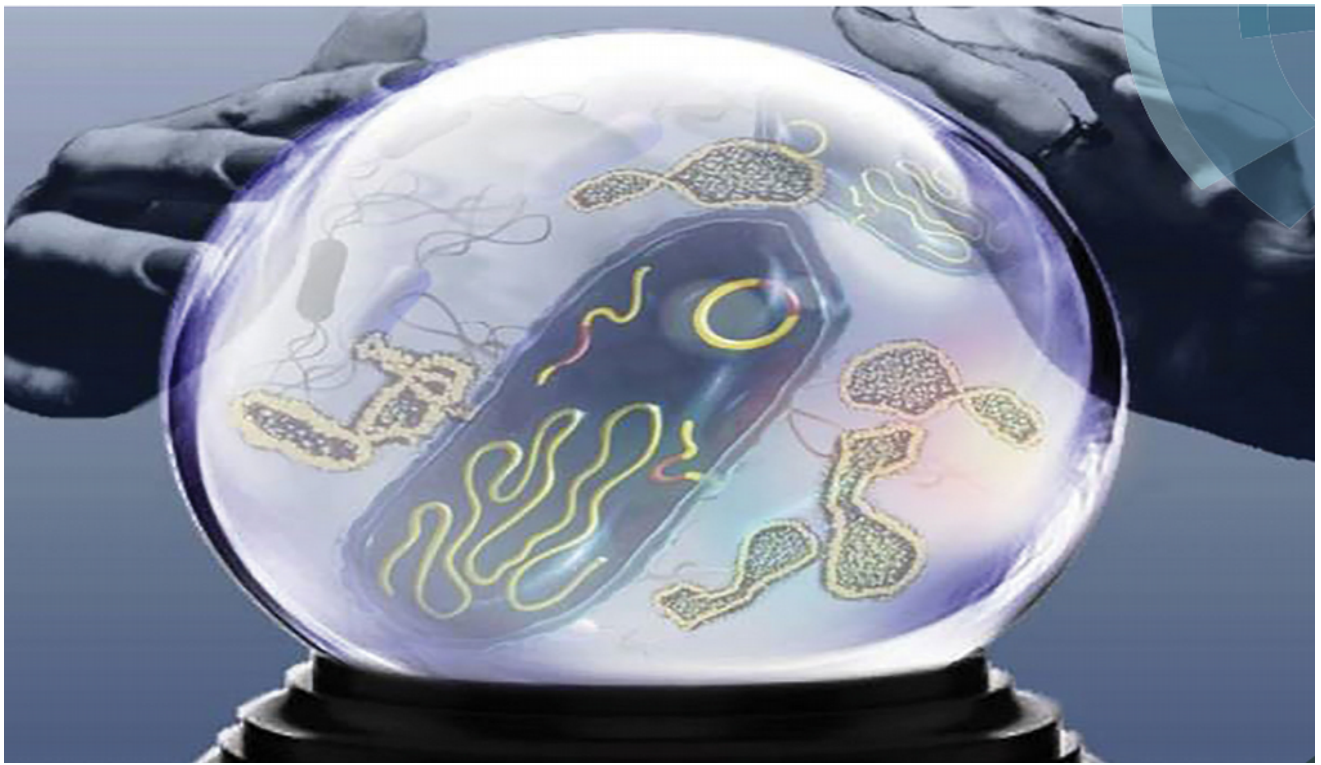
Multi-Omics Data Integration via Machine Learning

Omics Integration and Systems Biology course

Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden



@NikolayOskolkov

https://github.com/NikolayOskolkov

Image adapted from Molecular Omics, Issue 1, 2018

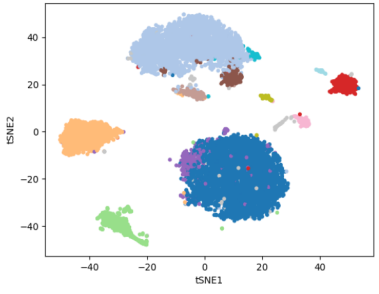
Brief introduction: who am I

2007 PhD in theoretical physics

2011 medical genetics at Lund University

2016 working at NBIS SciLifeLab, Sweden

Single cell



Biomedical data integration

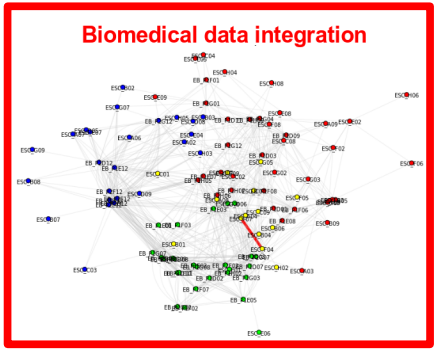
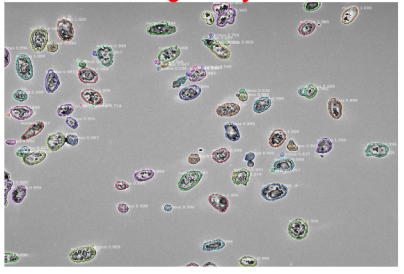
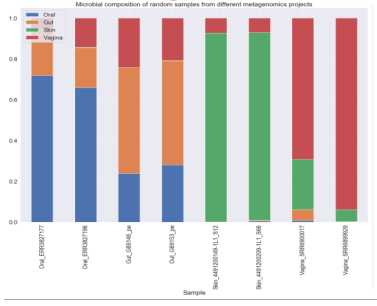


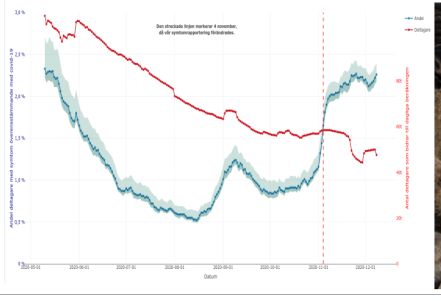
Image analysis



Metagenomics

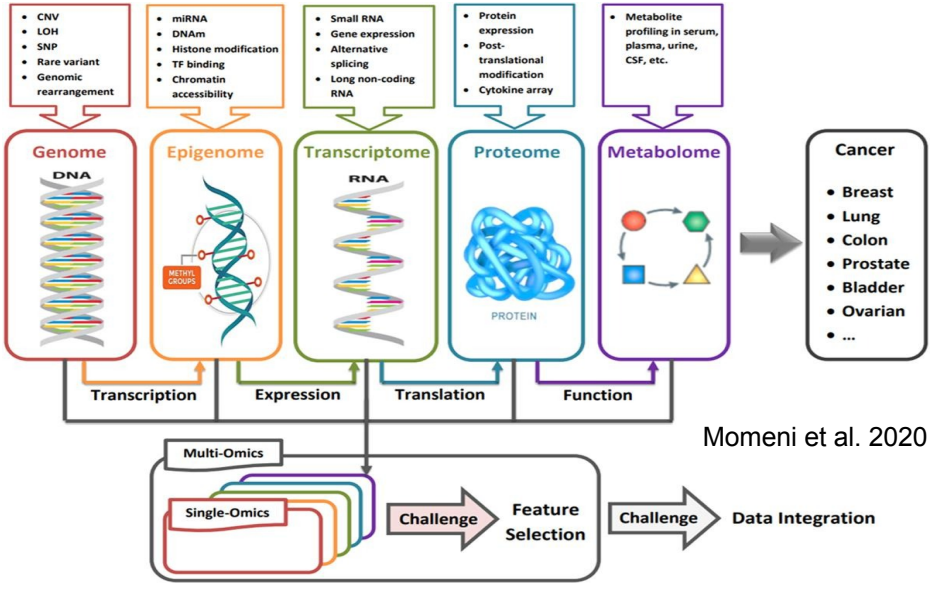


Epidemiology

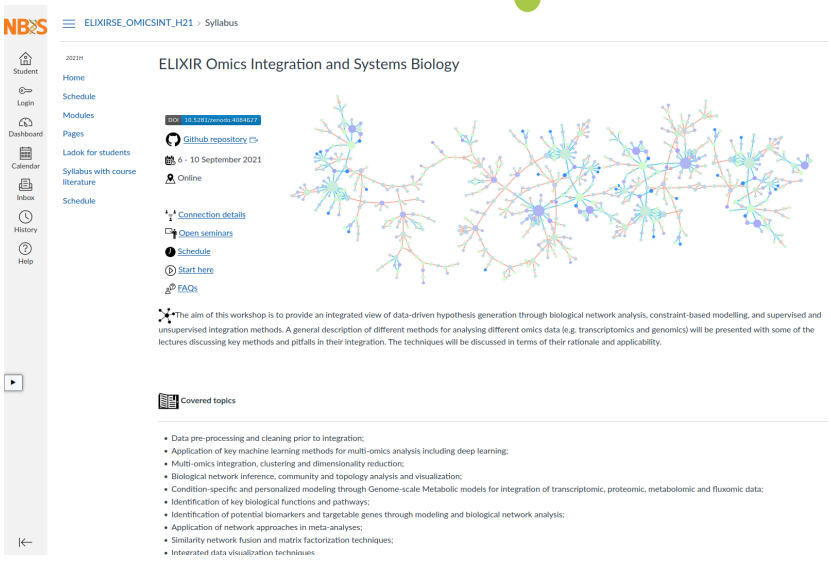


Ancient microbiome

Multi-Omics Begins: 2015 – until now



Momeni et al. 2020



https://github.com/NBISweden/workshop_omics_integration

Rönn et al., Scientific Reports 2024

BOUGHT BY:

