Multi-Omics Data Integration via Machine Learning

Oslo Bioinformatics Workshop Week 2022, Oslo, Norway, 9.12.2022
Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden

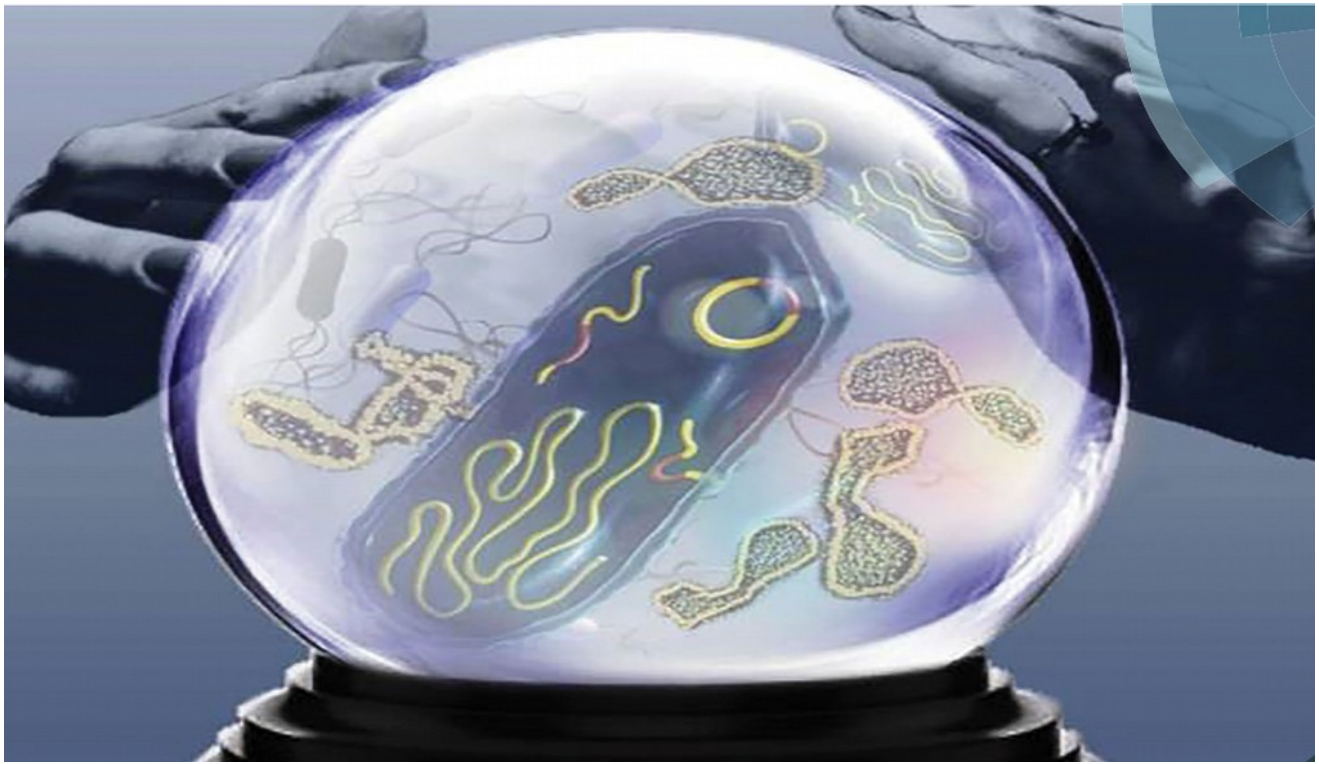


Image adapted from Molecular Omics, Issue 1, 2018

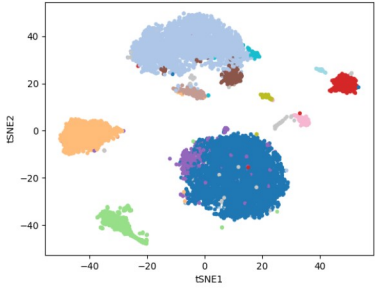
Brief introduction: who am I

2007 PhD in theoretical physics

2011 medical genetics at Lund University

2016 working at NBIS SciLifeLab, Sweden

Single cell



Biomedical data integration

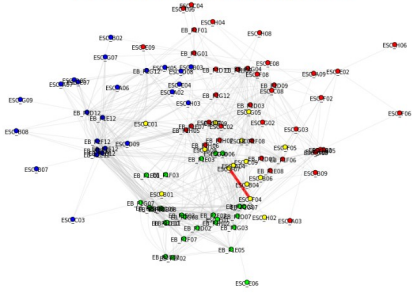
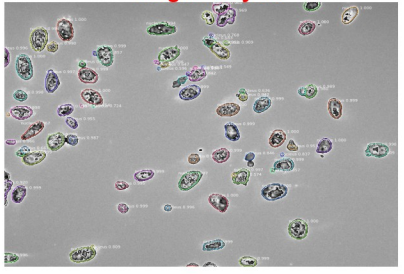
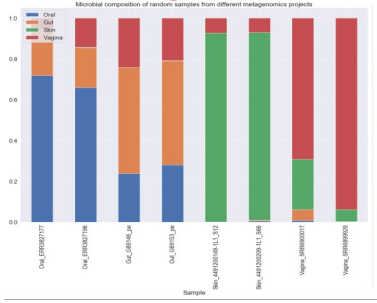


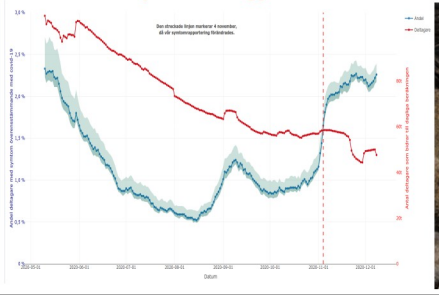
Image analysis



Metagenomics

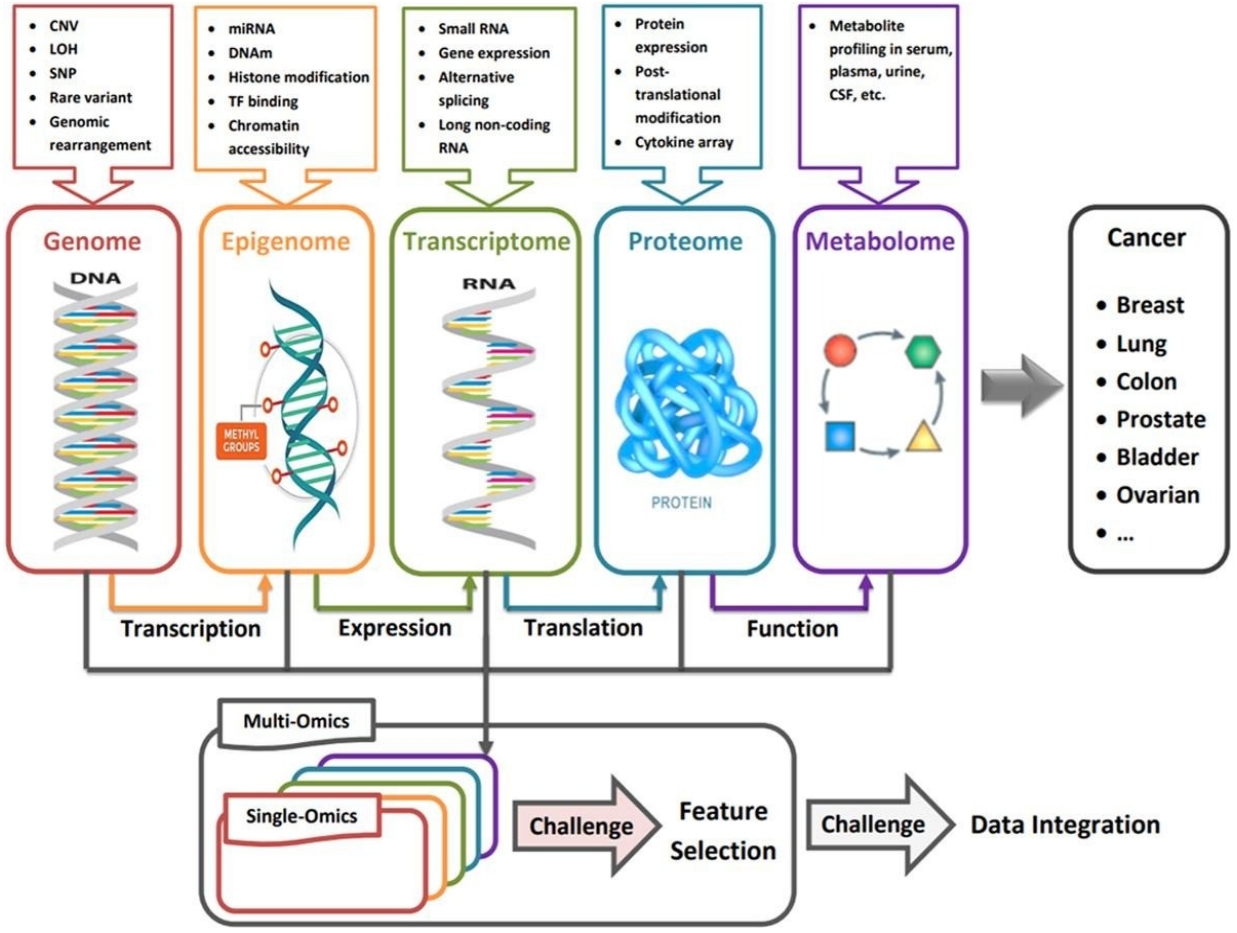


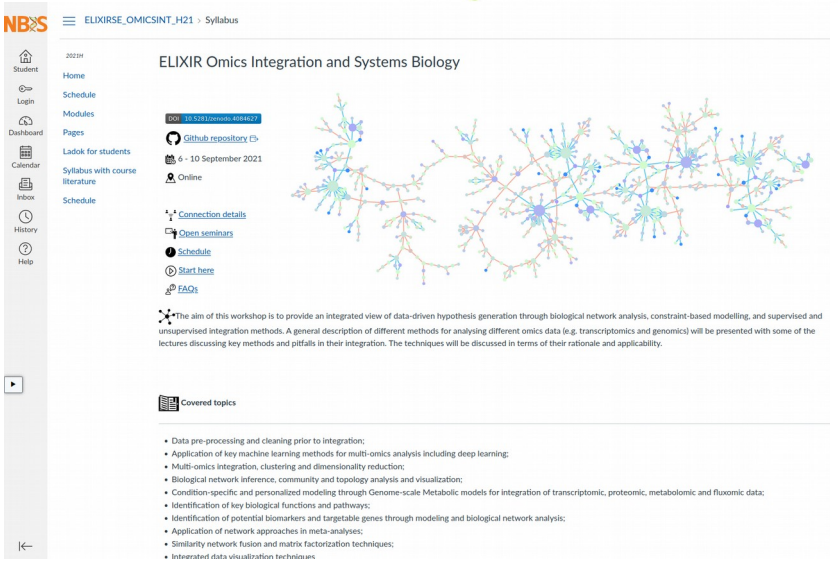
Epidemiology

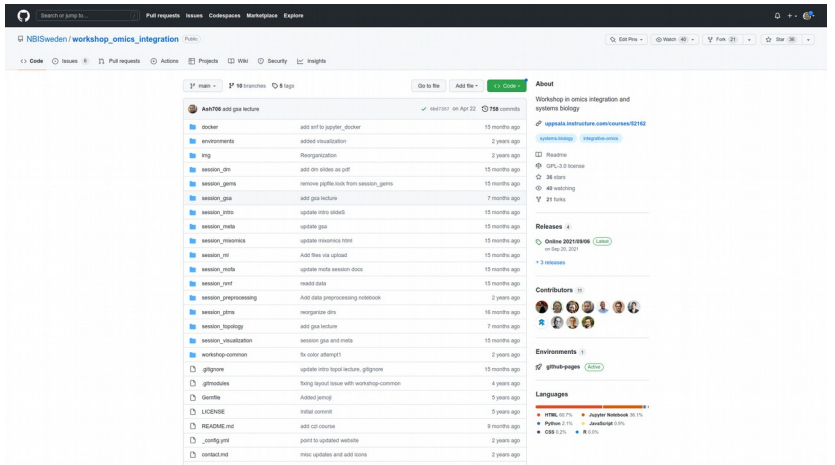


Ancient microbiome

Multi-Omics Data in Life Sciences







Momeni et al., J Biomed Inform. 2020

https://github.com/NBISweden/workshop_omics_integration

Introduction:
High Dimensional Biological Data

Biological Data are High Dimensional

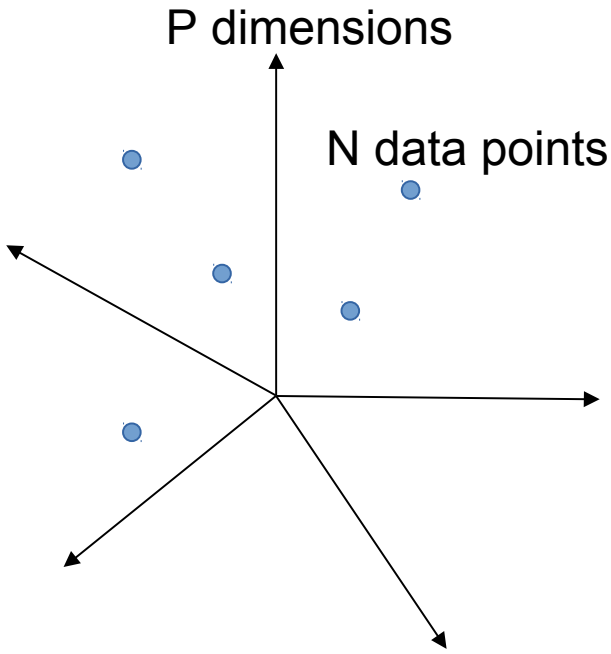
Statistical observations:
e.g. samples, cells etc.

