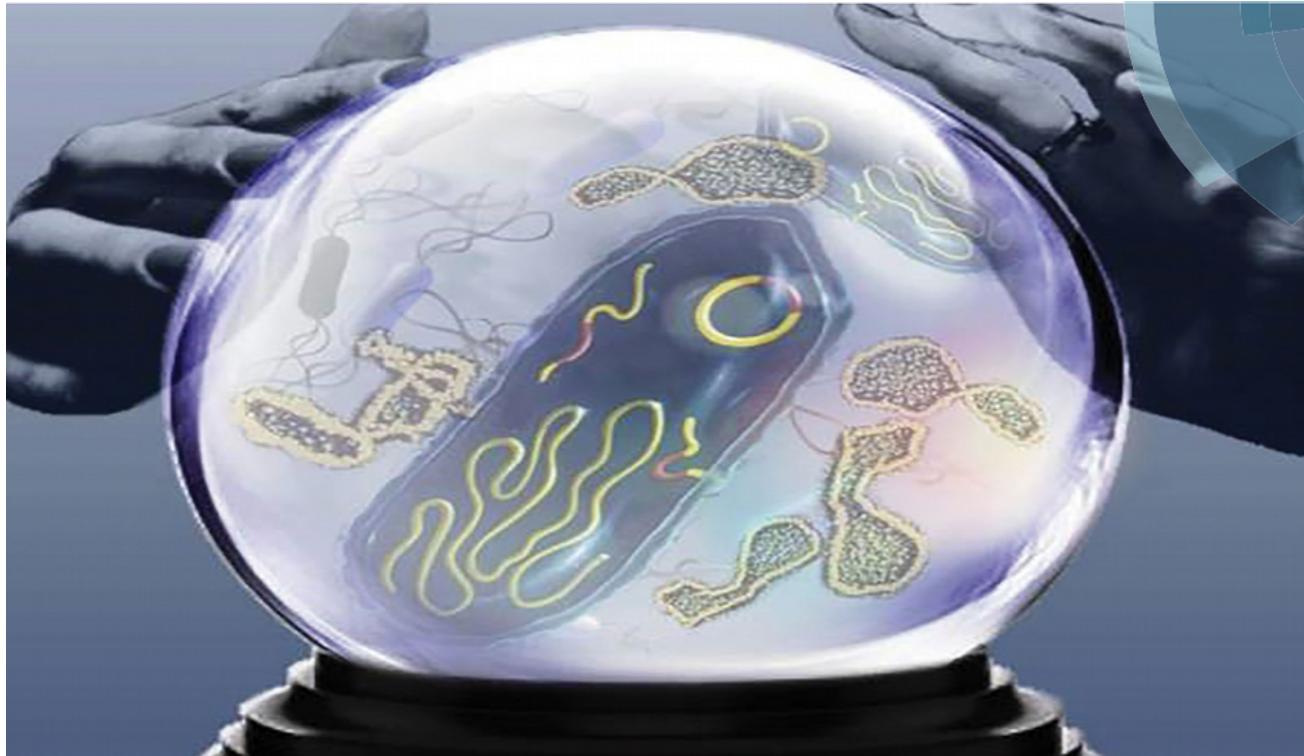


# Multi-Omics Data Integration via Machine Learning

Omics Integration and Systems Biology course

Nikolay Oskolkov, Lund University, NBIS SciLifeLab, Sweden



@NikolayOskolkov



**GitHub**

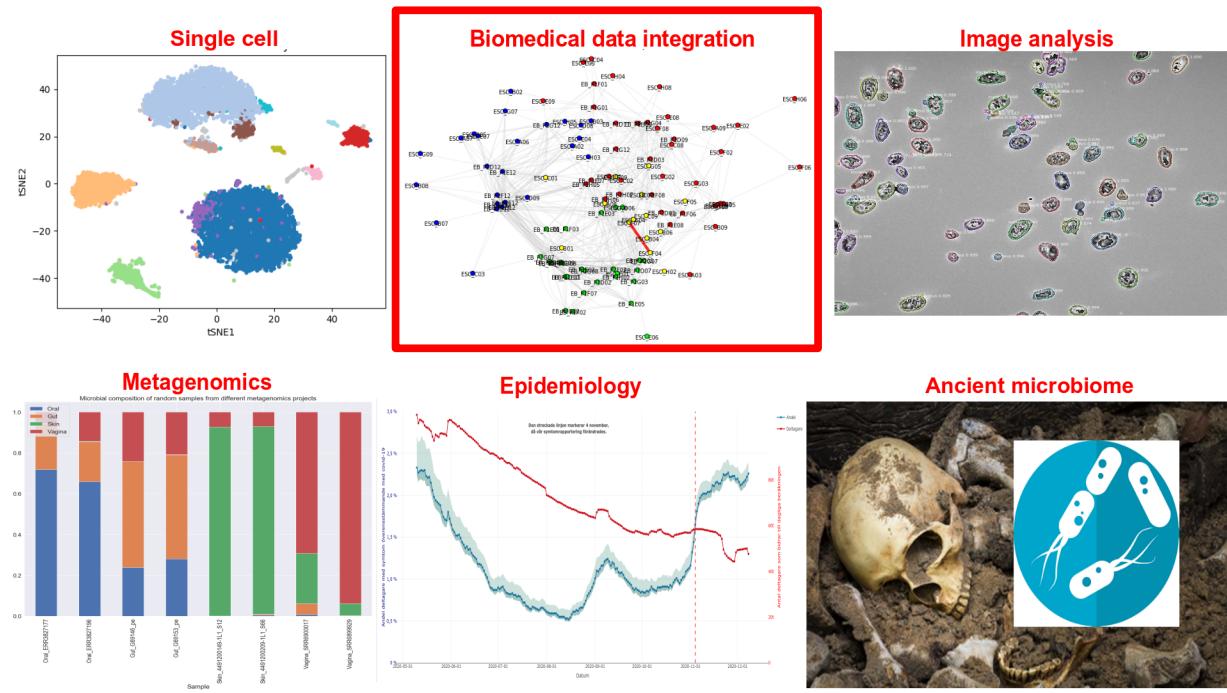
<https://github.com/NikolayOskolkov>

Image adapted from Molecular Omics, Issue 1, 2018

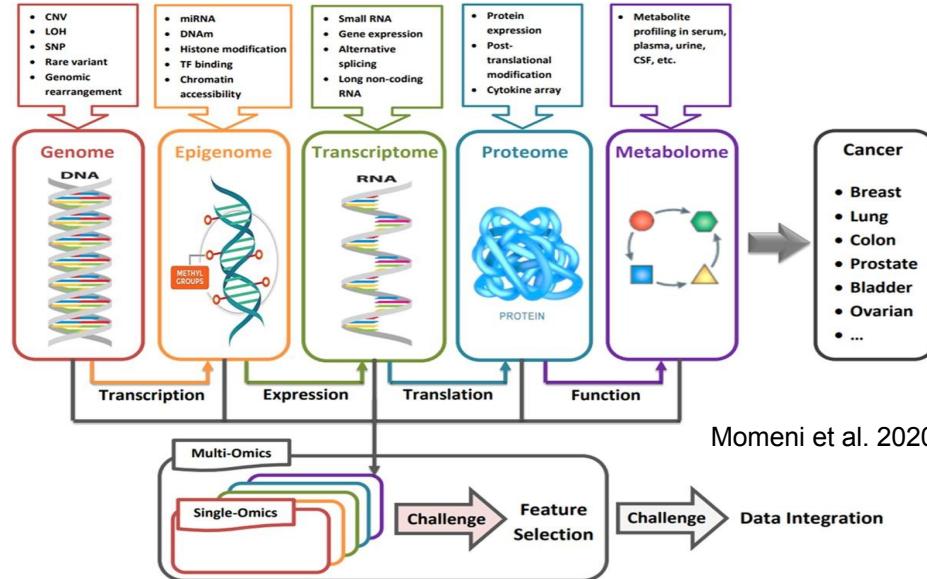
2007 PhD in theoretical physics

2011 medical genetics at Lund University

2016 working at NBIS SciLifeLab, Sweden



# Multi-Omics Begins: 2015 – until now



**scientific reports**

Explore content ▾ About the journal ▾ Publish with us ▾

nature > scientific reports > articles > article

Article | [Open access](#) | Published: 25 June 2024

## Predicting type 2 diabetes via machine learning integration of multiple omics from human pancreatic islets

Tina Rönn, Alexander Perfiliev, Nikolay Oskolkov & Charlotte Ling

*Scientific Reports* 14, Article number: 14637 (2024) | [Cite this article](#)

Rönn et al.,  
Scientific Reports 2024

The screenshot shows the ELIXIR Omics Integration and Systems Biology workshop syllabus page. The sidebar includes links for Student, Home, Login, Schedule, Modules, Pages, Ladek for students, Syllabus with course literature, and Schedule. The main content area displays a network graph of biological interactions and text about the workshop's aims and covered topics.

**ELIXIR Omics Integration and Systems Biology**

The aim of this workshop is to provide an integrated view of data-driven hypothesis generation through biological network analysis, constraint-based modelling, and supervised and unsupervised integration methods. A general description of different methods for analysing different omics data (e.g. transcriptomics and genomics) will be presented with some of the lectures discussing key methods and pitfalls in their integration. The techniques will be discussed in terms of their rationale and applicability.

