Introduction to Machine Learning

Plan for today

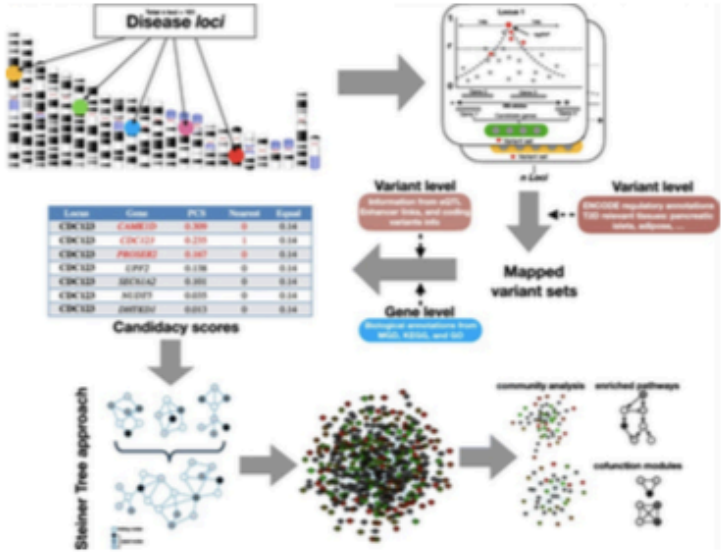
- Introduction to data analysis
- Data wrangling and getting you used to colab
- Unsupervised learning – PCA
- Unsupervised learning – exercises , other options

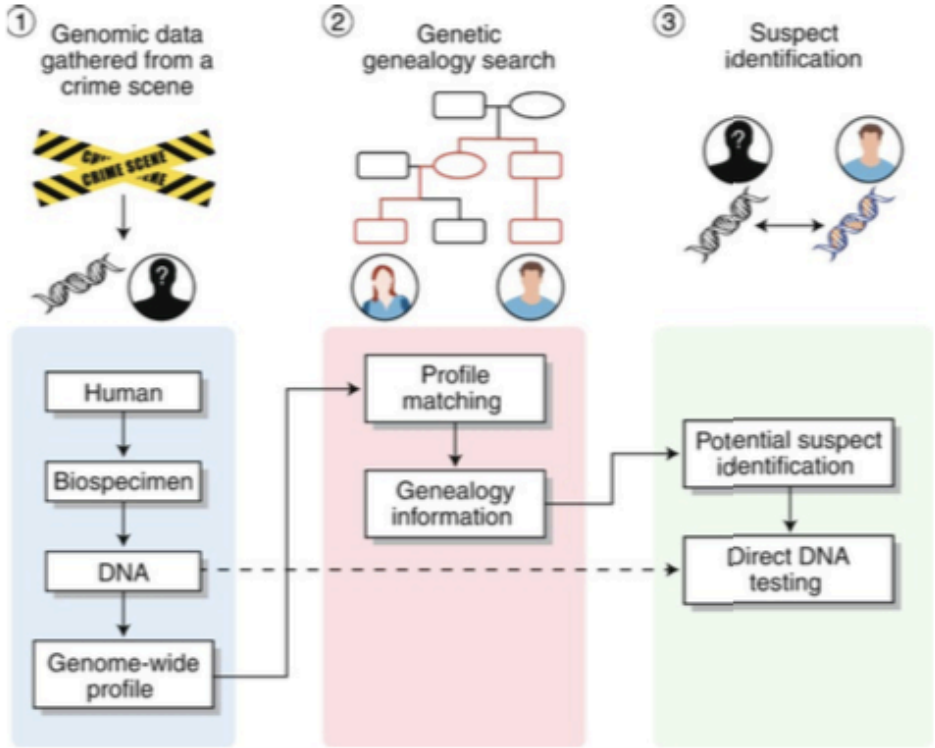

Input
Data
Black
box
analysis
Results/
Predictions

Parts of a study

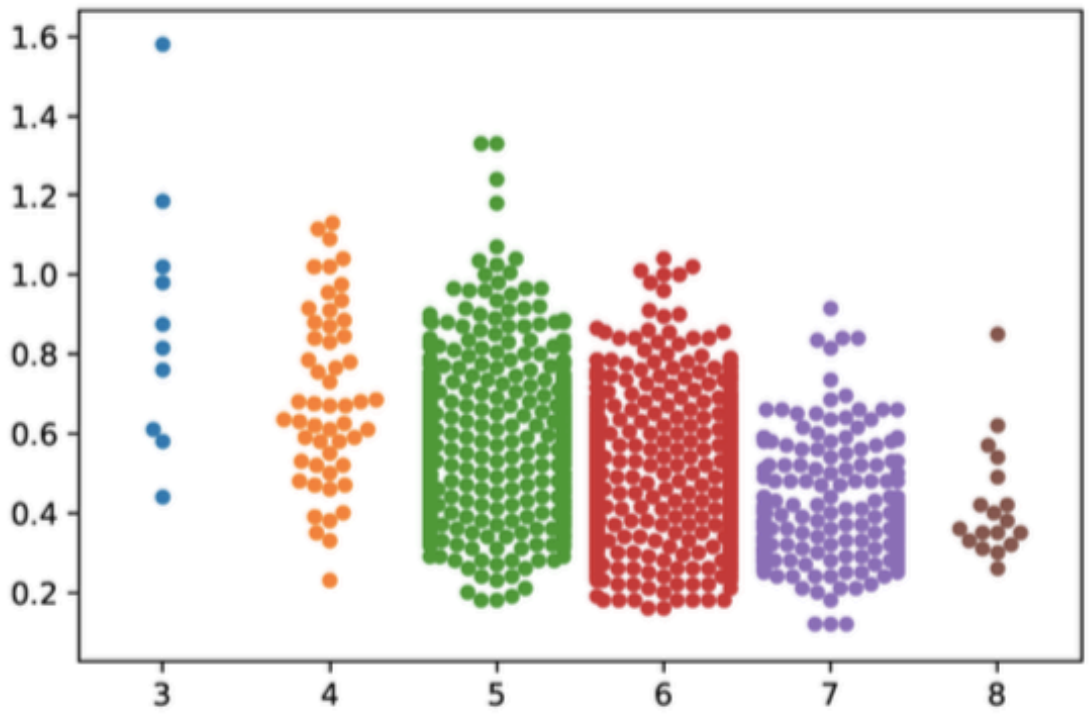
- Study design
- Collect data
- Clean data – exploratory analyses
- Analyse your data
  - Choose a technique
  - Evaluate the model
  - Tune parameters
  - Predict