N

P

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

Features: genes, proteins, microbes, metabolites etc.



High Dimensional Data:
P >> N

For a robust statistical analysis, one should properly “sample” the P-dimensional space, hence large sample size is required, N >> P

Types of Statistical Analysis

P is the number of features (genes, proteins, genetic variants etc.)
N is the number of observations (samples, cells, nucleotides etc.)

Biology / Biomedicine

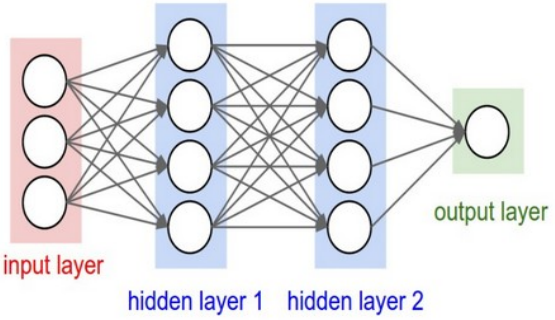
Bayesianism

P >> N

Frequentism

P ~ N

Deep Learning



P << N

Amount of Data

The Curve of Dimensionality

Ex.1

$$Y = \alpha + \beta X$$

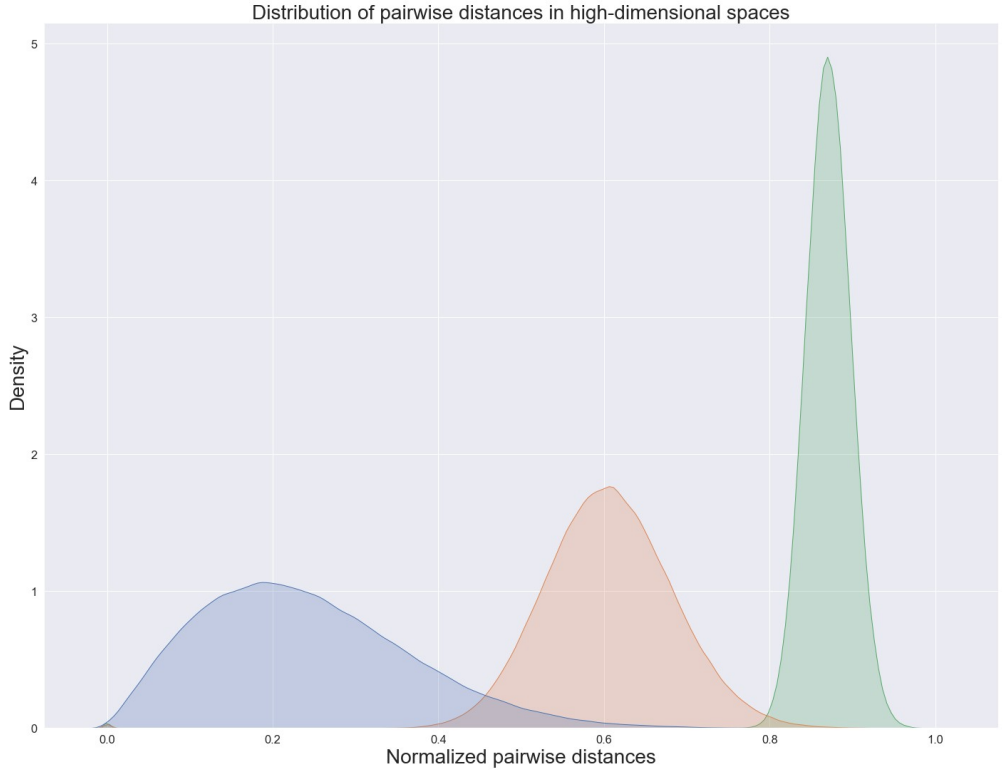
$$\beta = (X^T X)^{-1} X^T Y$$

$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \cdots \to \infty, \quad n << p$$

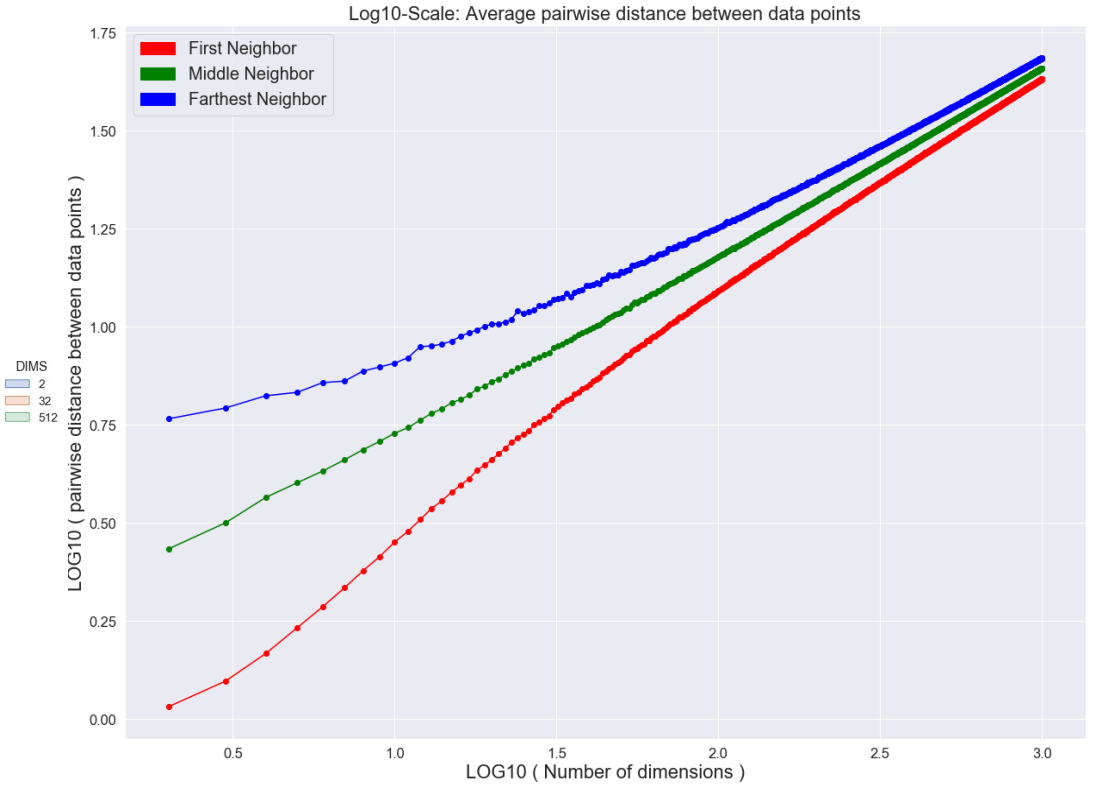
Ex.2 $E[\hat{\sigma}^2] = \frac{n-p}{n}\sigma^2$

Biased ML variance estimator in HD-space

Distances in High Dimensions



Data points become far from each other and equidistant in high dimensions



The differences between closest and farthest data point neighbours disappears in high-dimensional spaces: can't run cluster analysis

In high-dimensional space we can not separate cases and controls any more

Typical Dimensions of Omics Data

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

Metabolomics
N ≈ P

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

Proteomics
N ≈ P

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix}$$

Metagenomics
N ≈ P

← **manageable**

P

$$N \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Transcriptomics
N << P
(Single cell: N <= P)

challenging ↓

P

$$N \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Genomics
N <<<< P

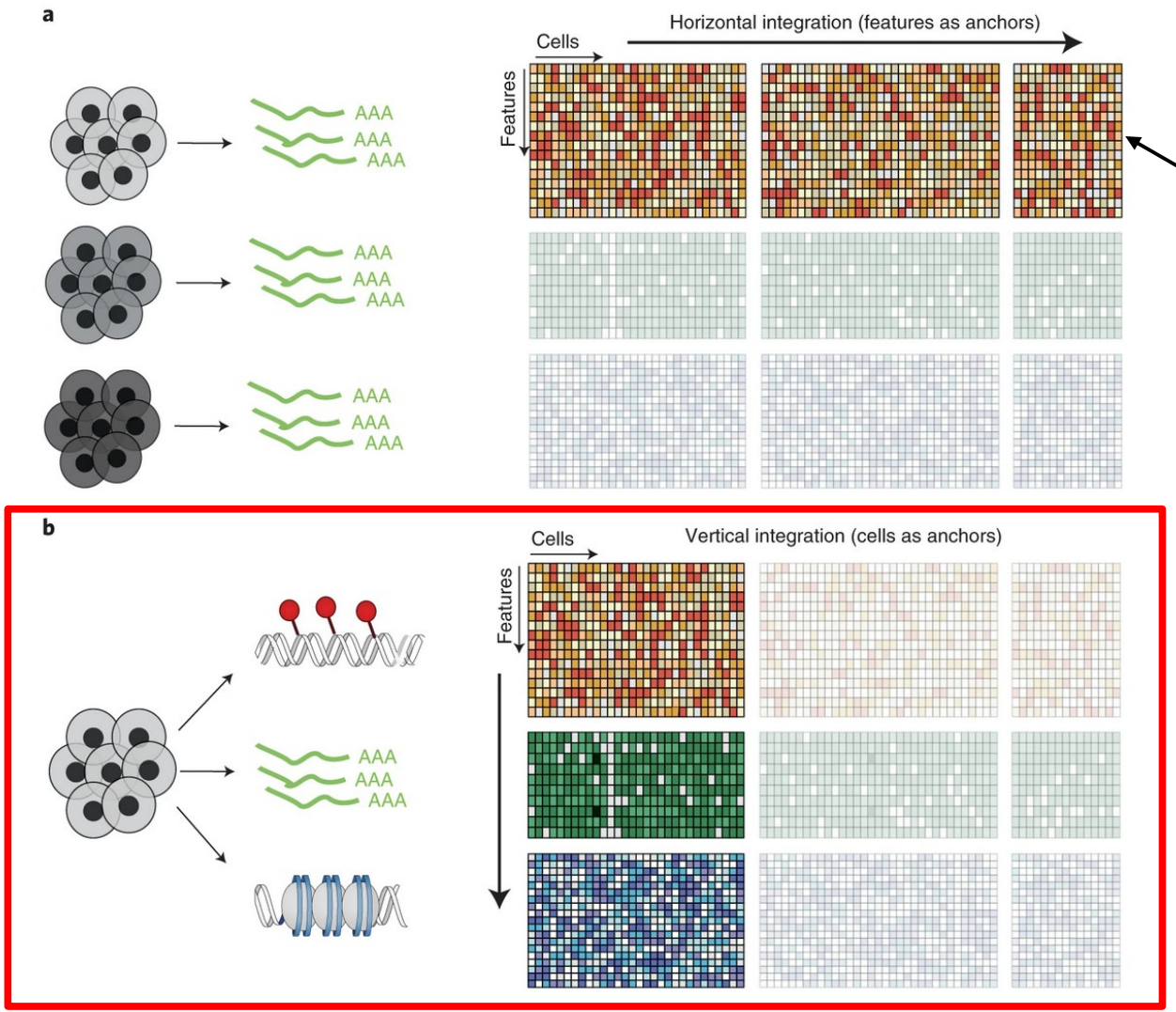
P

$$N \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{pmatrix}$$

Methylomics
N <<<< P

Multi-Omics Data Integration

Integration Across Features vs. Samples



Similar to batch correction problem

Here I will focus on integration across features

Argelaguet et al., 2021

Integrating Omics Across Features

Statistical observations:
e.g. samples, cells etc.

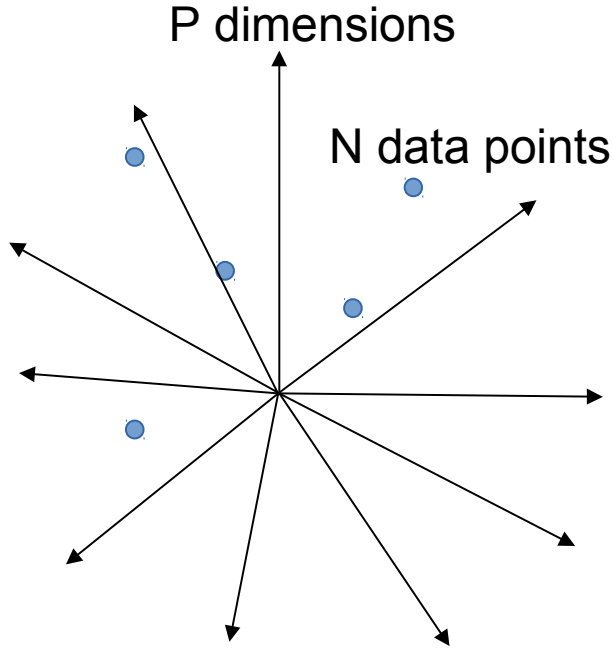
Features: genes, proteins, microbes, metabolites etc.