**Covered topics**

- Data pre-processing and cleaning prior to integration;
- Application of key machine learning methods for multi-omics analysis including deep learning;
- Multi-omics integration, clustering and dimensionality reduction;
- Biological network inference, community and topology analysis and visualization;
- Condition-specific and personalized modeling through Genome-scale Metabolic models for integration of transcriptomic, proteomic, metabolomic and fluxomic data;
- Identification of key biological functions and pathways;
- Identification of potential biomarkers and targetable genes through modeling and biological network analysis;
- Application of network approaches in meta-analyses;
- Similarity network fusion and matrix factorization techniques;
- Integrated data visualization techniques

[https://github.com/NBISweden/workshop\\_omics\\_integration](https://github.com/NBISweden/workshop_omics_integration)

**BOUGHT BY:**



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

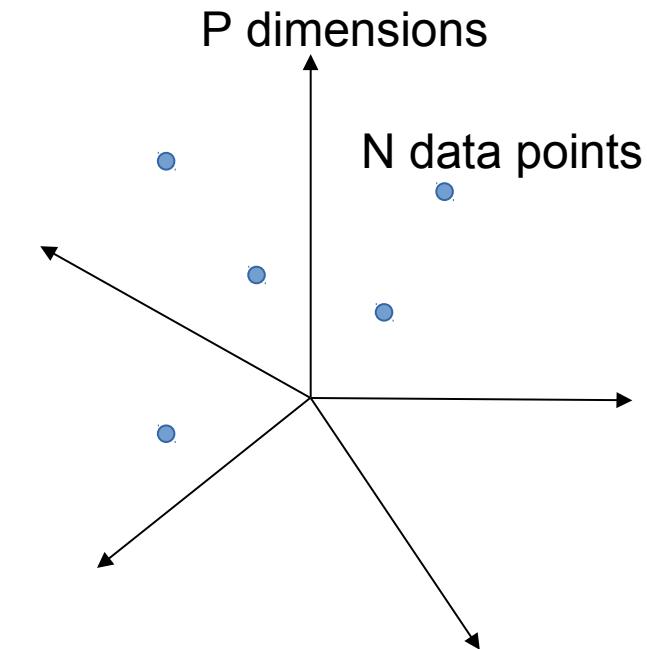
Category	Subcategory	Subsubcategory	Parameter	Value
1	1	1	1	1
1	1	1	2	2
1	1	2	1	3
1	1	2	2	4
1	1	2	3	5
1	1	2	4	6
1	1	2	5	7
1	1	2	6	8
1	1	2	7	9
1	1	2	8	10
1	1	2	9	11
1	1	2	10	12
1	1	2	11	13
1	1	2	12	14
1	1	2	13	15
1	1	2	14	16
1	1	2	15	17
1	1	2	16	18
1	1	2	17	19
1	1	2	18	20
1	1	2	19	21
1	1	2	20	22
1	1	2	21	23
1	1	2	22	24
1	1	2	23	25
1	1	2	24	26
1	1	2	25	27
1	1	2	26	28
1	1	2	27	29
1	1	2	28	30
1	1	2	29	31
1	1	2	30	32
1	1	2	31	33
1	1	2	32	34
1	1	2	33	35
1	1	2	34	36
1	1	2	35	37
1	1	2	36	38
1	1	2	37	39
1	1	2	38	40
1	1	2	39	41
1	1	2	40	42
1	1	2	41	43
1	1	2	42	44
1	1	2	43	45
1	1	2	44	46
1	1	2	45	47
1	1	2	46	48
1	1	2	47	49
1	1	2	48	50
1	1	2	49	51
1	1	2	50	52
1	1	2	51	53
1	1	2	52	54
1	1	2	53	55
1	1	2	54	56
1	1	2	55	57
1	1	2	56	58
1	1	2	57	59
1	1	2	58	60
1	1	2	59	61
1	1	2	60	62
1	1	2	61	63
1	1	2	62	64
1	1	2	63	65
1	1	2	64	66
1	1	2	65	67
1	1	2	66	68
1	1	2	67	69
1	1	2	68	70
1	1	2	69	71
1	1	2	70	72
1	1	2	71	73
1	1	2	72	74
1	1	2	73	75
1	1	2	74	76
1	1	2	75	77
1	1	2	76	78
1	1	2	77	79
1	1	2	78	80
1	1	2	79	81
1	1	2	80	82
1	1	2	81	83
1	1	2	82	84
1	1	2	83	85
1	1	2	84	86
1	1	2	85	87
1	1	2	86	88
1	1	2	87	89
1	1	2	88	90
1	1	2	89	91
1	1	2	90	92
1	1	2	91	93
1	1	2	92	94
1	1	2	93	95
1	1	2	94	96
1	1	2	95	97
1	1	2	96	98
1	1	2	97	99
1	1	2	98	100
1	1	2	99	101
1	1	2	100	102
1	1	2	101	103
1	1	2	102	104
1	1	2	103	105
1	1	2	104	106
1	1	2	105	107
1	1	2	106	108
1	1	2	107	109
1	1	2	108	110
1	1	2	109	111
1	1	2	110	112
1	1	2	111	113
1	1	2	112	114
1	1	2	113	115
1	1	2	114	116
1	1	2	115	117
1	1	2	116	118
1	1	2	117	119
1	1	2	118	120
1	1	2	119	121
1	1	2	120	122
1	1	2	121	123
1	1	2	122	124
1	1	2	123	125
1	1	2	124	126
1	1	2	125	127
1	1	2	126	128
1	1	2	127	129
1	1	2	128	130
1	1	2	129	131
1	1	2	130	132
1	1	2	131	133
1	1	2	132	134
1	1	2	133	135
1	1	2	134	136
1	1	2	135	137
1	1	2	136	138
1	1	2	137	139
1	1	2	138	140
1	1	2	139	141
1	1	2	140	142
1	1	2	141	143
1	1	2	142	144
1	1	2	143	145
1	1	2	144	146
1	1	2	145	147
1	1	2	146	148
1	1	2	147	149
1	1	2	148	150
1	1	2	149	151
1	1	2	150	152
1	1	2	151	153
1	1	2	152	154
1	1	2	153	155
1	1	2	154	156
1	1	2	155	157
1	1	2	156	158
1	1	2	157	159
1	1	2	158	160
1	1	2	159	161
1	1	2	160	162
1	1	2	161	163
1	1	2	162	164
1	1	2	163	165
1	1	2	164	166
1	1	2	165	167
1	1	2	166	168
1	1	2	167	169
1	1	2	168	170
1	1	2	169	171
1	1	2	170	172
1	1	2	171	173
1	1	2	172	174
1	1	2	173	175
1	1	2	174	176
1	1	2	175	177
1	1	2	176	178
1	1	2	177	179
1	1	2	178	180
1	1	2	179	181
1	1	2	180	182
1	1	2	181	183
1	1	2	182	184
1	1	2	183	185
1	1	2	184	186
1	1	2	185	187
1	1	2	186	188
1	1	2	187	189
1	1	2	188	190
1	1	2	189	191
1	1	2	190	192
1	1	2	191	193
1	1	2	192	194
1	1	2	193	195
1	1	2	194	196
1	1	2	195	197
1	1	2	196	198
1	1	2	197	199
1	1	2	198	200
1	1	2	199	201
1	1	2	200	202
1	1	2	201	203
1	1	2	202	204
1	1	2	203	205
1	1	2	204	206
1	1	2	205	207
1	1	2	206	208
1	1	2	207	209
1	1	2	208	210
1	1	2	209	211
1	1	2	210	212
1	1	2	211	213
1	1	2	212	214
1	1	2	213	215
1	1	2	214	216
1	1	2	215	217
1	1	2	216	218
1	1	2	217	219
1	1	2	218	220
1	1	2	219	221
1	1	2	220	222
1	1	2	221	223
1	1	2	222	224
1	1	2	223	225
1	1	2	224	226
1	1	2	225	227
1	1	2	226	228
1	1	2	227	229
1	1	2	228	230
1	1	2	229	231
1	1	2	230	232
1	1	2	231	233
1	1	2	232	234
1	1	2	233	235
1	1	2	234	236
1	1	2	235	237
1	1	2	236	238
1	1	2	237	239
1	1	2	238	240
1	1	2	239	241
1	1	2	240	242
1	1	2	241	243
1	1	2	242	244
1	1	2	243	245
1	1	2	244	246
1	1	2	245	247
1	1	2	246	248
1	1	2	247	249
1	1	2	248	250
1	1	2	249	251
1	1	2	250	252
1	1	2	251	253
1	1	2	252	254
1	1	2	253	255
1	1	2	254	256
1	1	2	255	257
1	1	2	256	258
1	1	2	257	259
1	1	2	258	260
1	1	2	259	261
1	1	2	260	262
1	1	2	261	263
1	1	2	262	264
1	1	2	263	265
1	1	2	264	266
1	1	2	265	267
1	1	2	266	268
1	1	2	267	269
1	1	2	268	270
1	1	2	269	271
1	1	2	270	272
1	1	2	271	273
1	1	2	272	274
1	1	2	273	275
1	1	2	274	276
1	1	2	275	277
1	1	2	276	278
1	1	2	277	279
1	1	2	278	280
1	1	2	279	281
1	1	2	280	282
1	1	2	281	283
1	1	2	282	284
1	1	2	283	285
1	1	2	284	286
1	1	2	285	287
1	1	2	286	288
1	1	2	287	290
1	1	2	288	291
1	1	2	289	292
1	1	2	290	293
1	1	2	291	294
1	1	2	292	295
1	1	2	293	296
1	1	2	294	297
1	1	2	295	298
1	1	2	296	299
1	1	2	297	300
1	1	2	298	301
1	1	2	299	302
1	1	2	300	303
1	1	2	301	304
1	1	2	302	305
1	1	2	303	306
1	1	2	304	307
1	1	2	305	308
1	1	2	306	309
1	1	2	307	310
1	1	2	308	311
1	1	2	309	312
1	1	2	310	313
1	1	2	311	314
1	1	2	312	315
1	1	2	313	316
1	1	2	314	317
1	1	2	315	318
1	1	2	316	319
1	1	2	317	320
1	1	2	318	321

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

Features: genes, proteins,  
microbes, metabolites etc.