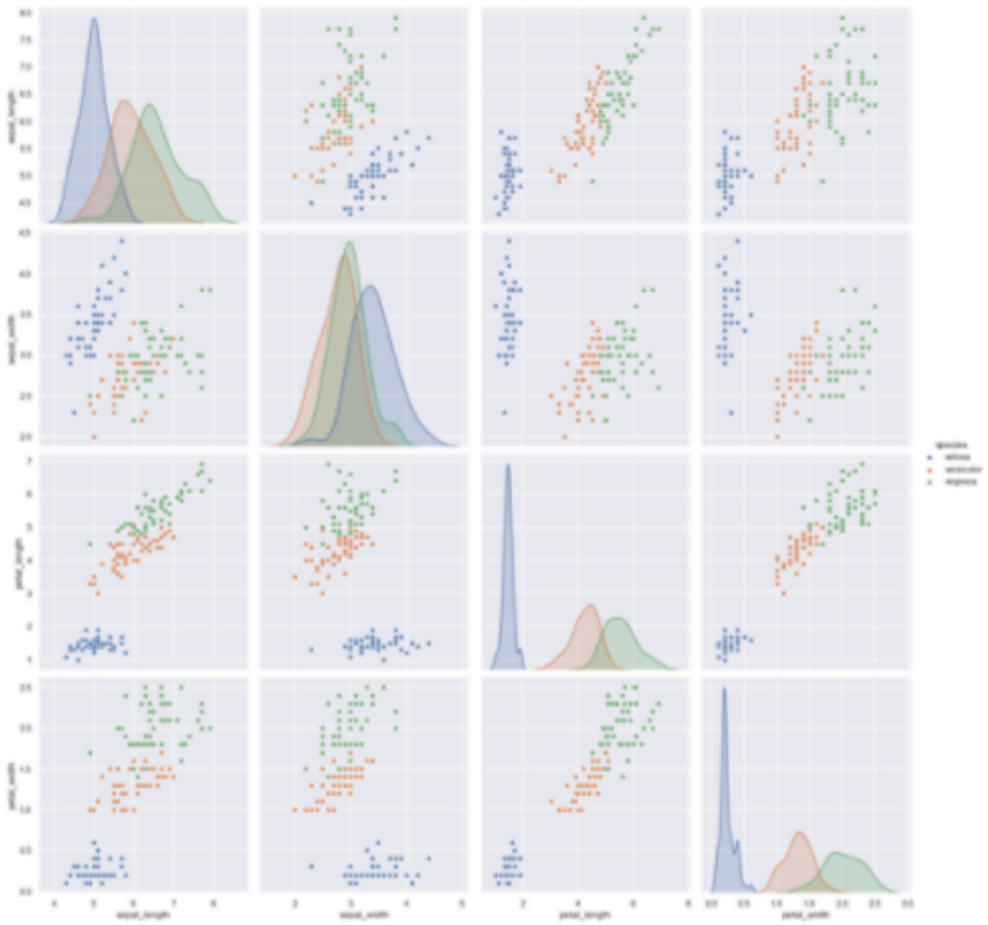
Collect data



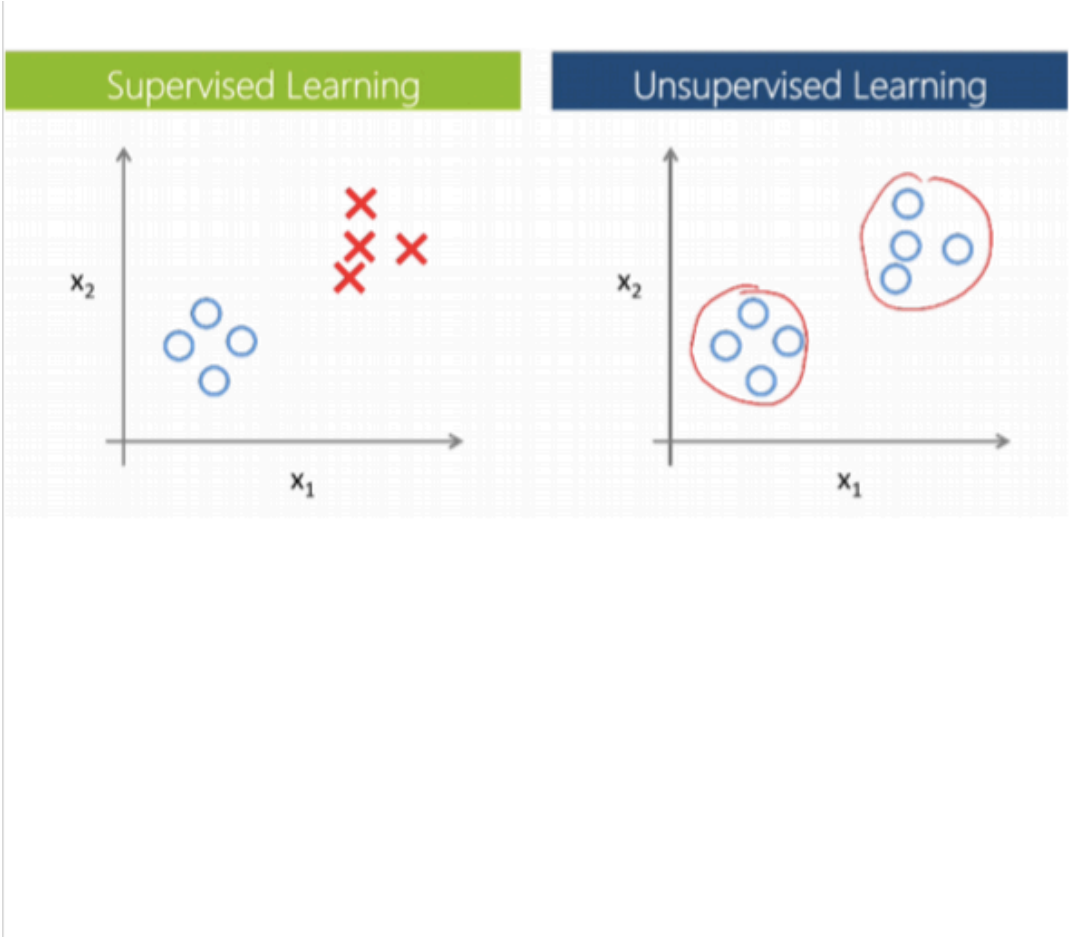


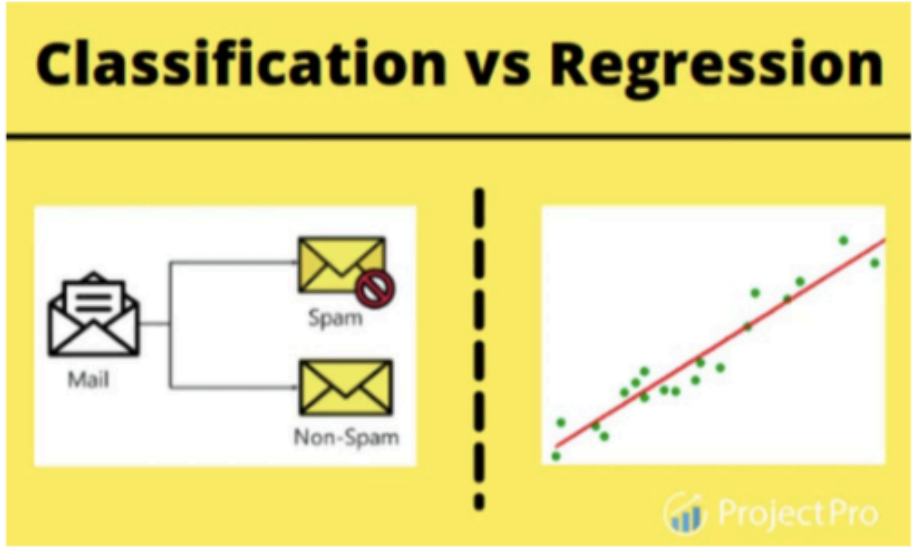
Cleaning/Exploring your data