CHAN ZUCKERBERG INITIATIVE

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

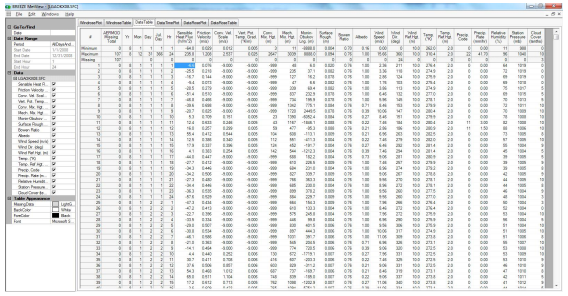
LEADING PROFESSIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY AND BIOINFORMATICS
CONNECTING, TRAINING, EMPOWERING, WORLDWIDE

Last run: February 2023, 94 applications

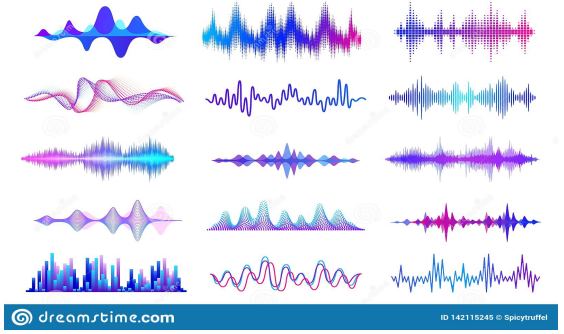
Introduction:
High Dimensional Biological Data

Various types of data around us

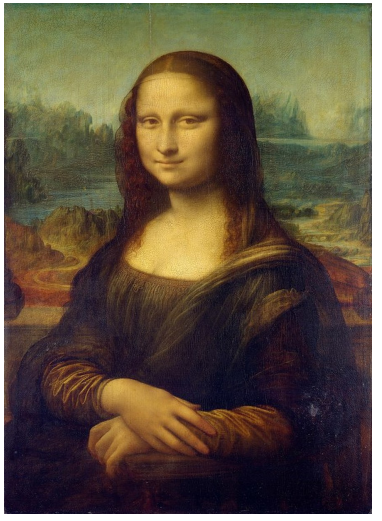
Tabular



Sound



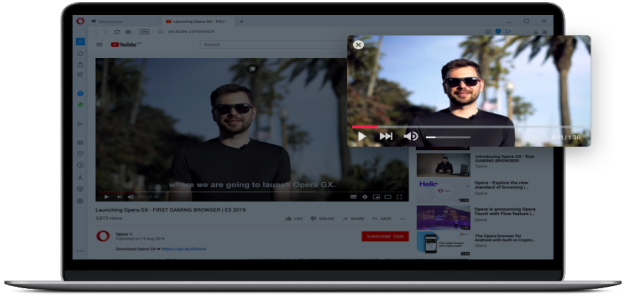
Image



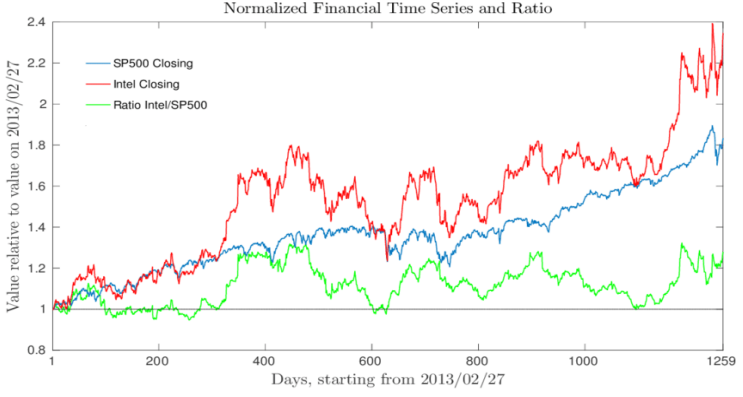
Text

DATA

Video



Time Series



Biological Data are High Dimensional

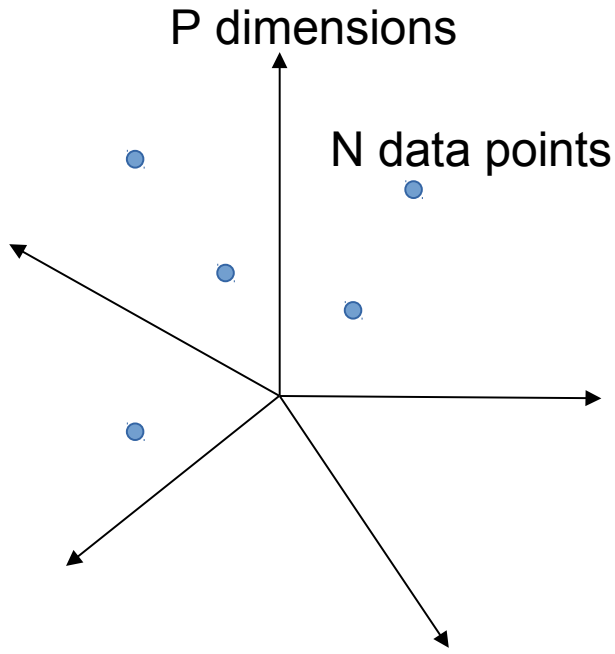
Statistical observations:
e.g. samples, cells etc.

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

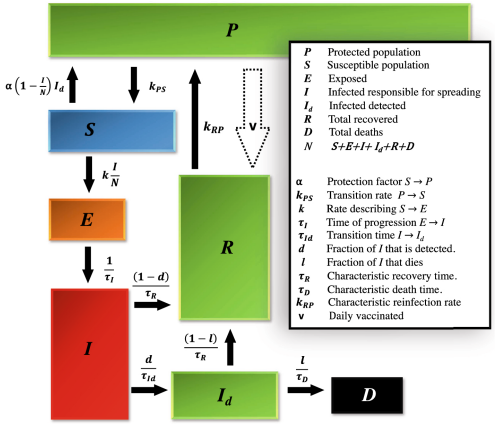
For a robust statistical analysis, one should properly “sample” the P-dimensional space, hence large sample size is required, N >> P

Some types of data analysis in Life Sciences

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

Biology / Biomedicine

Mathematical modeling



N ⟶ 0

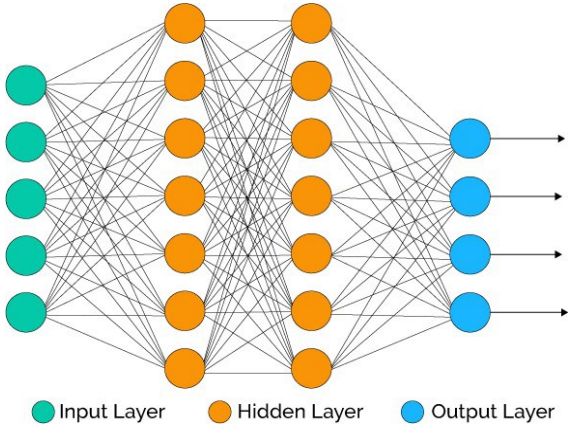
Bayesianism

N << P

Frequentism

N ≈ P

Machine Learning



N >> P

Hypothesis-driven ⟶ **Data-driven**

Amount of Data

The Curse of Dimensionality

Ex.1

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

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

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

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

Biased ML variance estimator in HD-space

Some peculiarities of Frequentist statistics

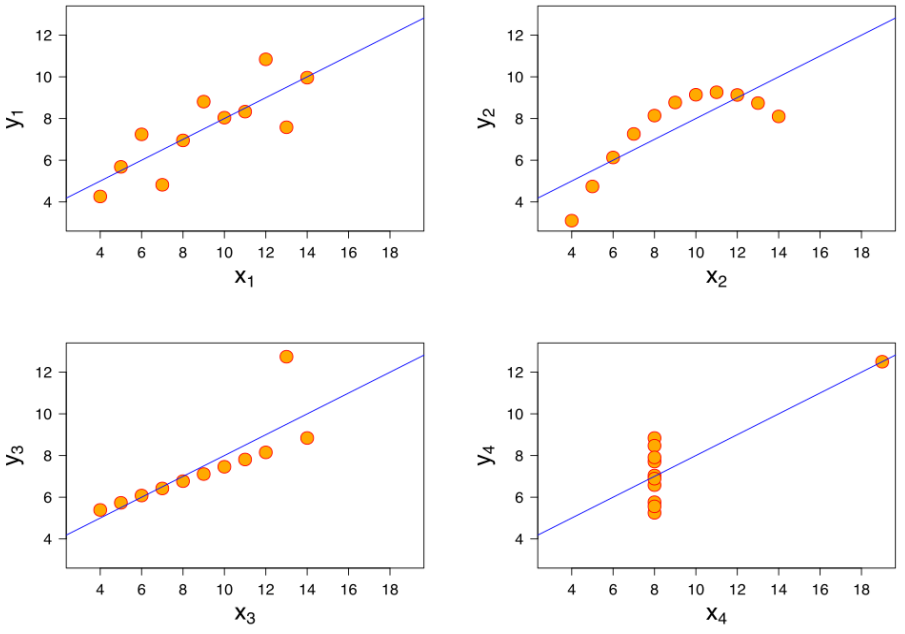
- Maximum likelihood based
- Focus on summary statistics
- Focus too much on p-values