**N**

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



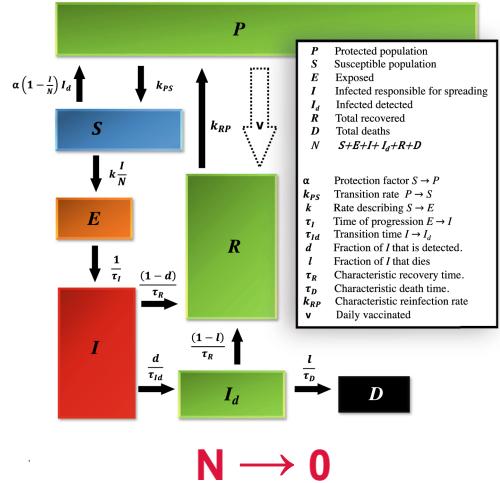
**High Dimensional Data:**  
**P >> N**

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

**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



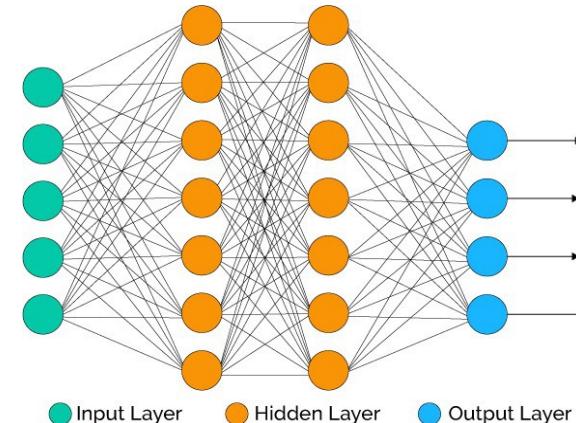
### Bayesianism



### Frequentism



### Machine Learning



### Hypothesis-driven

### The Curse of Dimensionality

### Amount of Data

Ex.1

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

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

$$(X^T X)^{-1} \sim \frac{1}{\det(X^T X)} \dots \rightarrow \infty, \quad n \ll p$$

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

Biased ML variance estimator in HD-space

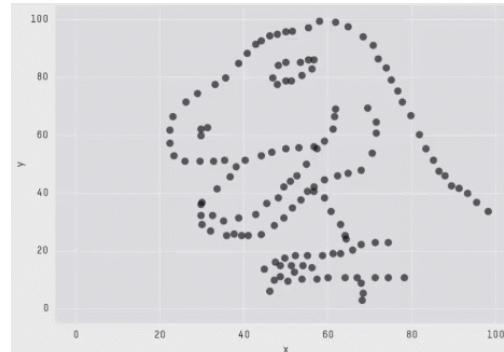
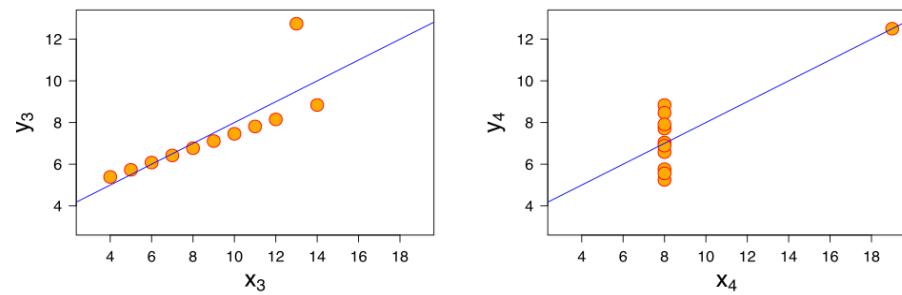
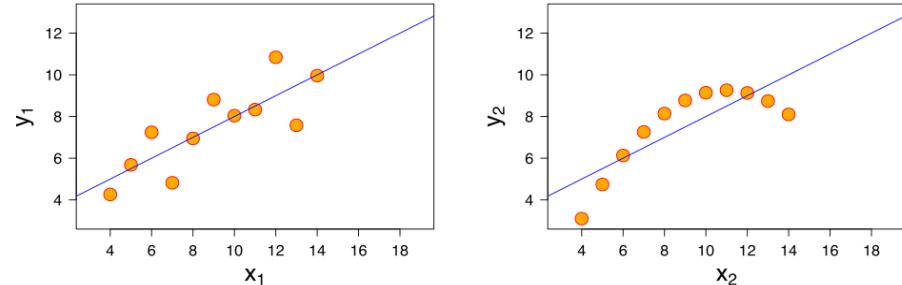
- Maximum likelihood based
- Focus on summary statistics
- Focus too much on p-values

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

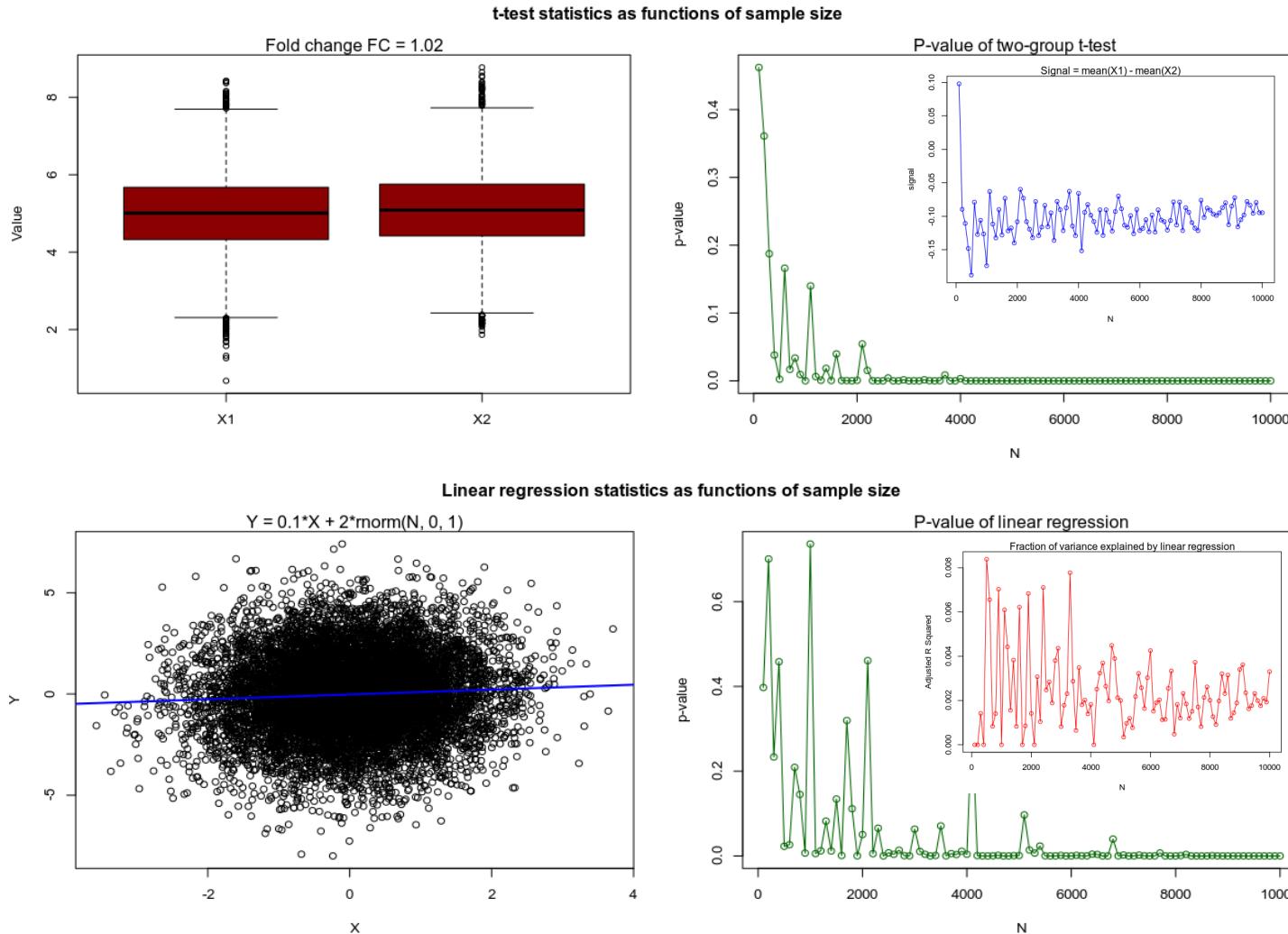
$$\frac{\partial L(x_i | \mu, \sigma^2)}{\partial \mu} = 0; \quad \frac{\partial L(x_i | \mu, \sigma^2)}{\partial \sigma^2} = 0$$

