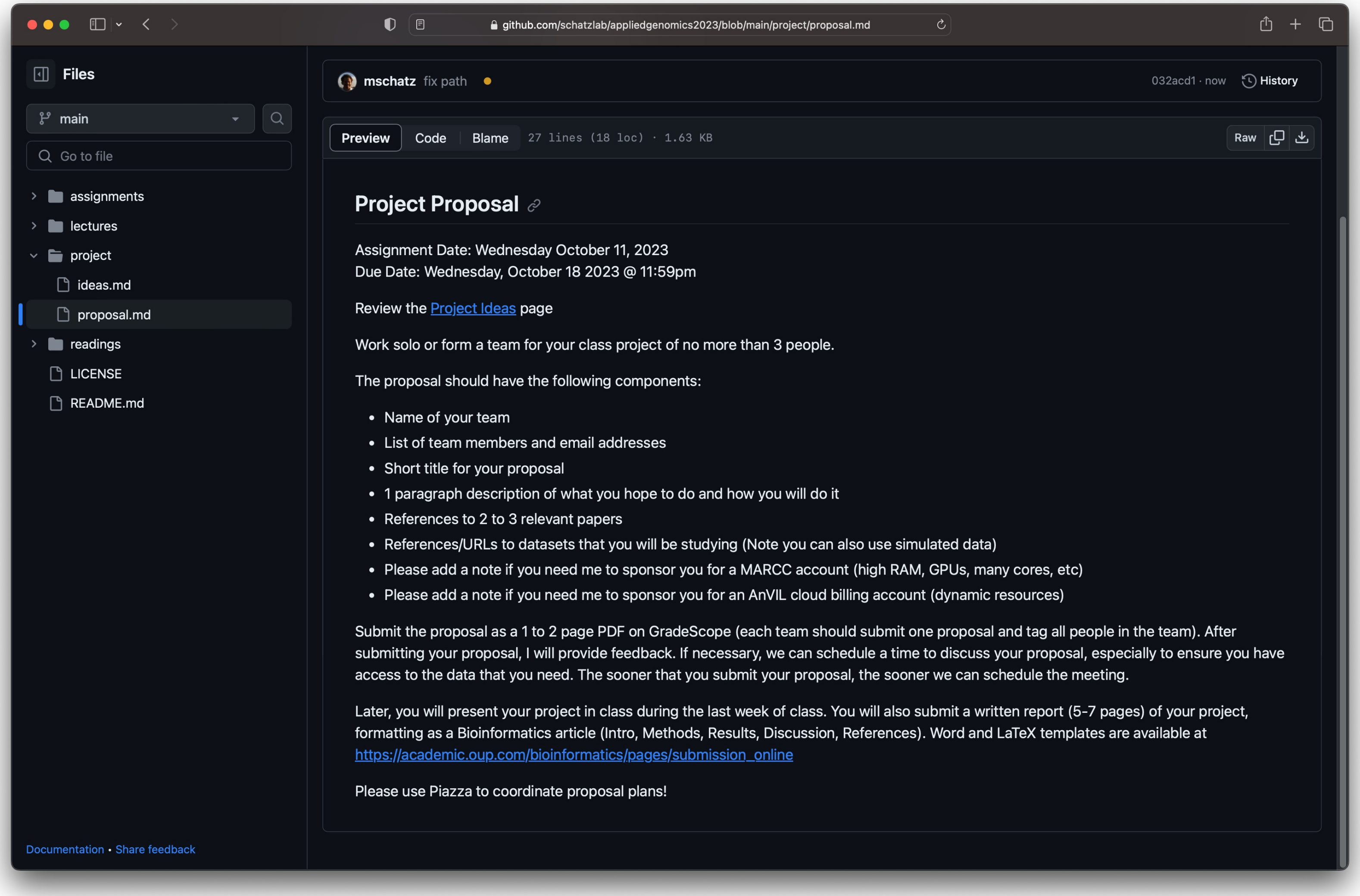
Annotation & ML

Michael Schatz

October 18, 2023

Lecture 15. Applied Comparative Genomics

Project Proposal
Due Wednesday Oct 18 by 11:59pm



Project Proposal

Assignment Date: Wednesday October 11, 2023
Due Date: Wednesday, October 18 2023 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project of no more than 3 people.

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc)
- Please add a note if you need me to sponsor you for an AnVIL cloud billing account (dynamic resources)

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team should submit one proposal and tag all people in the team). After submitting your proposal, I will provide feedback. If necessary, we can schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting.

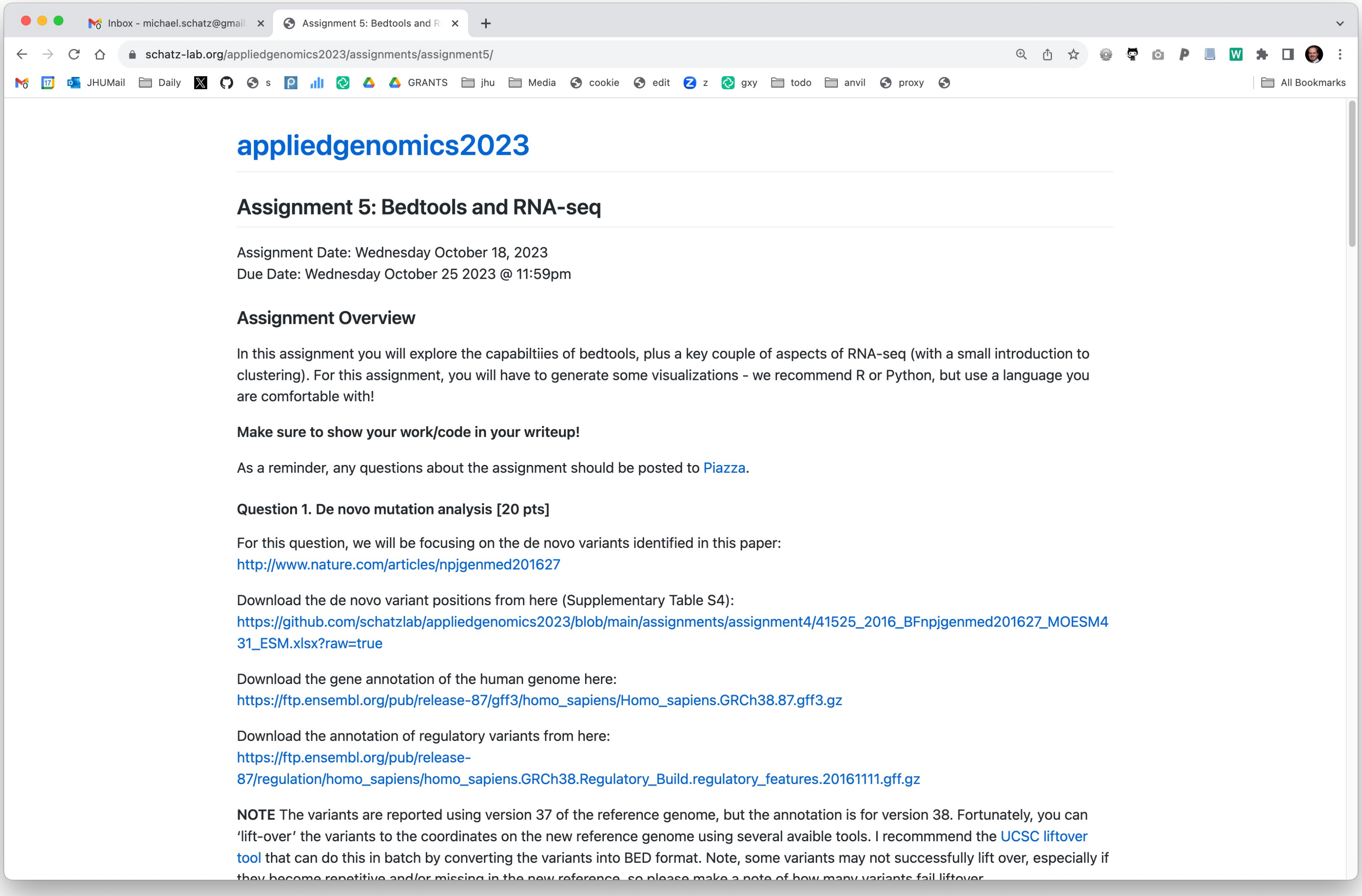
Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Please use Piazza to coordinate proposal plans!

