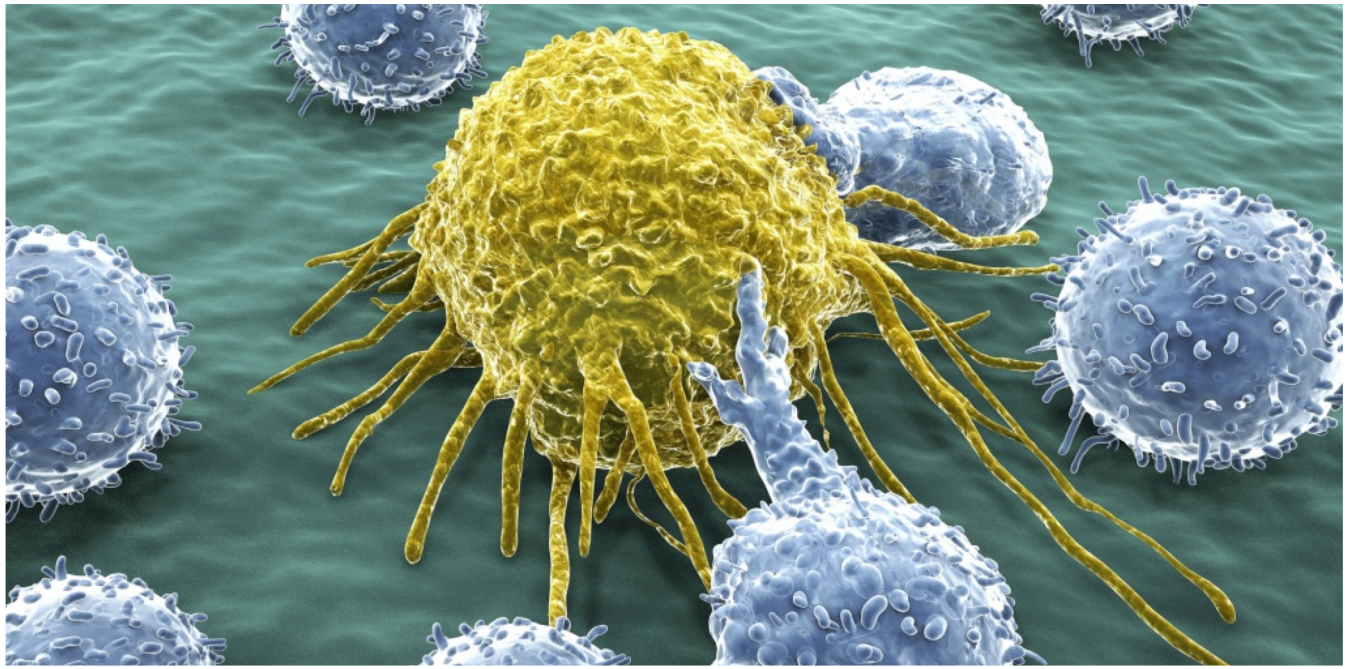
Finding Correlated Trends across Multiple Data Sets using Matrix Factorization

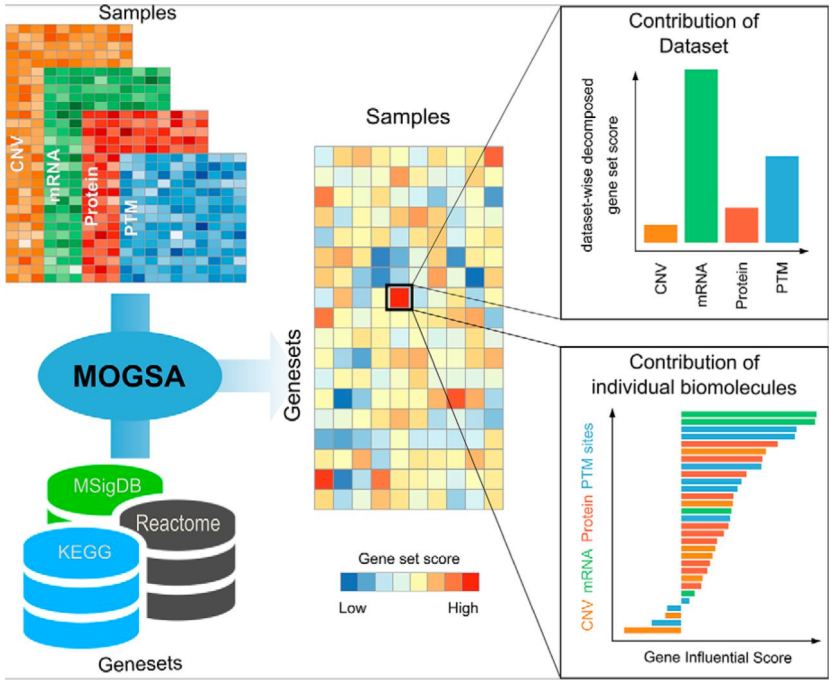


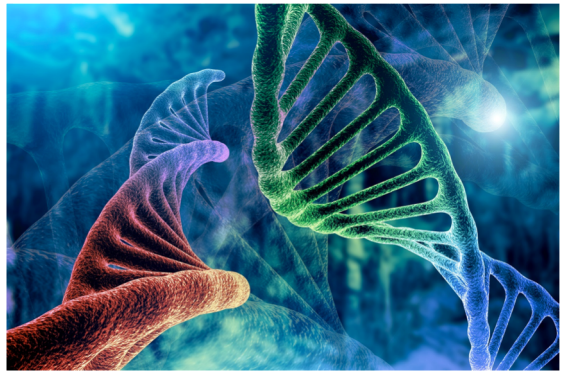
Aedín Culhane, PhD

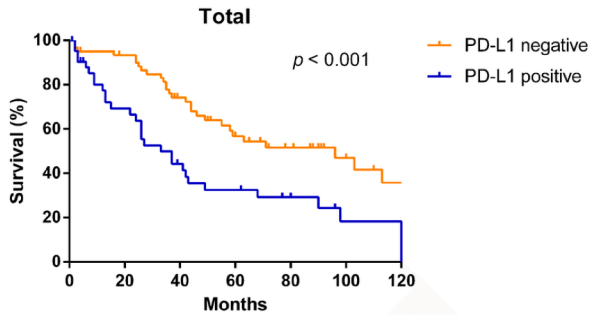
Twitter @AedinCulhane

Dana-Farber Cancer Institute | Department of Data Science

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH







A little about me

- PhD Scientist and PI in Computational Oncology
- Most coding, data analysis and methods development in R/Bioconductor
- Run Boston R/Bioconductor for genomics meetup group

meetup

- Co-Chair Bioc2020 http://bioc2020.bioconductor.org

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Dana-Farber Cancer Institute | Department of Data Science