N

P_1

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

Omic1: RNAseq



N

P_2

0	3	1	0	2	3	8	1	1	3
1	1	0	0	7	1	2	2	3	3
1	2	2	0	0	6	7	1	2	2
1	2	3	10	0	4	6	1	0	5
3	2	2	1	4	3	2	1	6	0
7	4	4	5	3	9	6	1	6	1
7	1	1	5	2	8	9	1	3	6
5	0	1	6	2	0	0	0	1	5
1	6	3	3	4	6	2	0	1	1
1	2	2	4	1	1	3	0	8	2

Omic2: BSseq

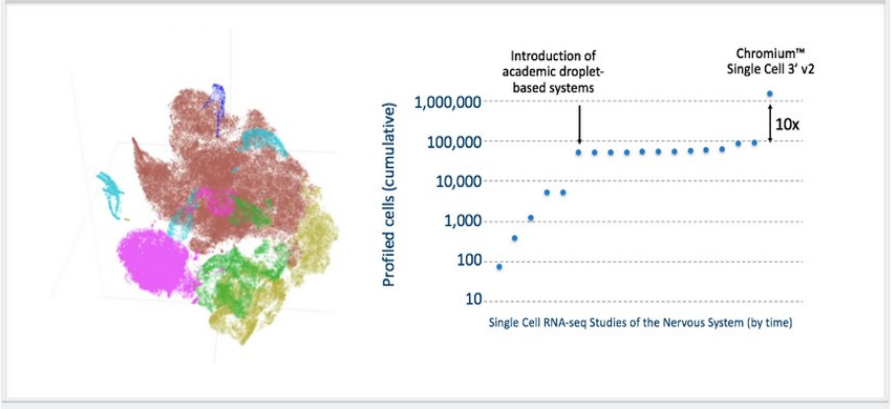
$P_1 + P_2 >> N$ integration across features leads to even more high-dimensional data

Big Data in Single Cell Genomics

CAREERS BLOG 10X UNIVERSITY

10X GENOMICS SOLUTIONS & PRODUCTS RESEARCH & APPLICATIONS EDUCATION & RESOURC

« Back to Blog ‹ Newer Article Older Article ›



Our 1.3 million single cell dataset is ready to download

0 KUDOS

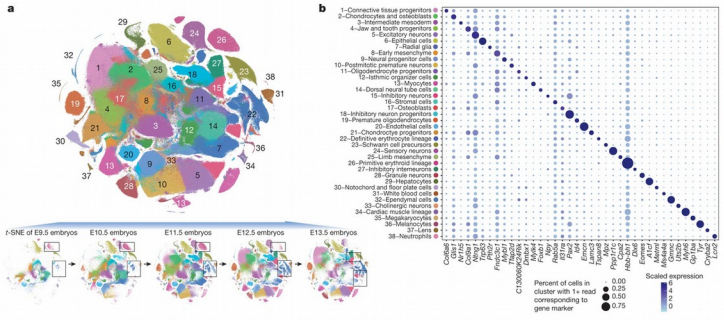
POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

At ASHG last year, we announced our 1.3 Million Brain Cell Dataset, which is, to date, the largest dataset published in the single cell RNA-sequencing (scRNA-seq) field. Using the Chromium™ Single Cell 3' Solution (v2 Chemistry), we were able to sequence and profile 1,308,421 individual cells from embryonic mice brains. Read more in our application note Transcriptional Profiling of 1.3 Million Brain Cells with the Chromium™ Single Cell 3' Solution.

MENU nature

Fig. 2: Identifying the major cell types of mouse organogenesis.

From: The single-cell transcriptional landscape of mammalian organogenesis



a, t-SNE visualization of 2,026,641 mouse embryo cells (after removing a putative doublet cluster), coloured by cluster identity (ID) from Louvain clustering (in **b**), and annotated on the basis of marker genes. The same t-SNE is plotted below, showing only cells from each stage (cell numbers from left to right: n = 151,000 for E9.5; 370,279 for E10.5; 602,784 for E11.5; 468,088 for E12.5; 434,490 for E13.5). Primitive erythroid (transient) and definitive erythroid (expanding) clusters are boxed. **b**, Dot plot showing expression of one selected marker gene per cell type. The size of the dot encodes the percentage of cells within a cell type in

BioTuring Solutions Resources

Explore 4,000,000 CELLS at ease with BIOTURING BROWSER

EXPLORER NOW

A next-generation platform to re-analyze published single-cell sequencing data

Single Cell Analysis

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September

by biomembers • August 30, 2019

Human Cell Atlas, single-cell data

We are glad to announce that we will upsize the current single-cell database in BioTuring Single-cell Browser to 5,500,000 cells this September. With this release, we will double the current number of publications indexed in BioTuring Single-cell Browser, and cross the number of cells hosted on available public single-cell data repositories like Human Cell Atlas (HCA) and Broad Institute's Single-cell Portal.

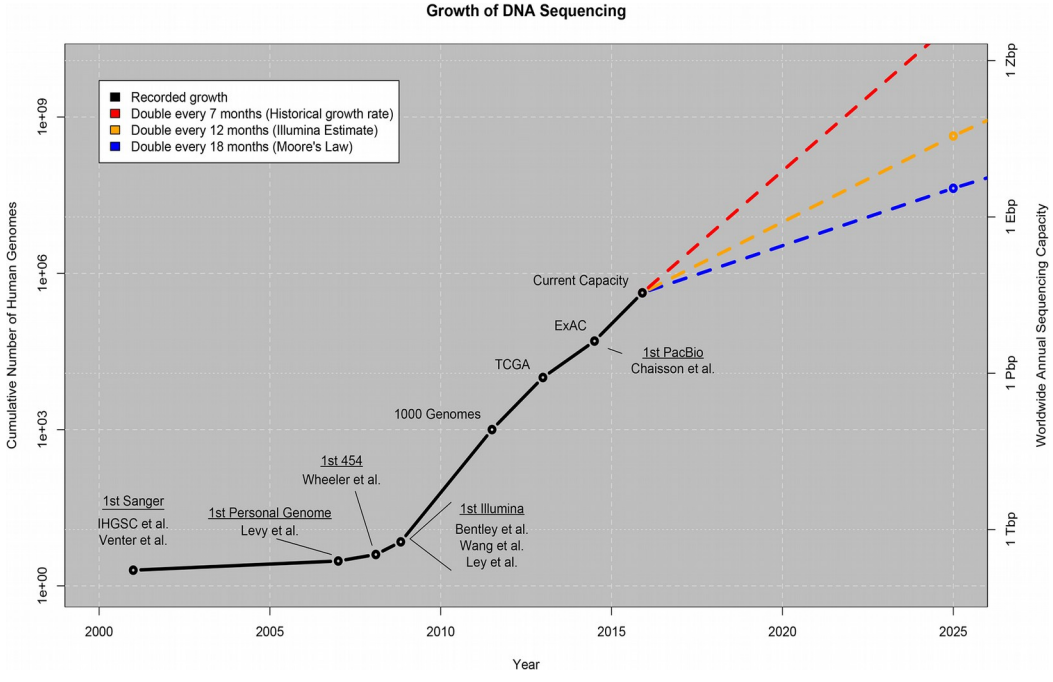
Search

RECENT POSTS

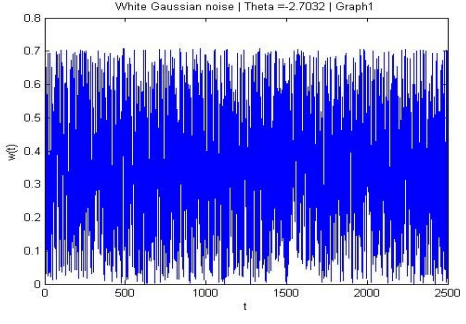
A new tool to interactively visualize single-cell objects (Seurat, Scanpy, SingleCellExperiments, ...)
September 26, 2019

5,500,000 cells will be indexed into BioTuring Single-cell Data Repository this September
August 30, 2019

Big in Size or Sample Size?



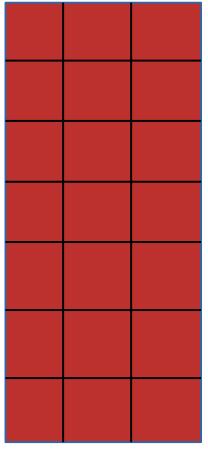
Stephens et al., (2015). Big Data: Astronomical or Genomical? PLoS Biology 13(7)