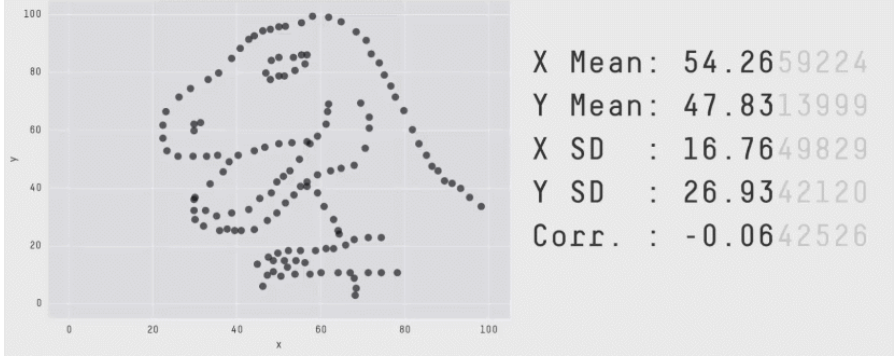
$$\mathrm{L}\left(\mathrm{x_i} \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{-\frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} (\mathrm{x_i} - \mu)^2}{2\sigma^2}}$$

$$\frac{\partial \mathrm{L}\left(\mathrm{x_i} \mid \mu, \sigma^2\right)}{\partial \mu} = 0; \quad \frac{\partial \mathrm{L}\left(\mathrm{x_i} \mid \mu, \sigma^2\right)}{\partial \sigma^2} = 0$$

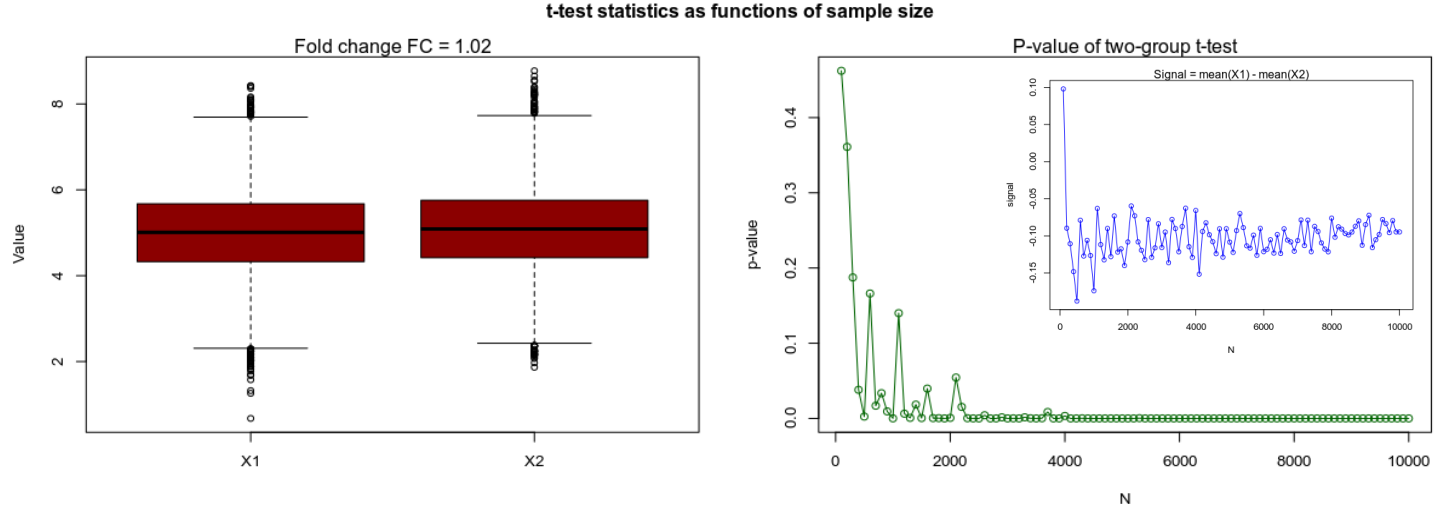
$$\mu = \frac{1}{N} \sum_{i=0}^{N} x_i \; - \text{mean estimator}$$

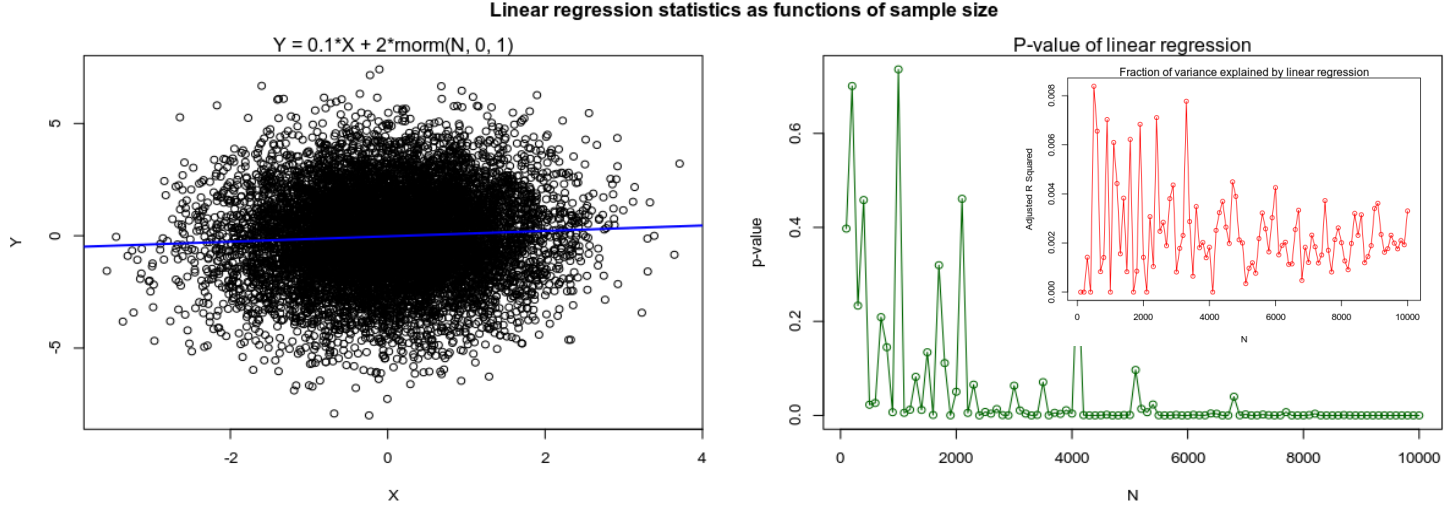
$$\sigma^2 = \frac{1}{N} \sum_{i=0}^{N} (x_i - \mu)^2 \; - \text{variance estimator}$$





Frequentist stats: too much focus on p-values





It is questionable whether p-value is the best metric for ranking features

Frequentist stats struggles with high-dimensional data

```r
n <- 20 # number of samples
p <- 2  # number of features / dimensions
Y <- rnorm(n)
X <- matrix(rnorm(n * p), n, p)
summary(lm(Y ~ X))
```

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0522 -0.6380  0.1451  0.3911  1.8829

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.14950    0.22949   0.651    0.523
X1          -0.09405    0.28245  -0.333    0.743
X2          -0.11919    0.24486  -0.487    0.633

Residual standard error: 1.017 on 17 degrees of freedom
Multiple R-squared:  0.02204,   Adjusted R-squared:  -0.09301
F-statistic: 0.1916 on 2 and 17 DF,  p-value: 0.8274
```

Going to higher dimensions →

```r
n <- 20 # number of samples
p <- 10 # number of features / dimensions
Y <- rnorm(n)
X <- matrix(rnorm(n * p), n, p)
summary(lm(Y ~ X))
```

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0255 -0.4320  0.1056  0.4493  1.0617

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.54916    0.26472   2.075   0.0679 .
X1           0.30013    0.21690   1.384   0.1998
X2           0.68053    0.27693   2.457   0.0363 *
X3          -0.10675    0.26010  -0.410   0.6911
X4          -0.21367    0.33690  -0.634   0.5417
X5          -0.19123    0.31881  -0.600   0.5634
X6           0.81074    0.25221   3.214   0.0106 *
X7           0.09634    0.24143   0.399   0.6992
X8          -0.29864    0.19004  -1.571   0.1505
X9          -0.78175    0.35408  -2.208   0.0546 .
X10          0.83736    0.36936   2.267   0.0496 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8692 on 9 degrees of freedom
Multiple R-squared:  0.6592,    Adjusted R-squared:  0.2805
F-statistic: 1.741 on 10 and 9 DF,  p-value: 0.2089
```

