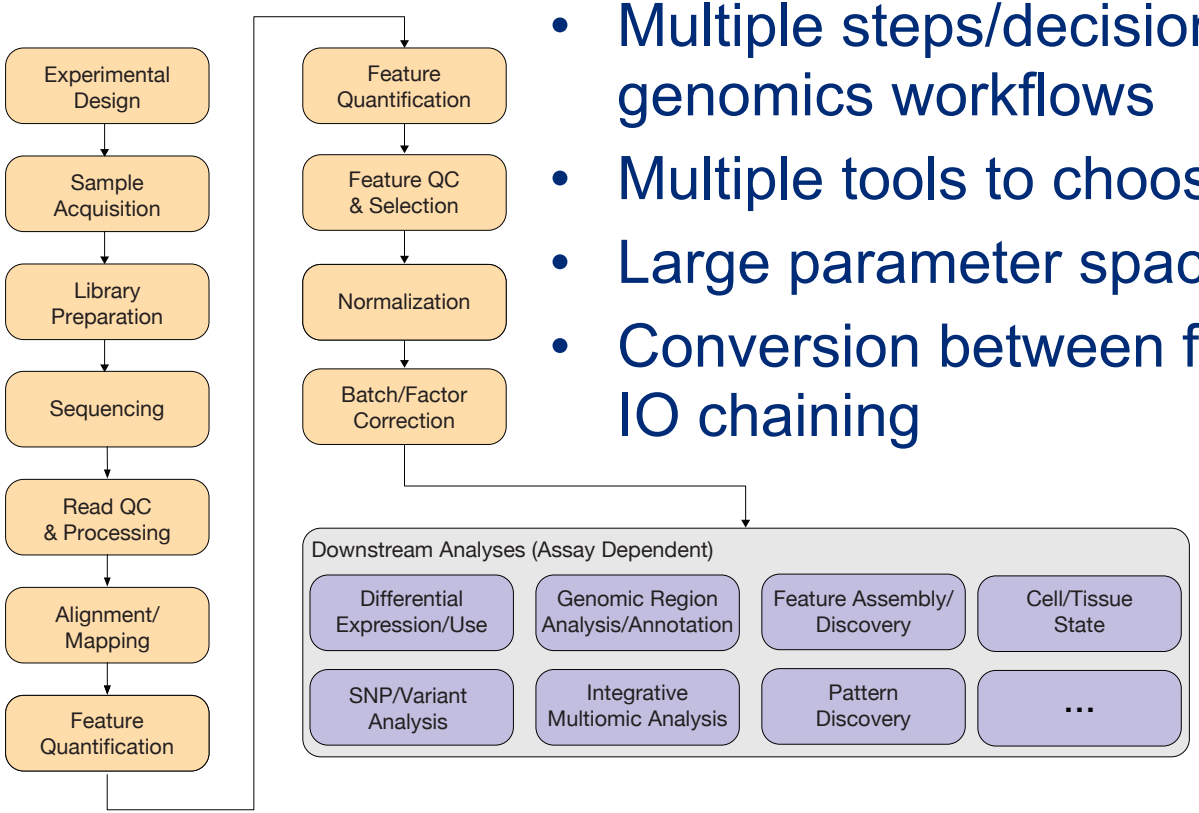
The theoretical workflow example



The Reality

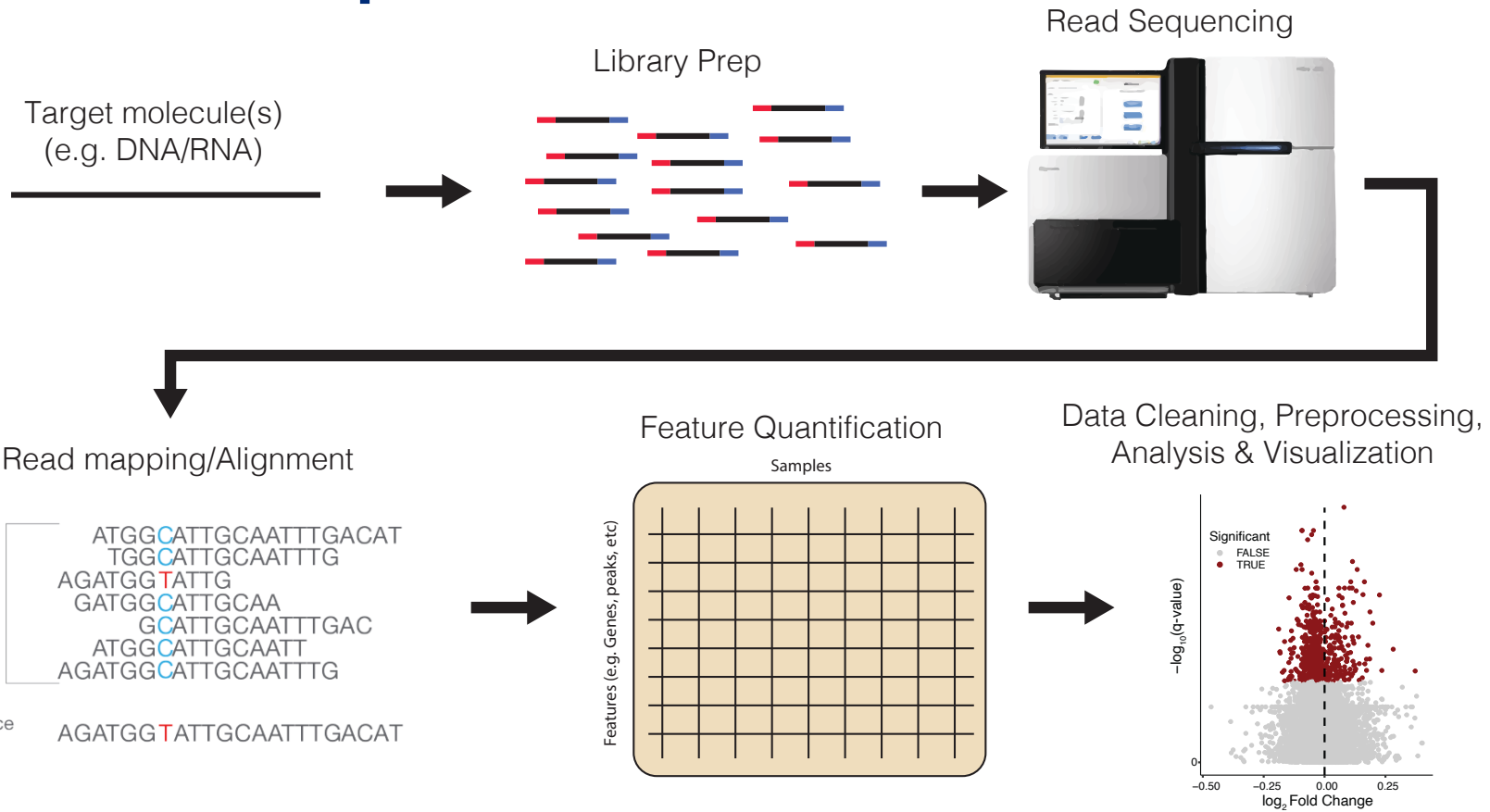


A (Grossly Oversimplified) Overview of *-omics Workflows



- Multiple steps/decisions in modern genomics workflows
- Multiple tools to choose from at each step
- Large parameter space for each tool
- Conversion between formats to facilitate IO chaining

Even more simplified



Sequencing is just a tool/assay

- Application still requires a hypothesis
  - There is still *room* for discovery-based questions, but the strongest studies still
- Many NGS assays are simply adaptations of ‘classical’ molecular biology assays to increase resolution and parallelization:
  - Diff RNA-Seq = high resolution, highly parallel Northern blot
  - ChIP-Seq = high resolution, highly parallel EMSA
  - Spatial Transcriptomics = highly parallel RNA in situ
- 100s of applications for nextgen sequencing of natural or exogenous nucleic acid sequences