A file with **White Noise** can also take a lot of disk space

Genomics / WGS: **Little Data**

$N_1 \sim 10^3$

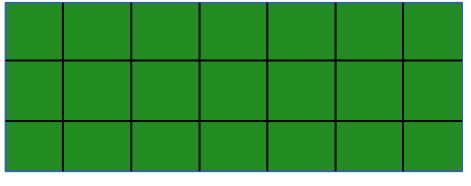


$P_1 \sim 10^6$

$N_1 * P_1 = N_2 * P_2 = 10^9$

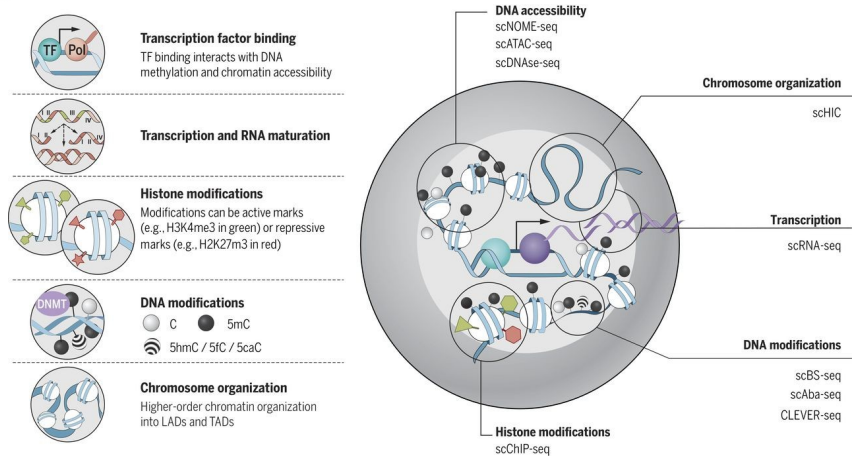
scRNAseq: **Big Data**

$N_2 \sim 10^6$

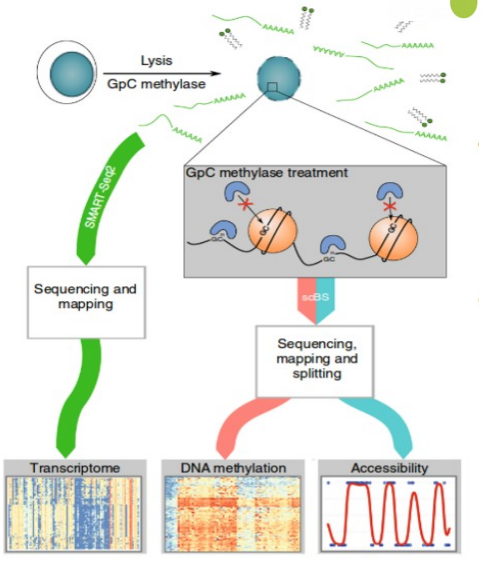


$P_2 \sim 10^3$

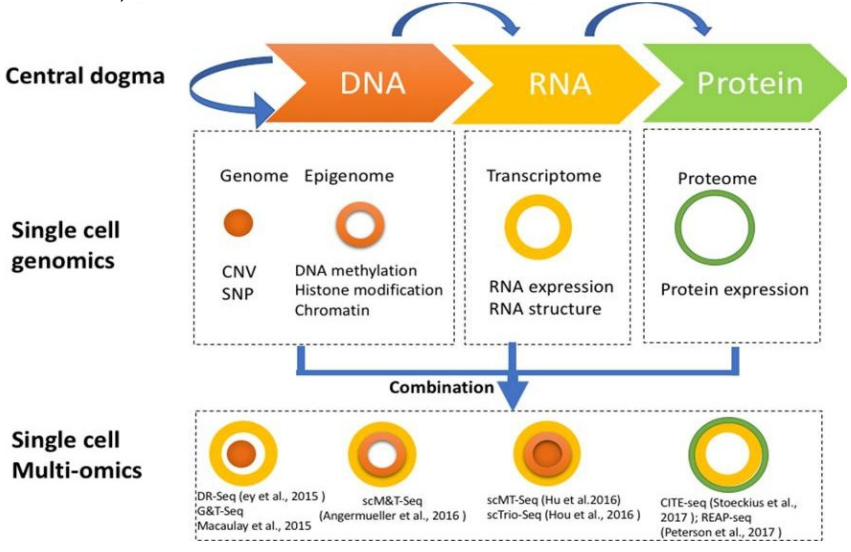
Multi-Omics in Single Cell Genomics



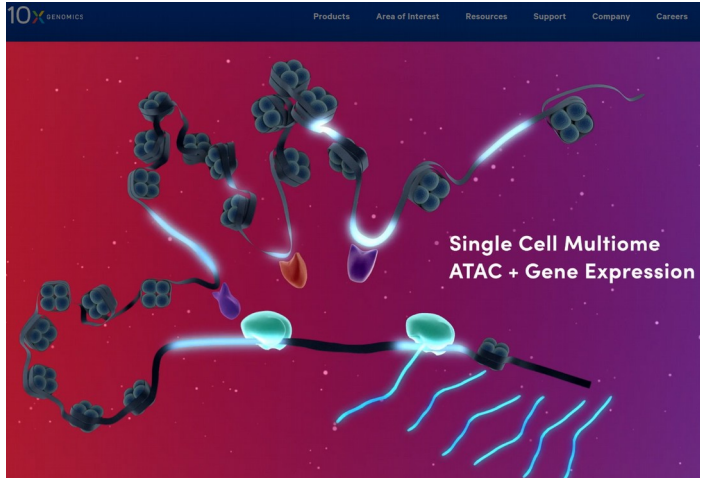
Kelsey et al., 2017, Science 358, 69-75



Clark et al., 2018, Nature Communications 9, 781



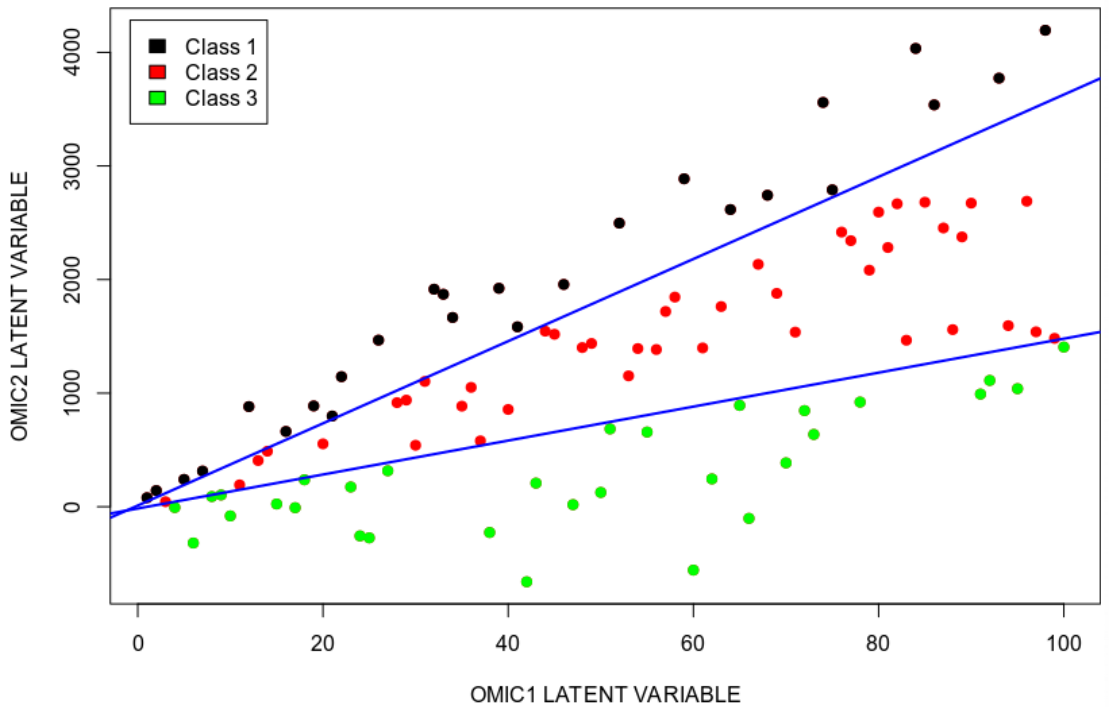
Hu et al., 2018, Frontier in Cell and Developmental Biology 6, 1-13

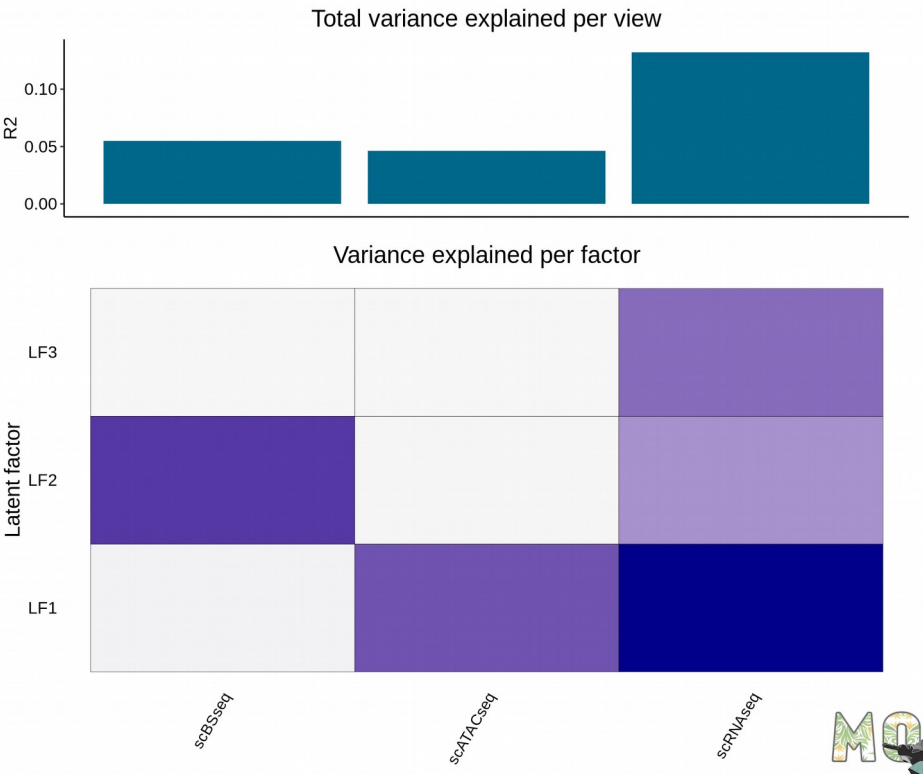


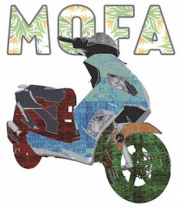
How to define and evaluate multi-Omics data integration?

What I mean by Omics Integration

Idea Behind OMICs Integration:
See Patterns Hidden in Individual OMICS

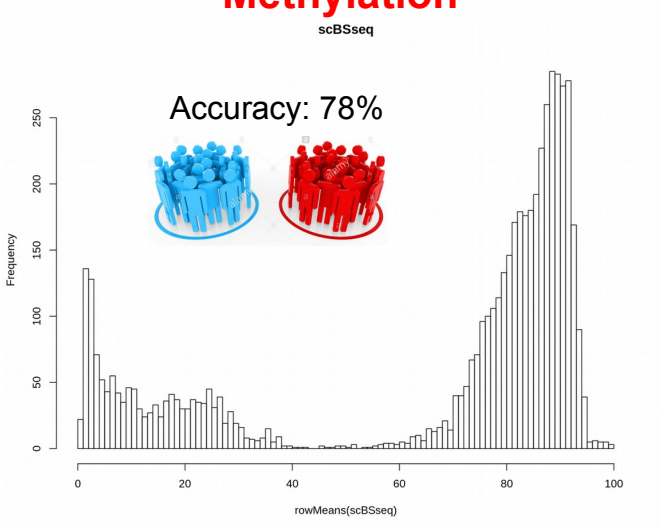




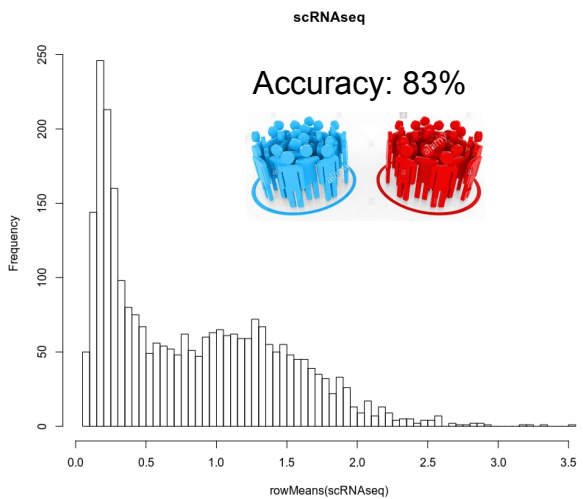


How I Evaluate Omics Integration

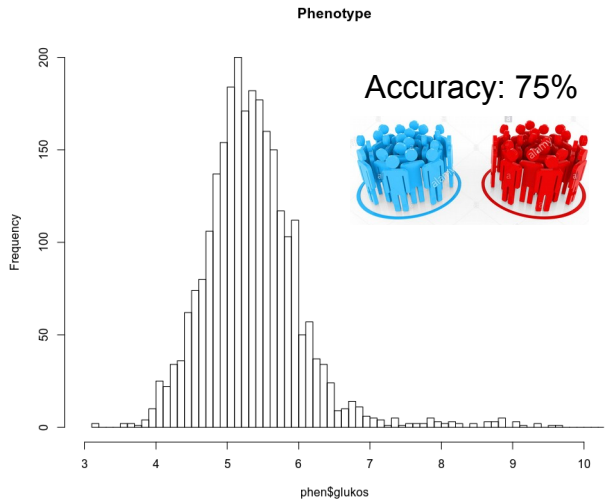
Methylation



Gene Expression



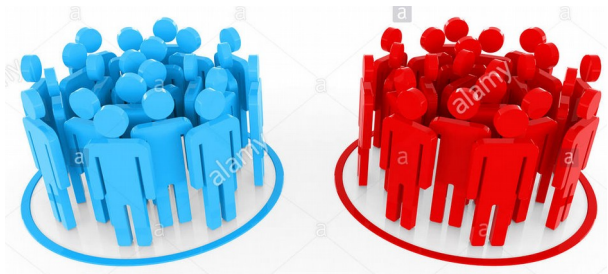
Clinical variable



1) Convert to common space:
Neural Networks, SNF, UMAP

2) Explicitly model distributions:
MOFA, Bayesian Networks

3) Extract common variation:
PLS, CCA, Factor Analysis

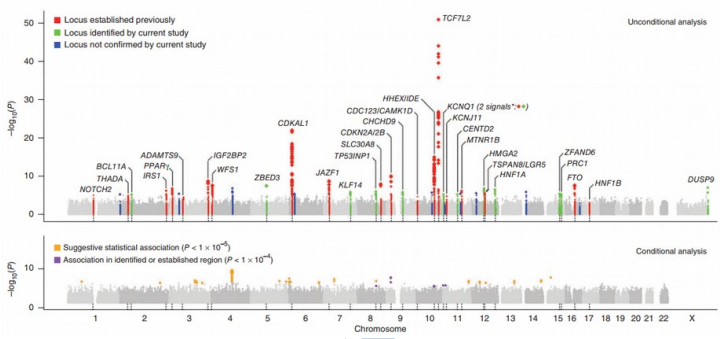


HEALTHY SICK

Data Integration
Accuracy: 96%

Prediction as a Criterion of Success

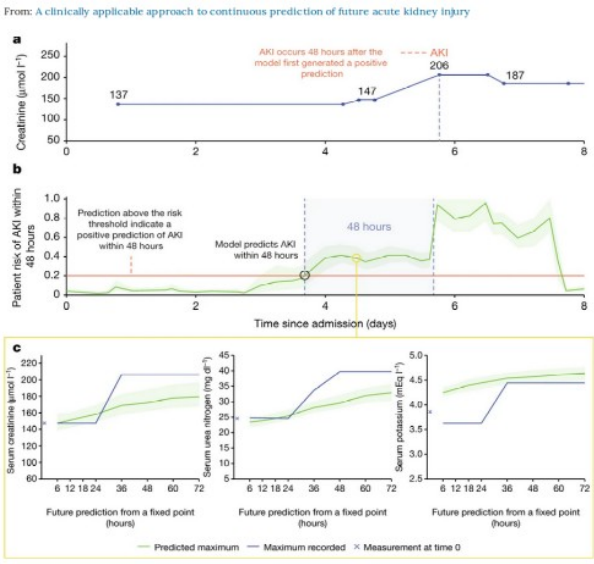
Statistics searches for candidates



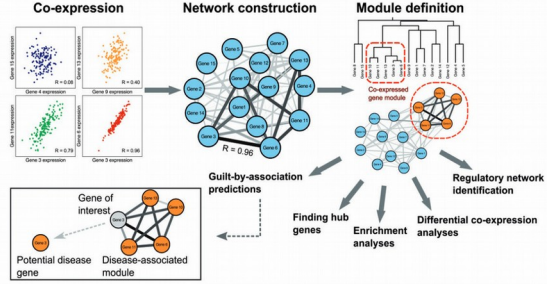
Consequence

B. Maher, Nature 456, 18-21 (2008)

Machine Learning optimizes prediction



Consequence



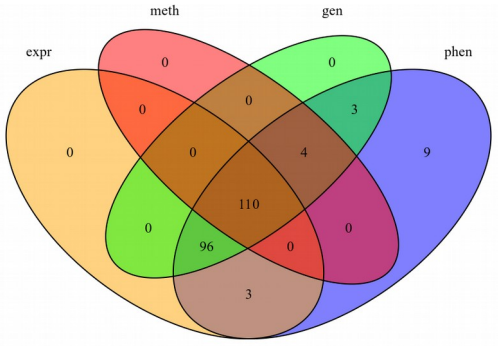
What Method of Integration to Select?