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

- Our research motivation
- PCA <-> SVD relationship
- Impact of data structure, preprocessing
- Matrix factorization for >1 datasets
- Layering annotation onto projection

Understanding the Complexity of disease

- Human disease occurs within a complex milieu of cells
- Multiple body systems are involved in disease response;
  immune, lymph, nervous, endocrine
- Disease response involves local and systemic signalling

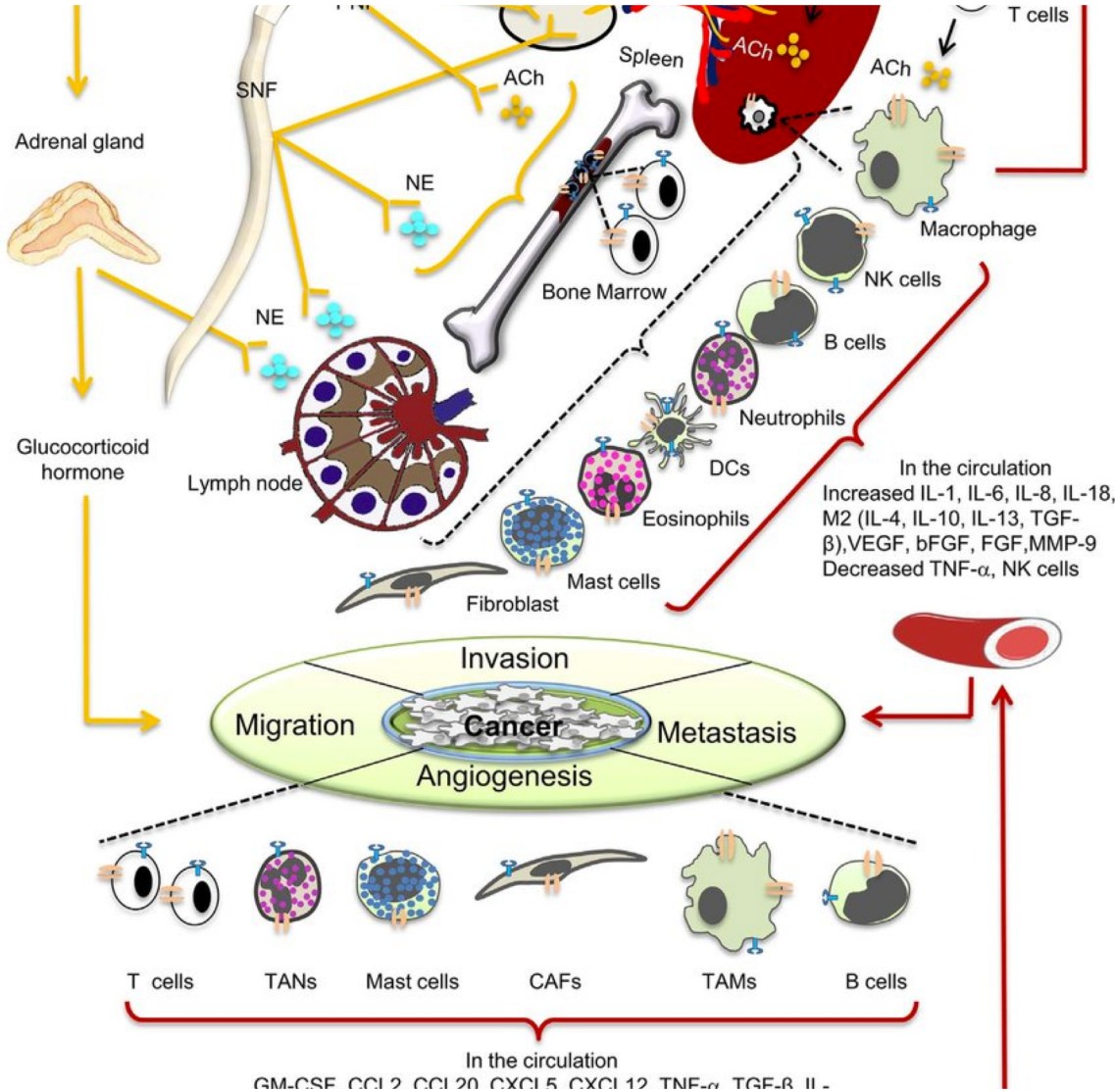
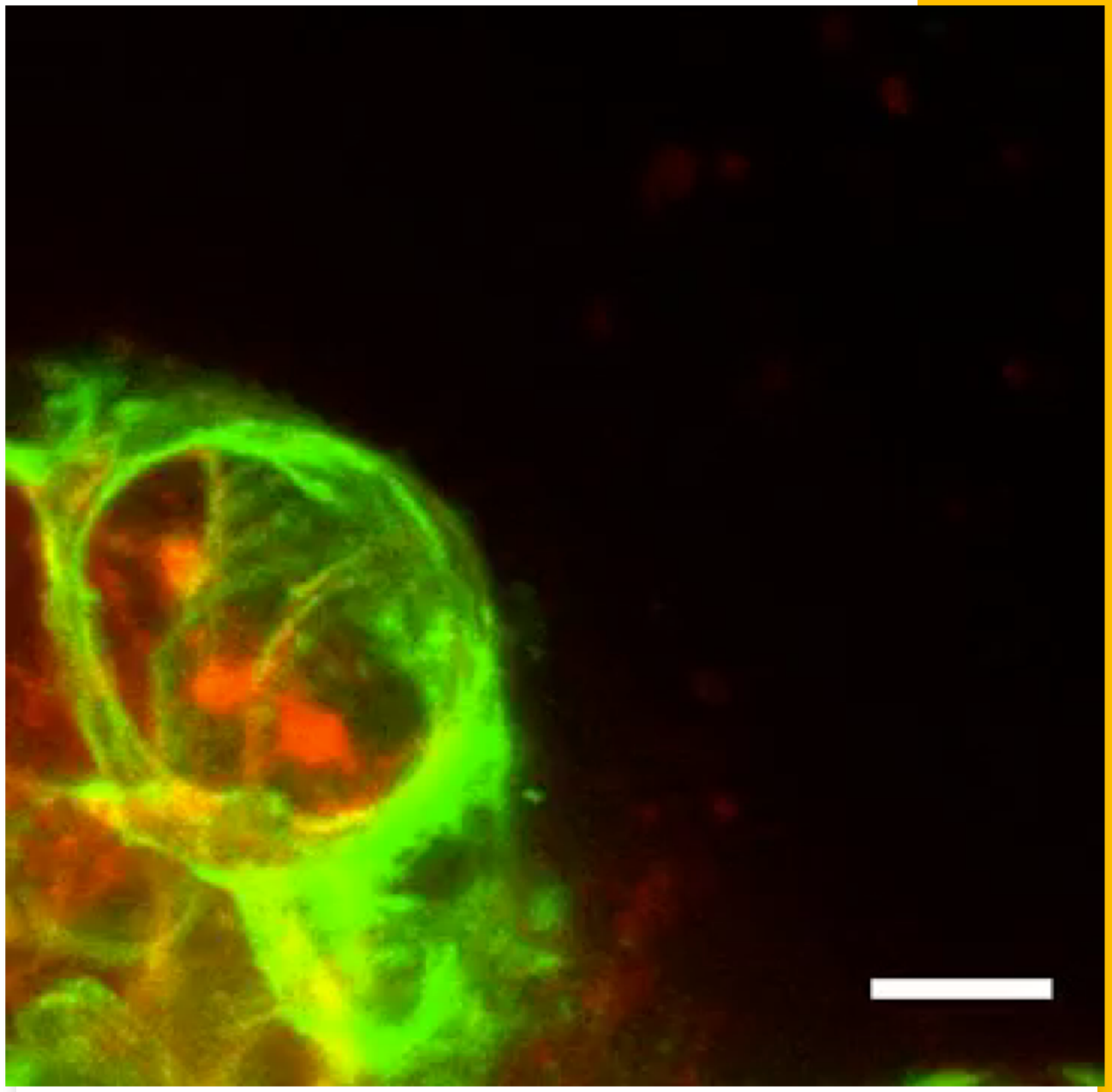