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

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



X Mean:	54.2659224
Y Mean:	47.8313999
X SD :	16.7649829
Y SD :	26.9342120
Corr. :	-0.0642526

**nature**

Explore content ▾ About the journal ▾ Publish with us ▾ Subscribe

View all journals Search Q Login ⓘ

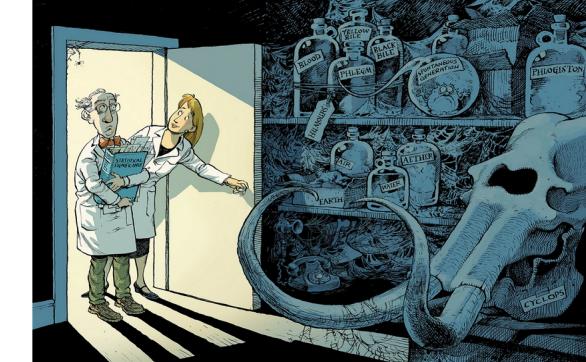
nature > comment > article

COMMENT | 20 March 2019

## Scientists rise up against statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

Valentin Amrhein ⓘ Sander Greenland & Blake McShane



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

```

1 n <- 20 # number of samples
2 p <- 2 # number of features / dimensions
3 Y <- rnorm(n)
4 X <- matrix(rnorm(n * p), n, p)
5 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 →

```

1 n <- 20 # number of samples
2 p <- 10 # number of features / dimensions
3 Y <- rnorm(n)
4 X <- matrix(rnorm(n * p), n, p)
5 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 →

```

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

```

Call:  
lm(formula = Y ~ X)

Residuals:

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

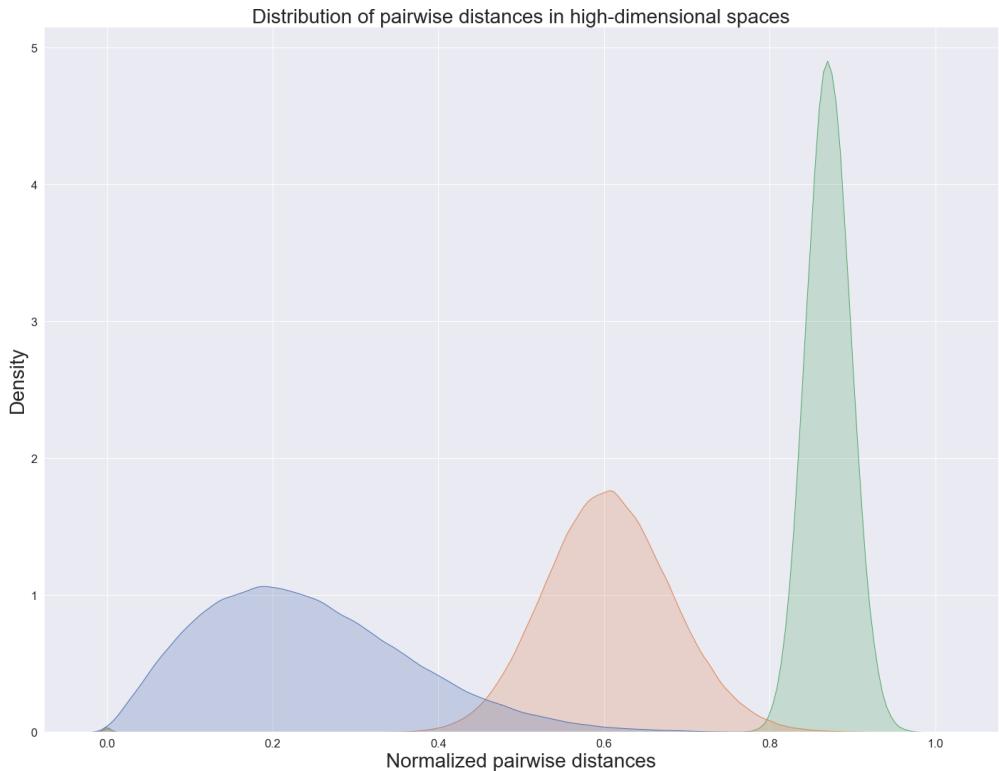
Coefficients: (1 not defined because of singularities)

	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	-1.7103	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

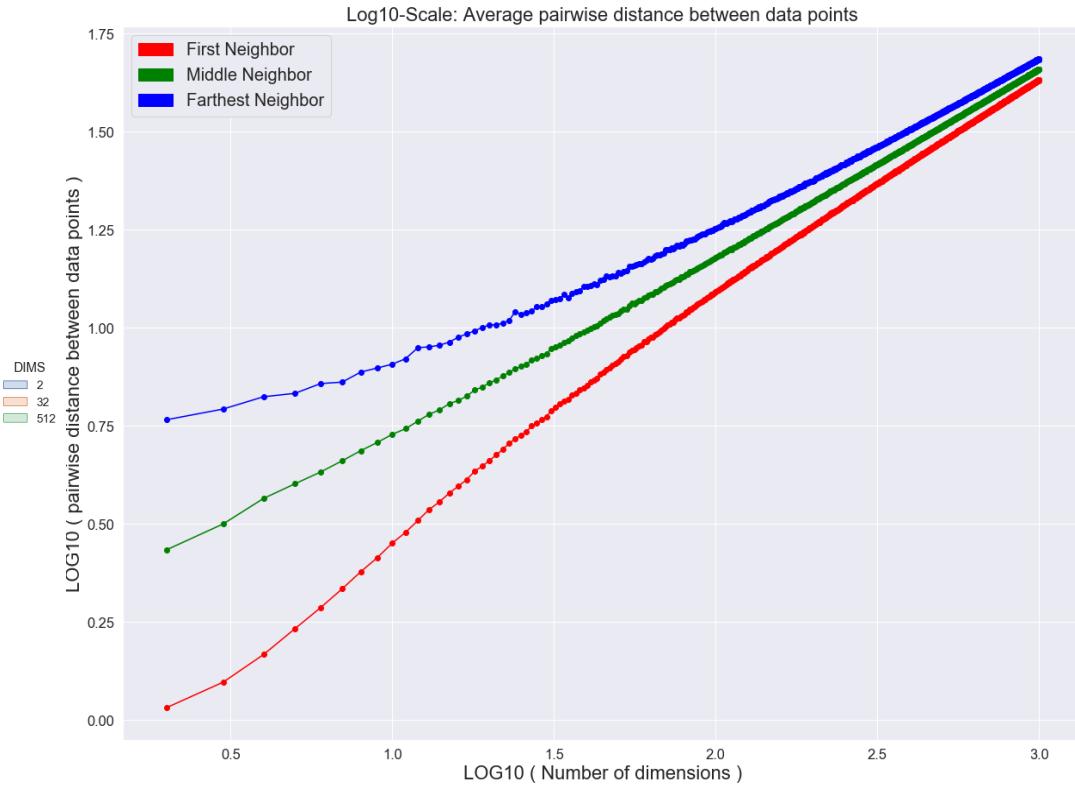
This is another way we face the Curse of Dimensionality in computational biology