	Linear	Non-Linear
Supervised	PLS / OPLS / mixOmics, LASSO / Ridge / Elastic Net	Neural Networks, Random Forest, Bayesian Networks
Unsupervised	Factor Analysis / MOFA	Autoencoder, SNF, UMAP, Clustering of Clusters

For Example:

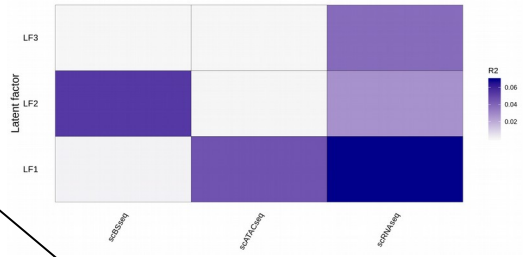
1) With ~100 samples it is a good idea to do **linear** Omics integration
2) T2D is a phenotype of interest, therefore **supervised** integration

Typical Workflow for Omics Integration



Data Set (4 Omics)
110 overlapping individuals

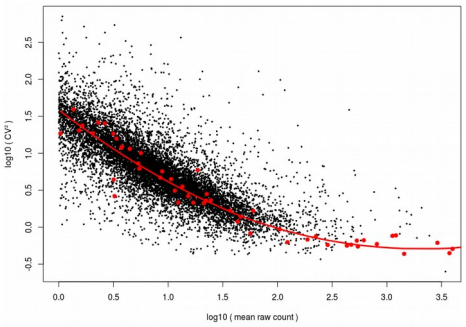
Check covariance



Train Set (n = 80)

Test Set (n = 30)

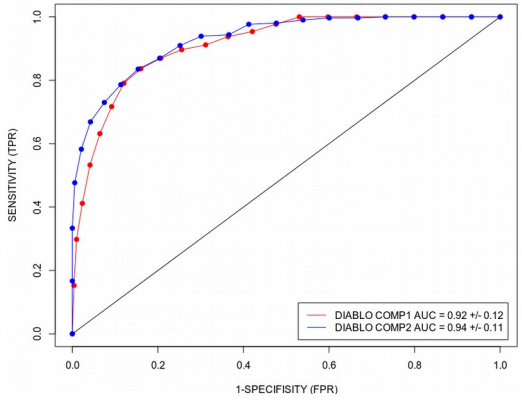
Unsupervised:
remove low-variance



Supervised:
LASSO

Feature Selection

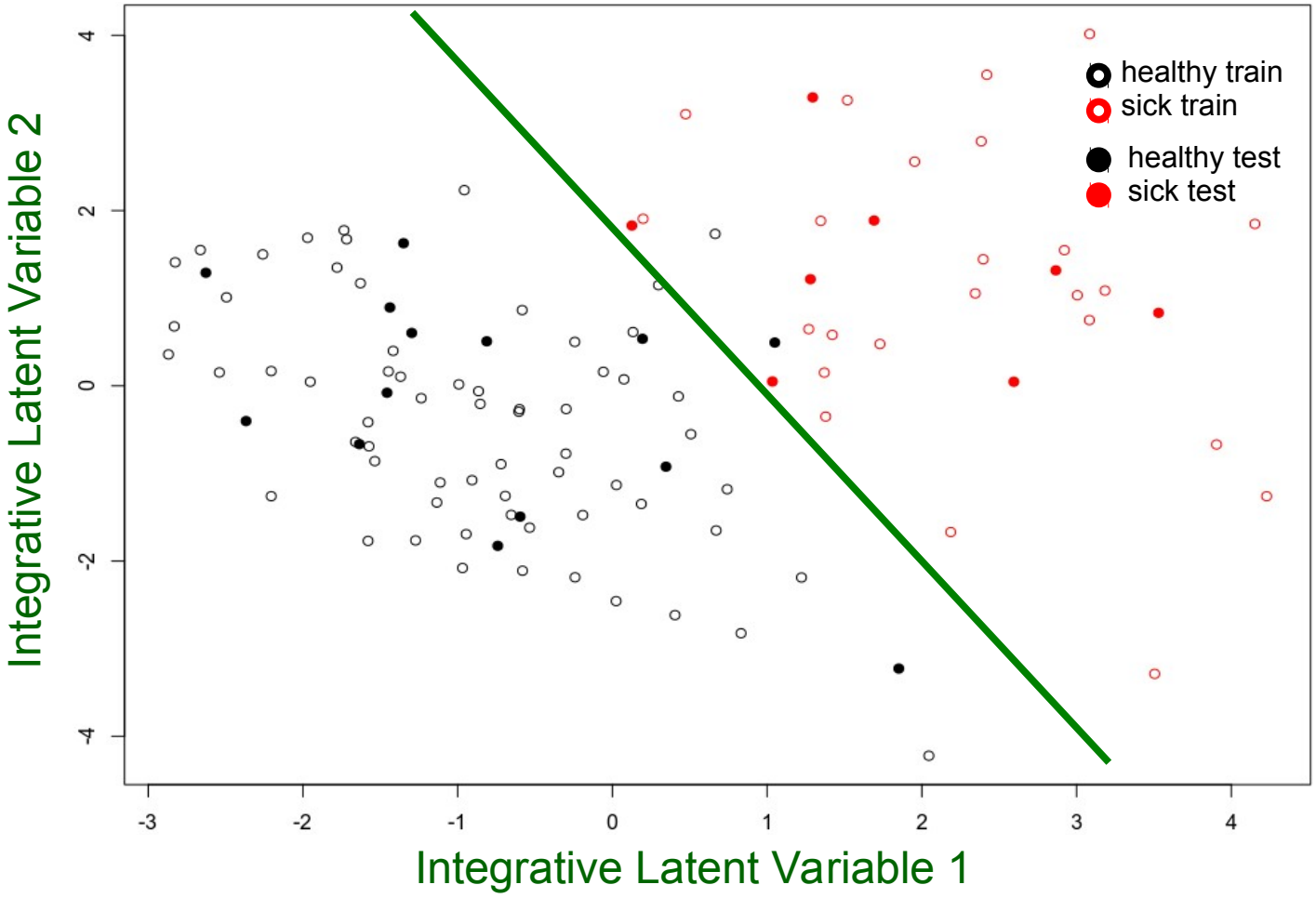
Evaluation



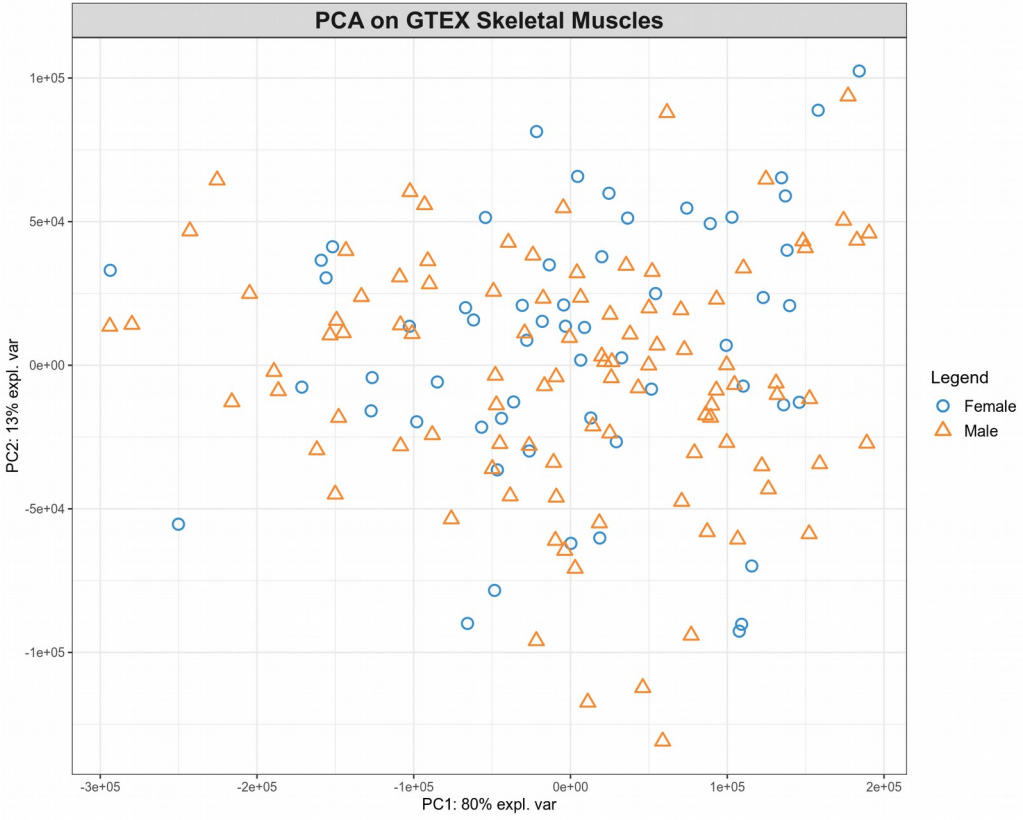
Omics Integration

Trained Model

Linear Separation: Decision Boundary



Supervised Univariate Feature Selection



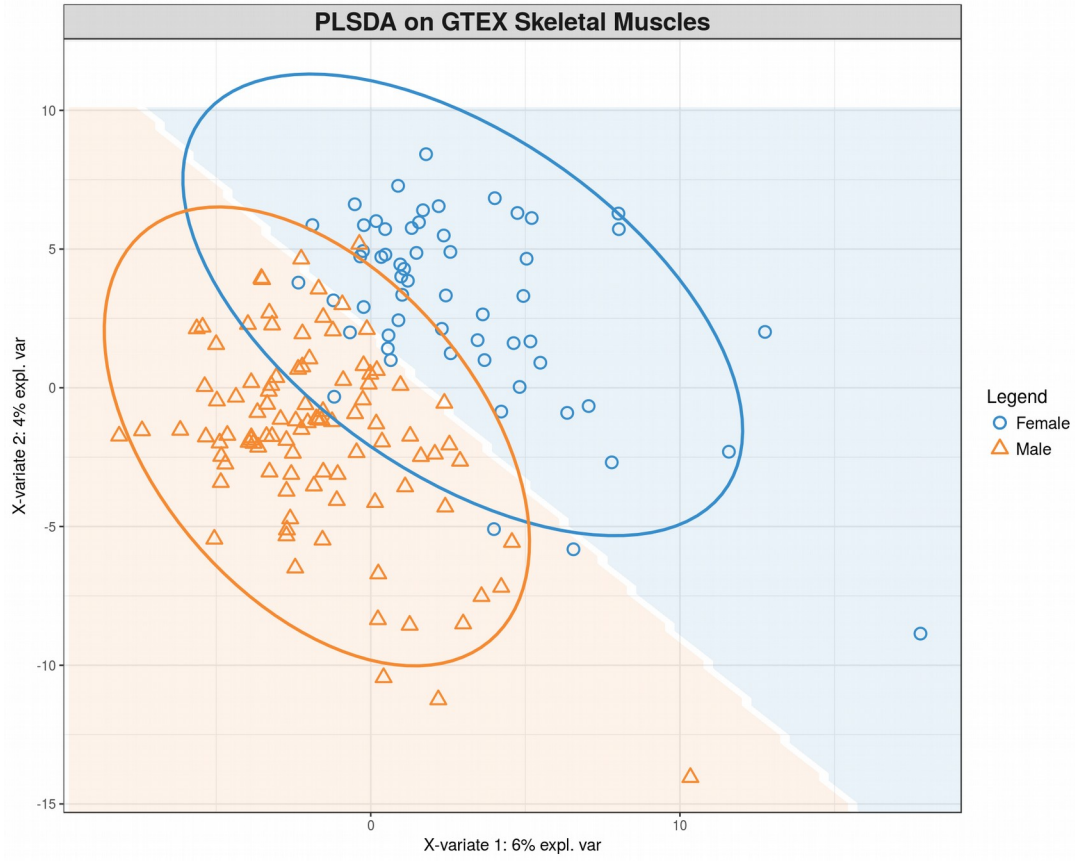

```r
rho <- vector()
p <- vector()
a <- seq(from=0, to=dim(X)[2], by=100)
for(i in 1:dim(X)[2])
{
  corr_output <- cor.test(X[,i], as.numeric(Y), method="spearman")
  rho <- append(rho,as.numeric(corr_output$estimate))
  p <- append(p,as.numeric(corr_output$p.value))
  if(isTRUE(i%in%a)==TRUE){print(paste("FINISHED ",i," FEATURES",sep=""))}
}
output <- data.frame(GENE=colnames(X), SPEARMAN_RHO=rho, PVALUE=p)
output$FDR <- p.adjust(output$PVALUE, method="fdr")
output <- output[order(output$FDR, output$PVALUE, -output$SPEARMAN_RHO), ]
head(output,10)
```

UnivarFeatureSelect.R hosted with ❤ by GitHub view raw

	GENE	SPEARMAN_RHO	PVALUE	FDR
256	ENSG00000184368.11_MAP7D2	-0.5730196	4.425151e-15	2.416132e-12
324	ENSG00000110013.8_SIAE	0.3403994	1.288217e-05	3.516833e-03
297	ENSG00000128487.12_SPECC1	-0.3003621	1.323259e-04	2.408332e-02
218	ENSG00000162512.11_SDC3	0.2945390	1.807649e-04	2.467441e-02
38	ENSG00000129007.10_CALML4	0.2879754	2.549127e-04	2.783647e-02
107	ENSG00000233429.5_HOTAIRM1	-0.2768054	4.489930e-04	4.085836e-02
278	ENSG00000185442.8_FAM174B	-0.2376098	2.731100e-03	2.130258e-01
421	ENSG00000234585.2_CCT6P3	-0.2322268	3.426233e-03	2.338404e-01
371	ENSG00000113312.6_TTC1	0.2284351	4.007655e-03	2.431310e-01
269	ENSG00000226329.2_AC005682.6	-0.2226587	5.064766e-03	2.523944e-01