Going to even higher dimensions →

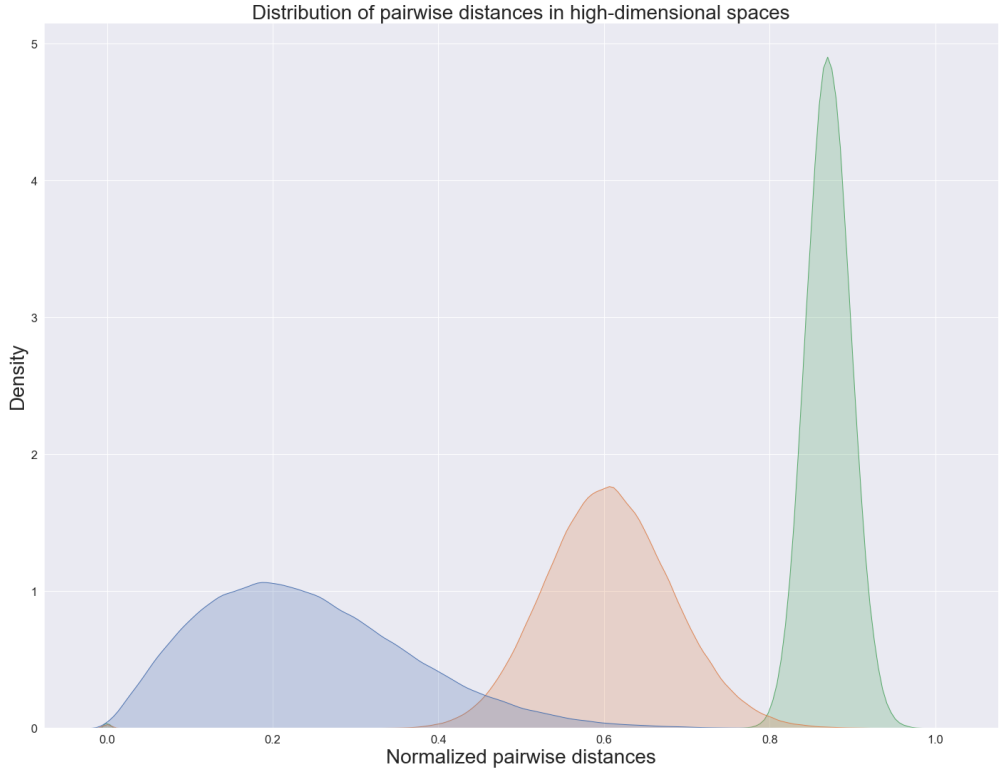
```r
n <- 20 # number of samples
p <- 20 # number of features / dimensions
Y <- rnorm(n)
X <- matrix(rnorm(n * p), n, p)
summary(lm(Y ~ X))
```

```
Call:
lm(formula = Y ~ X)

Residuals:
ALL 20 residuals are 0: no residual degrees of freedom!

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.34889        NaN     NaN      NaN
X1           0.66218        NaN     NaN      NaN
X2           0.76212        NaN     NaN      NaN
X3          -1.35033        NaN     NaN      NaN
X4          -0.57487        NaN     NaN      NaN
X5           0.02142        NaN     NaN      NaN
X6           0.40290        NaN     NaN      NaN
X7           0.03313        NaN     NaN      NaN
X8          -0.31983        NaN     NaN      NaN
X9          -0.92833        NaN     NaN      NaN
X10          0.18091        NaN     NaN      NaN
X11         -1.37618        NaN     NaN      NaN
X12          2.11438        NaN     NaN      NaN
X13         -1.75103        NaN     NaN      NaN
X14         -1.55073        NaN     NaN      NaN
X15          0.01112        NaN     NaN      NaN
X16         -0.50943        NaN     NaN      NaN
X17         -0.47576        NaN     NaN      NaN
X18          0.31793        NaN     NaN      NaN
X19          1.43615        NaN     NaN      NaN
X20               NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:    NaN
F-statistic:   NaN on 19 and 0 DF,  p-value: NA
```

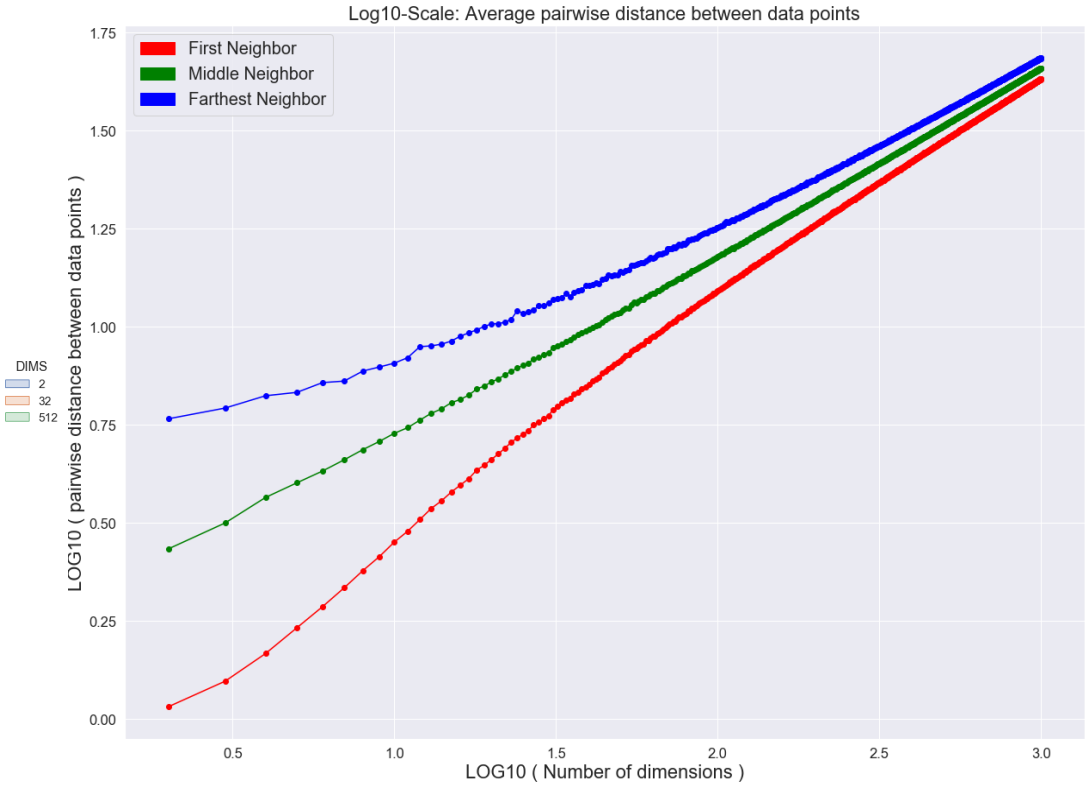
This is another way we face the Curse of Dimensionality in computational biology

More on the Curse of Dimensionality

Distances in High Dimensions



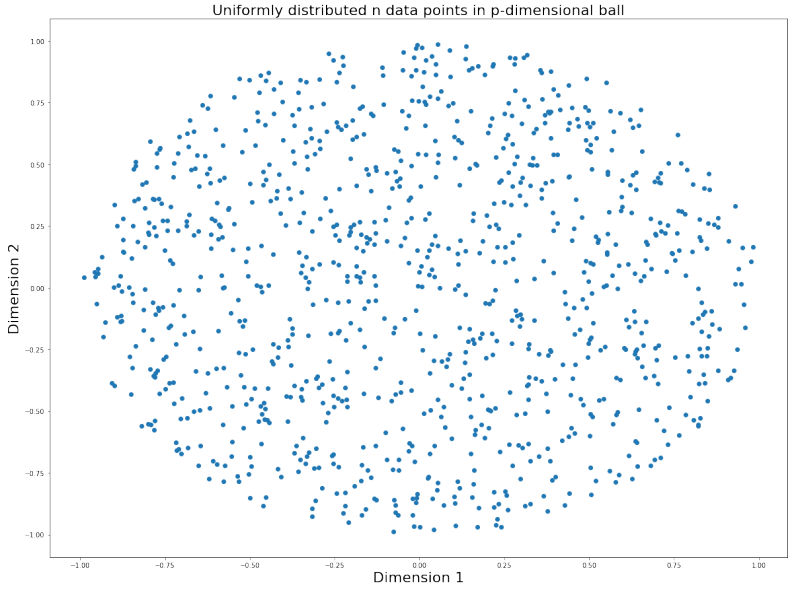
Data points become far from each other and equidistant in high dimensions

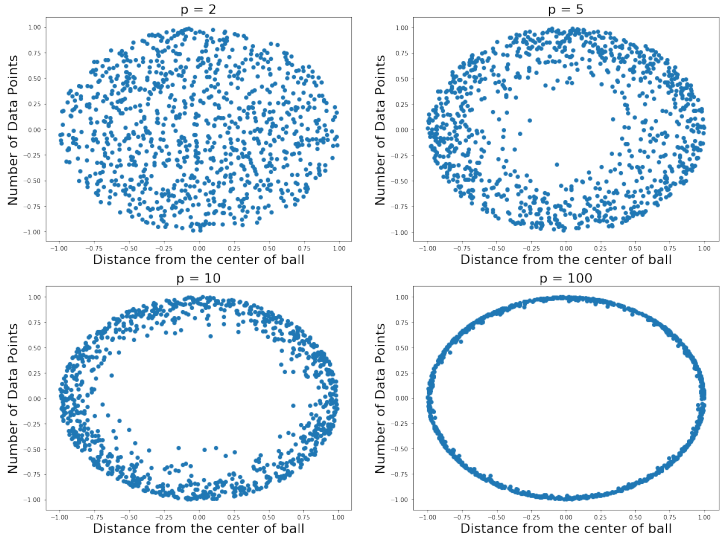


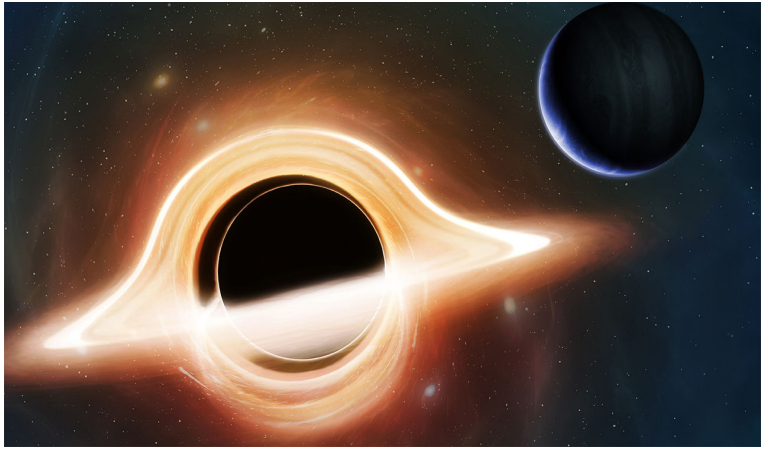
The differences between closest and farthest data point neighbours disappears in high-dimensional spaces: can't run cluster analysis