https://github.com/schatzlab/appliedgenomics2023/blob/main/project/proposal.md

Check Piazza for questions!

Assignment 5
Due: Wednesday Oct 25, 2023 by 11:59pm


Inbox - michael.schatz@gmail
Assignment 5: Bedtools and R
schatz-lab.org/appliedgenomics2023/assignments/assignment5/
JHUMail
Daily
s
GRANTS
jhu
Media
cookie
edit
z
gxy
todo
anvil
proxy
All Bookmarks

appliedgenomics2023

Assignment 5: Bedtools and RNA-seq

Assignment Date: Wednesday October 18, 2023
Due Date: Wednesday October 25 2023 @ 11:59pm

Assignment Overview

In this assignment you will explore the capabiltiies of bedtools, plus a key couple of aspects of RNA-seq (with a small introduction to clustering). For this assignment, you will have to generate some visualizations - we recommend R or Python, but use a language you are comfortable with!

Make sure to show your work/code in your writeup!

As a reminder, any questions about the assignment should be posted to Piazza.

Question 1. De novo mutation analysis [20 pts]

For this question, we will be focusing on the de novo variants identified in this paper: http://www.nature.com/articles/npjgenmed201627

Download the de novo variant positions from here (Supplementary Table S4): https://github.com/schatzlab/appliedgenomics2023/blob/main/assignments/assignment4/41525_2016_BFnpjgenmed201627_MOESM431_ESM.xlsx?raw=true

Download the gene annotation of the human genome here: https://ftp.ensembl.org/pub/release-87/gff3/homo_sapiens/Homo_sapiens.GRCh38.87.gff3.gz

Download the annotation of regulatory variants from here: https://ftp.ensembl.org/pub/release-87/regulation/homo_sapiens/homo_sapiens.GRCh38.Regulatory_Build.regulatory_features.20161111.gff.gz

NOTE The variants are reported using version 37 of the reference genome, but the annotation is for version 38. Fortunately, you can 'lift-over' the variants to the coordinates on the new reference genome using several avaible tools. I recommmend the UCSC liftover tool that can do this in batch by converting the variants into BED format. Note, some variants may not successfully lift over, especially if they become repetitive and/or missing in the new reference, so please make a note of how many variants fail liftover.

https://schatz-lab.org/appliedgenomics2023/assignments/assignment5/

Check Piazza for questions!

Clustering Refresher

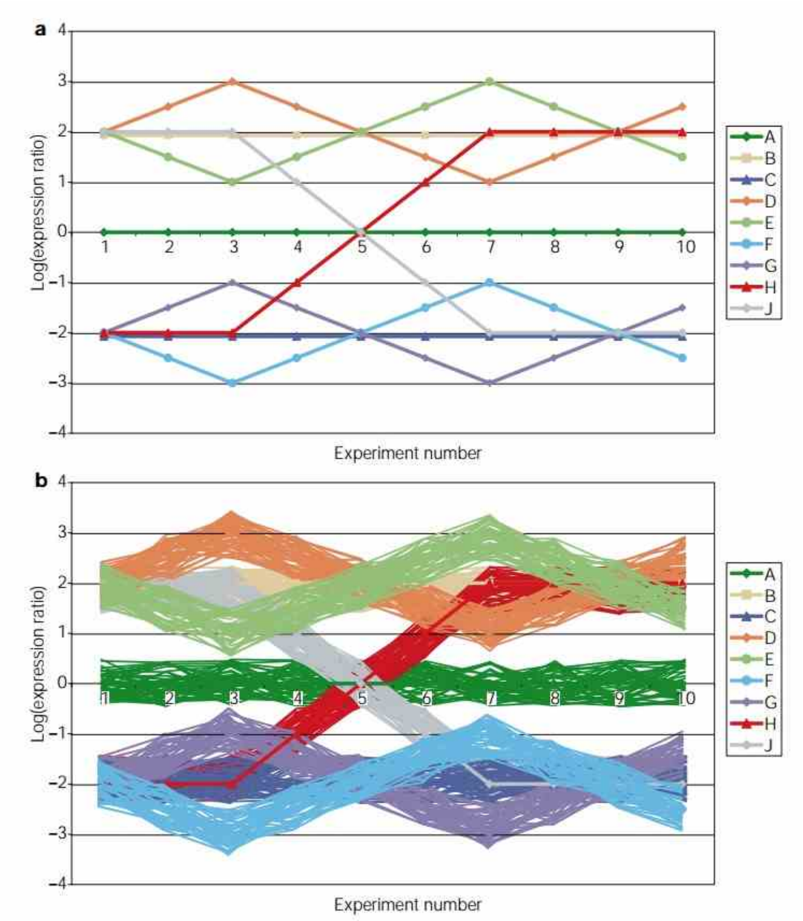
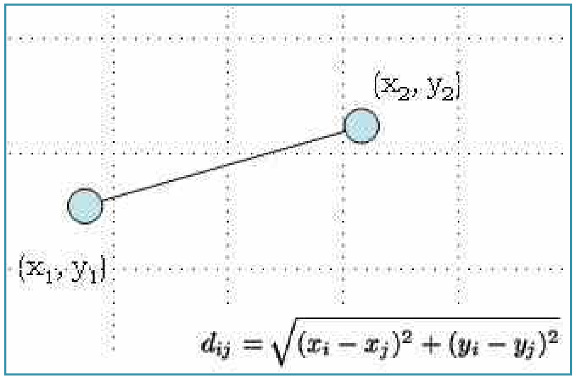
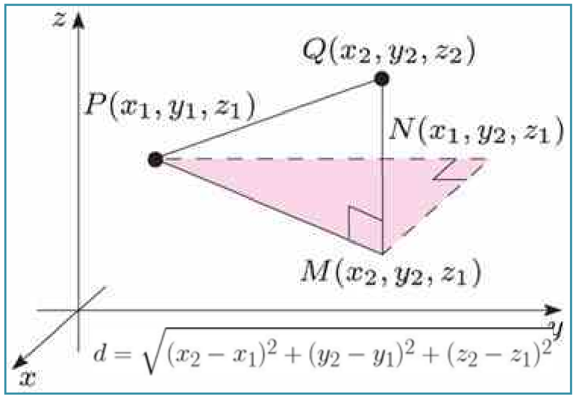


