Data analysis
Mayo-Illinois Computational Genomics Course
June 8, 2020
Introduction
Objective

To learn to learn to visualize, analyze genomic data
genomic data

Data analysis is the iterative process of advancing scientific theory using quantitative data.
To learn to learn to

Learn key concepts rather than specific implementations.

“I tell my students, ‘the language in which you’ll spend most of your working life hasn't been invented yet, so we can't teach it to you. Instead we have to give you the skills you need to learn new languages as they appear.’”

Brian Harvey
https://people.eecs.berkeley.edu/~bh/sicp.html
Genomic data analysis workflow

1. Experimental design
2. Quality control
3. Preprocessing
4. Analysis
5. Biological interpretation
Today’s topics

1. Experimental design
2. Quality control
3. Preprocessing
4. Analysis
   • Statistics
     • Framework
     • Methods
   • R
5. Biological interpretation

• Illustration

Statistics
Statistics provides a mathematical framework and a set of methods for drawing inferences from quantitative data.
Framework

**Theory**

Question → Formulation → Population distribution function → Inference (uncertainty) → Sampling (randomness) → Data → Interpretation → Answer

- **Question**
- **Formulation**
- **Theory**
- **Answer**
- **Inference (uncertainty)**
- **Sampling (randomness)**
- **Data**

**Framework**

- **Interpretation**
Methods

What they do:

• Supervised learning
  • Testing
  • Estimation
  • Prediction

• Unsupervised learning
  • Clustering
  • Dimension reduction

• Visualization

• Ex. Alzheimer’s disease GWAS
• Supervised (AD status)
  • Is a SNP associated with AD?
  • How is a SNP associated?
  • Given SNPs, predict AD status.

• Unsupervised
  • What are the subpopulations?
  • Reconstruct ancestry scores.
Methods

What they are:

- Classical
- Standard (introductory)
- Standard (advanced)
- New
Methods

How to choose: match analysis task with data structure


Rosner (2015)
R language

- R is a programming language for data analysis and visualization
- R language = **objects** (data) and **procedures** (manipulate objects)
- R expression = valid combination of objects and procedures
- R script = sequence of R expressions (at most one per line)
- R package = community-contributed R procedures and scripts

Example script:
```
file <- "~/data/GSM2818521_larva_counts_matrix.txt"
pandey <- read.table(file, header = TRUE)
dim(pandey)
```
R interpreter

An (R) interpreter is a computer program that:
1. Reads an (R) expression
2. Evaluates the expression
3. Prints the result
4. Loops

Ignores lines starting with # (comments)
RStudio

• RStudio is a development environment, i.e., a program that makes writing R scripts easier

• Panes
  1. Source
  2. Console
  3. Environment
  4. Output
RStudio customization
Example data

Current Biology

Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq

Highlights
- Establishment of a single-cell RNA-seq protocol for neurons in zebrafish
- Identification of 18 distinct habenular types
- Retention of neuronal types between larva and adult
- RNA-FISH and image registration to build a gene expression atlas

Authors
Shristi Pandey, Karthik Shekhar, Aviv Regev, Alexander F. Schier

Correspondence
p.shristi@gmail.com (S.P.), schier@fas.harvard.edu (A.F.S.)

In Brief
Pandey et al. use scRNA-seq to define more than a dozen different neuronal types in the zebrafish habenula. Cell types are retained between larva and adult.
Theory

What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism?

- What Is an Adult Cell Type, Really?
- Defining Cell Type Space
- Cellular Demographics, Recorded

Hans Clevers
Hubrecht Institute

Suzanne Rafelski
Allen Institute for Cell Science

Michael Elowitz
Caltech

The human body is home to hundreds of cell types. Some are rather unobtrusive; others,

Canonical cell types, e.g., muscle and nerve, were originally defined by the functions of

It seems to me that we are at the beginning of a paradigm shift on the issue of cell type.

“Characterize” cell types in larval habenula
Genomic data analysis workflow

1. Experimental design
2. Quality control
3. Preprocessing
4. Analysis
5. Biological interpretation
Prepare R

Analysis will be implemented using the R package Seurat v3.1.

library("Seurat")

## set random seed for reproducibility
set.seed(1)

s_obj <- CreateSeuratObject(pandey)
2. Quality control

s_obj <- PercentageFeatureSet(s_obj,
    pattern = "^MT-",
    col.name = "percent.mito")

VlnPlot(s_obj,
    features = c("nCount_RNA",
                  "nFeature_RNA",
                  "percent.mito"))

s_obj <- subset(s_obj, percent.mito <= 5 & nCount_RNA <= 2e4)
3. Preprocessing

s_obj <- NormalizeData(s_obj)
s_obj <- FindVariableFeatures(s_obj)
s_obj <- ScaleData(s_obj, vars.to.regress = c("nCount_RNA"))
s_obj <- RunPCA(s_obj)
Iteration 1: Question

What are the different cell types in the larval habenula?
Iteration 1: Analysis

Identify cell types

• Task:
  • Unsupervised
    • Clustering

• Method:
  • SNN clustering

\[
s_{obj} \gets \text{FindNeighbors}(s_{obj})
\]
\[
s_{obj} \gets \text{FindClusters}(s_{obj}, \text{resolution} = 0.5)
\]

Determines number of clusters
Iteration 1: Analysis

Visualize cell types

• Task:
  • Unsupervised
    • Dimension reduction
  • Visualization

• Method:
  • UMAP projection
  • Scatterplot

\[
s_{\text{obj}} \leftarrow \text{RunUMAP}(s_{\text{obj}}, \text{dims} = 1:20) \\
\text{DimPlot}(s_{\text{obj}})
\]
Iteration 1: Answer
Iteration 2: Question

Which genes are substantially differentially expressed between the cell types?

Formulation
Iteration 2: Analysis

Identify genes

• Task:
  • Supervised
    • Testing
    • Estimation

• Methods:
  • Wilcoxon, FDR
  • Mean

```r
markers = FindAllMarkers(s_obj, logfc.threshold = 1.5)

markers = markers[markers$p_val_adj <= 0.05]

head(markers)
```

“Substantial”

“Differential”
Iteration 2: Analysis

Visualize genes

- Task:
  - Visualization

- Methods:
  - Scatterplot

```r
FeaturePlot(s_obj, features = c("G0S2", "TP53I11B", "FXYD1")) +
  patchwork::plot_layout(ncol = 3)
```
Iteration 2: Answer
Iteration 3: Question

In what biological processes are the differentially expressed genes involved?

Formulation
Iteration 3: Analysis

Identify processes

• Task:
  • Supervised
  • Testing

• Methods:
  • Fisher’s exact test

```r
gene_names = unique(markers$gene)
length(gene_names)
cat(gene_names, sep = "\n")
```

https://david.ncifcrf.gov/
Iteration 3: Answer

https://david.ncifcrf.gov/
5. Biological interpretation?

Larval habenula cells are mostly distinguished by whether they are neurons or not (?).
Conclusions
Next steps

- [https://astrobiomike.github.io/about/](https://astrobiomike.github.io/about/):
  1. Fundamentals and concepts are important, not details.
  2. Don’t let yourself become paralyzed by options.
  3. Try to find a bioinformatics community to be a part of.
  4. Good documentation is for science, you, and the community.
  5. Be aware that you will often need to let some things go.

- Simple statistical methods can go a long way:
  枯れた技術の水平思考
Thank you!