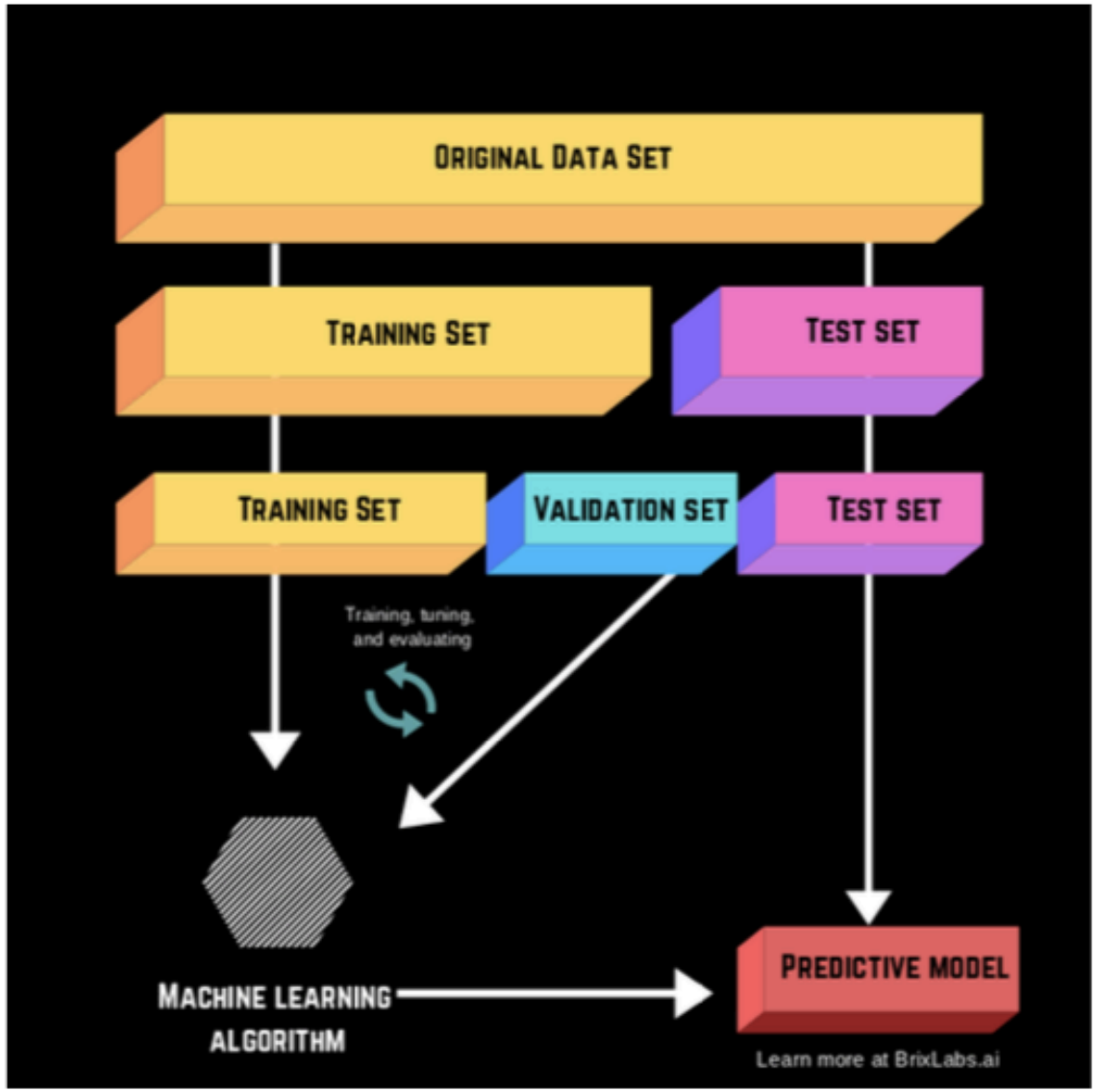


Model/Technique: Select a question to answer





Train, validate, test, repeat



Splitting your dataset: AVOID Overfitting!