# More on the Curse of Dimensionality

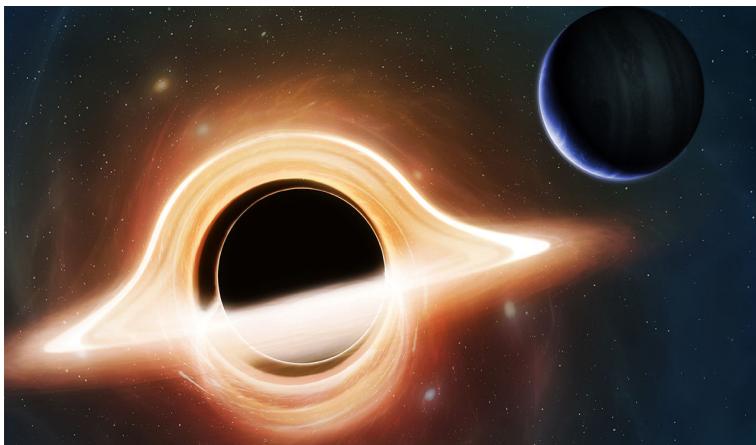
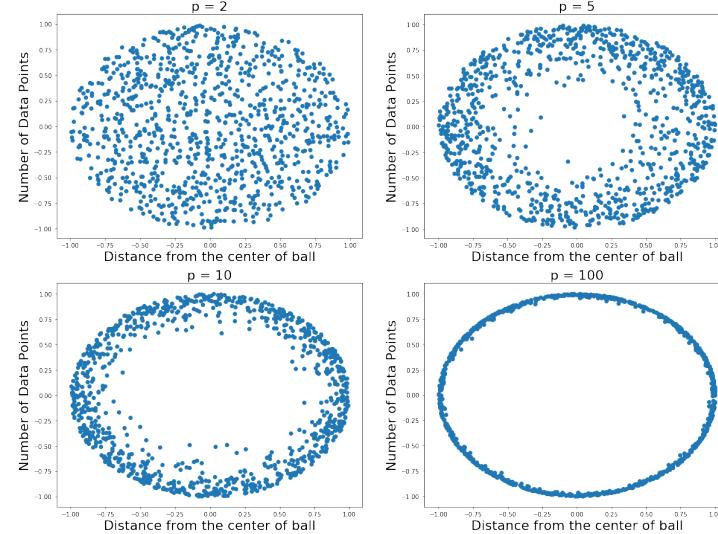
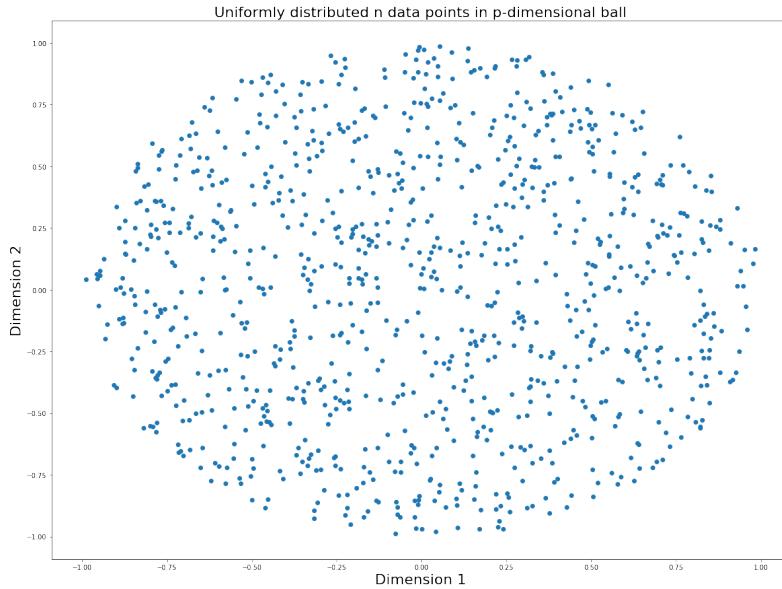


Data points become far from each other and equidistant in high dimensions

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



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



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

# The curse(s) of dimensionality

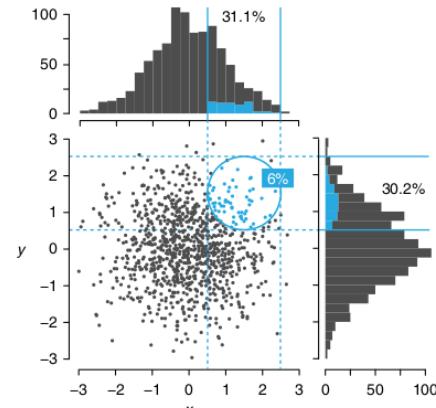
There is such a thing as too much of a good thing.

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'<sup>1</sup> (CoD) and manifests itself in a variety of ways. This month, we discuss four important problems of dimensionality as it applies to data sparsity<sup>1,2</sup>, multicollinearity<sup>3</sup>, multiple testing<sup>4</sup> and overfitting<sup>5</sup>. 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  $n$

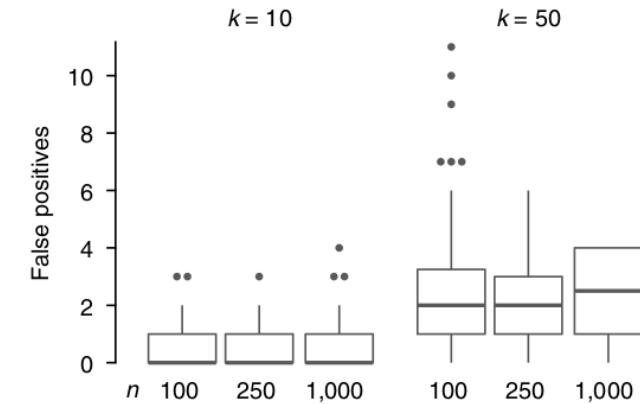


**Fig. 1 | Data tend to be sparse in higher dimensions.** Among 1,000 ( $x, y$ ) points in which both  $x$  and  $y$  are normally distributed with a mean of 0 and s.d.  $= 1$ , only 6% fall within  $\sigma$  of  $(x, y) = (1.5, 1.5)$  (blue circle). However, when the data are projected into a lower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

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  $n$  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

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

$N \approx P$

# Metabolomics

## N ≈ P

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \quad \text{Proteomics} \quad N \approx P$$

# Proteomics N ≈ P

$$\begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \text{Metagenomics} \\ N \approx P$$

# Metagenomics N ≈ P

— manageable

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

# Transcriptomics

## N << P

(Single cell: N <= P)

# challenging

$$N \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 \end{pmatrix} \textcolor{red}{P} \quad \textcolor{red}{Genomics} \\ \textcolor{red}{N} \llll P$$

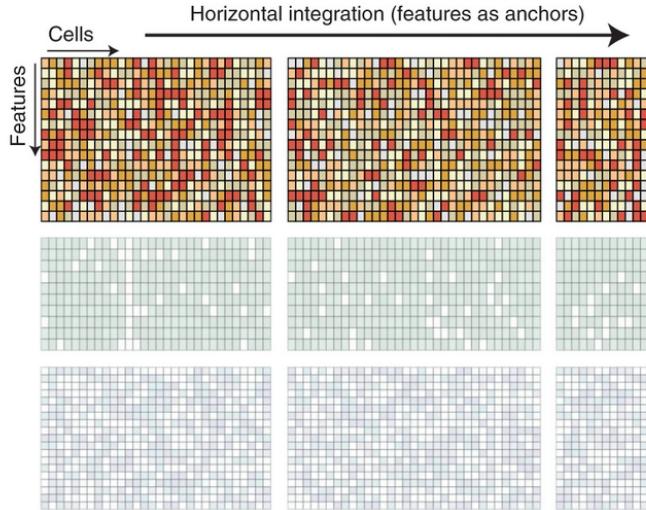
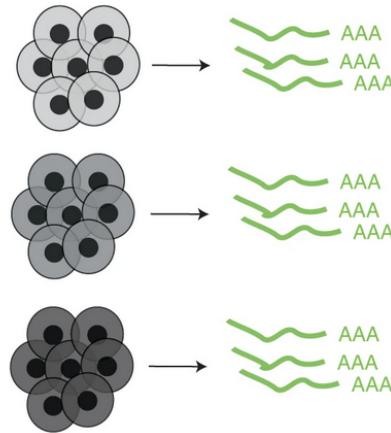
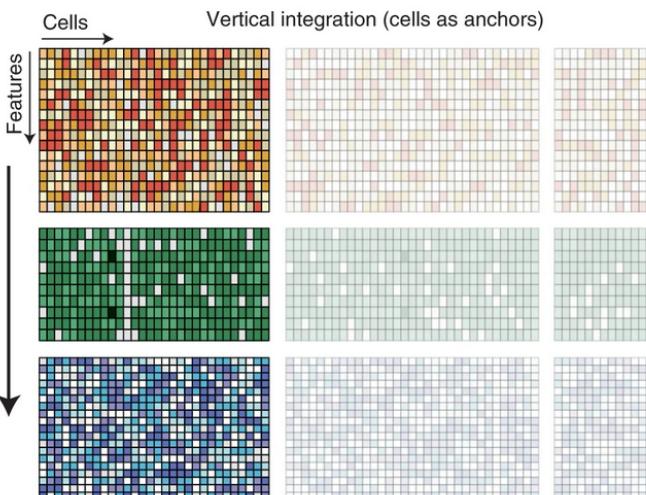
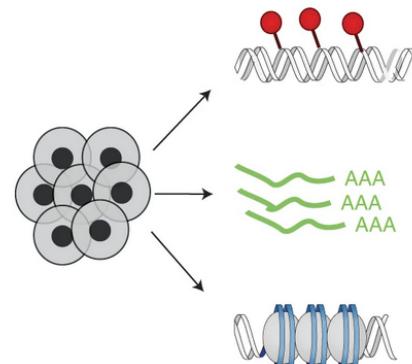
# Genomics

N <<< P

$$N \left( \begin{array}{cccccccccccccccccccccccc} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \end{array} \right) \text{Methylomics}$$

# Methylomics N <<< P

# Multi-Omics Data Integration

**a****b**

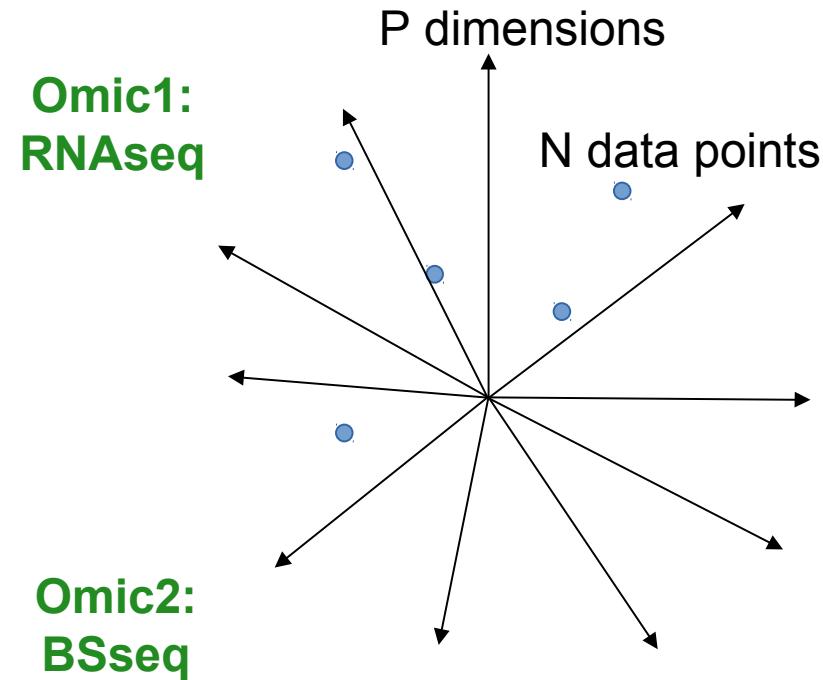
Statistical observations:  
e.g. samples, cells etc.