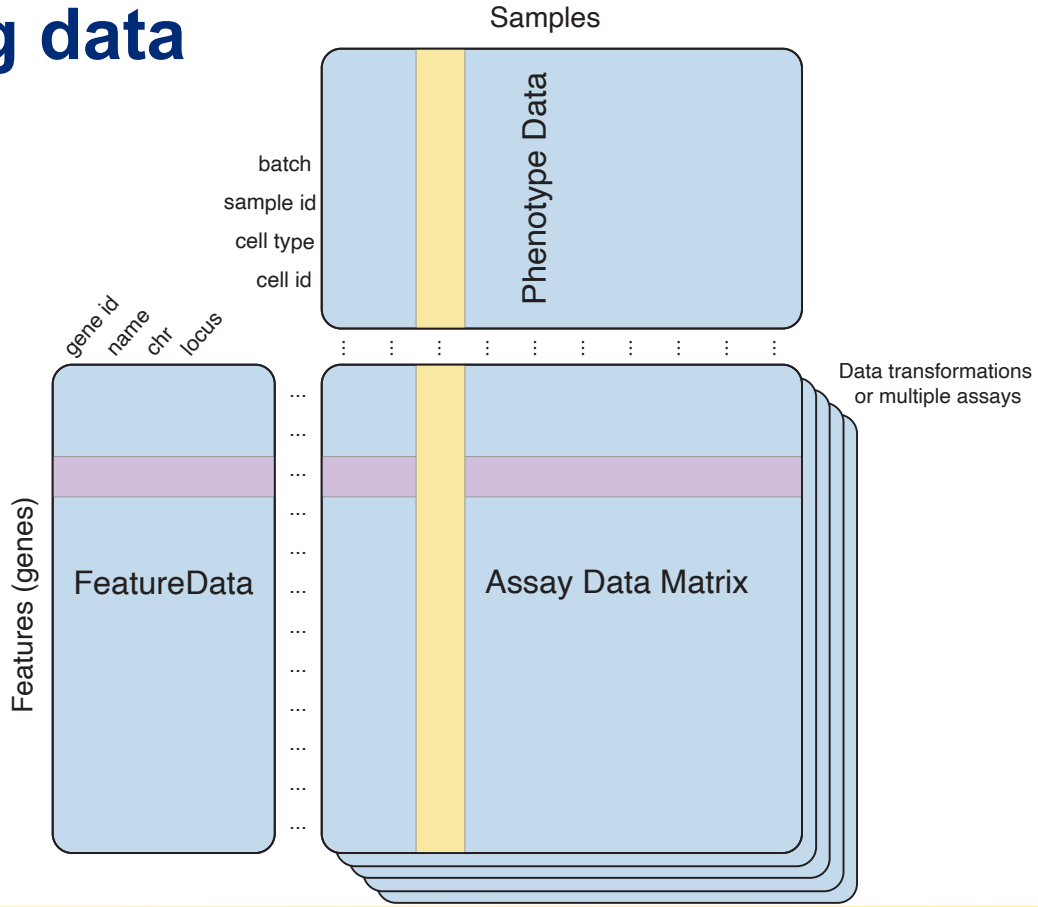
Computational Biology *is* Experimental Biology

- Do not ***always*** conflate high-throughput sequencing with ‘big data’-style science
  - Hypothesis-driven vs discovery science
- Important to define the question and understand the hypotheses to be tested
- Nature of high-throughput sequencing experiments allows for testing multiple hypotheses
  - e.g. differential analysis is equivalent to multiple independent pairwise tests between conditions (1 per feature)

NGS Sequencing Applications (Abridged)

- DNA
  - Whole Genome/Exome Sequencing
  - Targeted Sequencing
  - Hybridization Capture
  - Variant Identification
- RNA
  - Whole Transcriptome Sequencing
  - Gene expression profiling
  - Splicing & RNA dynamics
  - RNA modification analysis
- Epigenomics
  - ChIP-Seq
  - Methyl-Seq
  - Other DNA base modifications