Figure 2 | **A synthetic gene-expression data set.** This data set provides an opportunity to evaluate how various clustering algorithms reveal different features of the data. **a** | Nine distinct gene-expression patterns were created with $\log_2$(ratio) expression measures defined for ten experiments. **b** | For each expression pattern, 50 additional genes were generated, representing variations on the basic patterns.

Euclidean Distance



$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

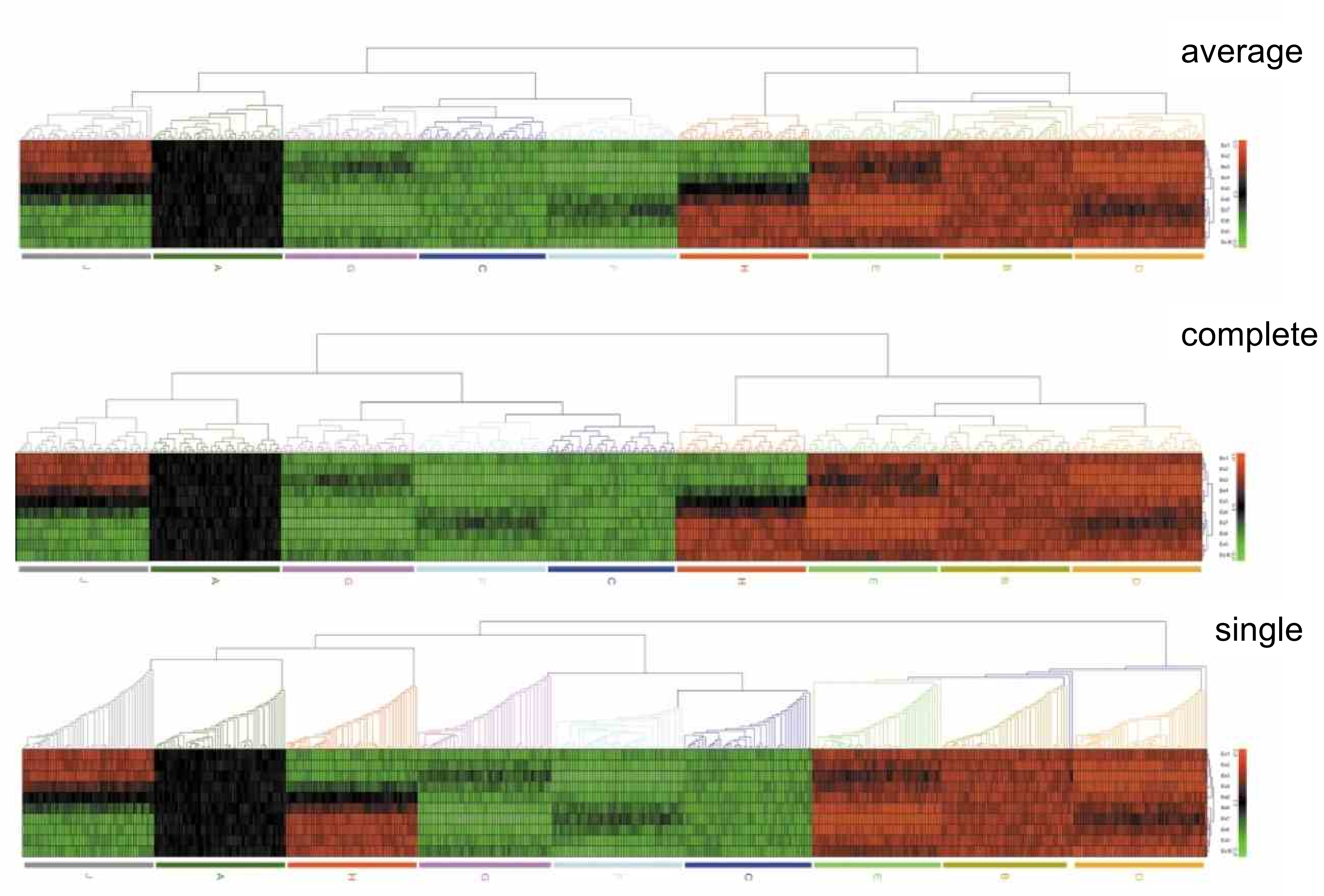


$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

Computational genetics: Computational analysis of microarray data
Quackenbush (2001) *Nature Reviews Genetics.* doi:10.1038/35076576