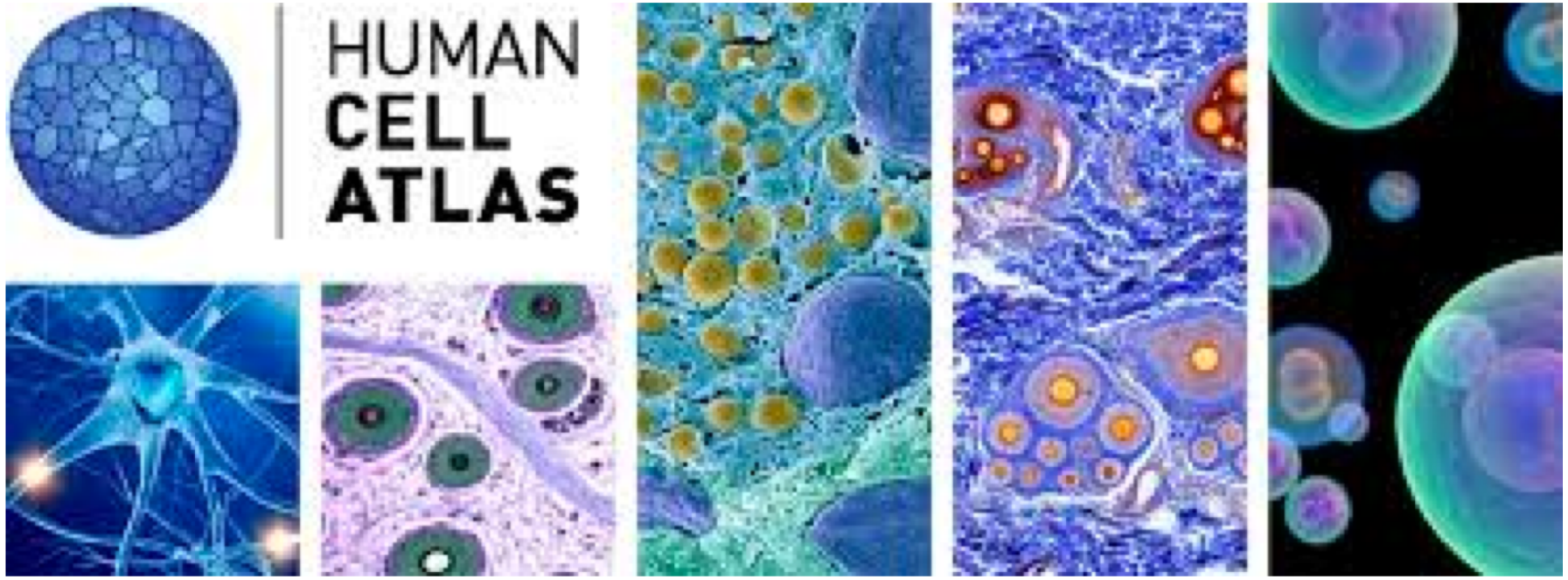
Figure from Kuol et al J Neuroimmunol. 2018 Feb 15;315:15-23

Cells Talk

We need to understand their conversations



Video from Sirka, O.K., Shamir, E.R., Ewald, A.J., 2018. Journal of Cell Biology 217, 3368–3381.

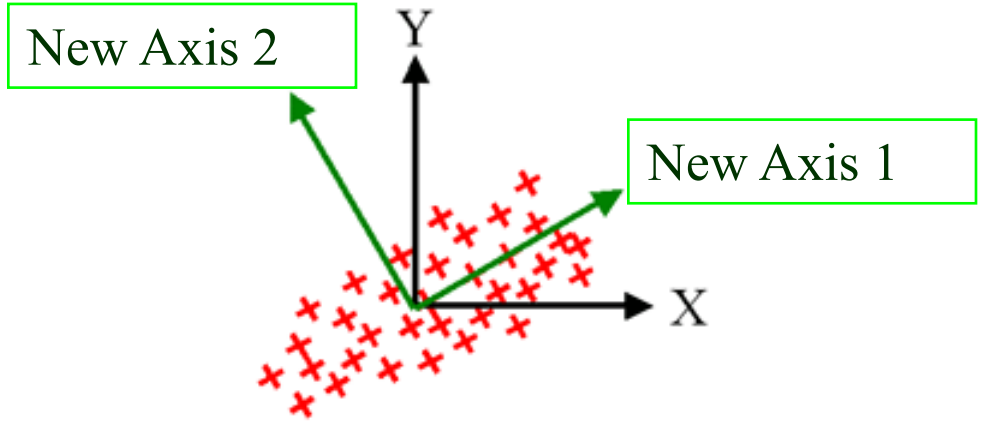


Goal: Create a Human Cell Atlas

catalog and map of all cell types to the location within tissues and within the body; temporal, spatial, development

Chan Zuckerberg Initiative

Matrix Factorization or Dimension reduction methods, including PCA are commonly used and is well suited to finding known & unknown (latent) patterns in large data