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

High-dimensional data make black holes







High-dimensional data can be viewed as having a **“hole in the middle”**, hence the concept of mean / centroid loses its validity, hence we can’t use Gaussian distribution

Literature on the Curse of Dimensionality

POINTS OF SIGNIFICANCE

The curse(s) of dimensionality

There is such a thing as too much of a good thing.

Naomi Altman and Martin Krzywinski

We generally think that more information is better than less. However, in the 'big data' era, the sheer number of variables that can be collected from a single sample can be problematic. This embarrassment of riches is called the 'curse of dimensionality'[1] (CoD) and manifests itself in a variety of ways. This month, we discuss four important problems of dimensionality as it applies to data sparsity[1,2], multicollinearity[3], multiple testing[4] and overfitting[5]. These effects are amplified by poor data quality, which may increase with the number of variables.

Throughout, we use n to indicate the sample size from the population of interest and p to indicate the number of observed variables, some of which may have missing values for some samples. For example, we may have $n = 1{,}000$ subjects and $p = 200{,}000$ single-nucleotide polymorphisms (SNPs).

First, as the dimensionality p increases, the 'volume' that the samples may occupy grows rapidly. We can think of each of the n

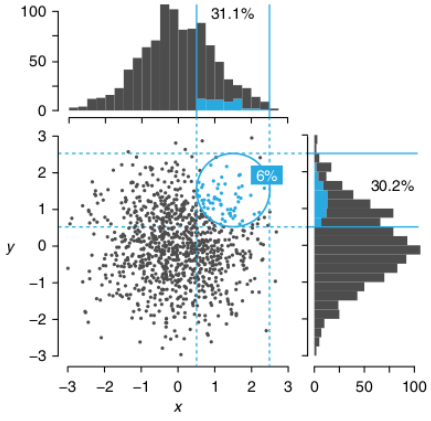


Fig. 1 | Data tend to be sparse in higher dimensions. Among 1,000 (x, y) points in which both x and y are normally distributed with a mean of 0 and s.d. $\sigma = 1$, only 6% fall within σ of $(x, y) = (1.5, 1.5)$ (blue circle). However, when the data are projected into a lower dimension—shown by histograms—about 30% of the points (all bins within blue solid lines) are within σ of 1.5. Blue

A and 100 to have the minor allele a. If we tabulate on two SNPs, A and B, we will expect only ten samples to exhibit both minor alleles with genotype ab. With SNPs A, B and C, we expect only one sample to have genotype abc, and with four or more SNPs, we expect empty cells in our table. We need a much larger sample size to observe samples with all the possible genotypes. As p increases, we may quickly find that there are no samples with similar values of a predictor.

Even with just five SNPs, our ability to predict and classify the samples is impeded because of the small number of subjects that have similar genotypes. In situations where there are many gene variants, this effect is exacerbated, and it may be very difficult to find affected subjects with similar genotypes and hence to predict or classify on the basis of genetic similarity.

If we treat the distance between points (e.g., Euclidian distance) as a measure of similarity, then we interpret greater distance as greater dissimilarity. As p increases, this

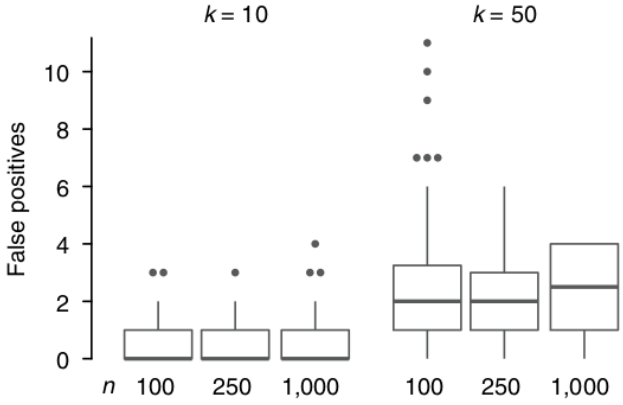


Fig. 3 | The number of false positives increases with each additional predictor. The box plots show the number of false positive regression-fit P values (tested at $\alpha = 0.05$) of 100 simulated multiple regression fits on various numbers of samples ($n = 100$, 250 and 1,000) in the presence of one true predictor and $k = 10$ and 50 extraneous uncorrelated predictors. Box plots show means (black center lines), 25th and 75th percentiles (box edges), and minimum and maximum values (whiskers). Outliers (dots) are jittered.

Correcting for multiple testing does not solve the problem of too many false-positive hits

Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018 Jun;15(6):399-400. doi: 10.1038/s41592-018-0019-x. PMID: 29855577.

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

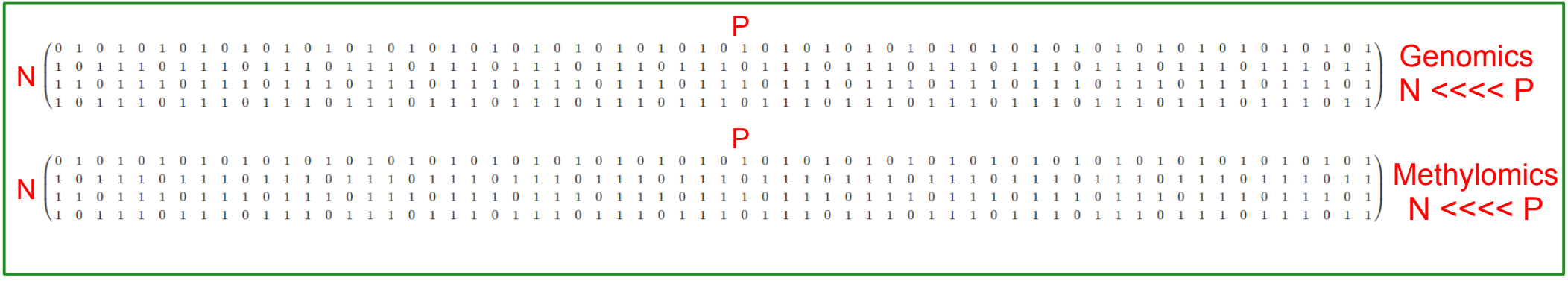
← **manageable**

P

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

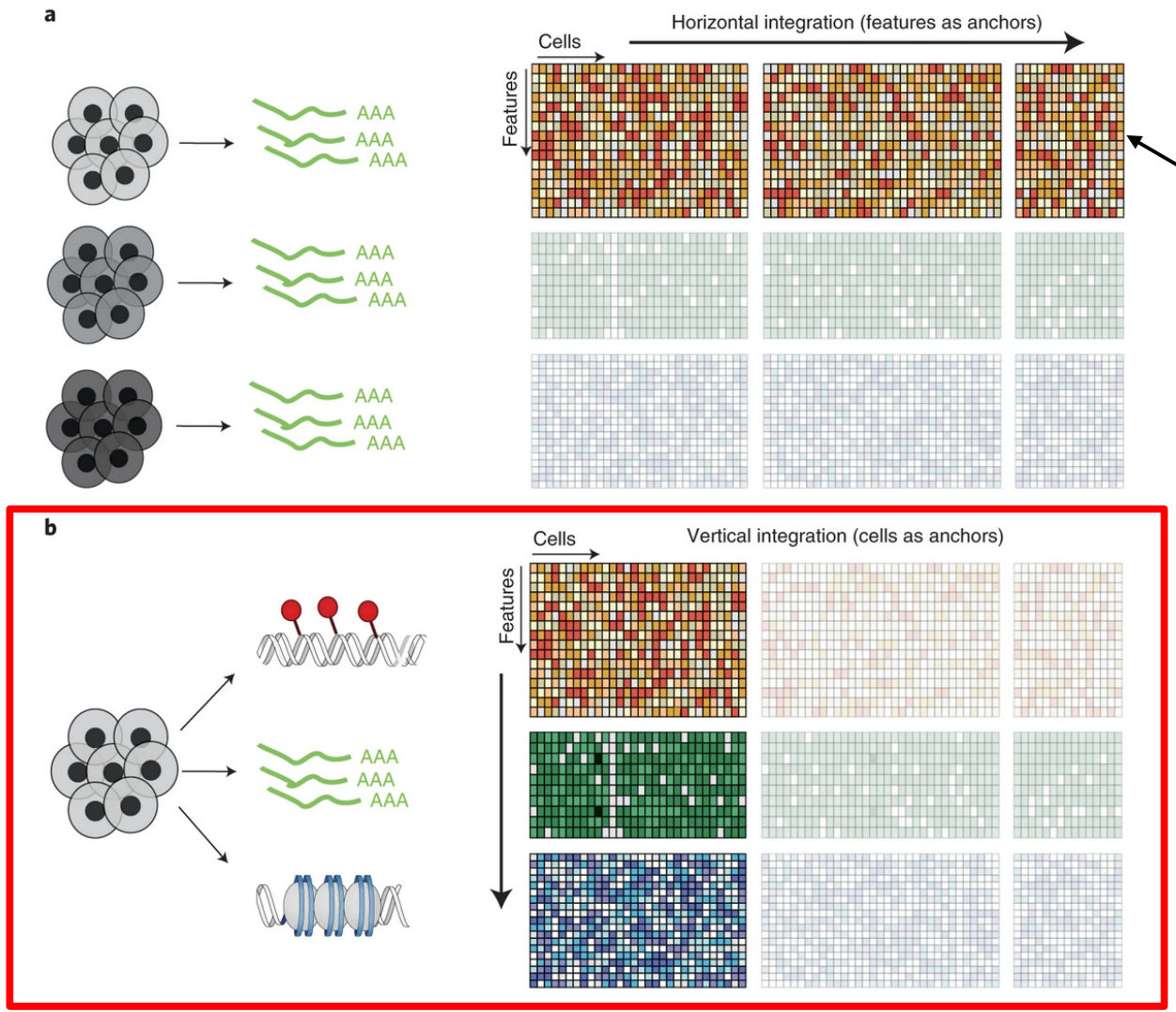
Transcriptomics
N << P
(Single cell: N <= P)