Hierarchical Clustering

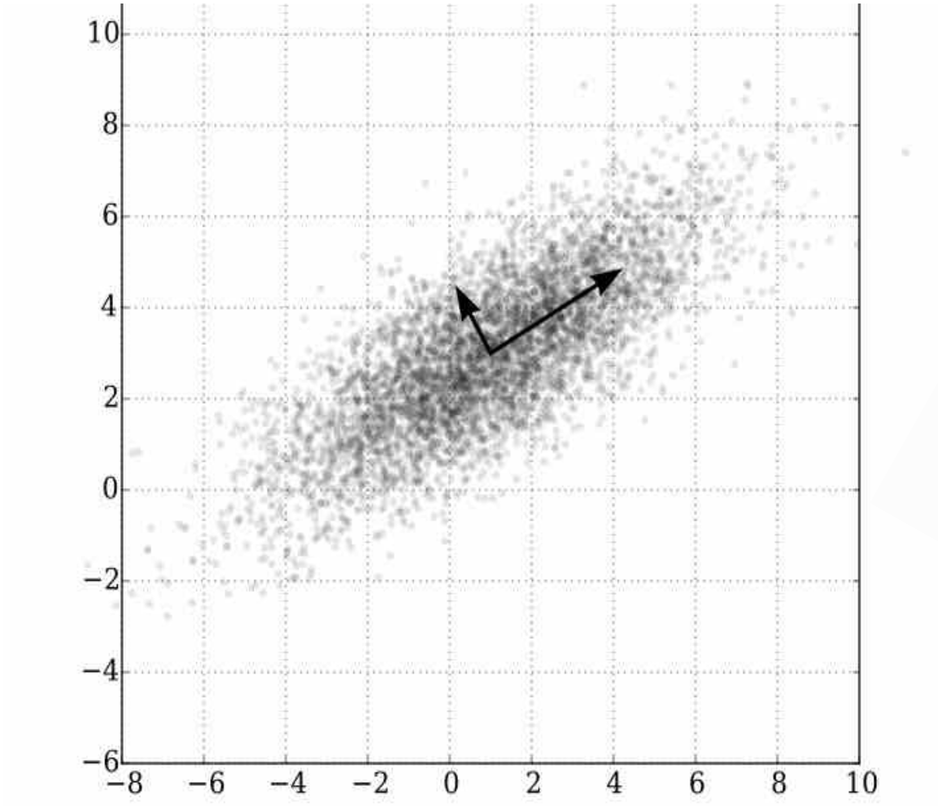


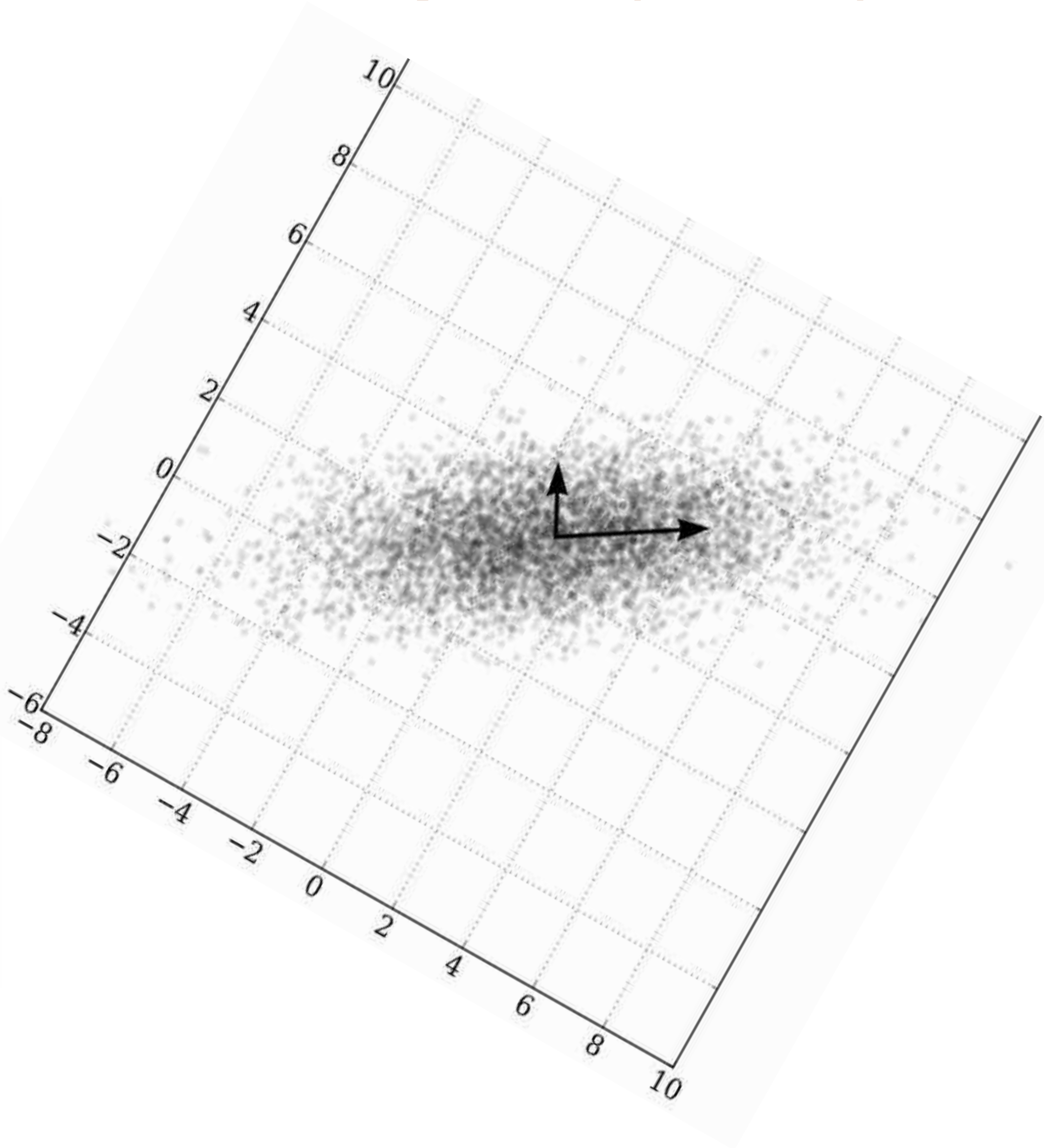
average

complete

single

Principle Components Analysis (PCA)





PC1: “New X”- The dimension with the most variability
PC2: “New Y”- The dimension with the second most variability

Principle Components Analysis (PCA)