Reduce the data matrix to a small number of linear vectors that explain most of the variance in the data

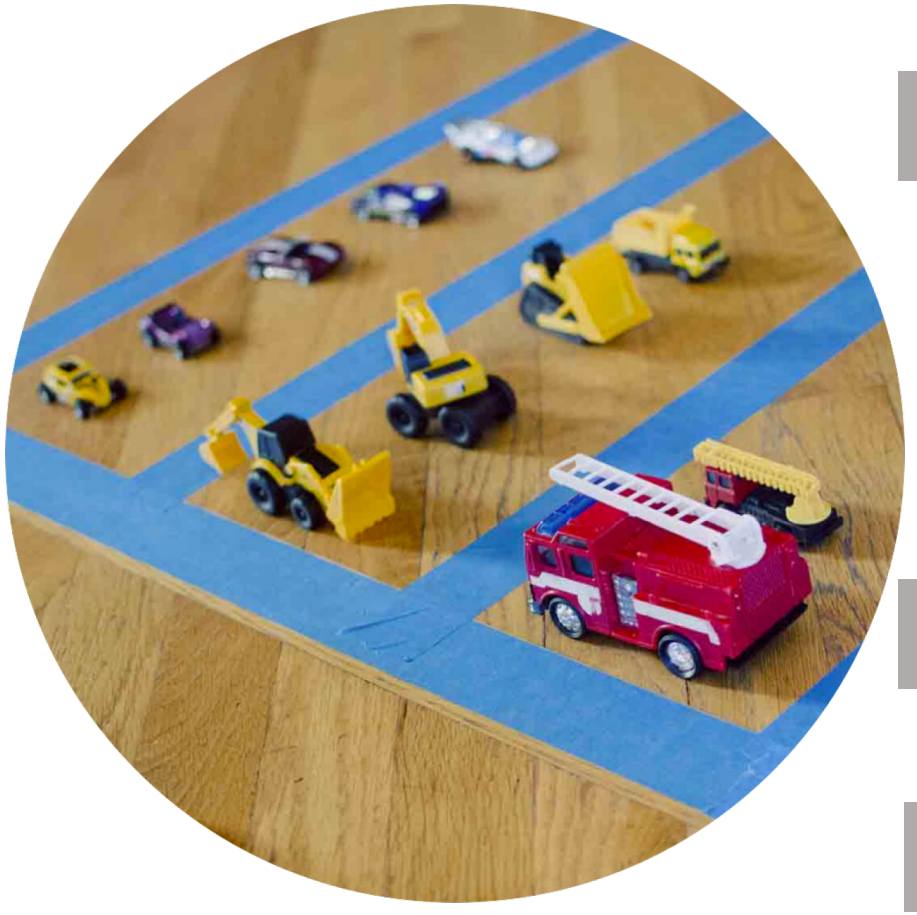
Dimension reduction is indispensable in large scale data analysis

Ordination

Dimension Reduction

Factor Analysis

Eigen analysis



Latent variable analysis

Matrix Factorization

Principal Component Analysis

Wavelet Decomposition

Spectral analysis

It is the basis of many classification, machine learning, recommender systems, analysis pipelines

Clustering, Classification

Exploratory Data analysis

Machine Learning

Lineage Reconstruction, Trajectory analysis

Dimension Reduction

Visualization t-SNE, UMAP

Also see Sun, S., Zhu, J., Ma, Y. *et al.* . *Genome Biol* **20,** 269 (2019)

In the original t-SNE article, PCA is step 1

Journal of Machine Learning Research 9 (2008) 2579-2605 Submitted 5/08; Revised 9/08; Published 11/08

Visualizing Data using t-SNE

Laurens van der Maaten LVDMAATEN@GMAIL.COM
TiCC
Tilburg University
P.O. Box 90153, 5000 LE Tilburg, The Netherlands

Geoffrey Hinton HINTON@CS.TORONTO.EDU
Department of Computer Science
University of Toronto
6 King's College Road, M5S 3G4 Toronto, ON, Canada

"In all of our experiments, we start by using PCA to reduce the dimensionality of the data to 30.

This speeds up the computation of pairwise distances between the datapoints and suppresses some noise without severely distorting the interpoint distances"