challenging



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

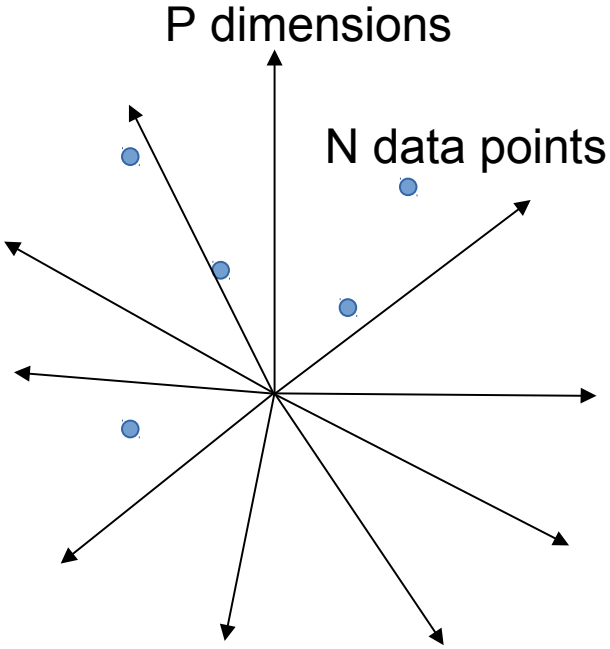
Omic1: RNAseq

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



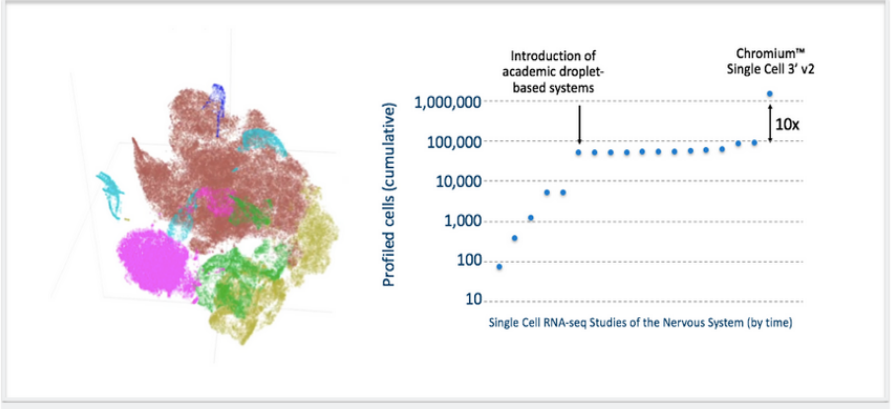
$P_1 + P_2 >> N$ integration across features leads to even more high-dimensional data

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

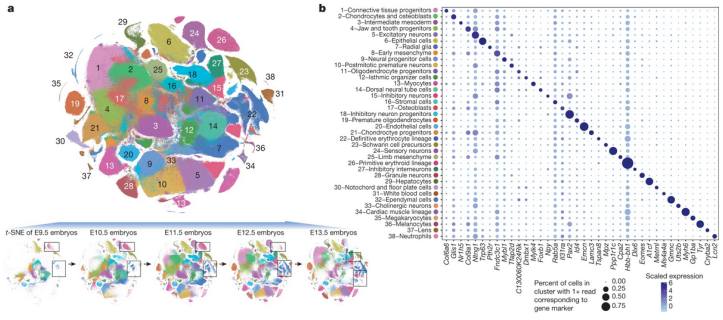
POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

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

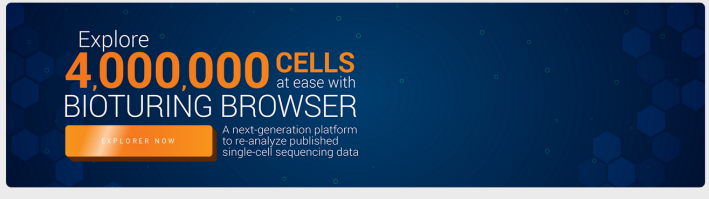
Fig. 2: Identifying the major cell types of mouse organogenesis.

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



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.

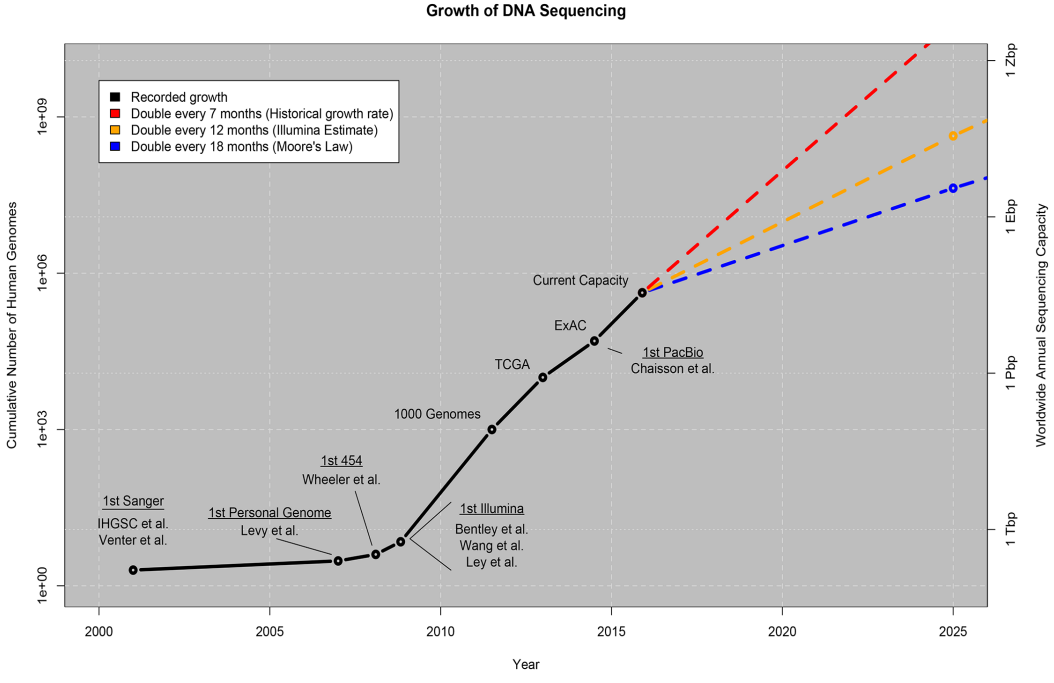
Search

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

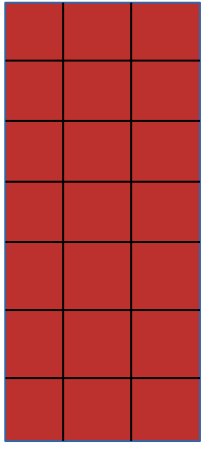
Big in Size or Sample Size?