Features: genes, proteins,  
microbes, metabolites etc.

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

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

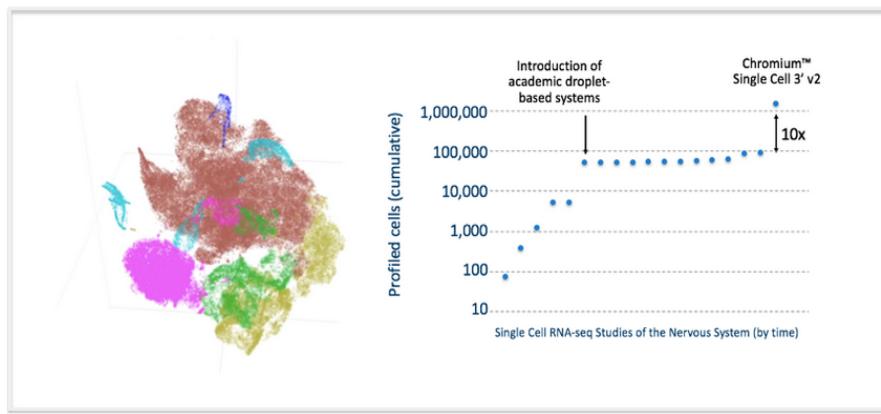
$P_1 + P_2 \gg N$  integration across features leads to even more high-dimensional data



# Big Data in Single Cell Genomics

CAREERS BLOG 10X UNIVERSITY

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

[« Back to Blog](#)
[< Newer Article](#) [Older Article >](#)


Our 1.3 million single cell dataset is ready 0 KUDOS



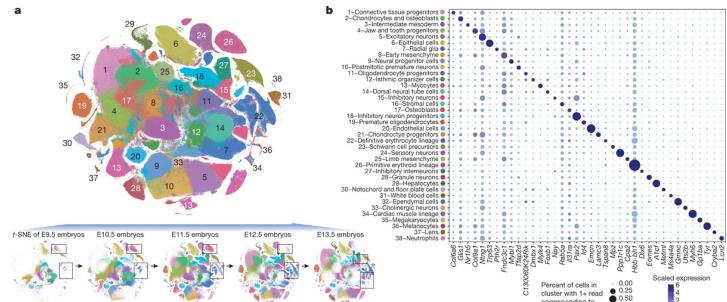
POSTED BY: grace-10x, on Feb 21, 2017 at 2:28 PM

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 which the gene is expressed. The color scale indicates the scaled expression level of the marker gene. The legend lists 38 cell types, numbered 1 to 38.

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 Search

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

RECENT POSTS

by @biometers • August 30, 2019

f | t | in

Human Cell Atlas, single-cell data

We are glad to announce that we will upscale 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.

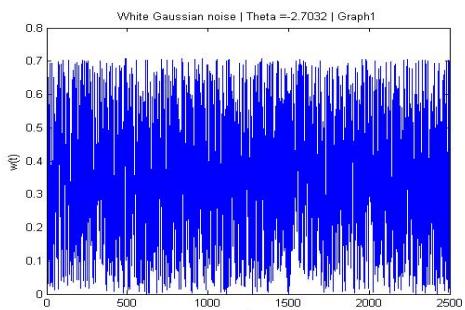
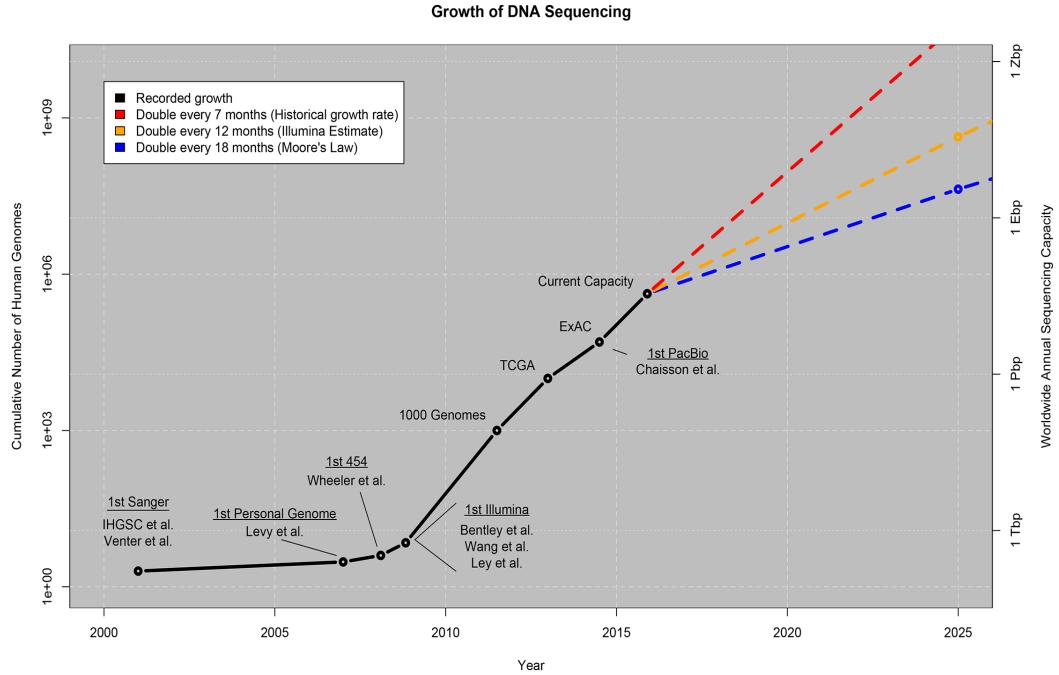
A new tool to interactively visualize single-cell objects (Seurat, Scanpy, SingleCellExperiments,...)

September 26, 2019

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

August 30, 2019

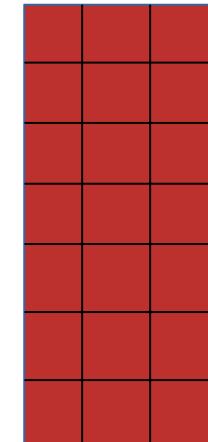
# Big in Size or Sample Size?