A common data structure for high-throughput sequencing data



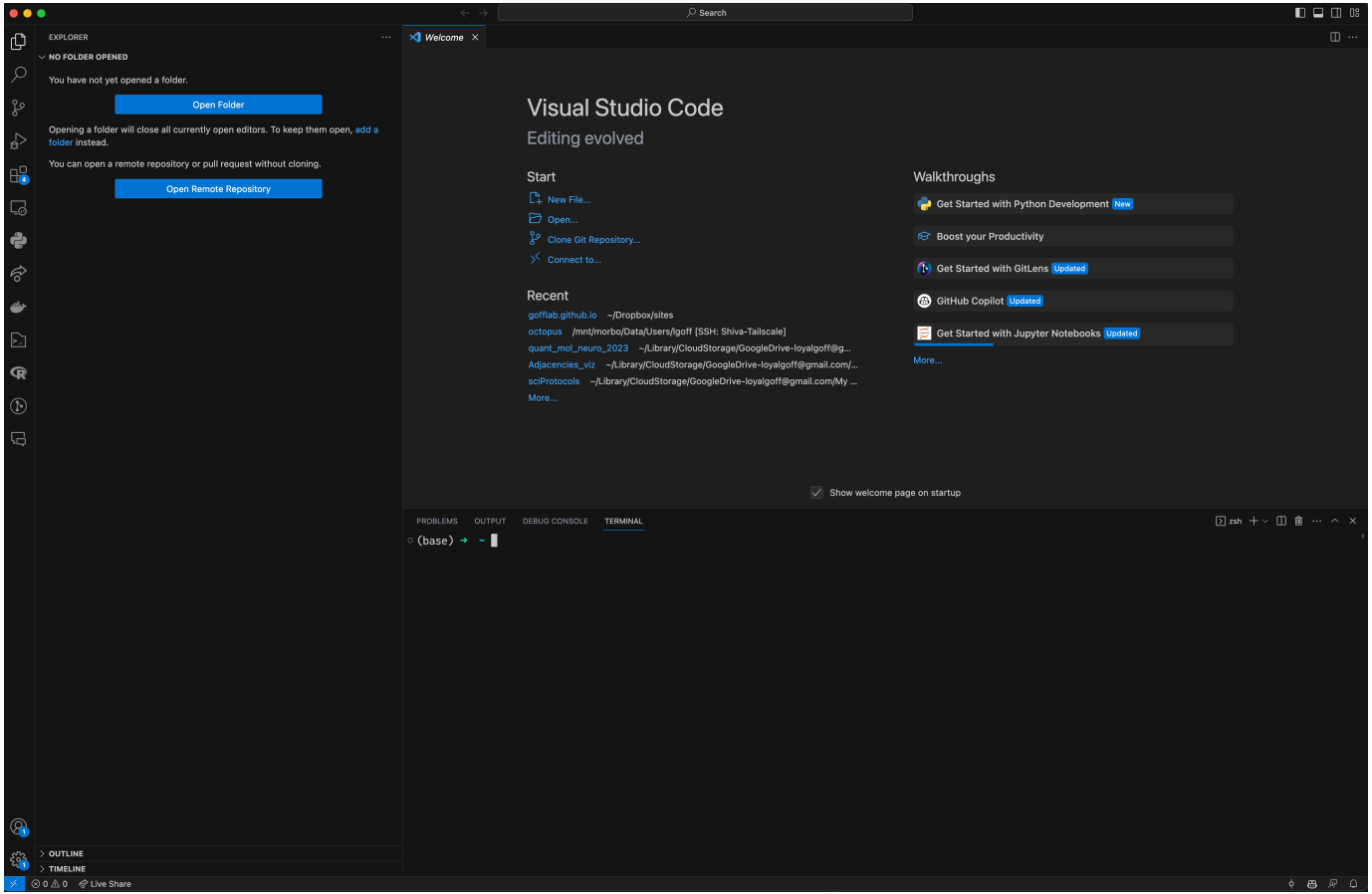
External resources

- https://portlandpress.com/biochemist/article/43/6/58/229924/Beginner-s-guide-to-next-generation-sequencing
- https://www.illumina.com/areas-of-interest/neurogenomics.html
-

JOHNS HOPKINS
SCHOOL of MEDICINE

INTRO TO VSCODE, TERMINAL, AND PYTHON

VSCode - Integrated Development Environment (IDE)



JOHNS HOPKINS
SCHOOL of MEDICINE

PROBLEM SET 1

Questions?