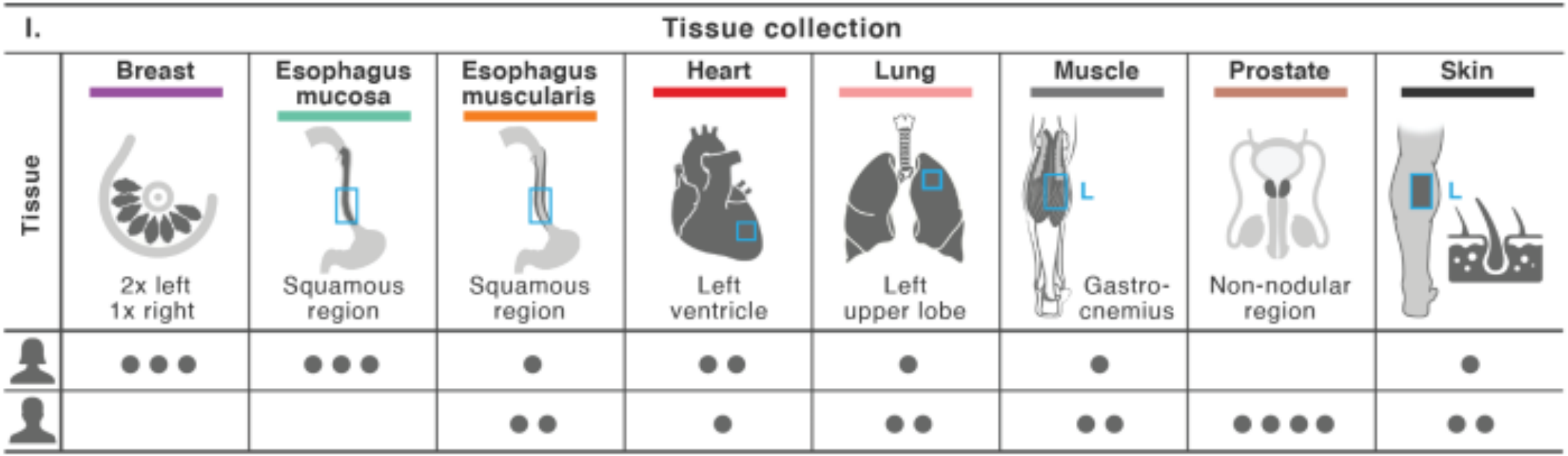
- Train your model parameters
- Validate your model

Test your model on an independent dataset

Real life example

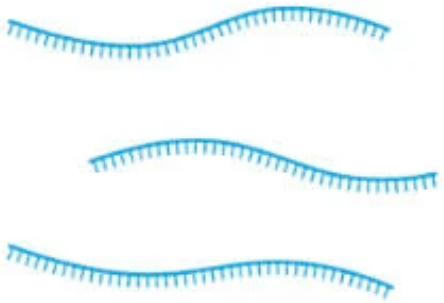
- We are going to look at these steps in a real life example
  - Gene expression studies from multiple tissues in humans: GTEx dataset