Test feature-by-feature for correlation with phenotype of interest: aka DGE

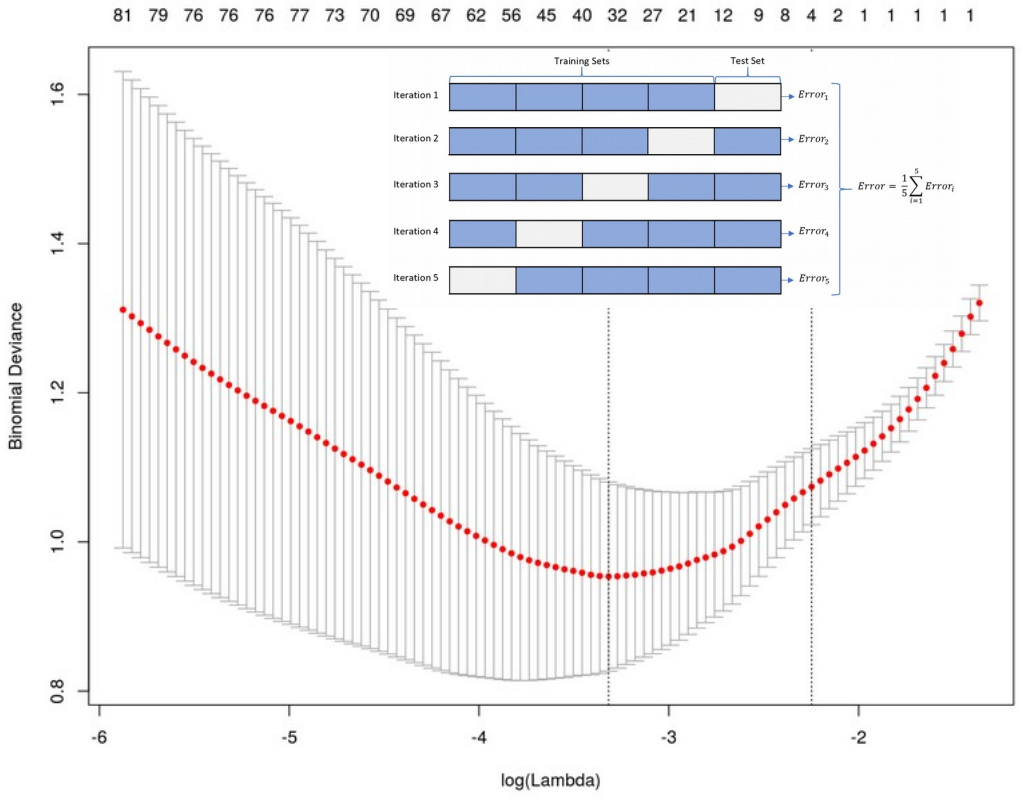
Supervised Multivariate Feature Selection



$$Y = \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

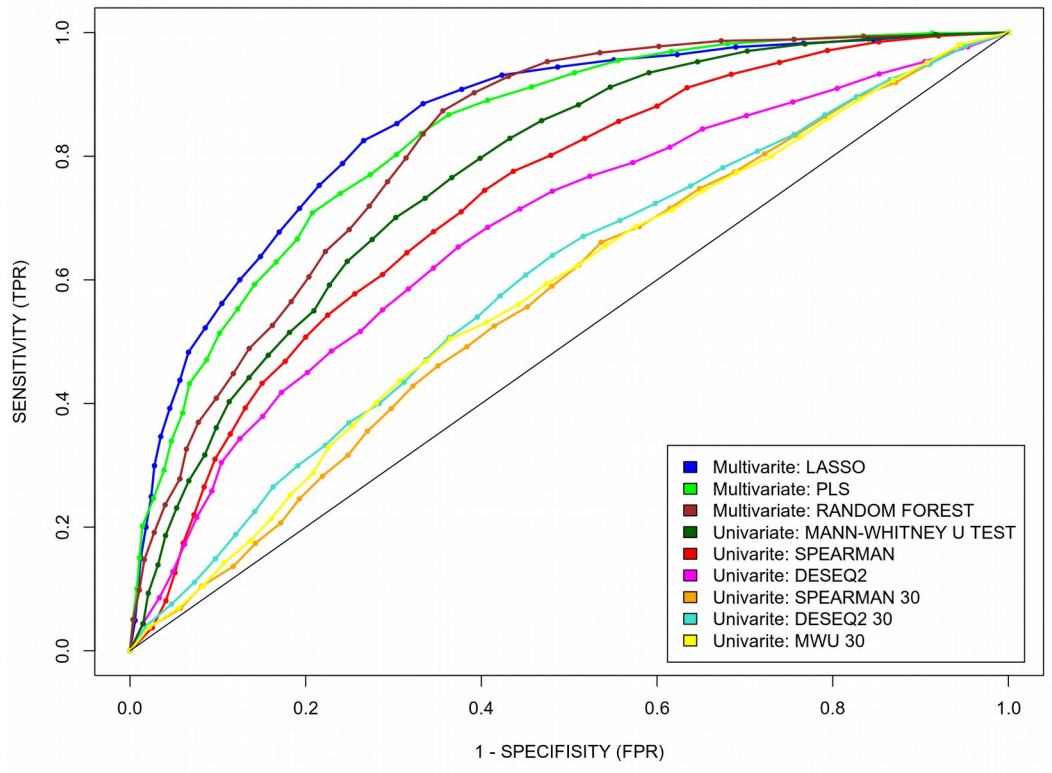
$$\text{OLS} = (y - \beta_1 X_1 - \beta_2 X_2)^2$$

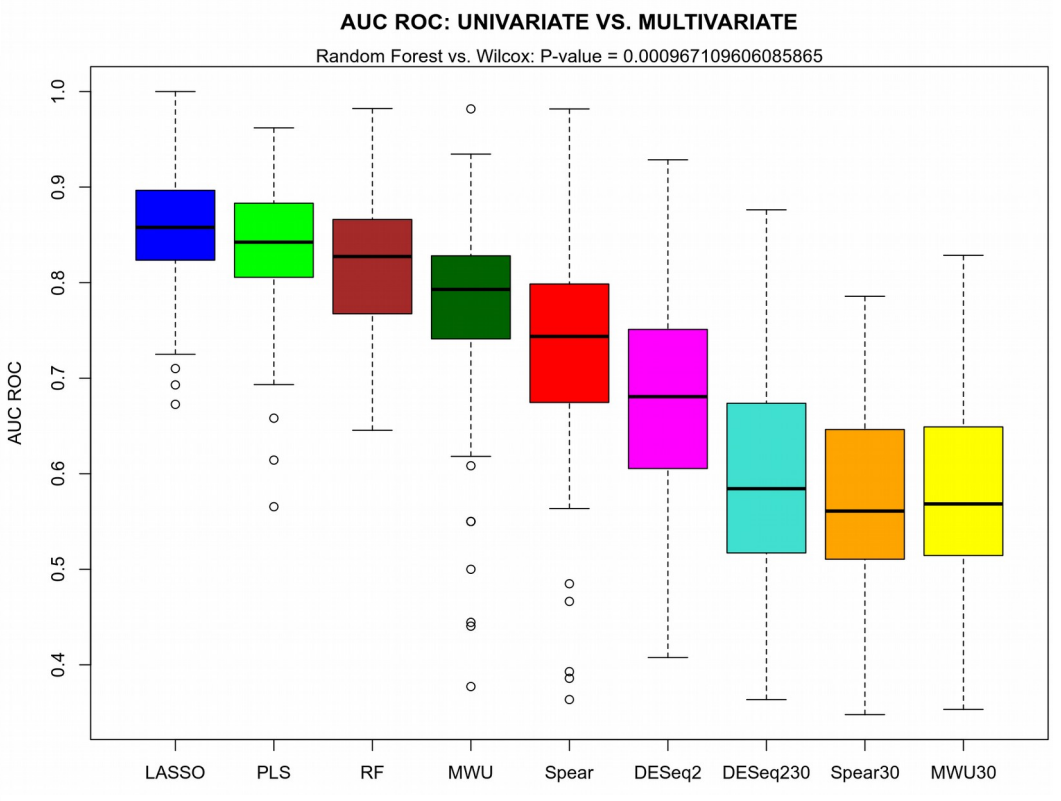
$$\text{Penalized OLS} = (y - \beta_1 X_1 - \beta_2 X_2)^2 + \lambda(|\beta_1| + |\beta_2|)$$



Select groups of features correlated with phenotype of interest: PLS, LASSO, LDA

Univariate vs. Multivariate Prediction

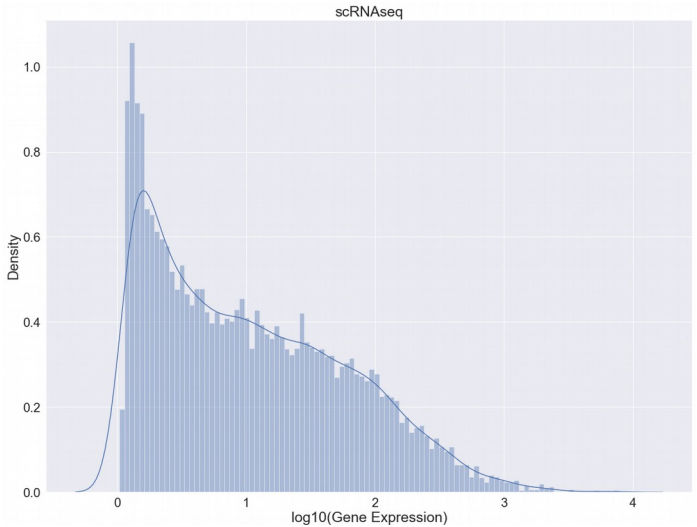


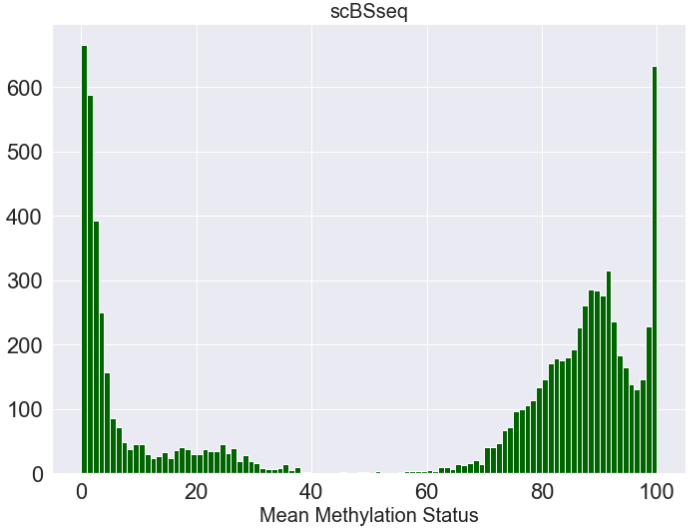


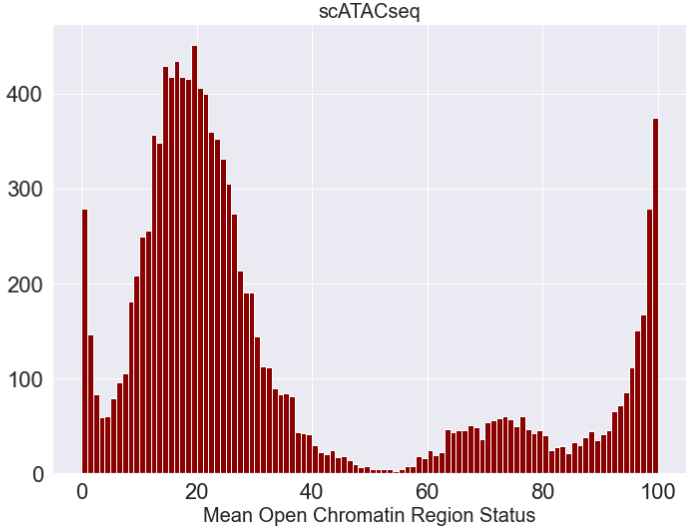
Multivariate feature selection generally has higher predictive capacity compared to univariate feature selection

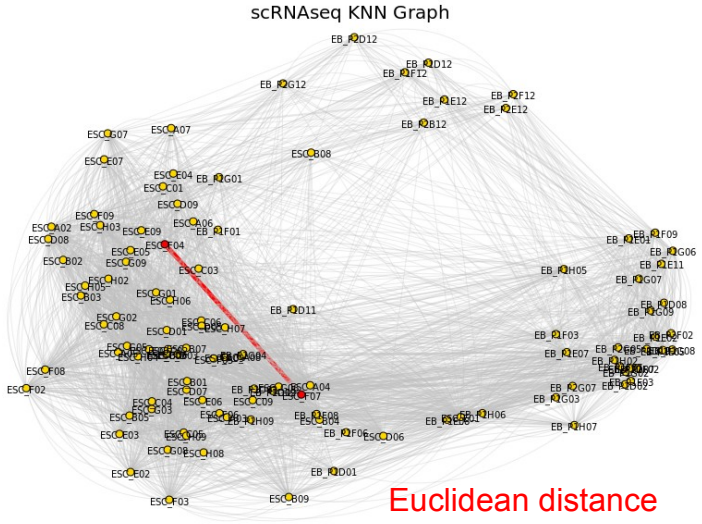
How Does This Work in Practice?
(an example from unsupervised integration)