PCA initialization improves t-SNE

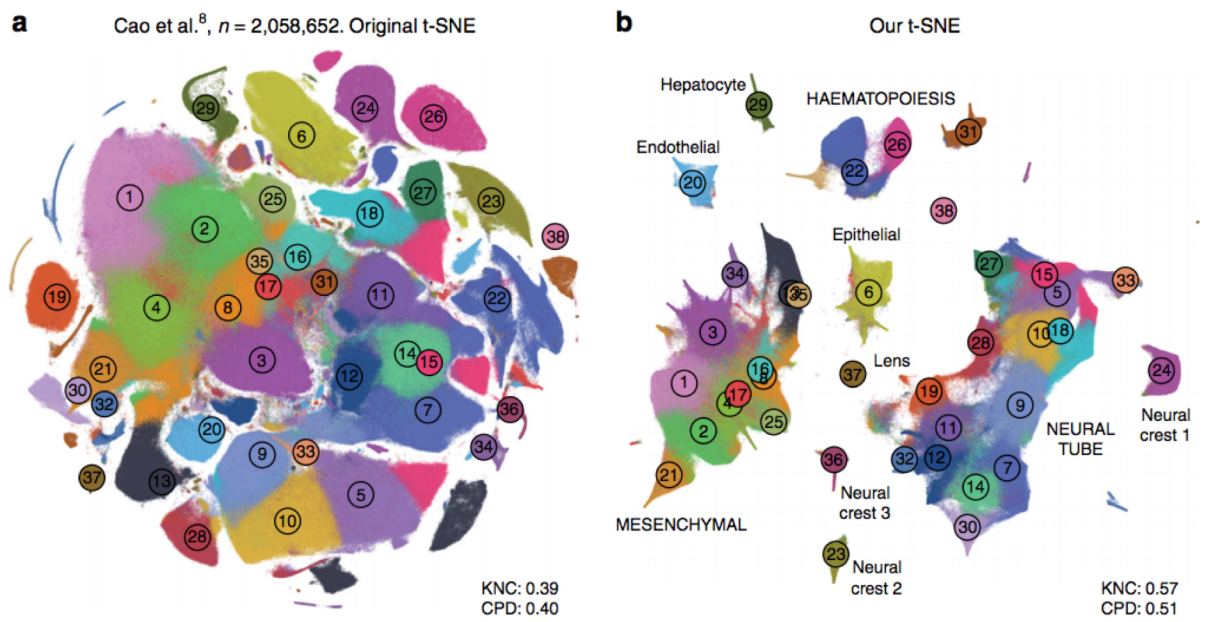


Fig. 9 Cao et al. data set. Sample size $n = 2{,}058{,}652$. Cluster assignments and cluster colours are taken from the original publication[8]. **a** T-SNE embedding from the original publication. The authors ran t-SNE in `scanpy` with default settings, i.e. with random initialisation, perplexity 30, and learning rate 1000. For cluster annotations, see original publication. **b** T-SNE embedding produced with our pipeline for large data sets: a random sample of 25,000 cells was embedded using PCA initialisation, learning rate 25,000/12, and perplexity combination of 30 and 250; all other cells were positioned on resulting embedding and this was used to initialise t-SNE with learning rate 2,058,652/12, perplexity 30, and exaggeration 4. Labels correspond to the ten developmental trajectories identified in the original publication. Labels in capital letters denote trajectories consisting of multiple clusters. 32,011 putative doublet cells are not shown in either panel.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13056-x OPEN

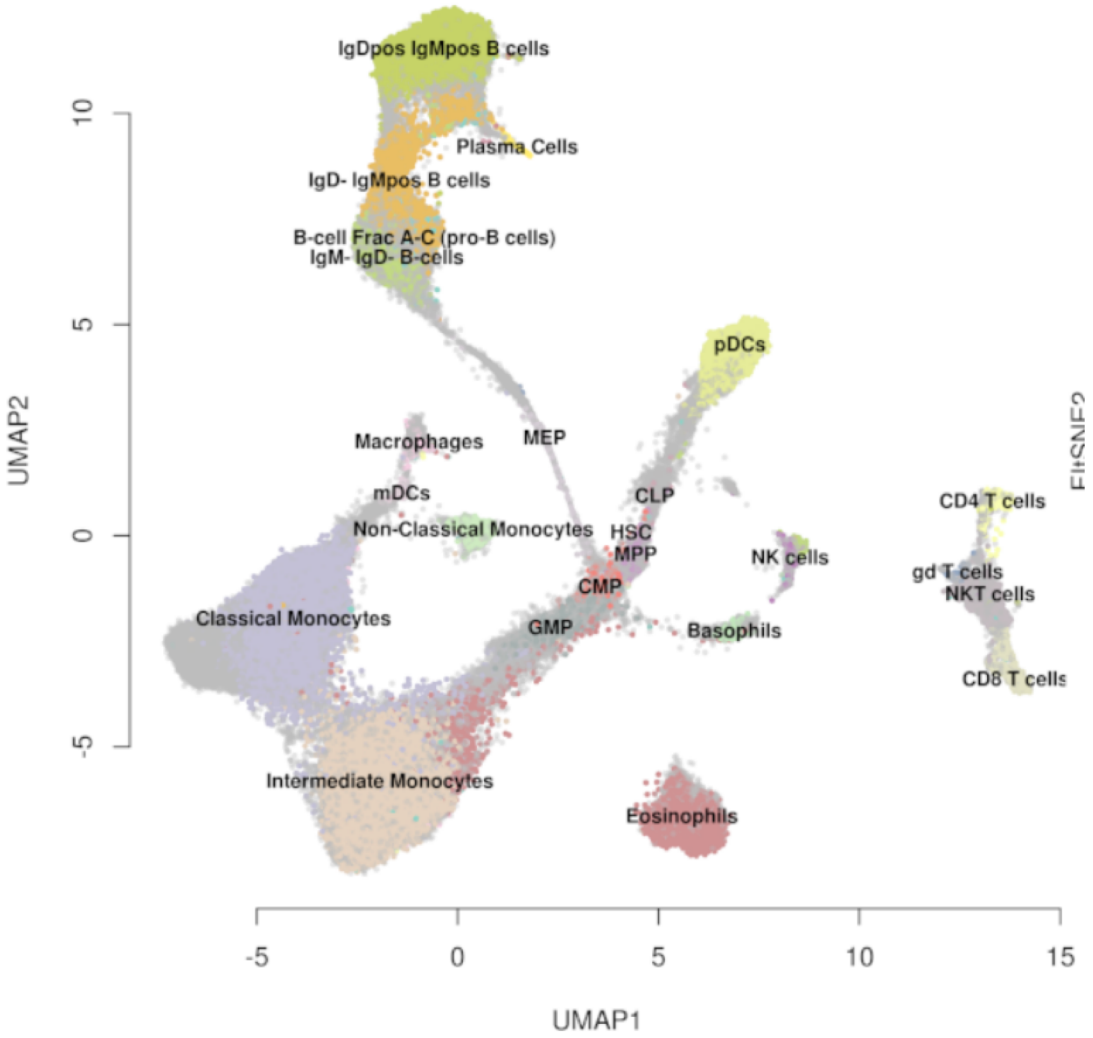
The art of using t-SNE for single-cell transcriptomics

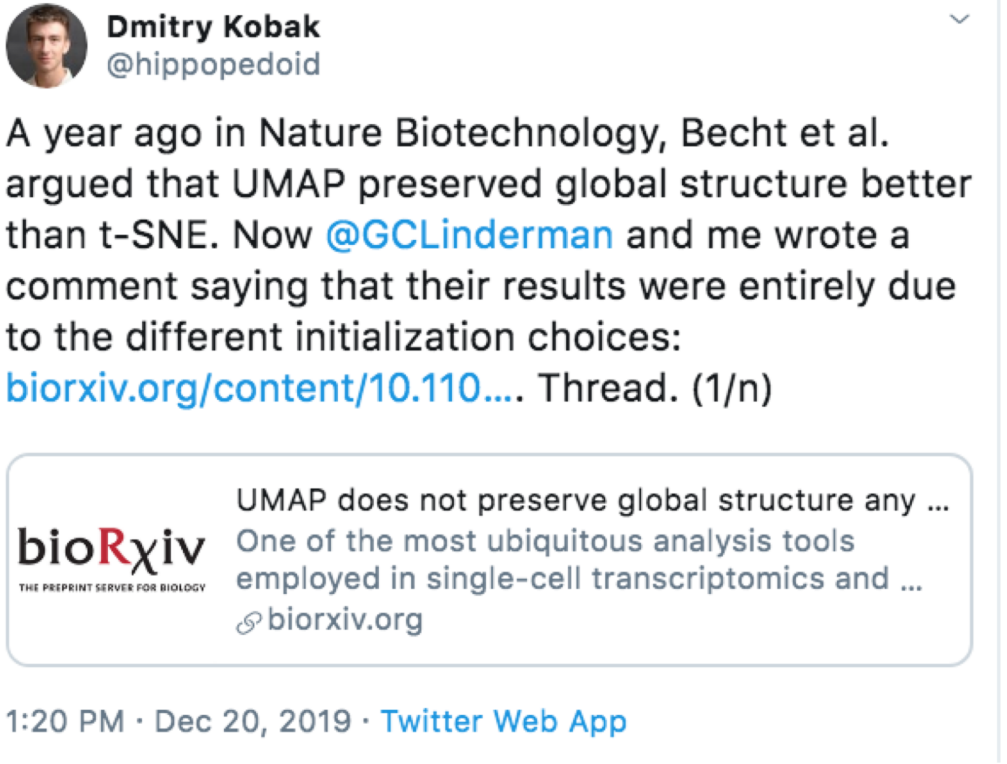
Dmitry Kobak[1*] & Philipp Berens[1,2,3,4*]