RNA-Seq to collect expression



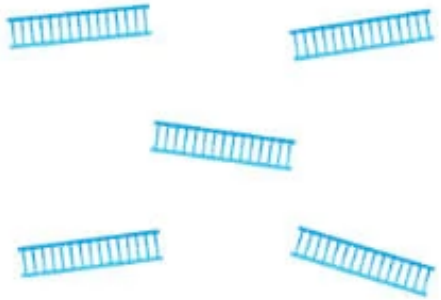
RNA Sequencing

1 Isolate RNA from samples

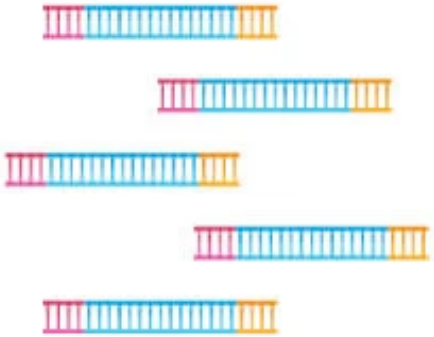


2 Fragment RNA into short segments

3 Convert RNA fragments into cDNA

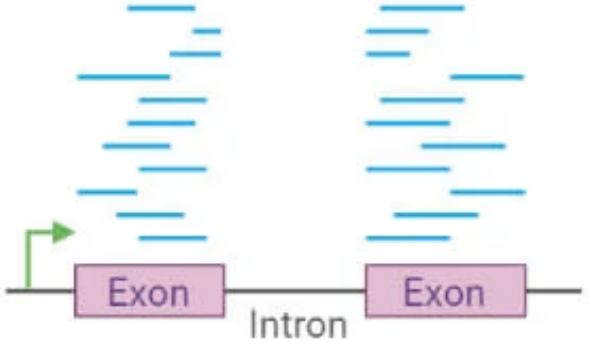


4 Ligate sequencing adapters and amplify

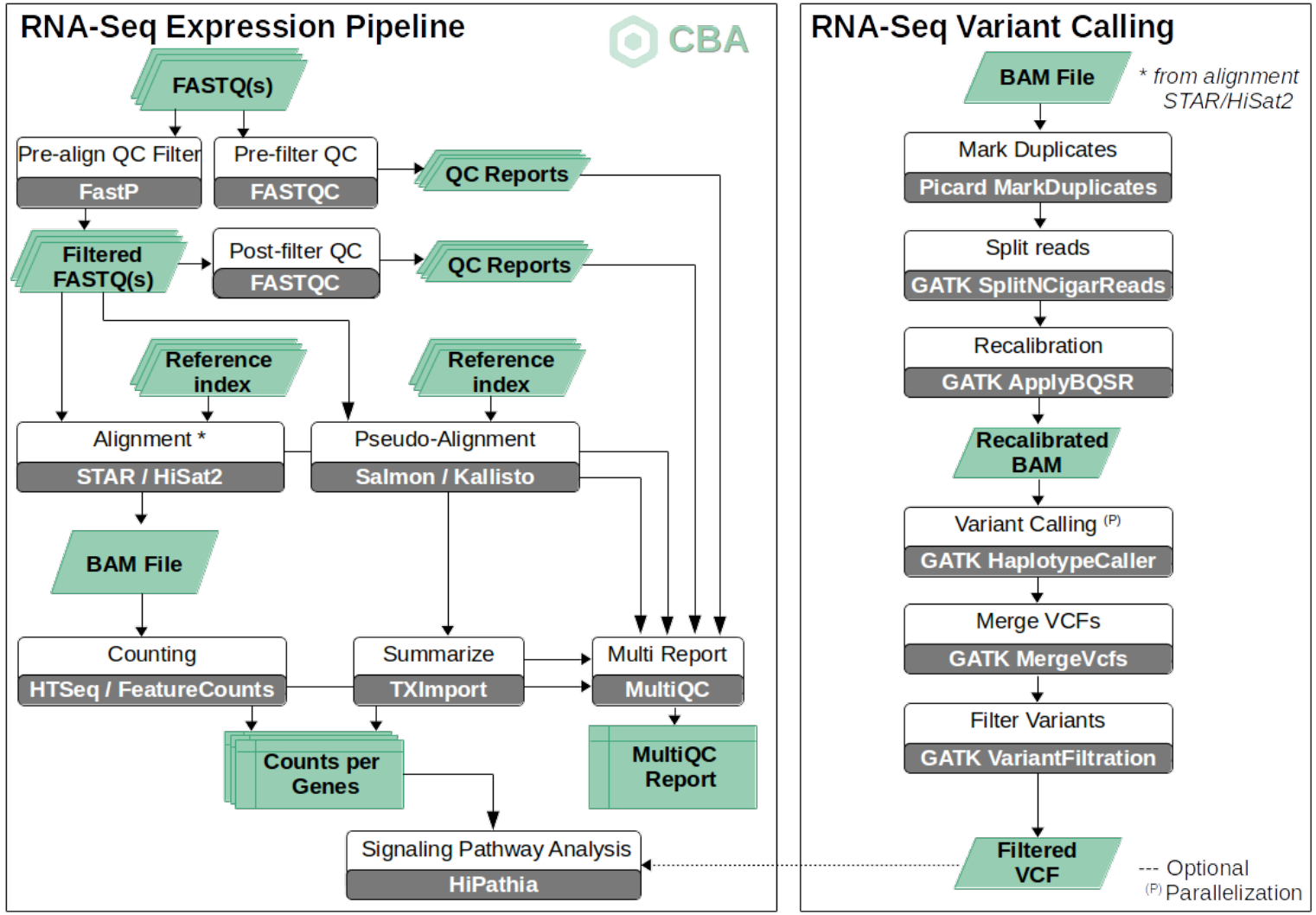


5 Perform NGS sequencing

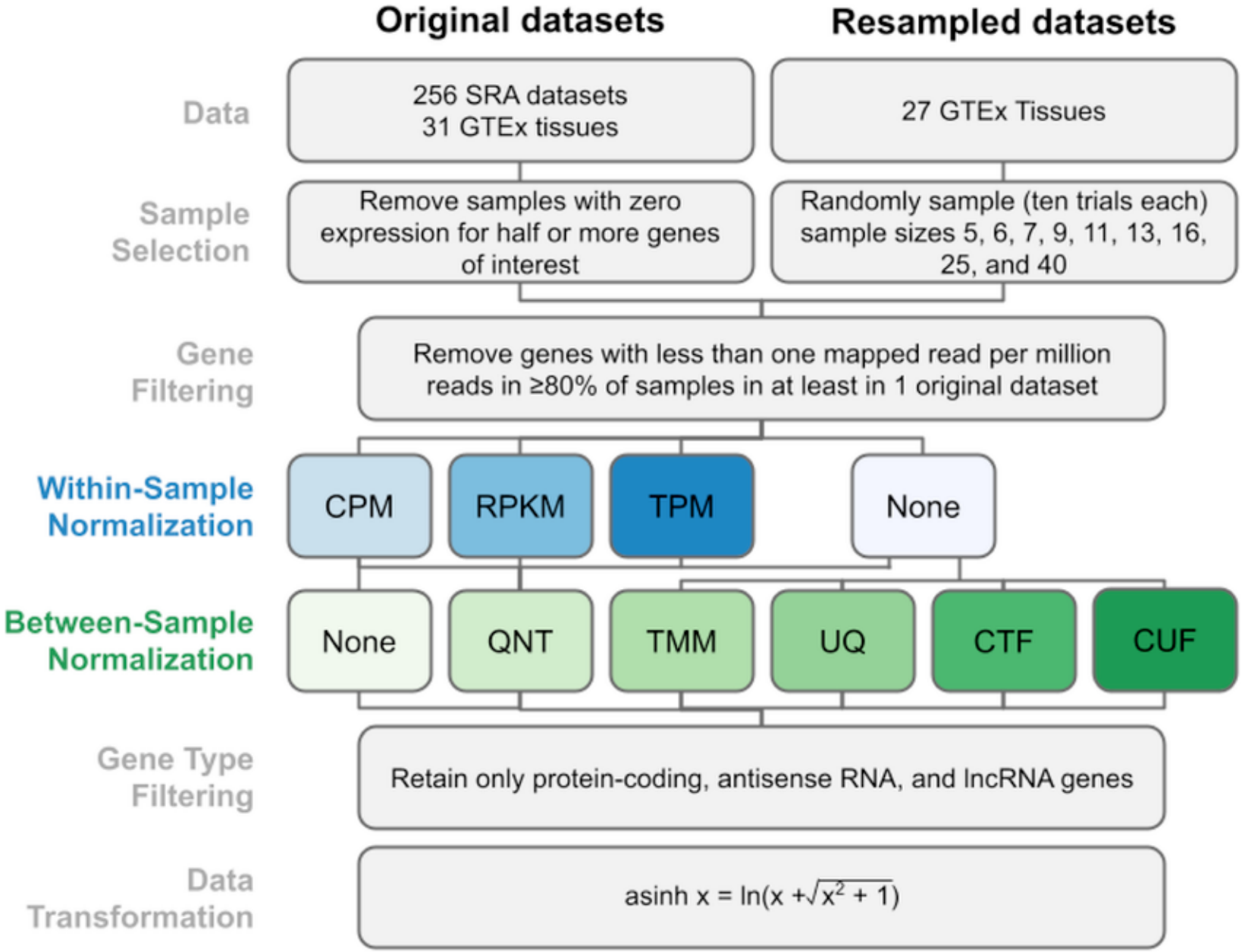
6 Map sequencing reads to the transcriptome/genome