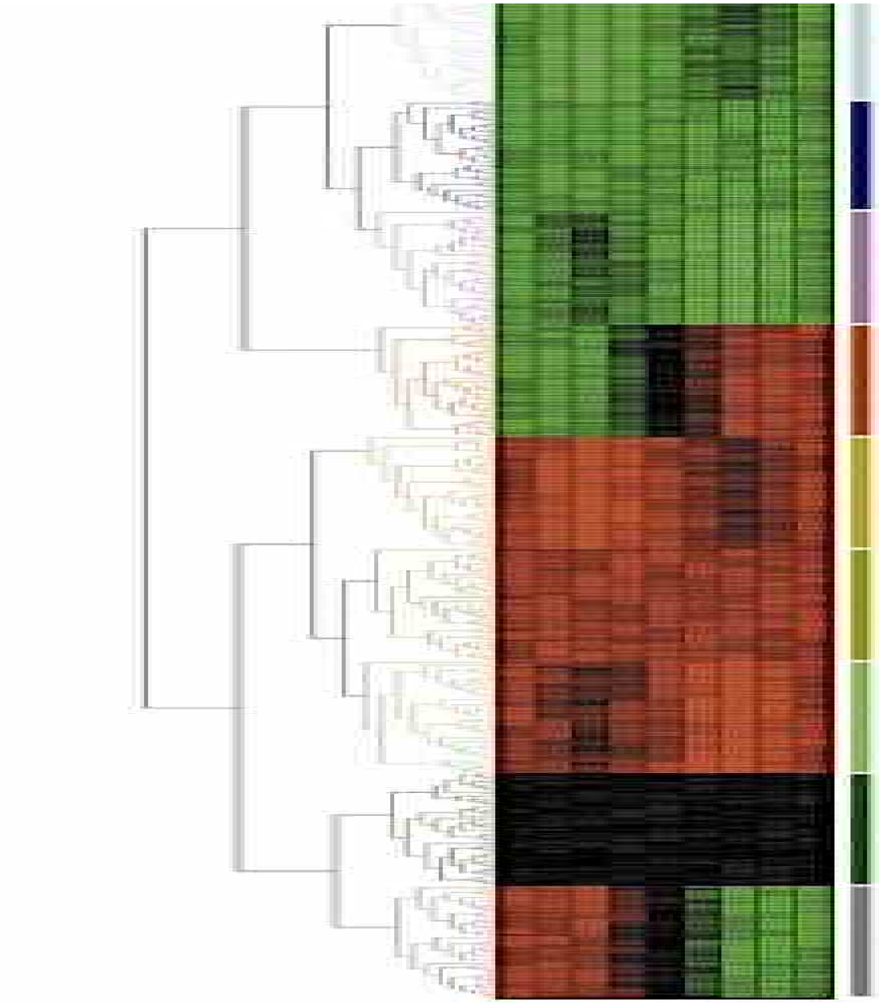
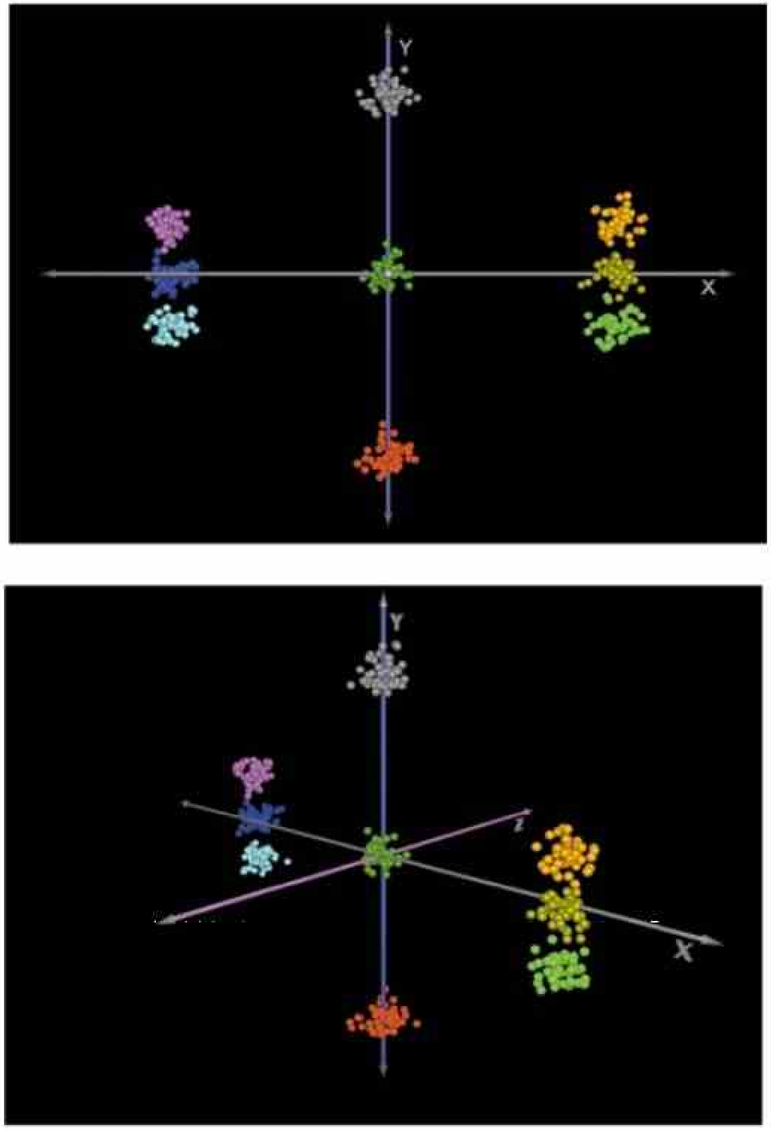
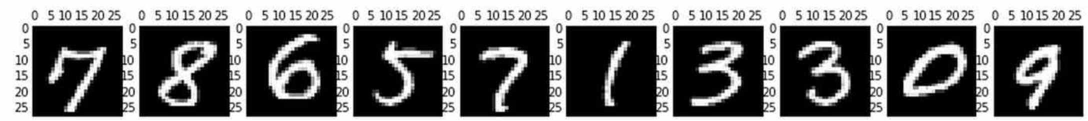
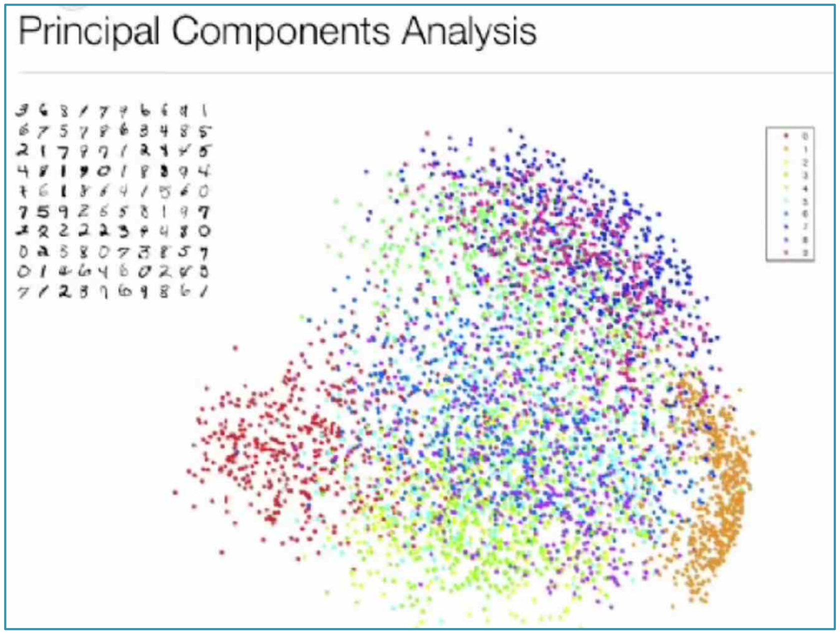