Single-cell transcriptomics yields ever growing data sets containing RNA expression levels for thousands of genes from up to millions of cells. Common data analysis pipelines include a dimensionality reduction step for visualising the data in two dimensions, most frequently performed using t-distributed stochastic neighbour embedding (t-SNE). It excels at revealing local structure in high-dimensional data, but naive applications often suffer from severe shortcomings, e.g. the global structure of the data is not represented accurately. Here we describe how to circumvent such pitfalls, and develop a protocol for creating more faithful t-SNE visualisations. It includes PCA initialisation, a high learning rate, and multi-scale similarity kernels; for very large data sets, we additionally use exaggeration and downsampling-based initialisation. We use published single-cell RNA-seq data sets to demonstrate that this protocol yields superior results compared to the naive application of t-SNE.

"We showed that using informative initialisation (such as PCA initialisation, or downsampling-based initialisation) can substantially improve the global structure of the final embedding because it survives through the optimisation process"

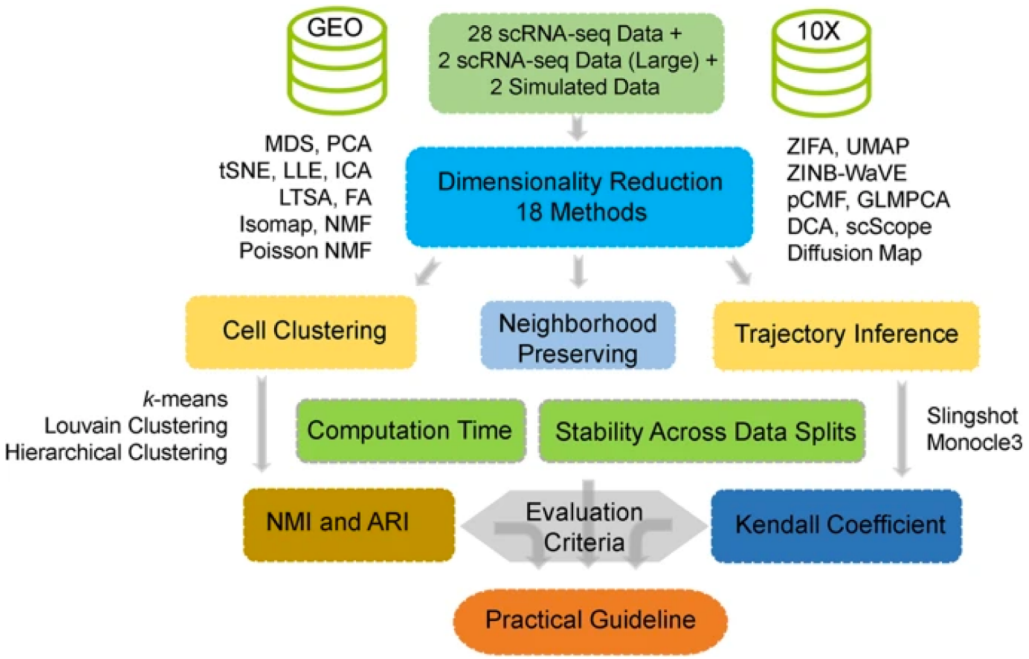
t-SNE with PCA embedding is as performant as UMAP