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

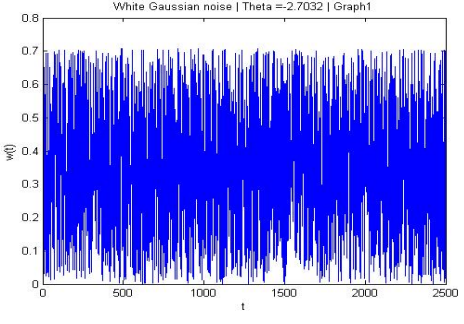
Genomics / WGS: **Little Data**

$N_1 \sim 10^3$



$P_1 \sim 10^6$

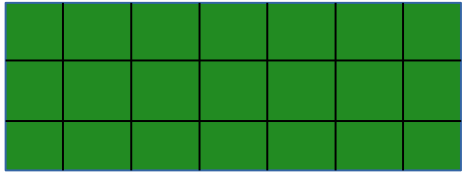
$N_1 * P_1 = N_2 * P_2 = 10^9$



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

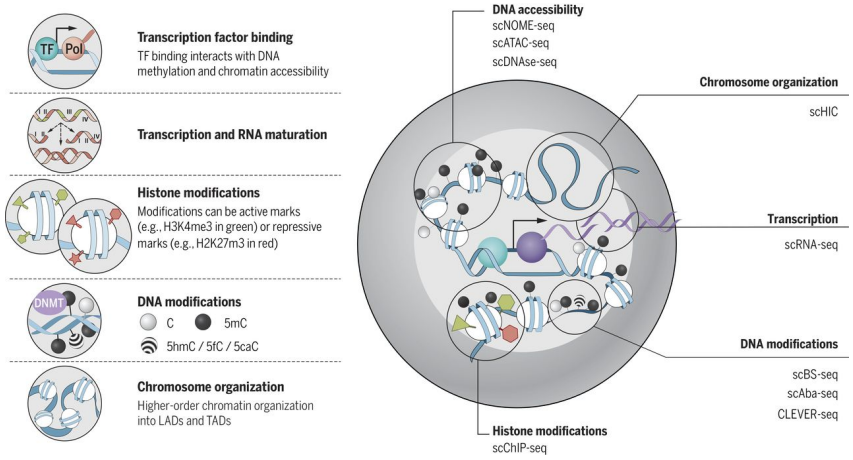
scRNAseq: **Big Data**

$N_2 \sim 10^6$

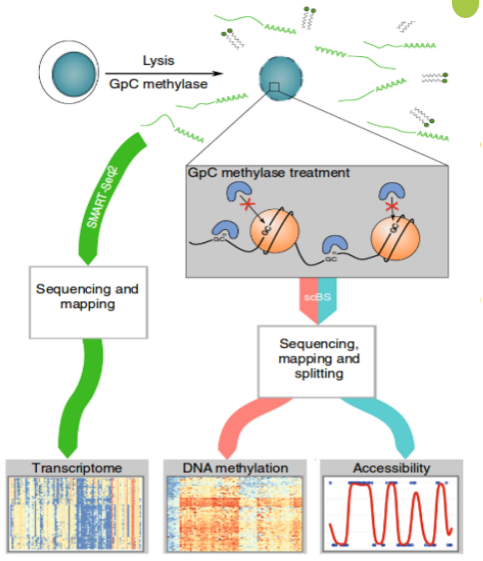


$P_2 \sim 10^3$

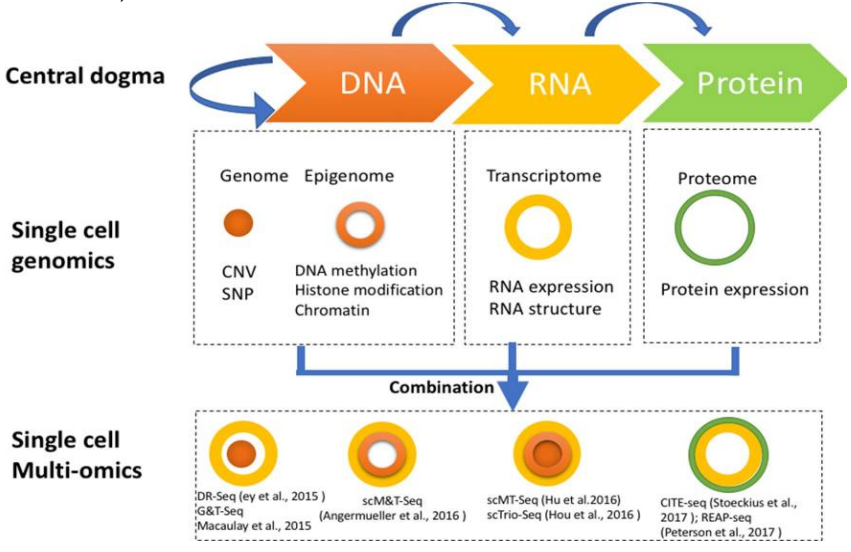
Multi-Omics in Single Cell Genomics



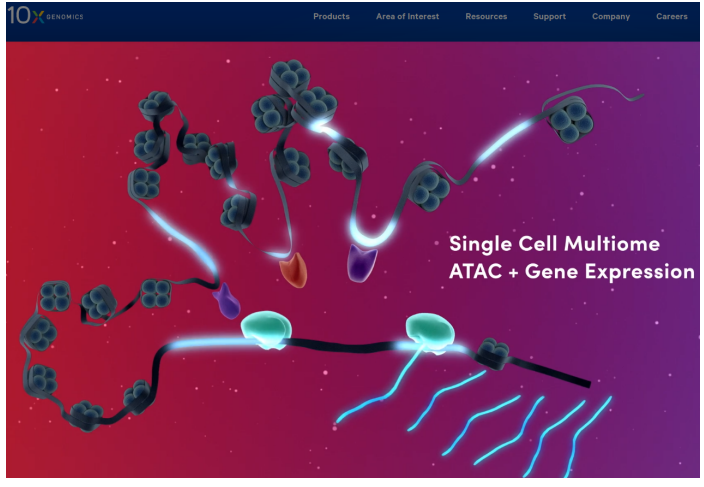
Kelsey et al., 2017, Science 358, 69-75



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?