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

Genomics / WGS: Little Data

$$N_1 \sim 10^3$$

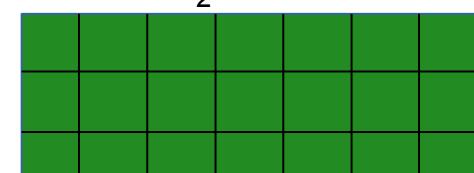


$$P_1 \sim 10^6$$

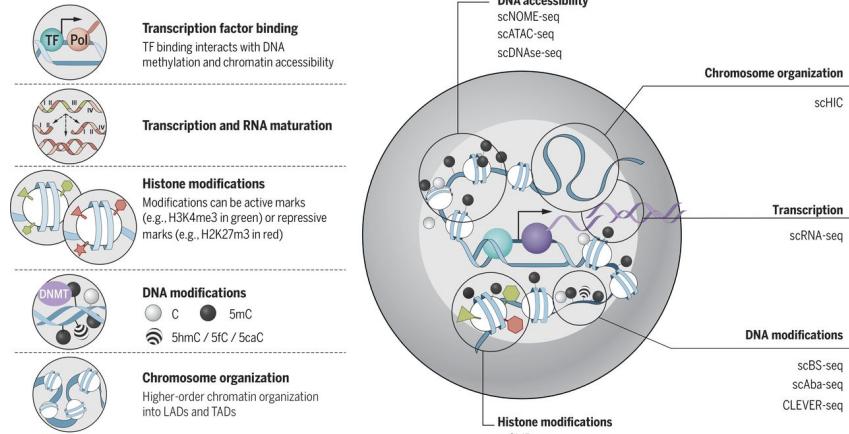
$$N_1 * P_1 = N_2 * P_2 = 10^9$$

scRNAseq: Big Data

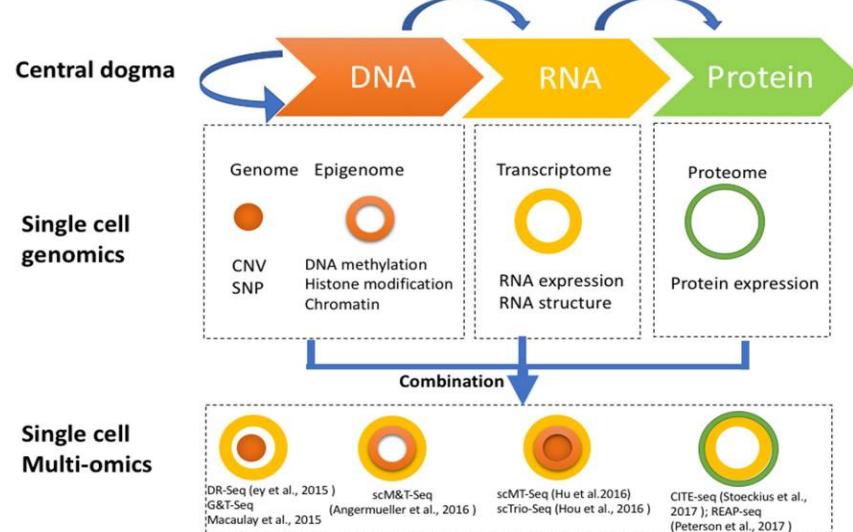
$$N_2 \sim 10^6$$



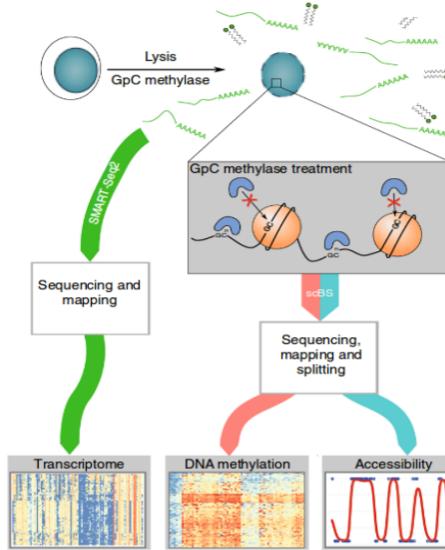
$$P_2 \sim 10^3$$



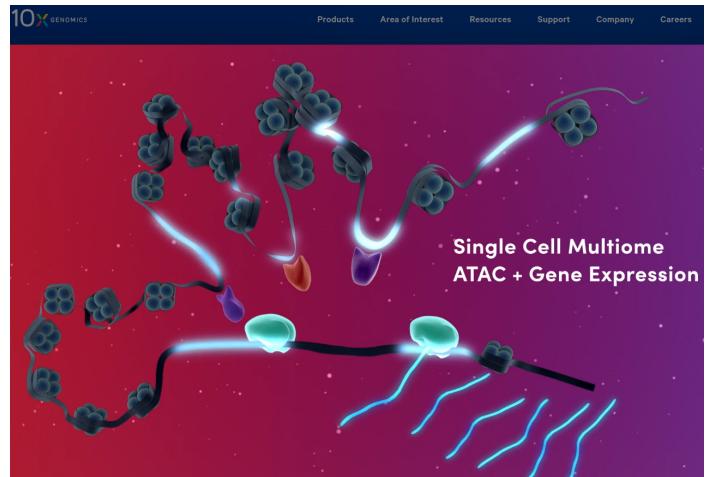
Kelsey et al., 2017, Science 358, 69-75



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



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

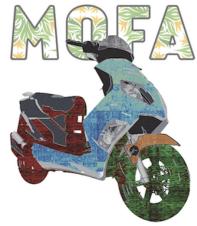
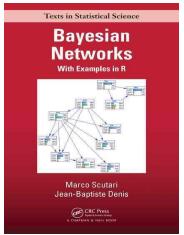


# How to define and evaluate multi-Omics data integration?

OnPLS

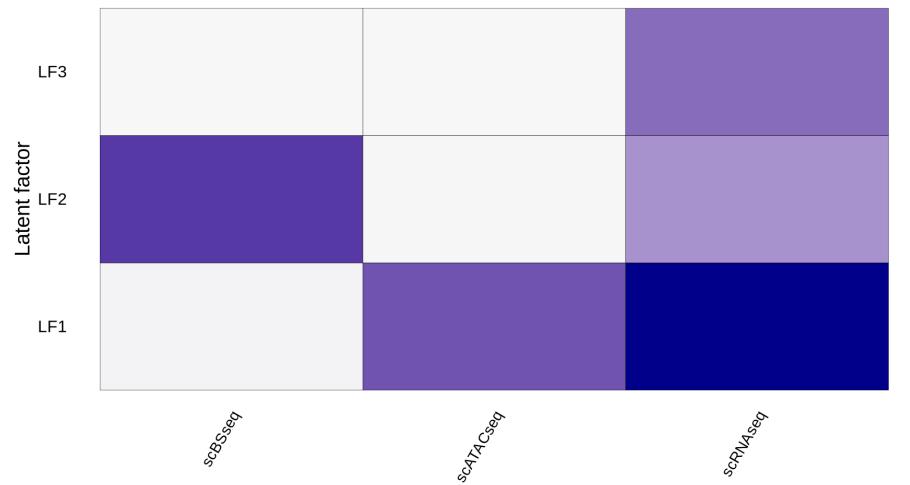
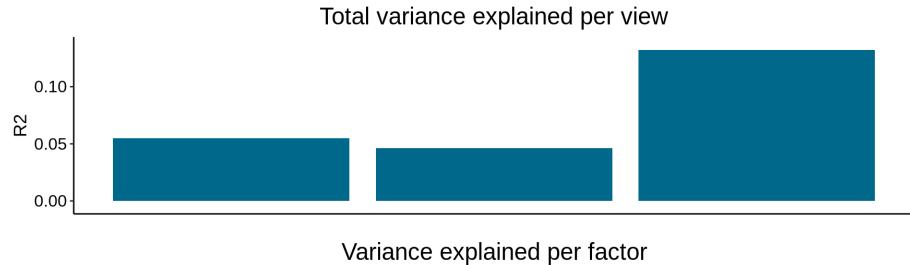
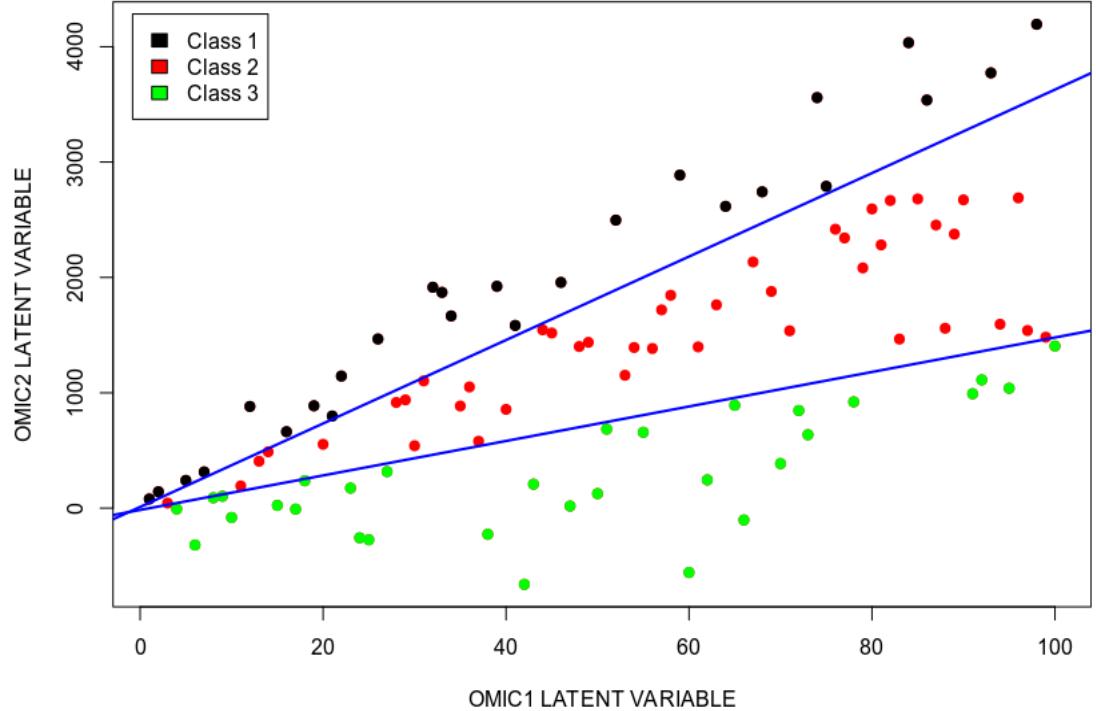
JIVE