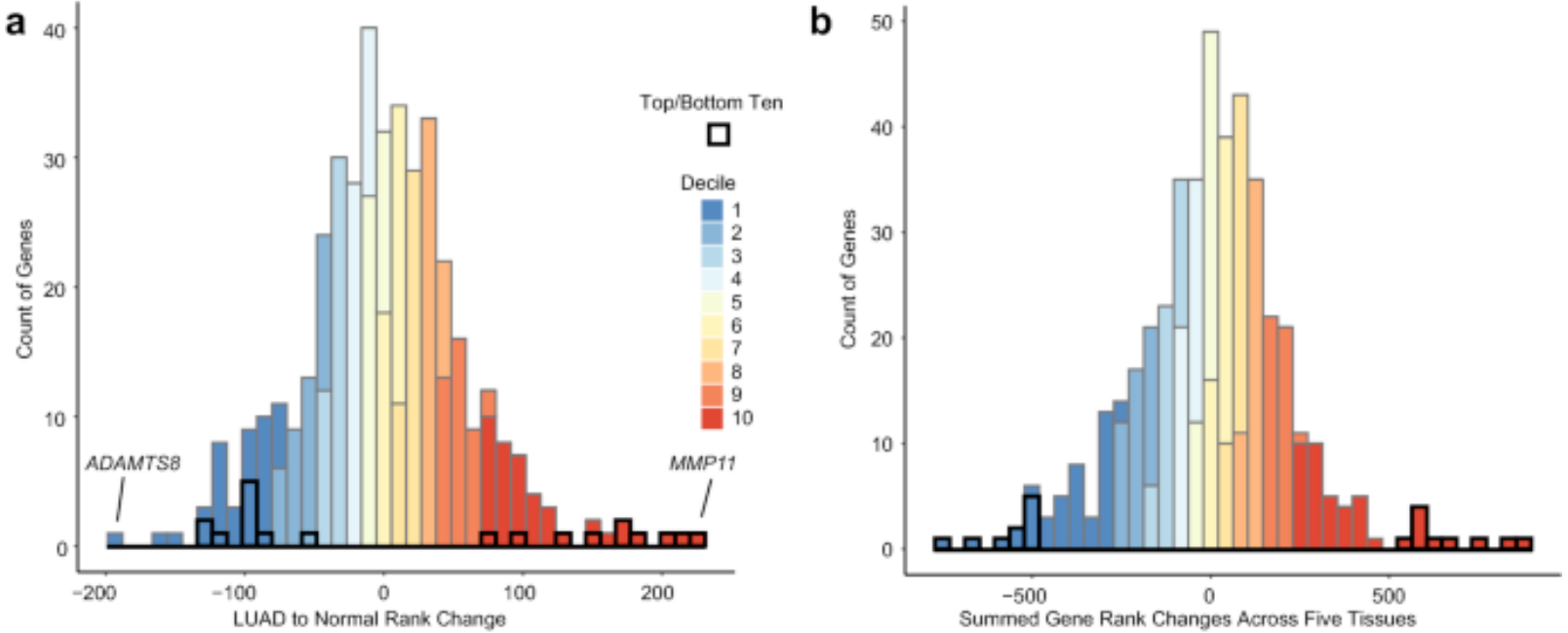
Initial processing - bioinformatics

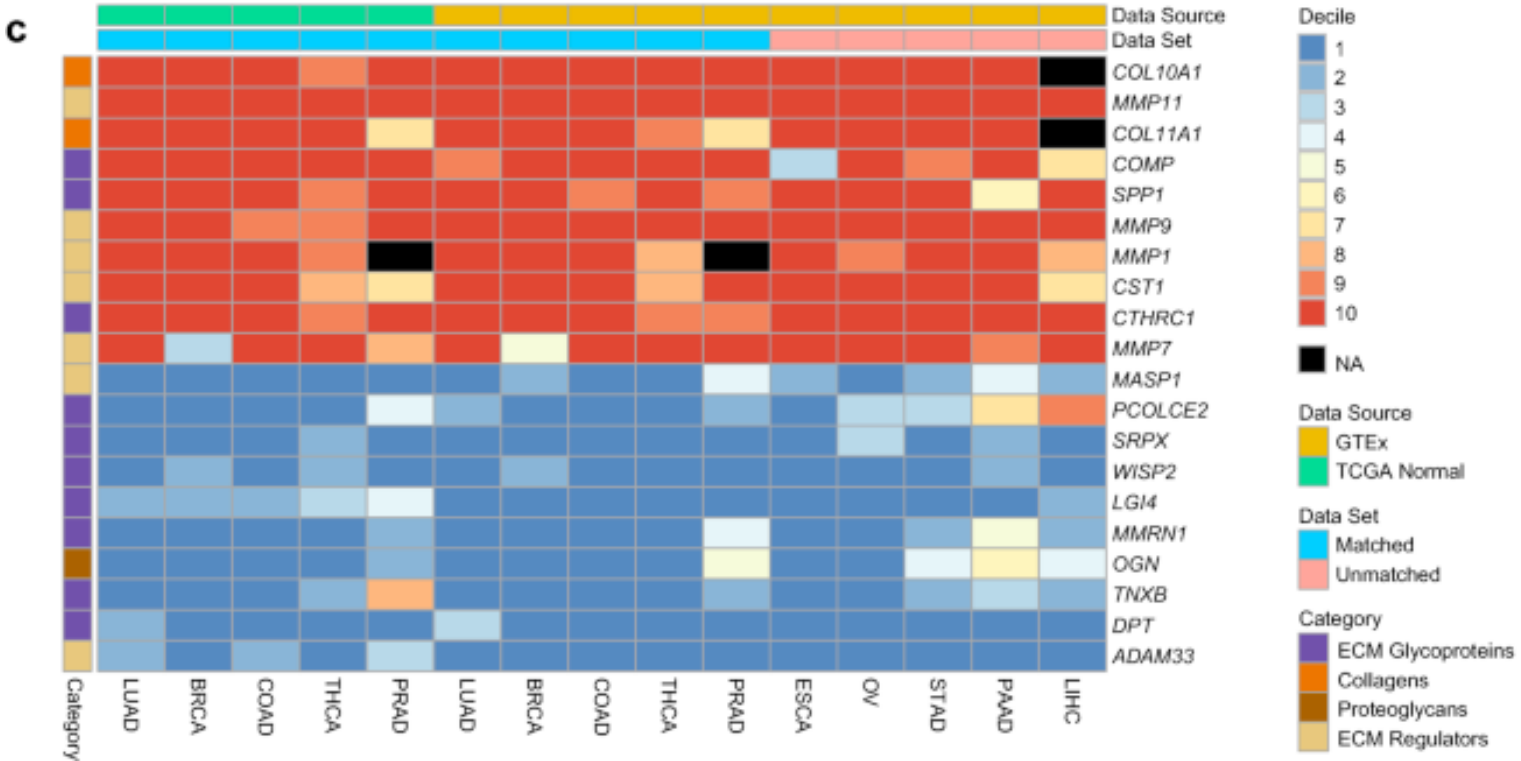


Data cleaning



Data exploration