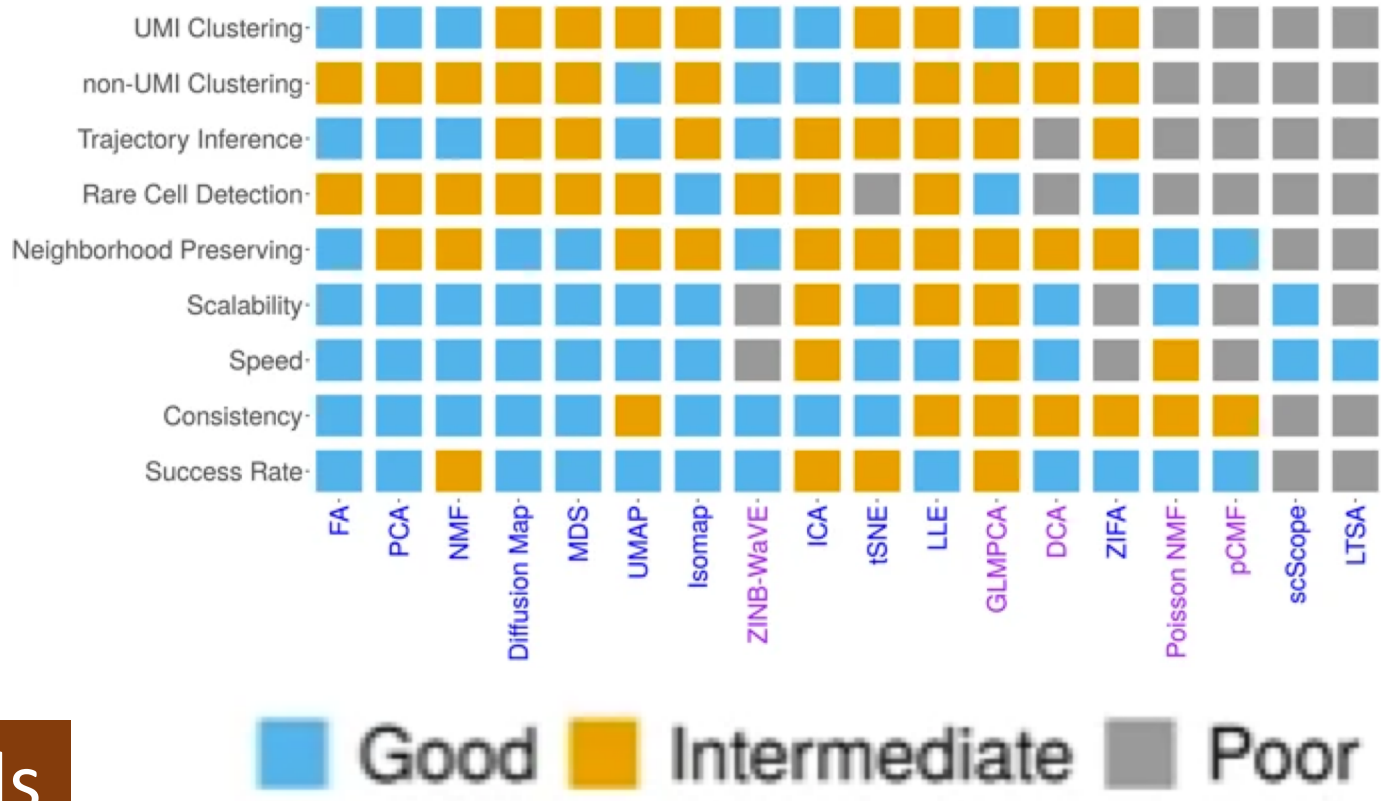




t-SNE with PCA initialization produces more meaningful embeddings and performs comparably to UMAP

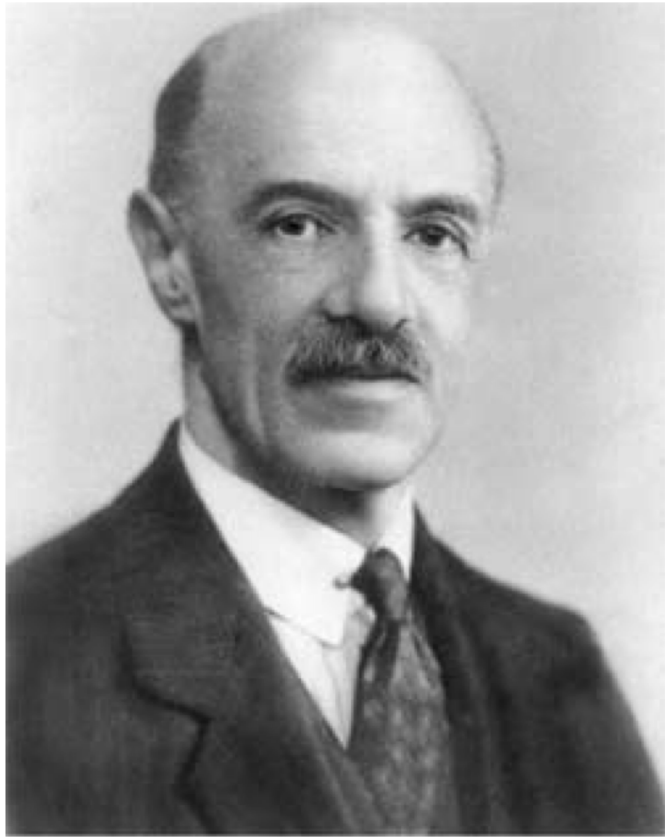
Assessment of the Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis





Comparison of 18 Methods

Sun, S., Zhu, J., Ma, Y. *et al.* . *Genome Biol* **20,** 269 (2019)



Charles Spearman (1863- 1945)

Grote Professor of the Philosophy of Mind and Logic, University College London

1904- Factor Analysis

"GENERAL INTELLIGENCE," OBJECTIVELY DETERMINED AND MEASURED.

By C. SPEARMAN.

THE PROOF AND MEASUREMENT OF ASSOCIATION BETWEEN TWO THINGS.

By C. SPEARMAN.

As example, we will take Pearson's chief line of investigation, Collateral Heredity, at that point where it comes into closest contact with our own topic, Psychology. Since 1898 he has, with government sanction and assistance, been collecting a vast number of data as to the amount of correspondence existing between brothers. A preliminary calculation, based in each case upon 800 to 1,000 pairs, led, in 1901, to the publication of the following momentous results:

COEFFICIENTS OF COLLATERAL HEREDITY.

Correlation of Pairs of Brothers.

PHYSICAL CHARACTERS.		MENTAL CHARACTERS.	
(Family Measurements.)		(School Observations.)	
Stature	0.5107	Intelligence	0.4559
Forearm	0.4912	Vivacity	0.4702
Span	0.5494	Conscientiousness	0.5929
Eye-color	0.5169	Popularity	0.5044
(School Observations.)		Temper	0.5068
Cephalic index	0.4861	Self-consciousness	0.5915
Hair-color	0.5452	Shyness	0.5281
Health	0.5203		
Mean	0.5171	Mean	0.5214

Dealing with the means for physical and mental characters, we are forced to the perfectly definite conclusion, *that the mental characters in man are inherited in precisely the same manner as the physical.*[1] Our mental and moral nature is, quite as much as our physical nature, the outcome of hereditary factors.

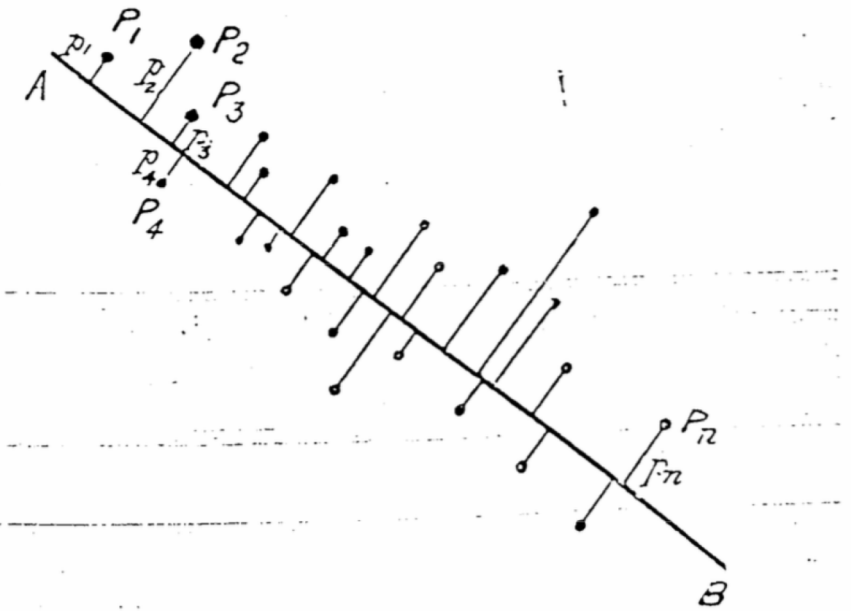
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