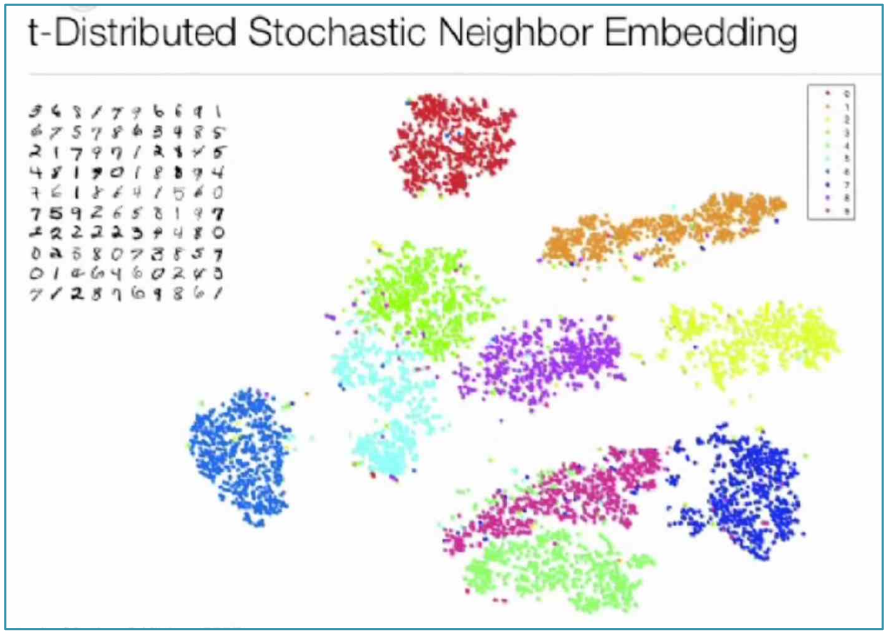
Figure 4 | **Principal component analysis.** The same demonstration data set was analysed using **a** | hierarchical (average-linkage) clustering and **b** | principal component analysis using Euclidean distance, to show how each treats the data, with genes colour coded on the basis of hierarchical clustering results for comparison.



PCA and t-SNE




Principal Components Analysis

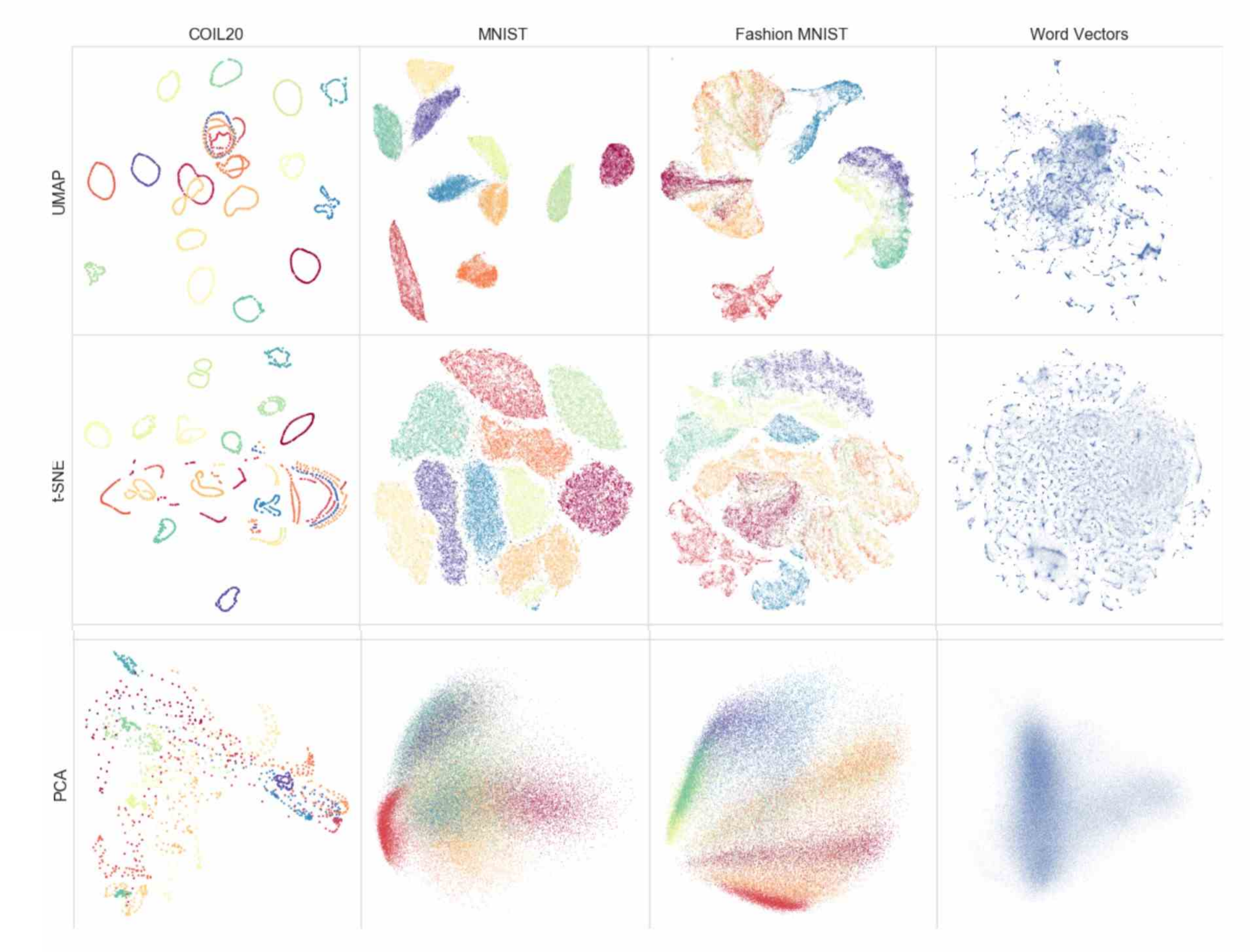

t-Distributed Stochastic Neighbor Embedding

t-distributed Stochastic Neighborhood Embedding

- Non-linear dimensionality reduction technique: distances are only locally meaningful
- Rather than Euclidean distances, for each point fits a Gaussian kernel to fit the nearest N neighbors (perplexity) that define the probabilities that two points should be close together
- Using an iterative spring embedding system to place high probability points nearby