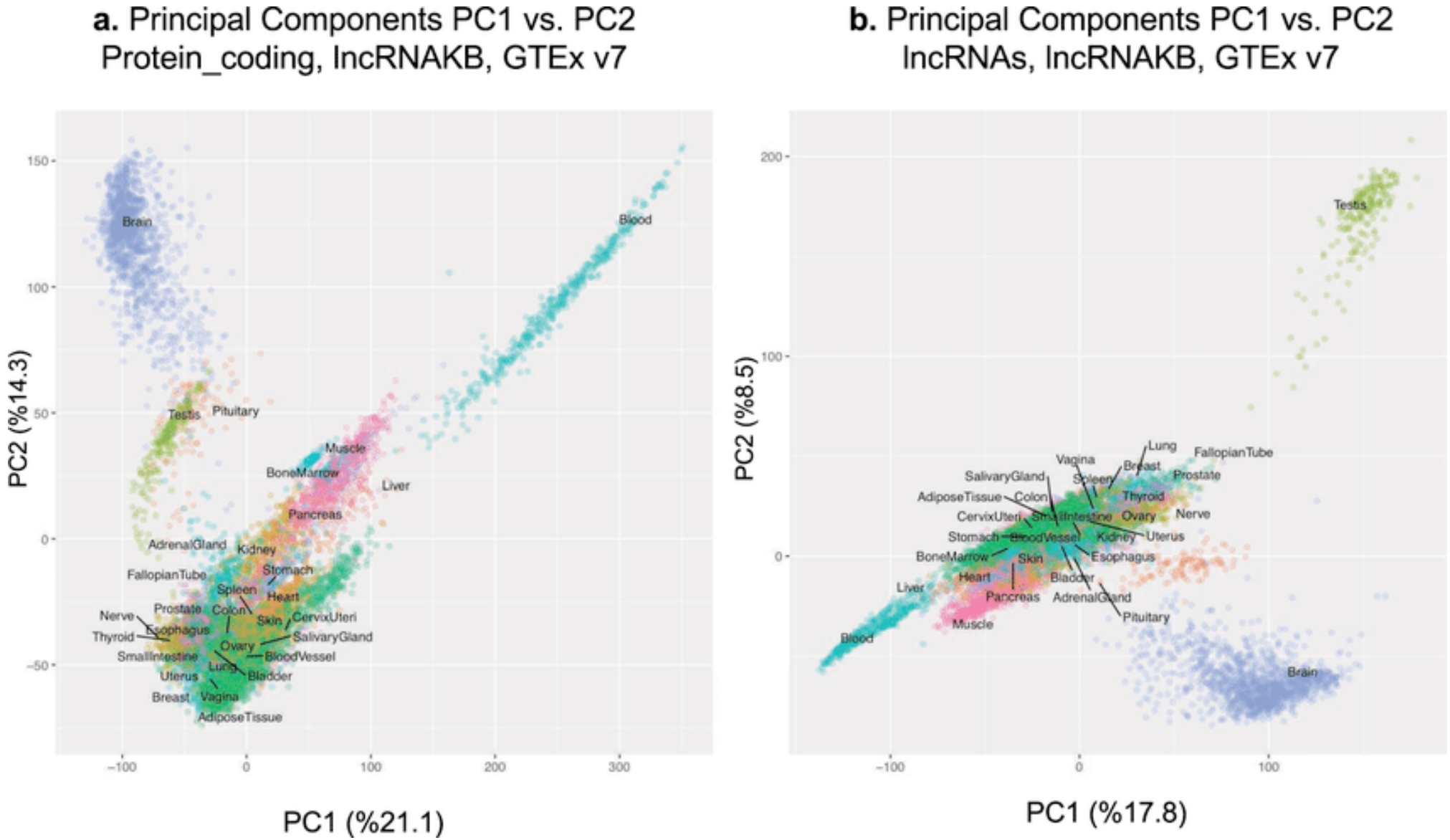


Choose your technique

- Pick the question you want to answer

Does expression show tissue specificity?