Single Cell Multi-Omics: scNMTseq

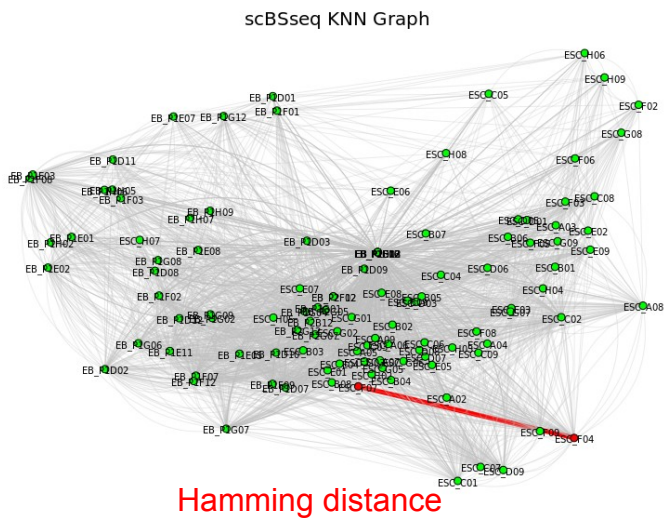




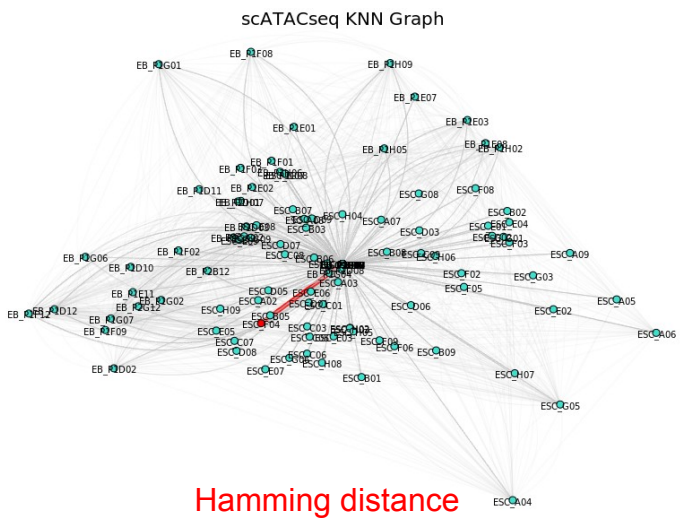




Euclidean distance



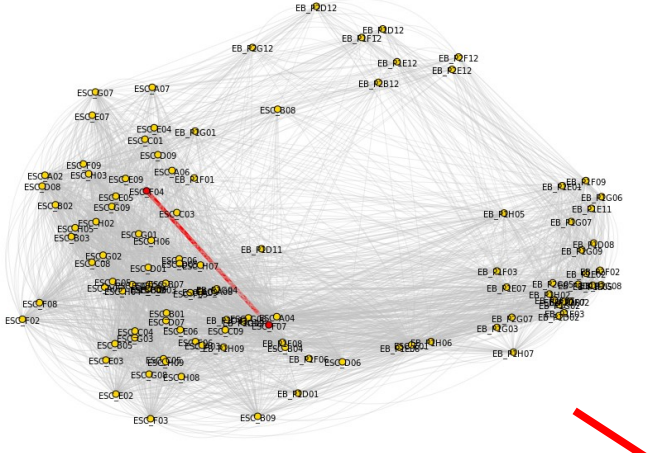
Hamming distance



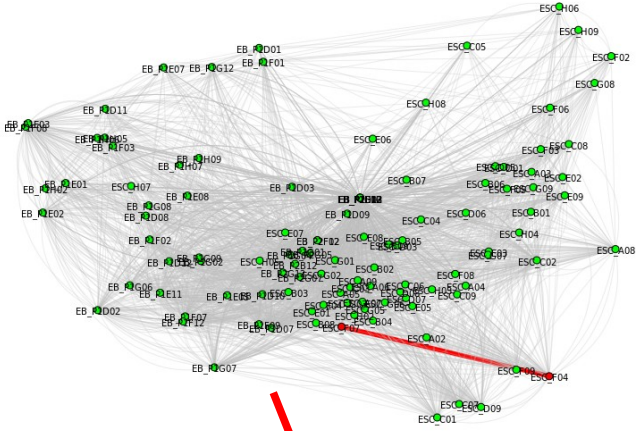
Hamming distance

Graph Intersection Method

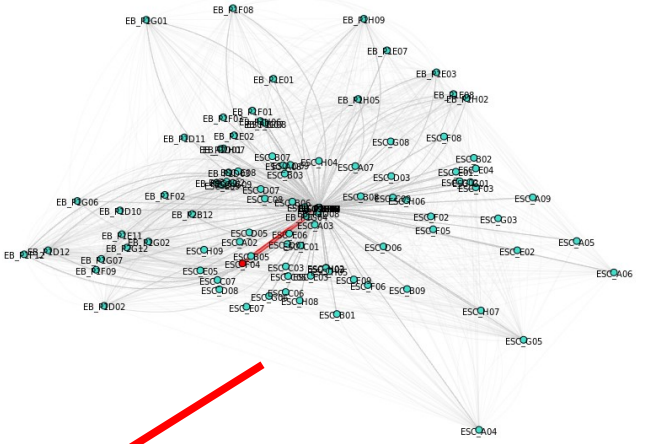
scRNAseq KNN Graph



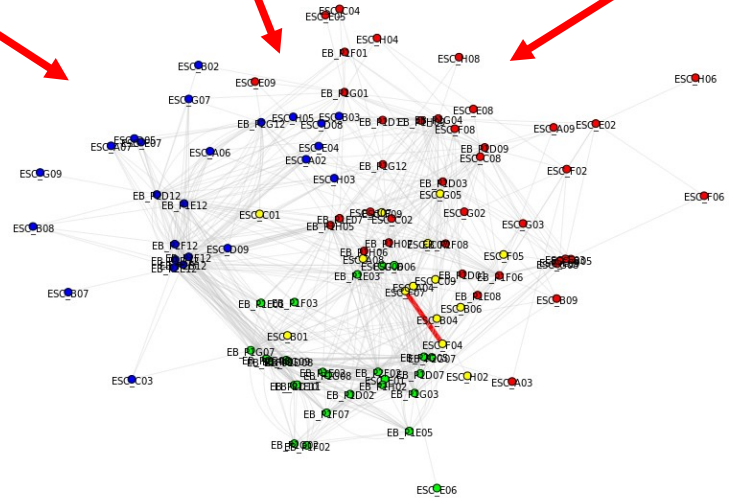
scBSseq KNN Graph



scATACseq KNN Graph

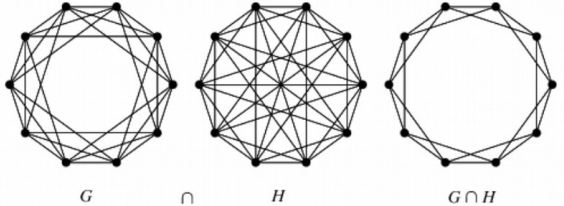


Consensus Graph



Graph Intersection

DOWNLOAD Wolfram Notebook

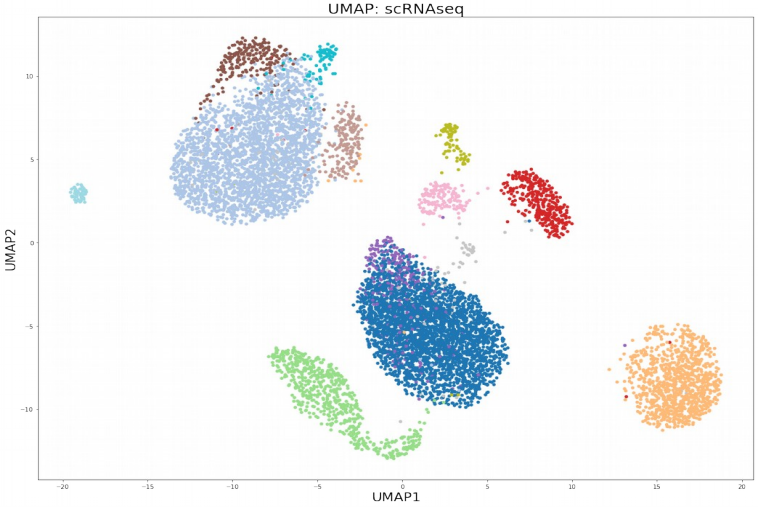


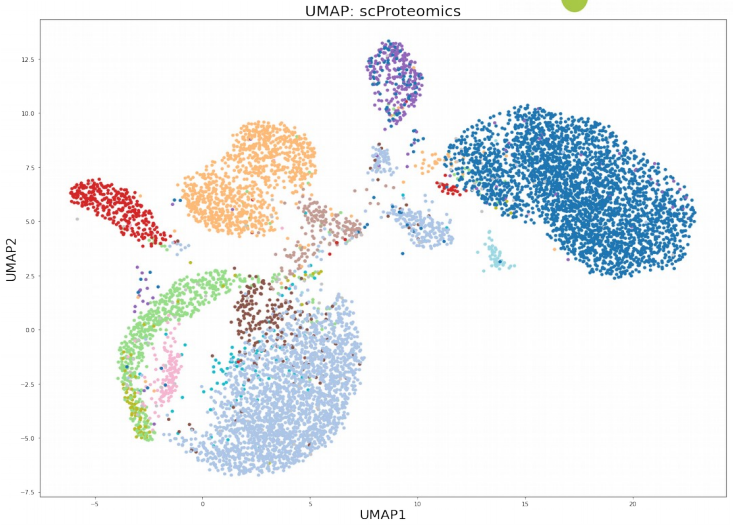
G ∩ H $G \cap H$

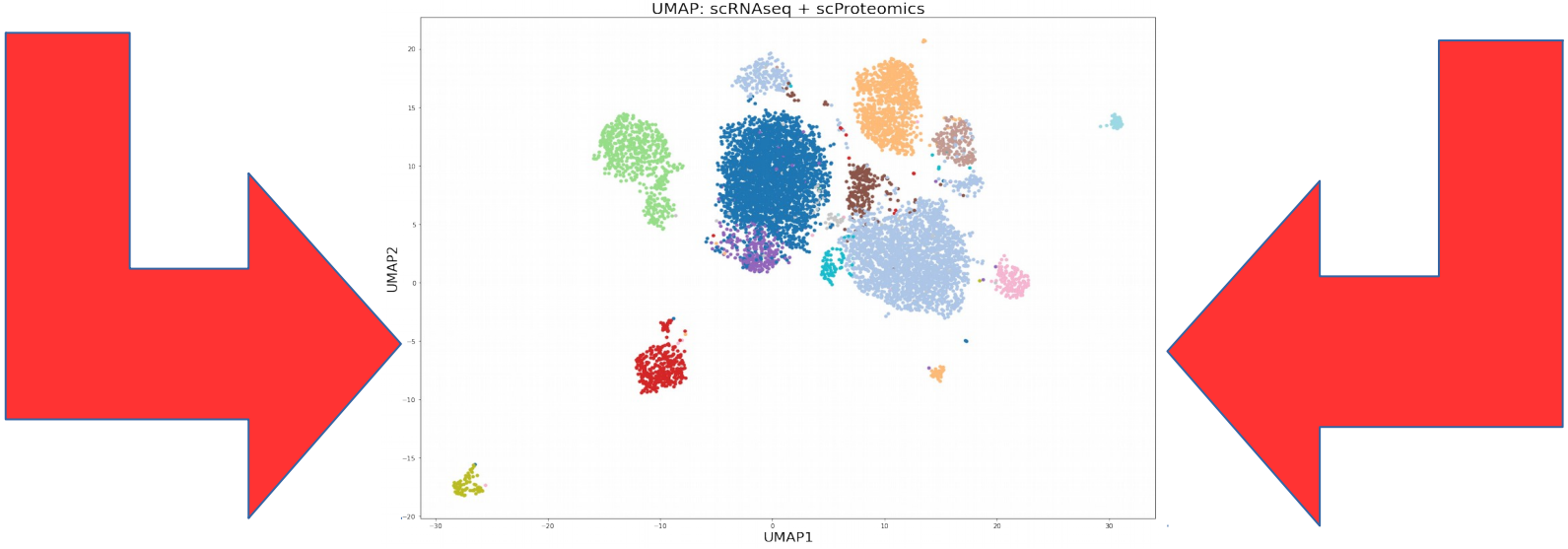
Let S be a set and $F = \{S_1, \ldots, S_p\}$ a nonempty family of distinct nonempty subsets of S whose union is $\bigcup_{i=1}^{p} S_i = S$. The intersection graph of F is denoted $\Omega(F)$ and defined by $V(\Omega(F)) = F$, with S_i and S_j adjacent whenever $i \neq j$ and $S_i \cap S_j \neq \emptyset$. Then a graph G is an intersection graph on S if there exists a family F of subsets for which G and $\Omega(F)$ are isomorphic graphs (Harary 1994, p. 19). Graph intersections can be computed in the Wolfram Language using `GraphIntersection[g, h]`.