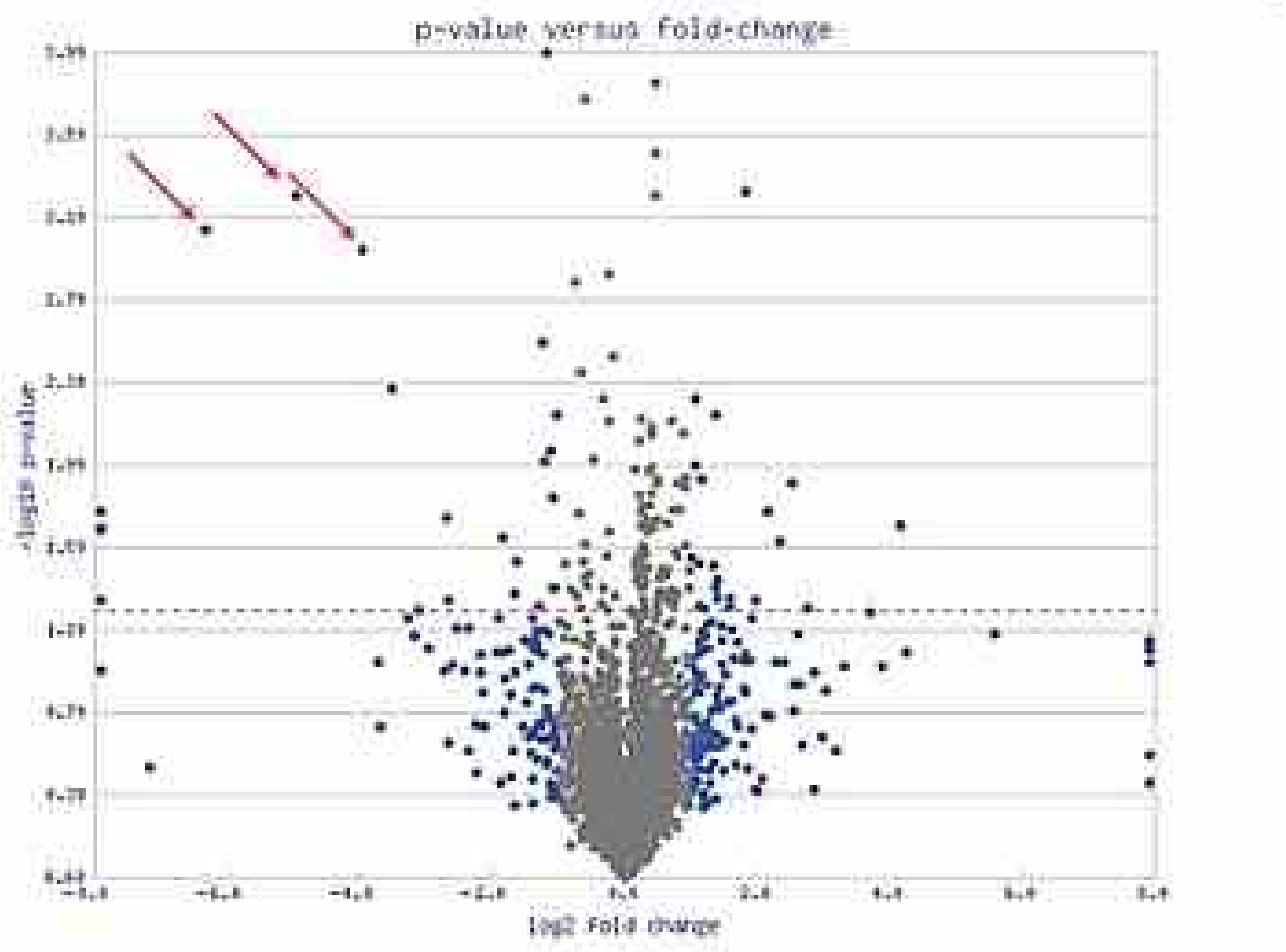
Visualizing Data Using t-SNE
https://www.youtube.com/watch?v=RJVL80Gg3lA

UMAP



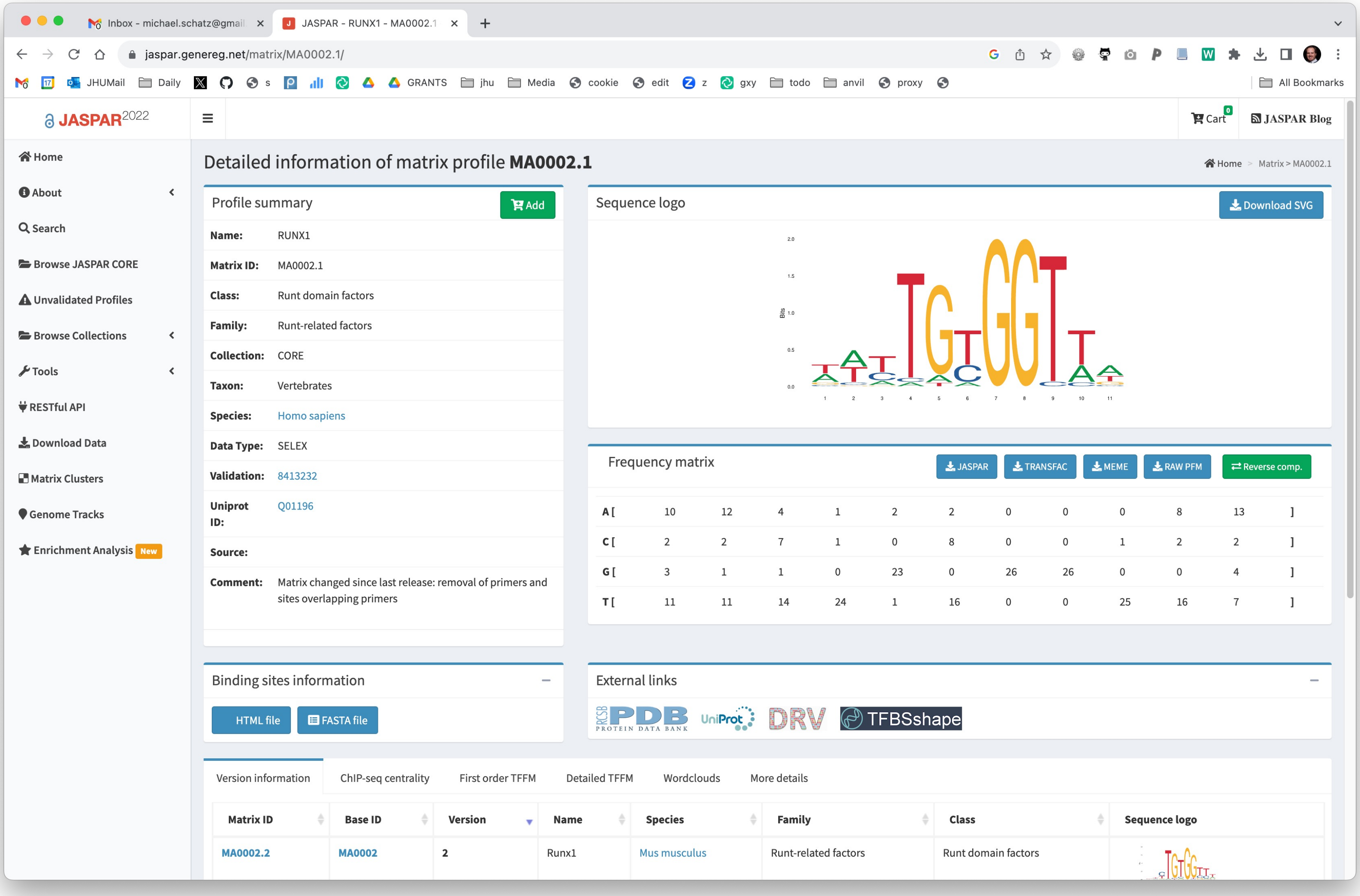
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction
McInnes et al (2018) arXiv. 1802.03426
https://www.youtube.com/watch?v=nq6iPZVUxZU
https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668