DISCO



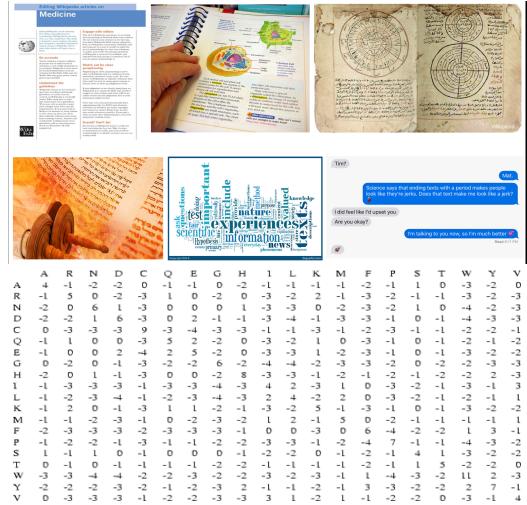
## Clustering of Clusters

Idea Behind Omics Integration:  
See Patterns Hidden in Individual Omics



# How I Evaluate Omics Integration

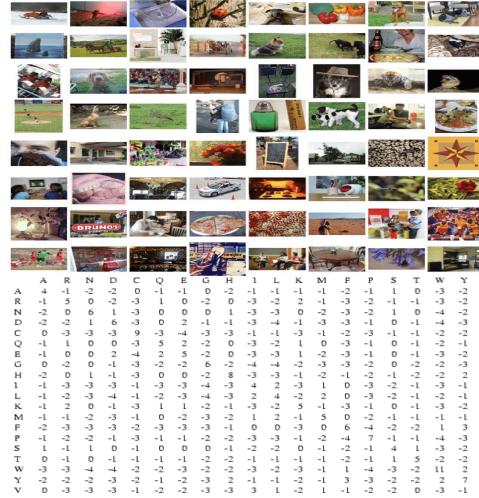
**TEXT (78%)**



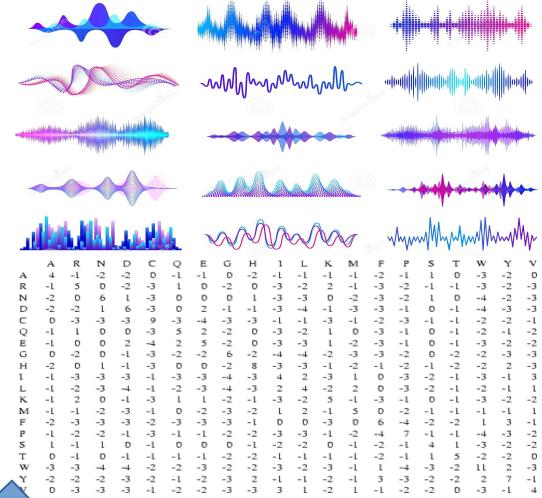
**Predict Facebook user interests**



**IMAGE (83%)**



**SOUND (75%)**



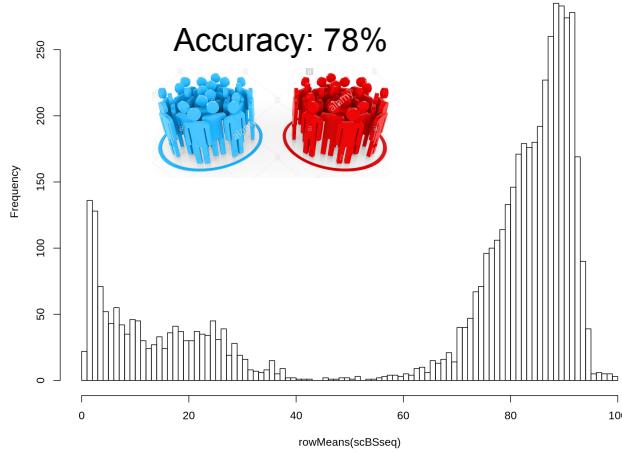
**Data Integration Accuracy: 96%**

# How I Evaluate Omics Integration

## Methylation

scBSseq

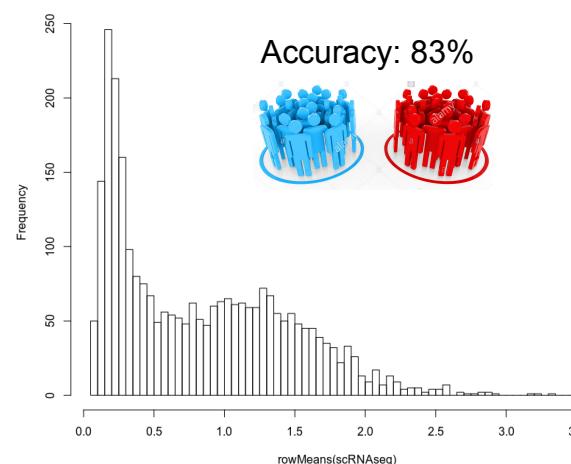
Accuracy: 78%

## Gene Expression

scRNAseq

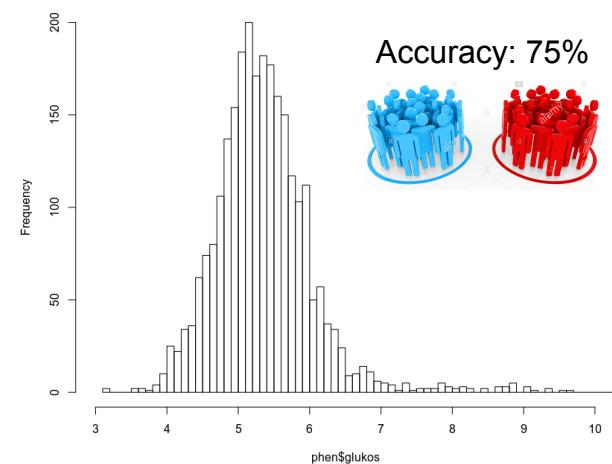
Accuracy: 83%

## Clinical variable

Phenotype

Accuracy: 75%

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

**2) Explicitly model distributions:**  
MOFA, Bayesian Networks

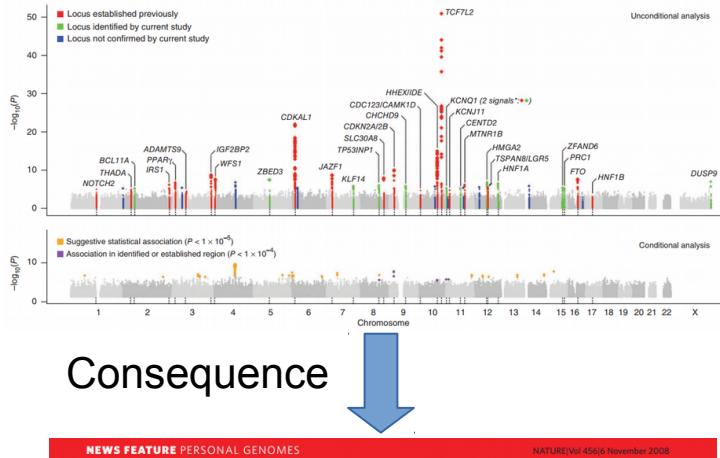
**3) Extract common variation:**  
PLS, CCA, Factor Analysis



**Data Integration Accuracy: 96%**

# Prediction as a Criterion of Success

## Statistics searches for candidates



**The case of the missing heritability**

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

## Machine Learning optimizes prediction



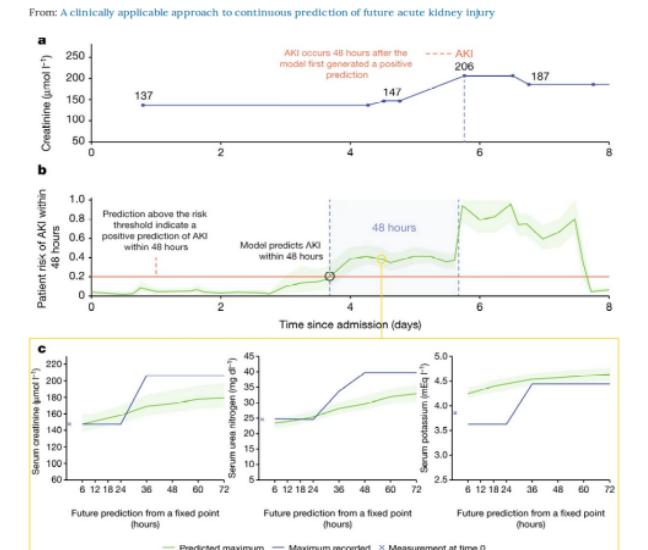
### A clinically applicable approach to continuous prediction of future acute kidney injury

Nenad Tomasev, Kavir Glorot, [...] Shahril Mohamed

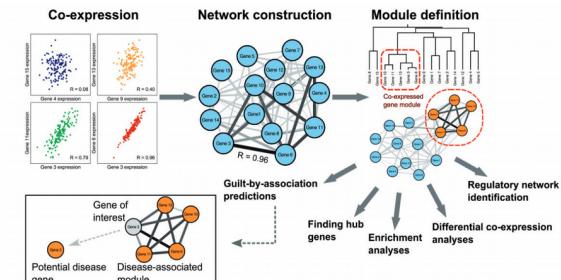
Nature 572, 116-119 (2019) | Download Citation ▾

#### Abstract

The early prediction of deterioration could have an important role in supporting healthcare professionals, as an estimated 11% of deaths in hospital follow a failure to promptly recognize and treat deteriorating patients<sup>1</sup>. To achieve this goal requires predictions of patient risk that are continuously updated and accurate, and delivered at an individual level with sufficient context and enough time to act. Here we describe a deep learning approach for the continuous risk prediction of future deterioration in patients, building on recent work that models adverse events from electronic health records<sup>2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17</sup> and using acute kidney injury – a common and potentially life-threatening condition<sup>18</sup> – as an exemplar. Our model was developed on a large, longitudinal dataset of electronic health records that cover diverse



## Consequence



- 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



*Knut och Alice  
Wallenbergs  
Stiftelse*



**LUNDS  
UNIVERSITET**