Keep edges consistently present across the Omics

UMAP for Data Integration: CITEseq

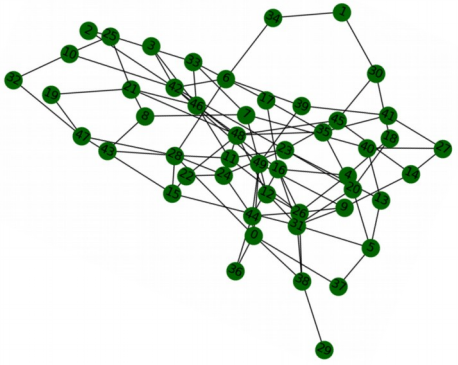






UMAP constructs high-dimensional graph

1) Construct high-dimensional graph



Original data

$X_{n,m}$, Perp

Compute high dimensional affinities

$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\Sigma_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)} \quad (1)$$

Construct matrix P by

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

$P_{n,n}$

Sample initial solution Y from $N(0, 10^{-4}I)$

$Y_{n,n}$

Compute low dimensional affinities

Optimize using gradient descent

Text

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\Sigma_{k\neq l}(1 + ||y_k - y_l||^2)^{-1}} \quad (2)$$

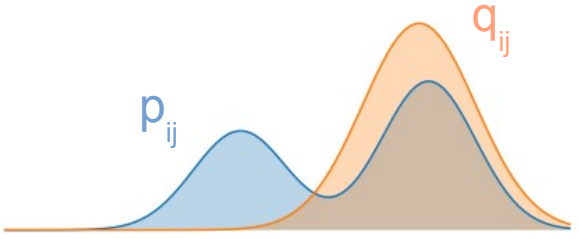
$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (3)$$

$Q_{n,n}$

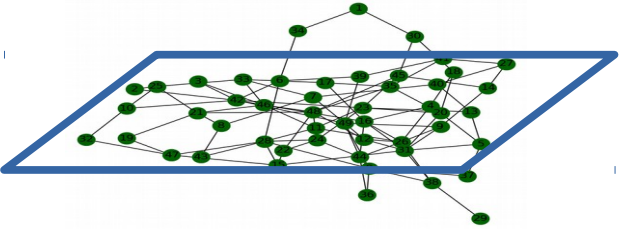
Construct matrix Q

Compute cost function

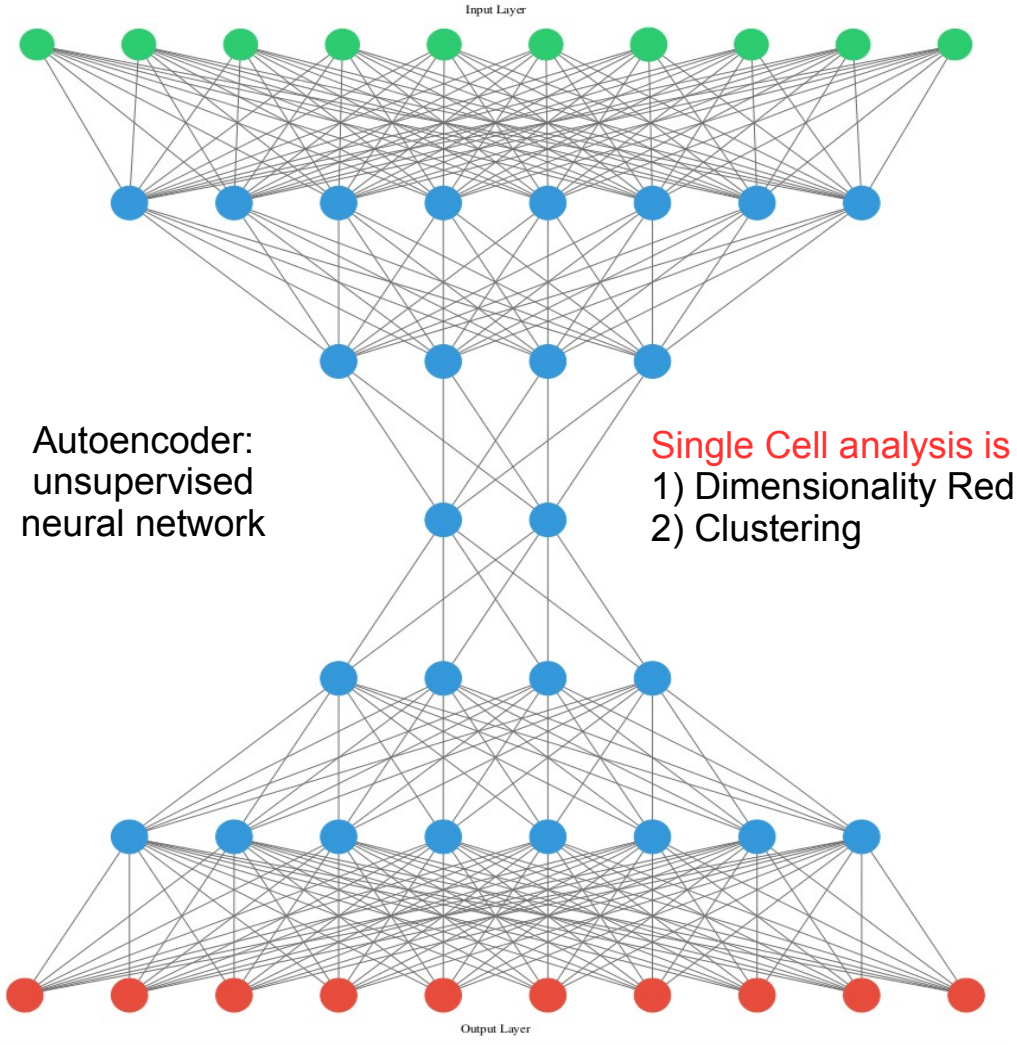
3) Collapse the graphs together



2) Construct low-dimensional graph

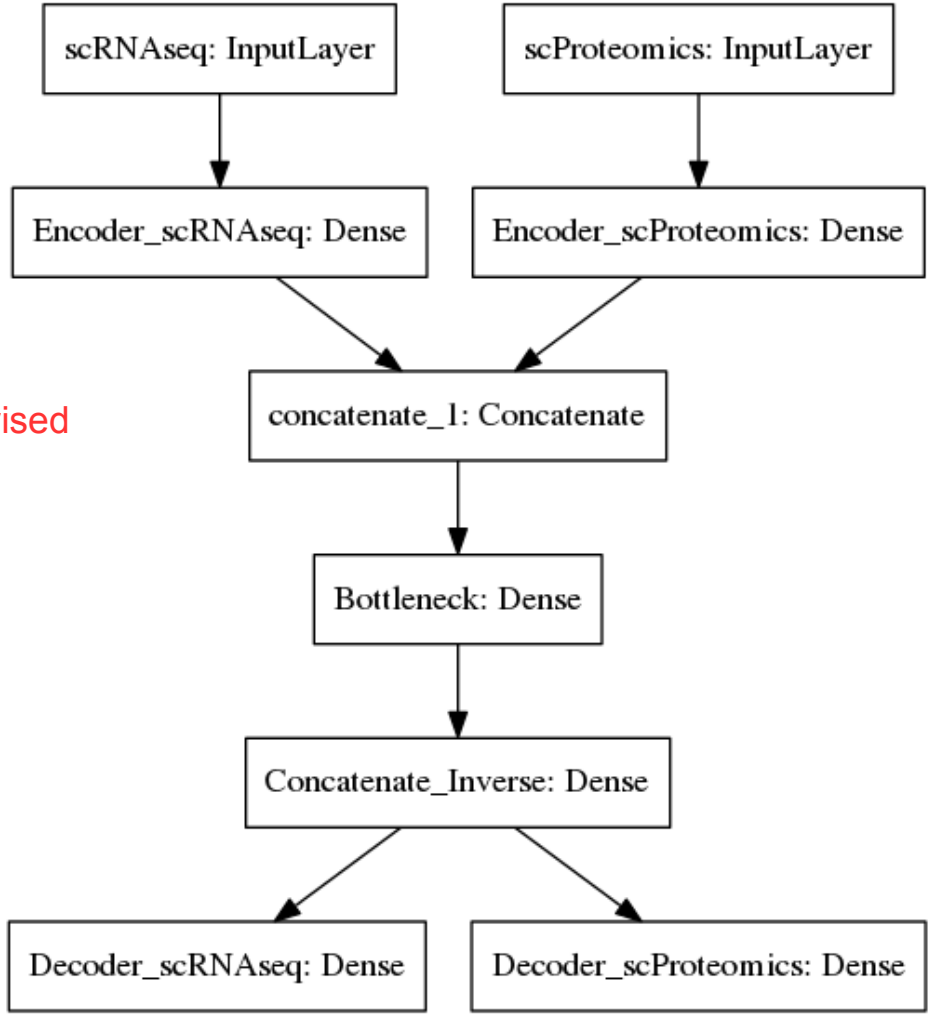


Data Integration via Neural Networks

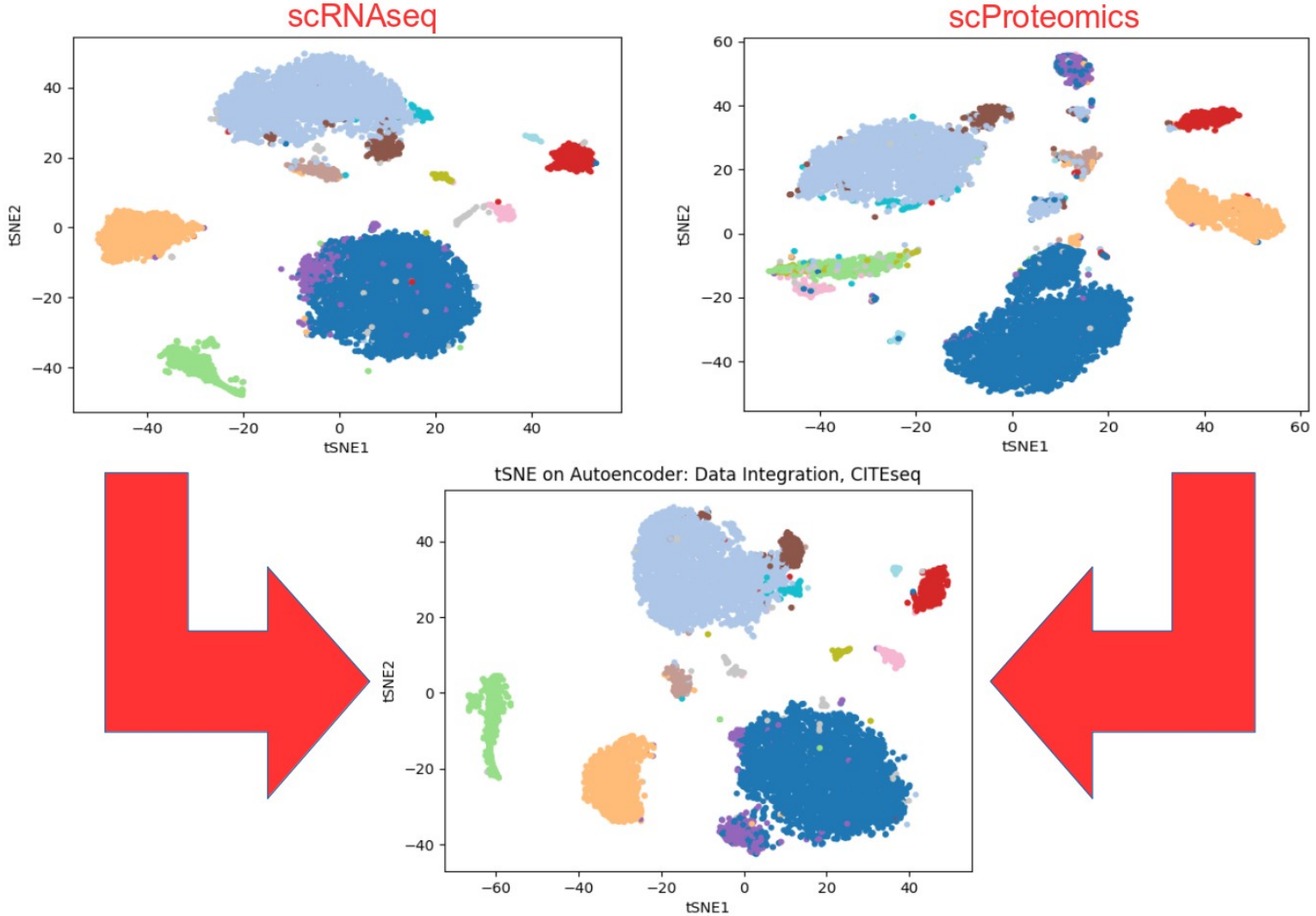


Autoencoder:
unsupervised
neural network

Single Cell analysis is unsupervised
1) Dimensionality Reduction
2) Clustering



Autoencoder for Data Integration: CITEseq



National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

LUNDS UNIVERSITET

Practical Demonstration

https://github.com/NikolayOskolkov/OsloBioinfoWeek2022