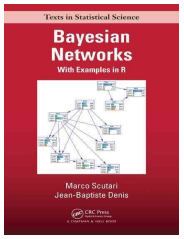
What I mean by Omics Integration

OnPLS

JIVE

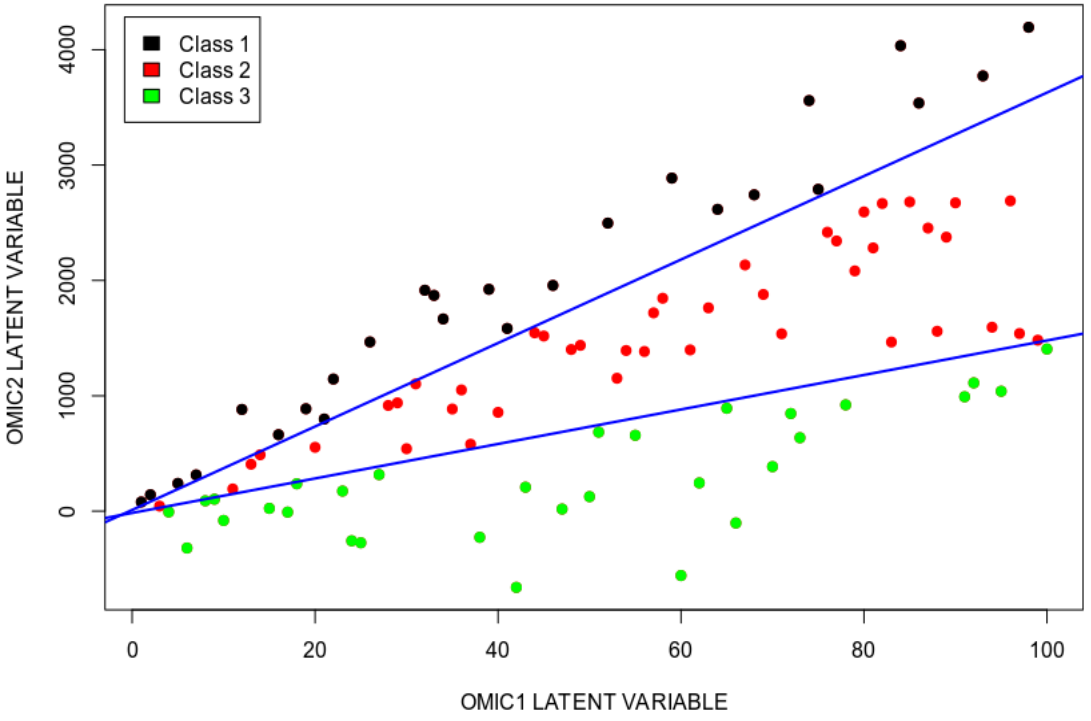
DISCO

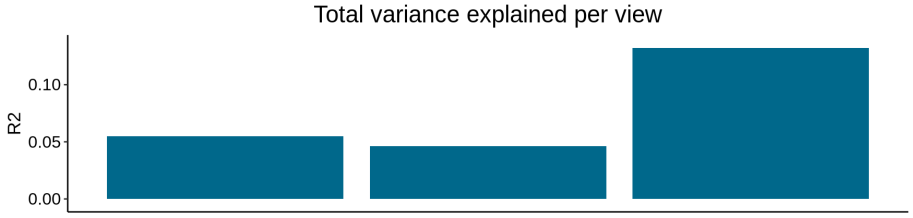
mixOmics

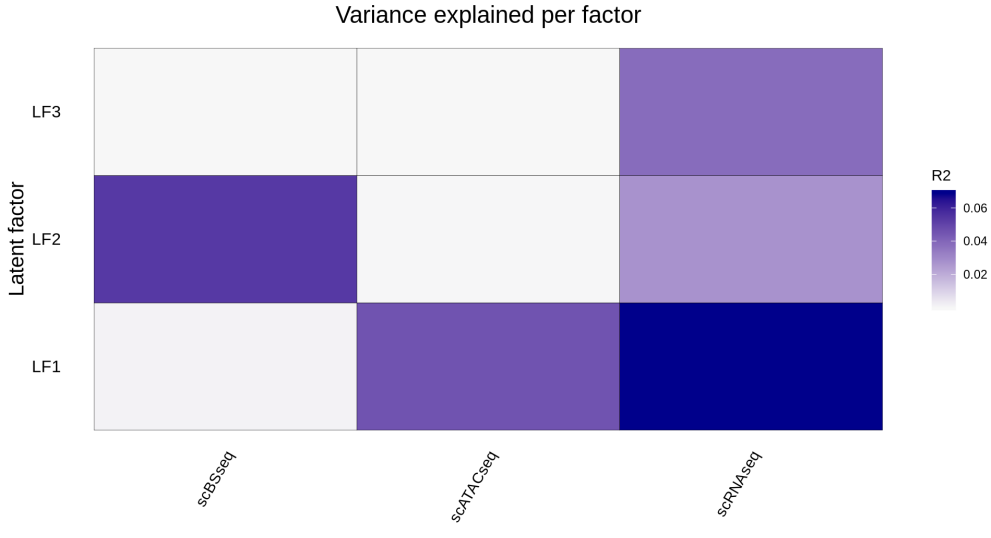


Clustering of Clusters

Idea Behind Omics Integration:
See Patterns Hidden in Individual Omics



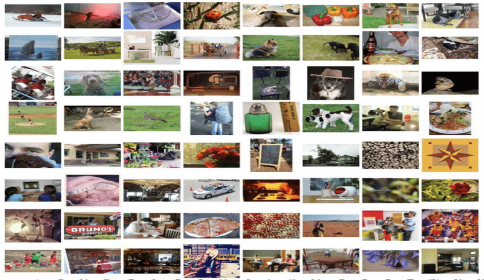




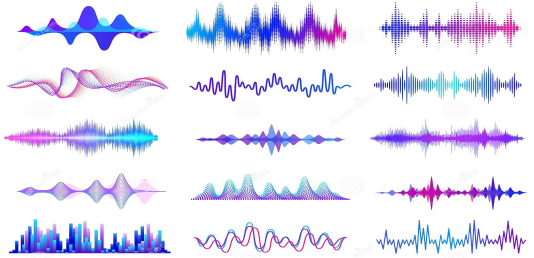
How I Evaluate Omics Integration

TEXT (78%)

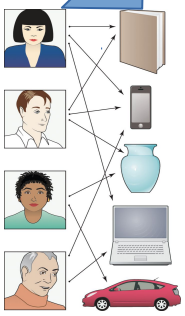
IMAGE (83%)



SOUND (75%)



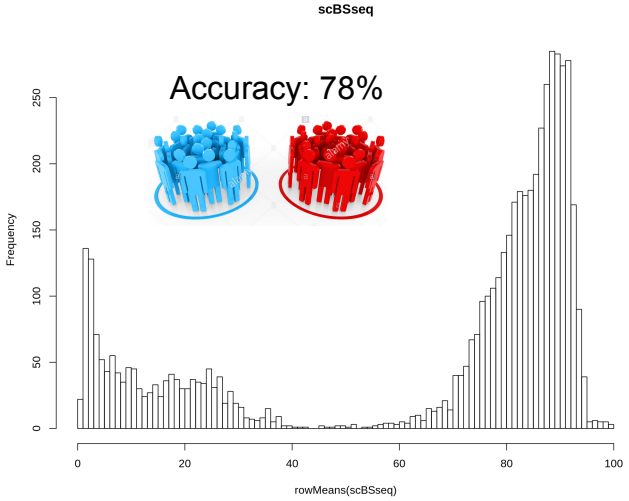
Predict Facebook user interests



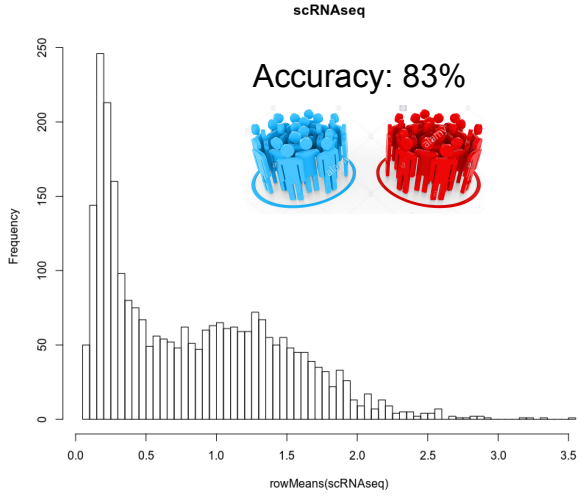
Data Integration Accuracy: 96%

How I Evaluate Omics Integration

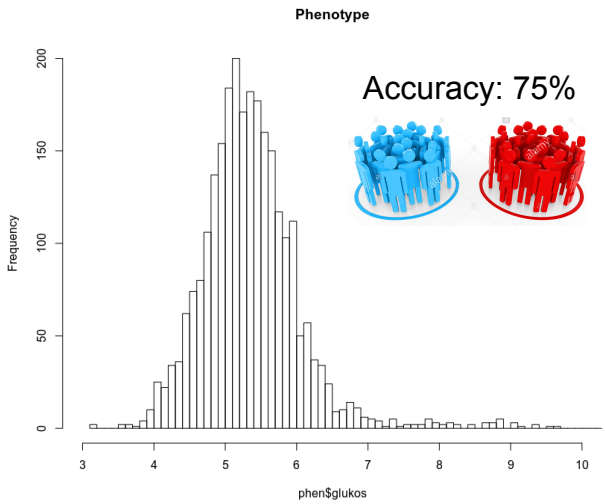
Methylation



Gene Expression



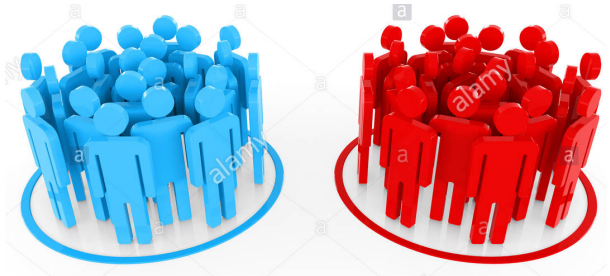
Clinical variable



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

2) Explicitly model distributions:
MOFA, Bayesian Networks

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

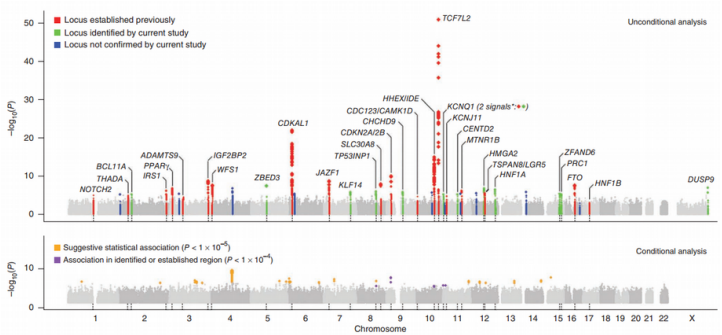


HEALTHY SICK

Data Integration
Accuracy: 96%

Prediction as a Criterion of Success

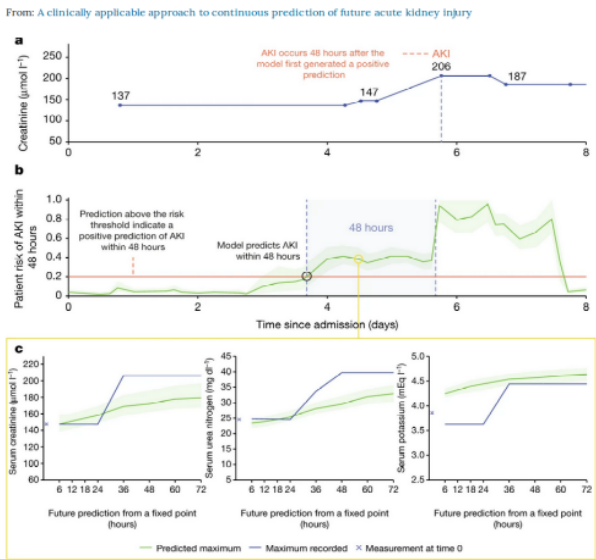
Statistics searches for candidates



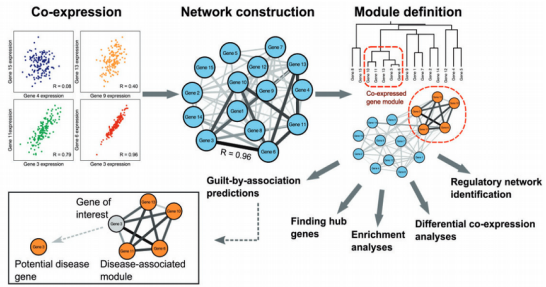
Consequence

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

Machine Learning optimizes prediction



Consequence



Take Home Messages

1) Biological data are high-dimensional and notoriously difficult to analyze

2) Integration across Omics is often sensitive to the Curse of Dimensionality

3) Integrating across Omics we expect to discover novel patterns in the data

4) Increased prediction accuracy is an indication of successful data integration

5) Single cell Omics are promising for integration in terms of statistical power

National Bioinformatics Infrastructure Sweden (NBIS)

Knut och Alice Wallenbergs Stiftelse

Vetenskapsrådet

Lunds Universitet