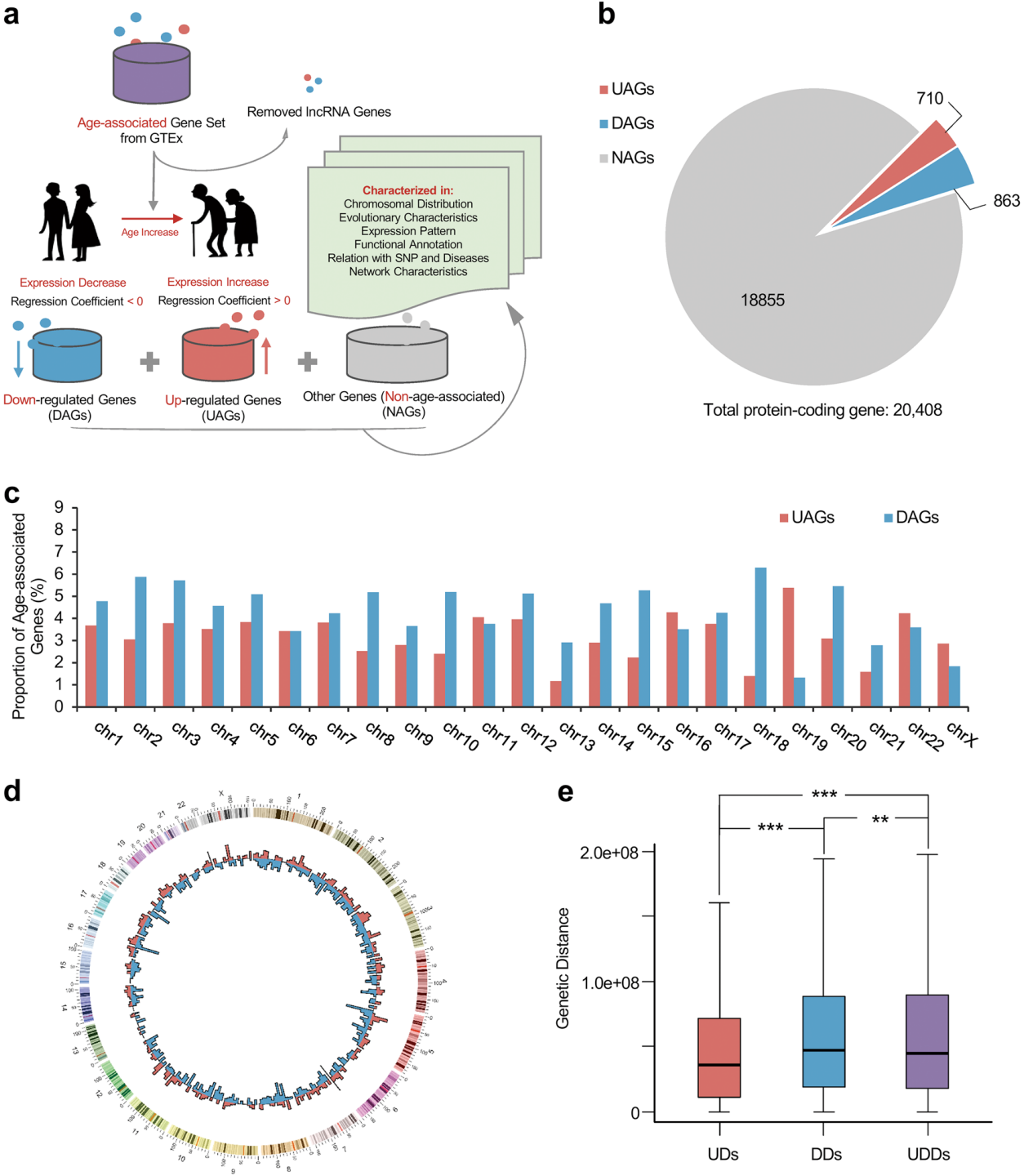
Unsupervised learning: PCA



Another question?

- Do aging related genes show differences in expression across tissues?

Differential expression – supervised learning (regression)



A technique for every study

- Your dataset is unique
- Choose your methods based on your questions
- Explore a bunch of methods to see what is useful and relevant for your study
  - Remember the assumptions

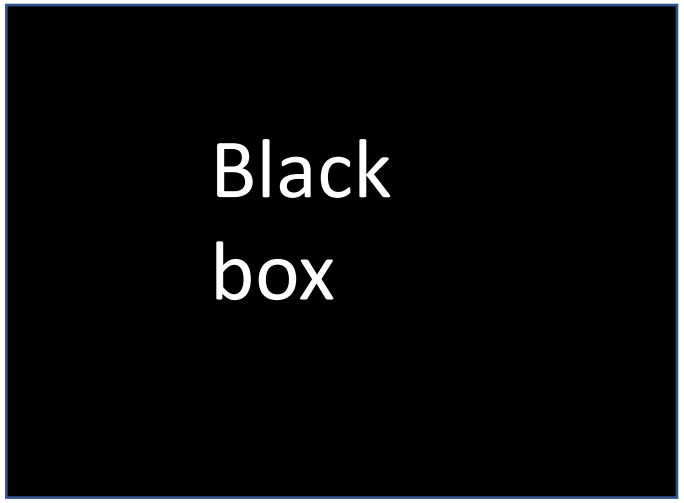
Deconstructing the black box of analysis

What does the black box do?

- Transform the data in some way
- Your question dictates what aspect of the data you want to understand/preserve and which technique is useful
  - Structure in the data – clustering, PCA …
  - Relation between two variables – regression

Deconstructing the black box

Data



**Results/
Predictions**

Deconstructing the black box

Data

Define loss/cost function
- Function(Data, parameters)

Estimate parameters to minimize cost function, subject to constraints

Results/ Predictions

Deconstructing the black box

Define loss/cost function
- Function(Data, parameters)

Estimate parameters to minimize cost function, subject to constraints

Cost function examples:
Classification- mismatch rate
Regression- estimation error
Clustering- within-cluster heterogeneity