Volcano Plot



https://en.wikipedia.org/wiki/Volcano_plot_%28statistics%29

JASPAR Database



Detailed information of matrix profile **MA0002.1**

Profile summary

Name:	RUNX1
Matrix ID:	MA0002.1
Class:	Runt domain factors
Family:	Runt-related factors
Collection:	CORE
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	SELEX
Validation:	8413232
Uniprot ID:	Q01196
Source:	
Comment:	Matrix changed since last release: removal of primers and sites overlapping primers

Sequence logo

Frequency matrix

A [	10	12	4	1	2	2	0	0	0	8	13	]
C [	2	2	7	1	0	8	0	0	1	2	2	]
G [	3	1	1	0	23	0	26	26	0	0	4	]
T [	11	11	14	24	1	16	0	0	25	16	7	]

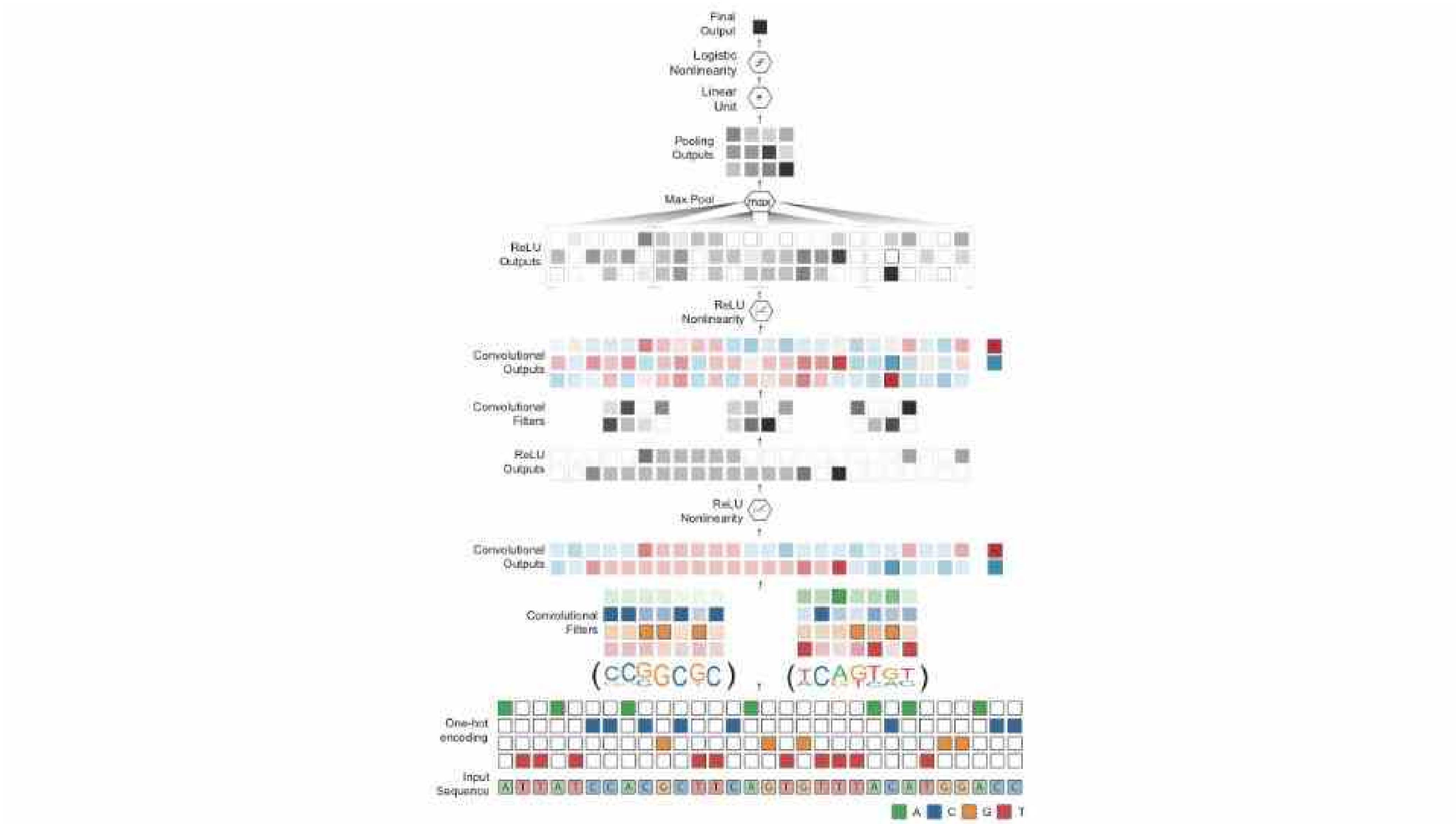
Binding sites information

External links

Matrix ID	Base ID	Version	Name	Species	Family	Class	Sequence logo
MA0002.2	MA0002	2	Runx1	Mus musculus	Runt-related factors	Runt domain factors	

https://jaspar.genereg.net/matrix/MA0002.1/

ML with Strings



One hot encoding to sequence classification

https://kundajelab.github.io/dragonn/tutorials.html