$$U = S(p^2) = \text{a minimum.}$$

If y were the dependent variable, we should have made

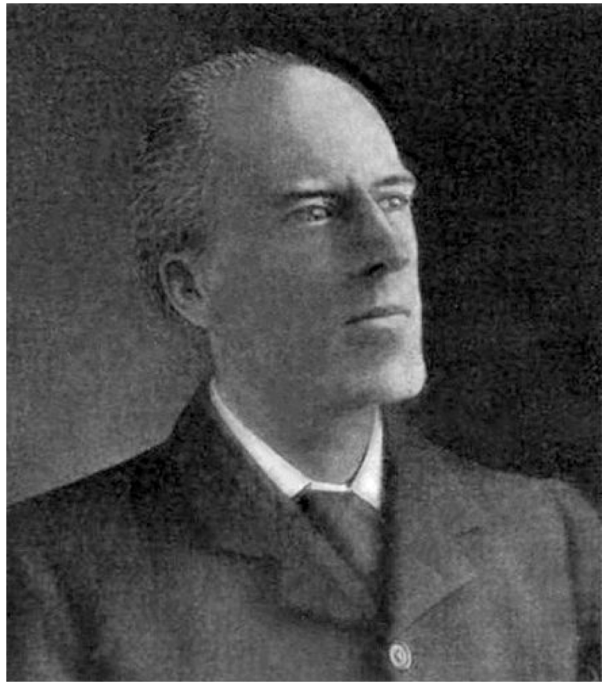
$$S(y' - y)^2 = \text{a minimum}$$

(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



Now clearly $U = S(p^2)$ is the moment of momentum, the second moment of the system of points, supposed equally loaded, about the line A B. But the second moment of a system about a series of parallel lines is always least for the

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* **2**:559-572.



That the line which fits best a system of n points in q-fold space passes through the centroid of the system and coincides in direction with the least axis of the ellipsoid of residuals.

Karl Pearson (1857 -1936)

Pearson was the Galton Professor of Eugenics at University College, London (UCL)

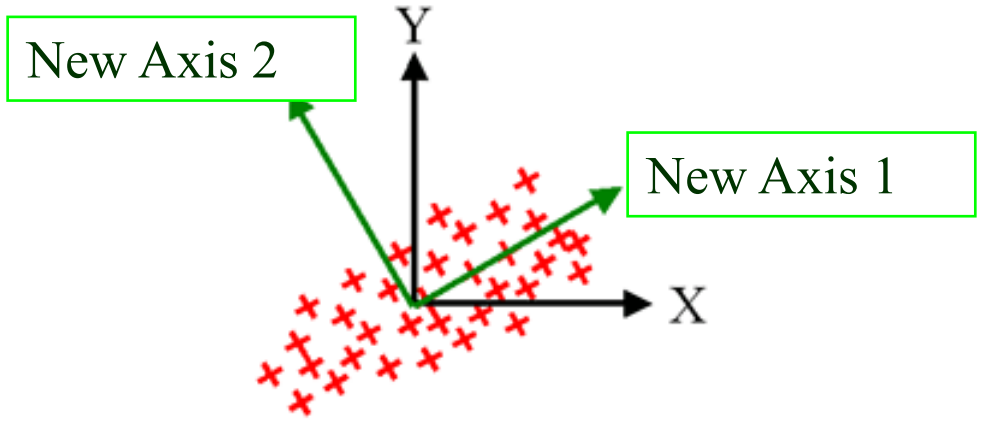
1901- PCA

Classical Dimension Reduction Matrix Factorization approaches

- Principal component analysis (PCA)
  - Correlation –based PCA
  - Covariance –based PCA (less common)

- Nonmetric multidimensional scaling (NMDS, MDS)
- Correspondence analysis (COA or CA)
- Principal co-ordinate analysis (PCoA)

Each perform a different pre-processing on the data data & uses SVD to reduce data to lower dimension that explain the “patterns” in the data



PC1 will explain the greatest proportion of the variance in the data.

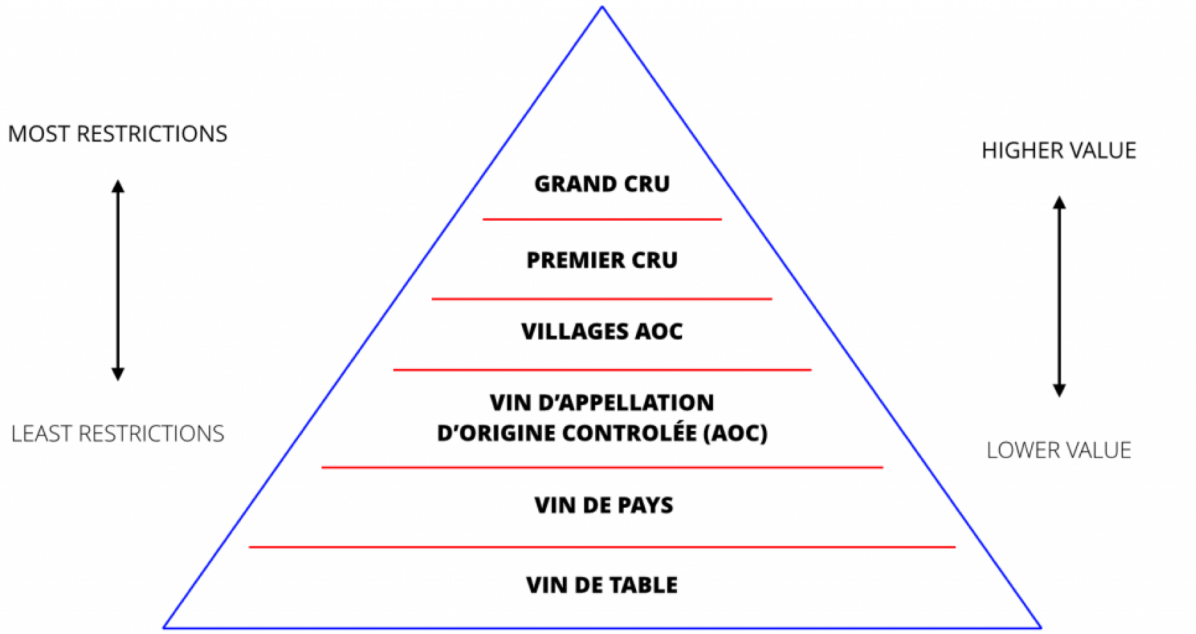
The second new axis, PC2 will be orthogonal, and will explain the next largest amount of variance

Stein-O'Brien GL, et al., Enter the Matrix: Factorization Uncovers Knowledge from Omics. 2018 *Trends in Genetics* DOI: (10.1016/j.tig.2018.07.003)

Meng & Zeleznik et al., (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics, 17(4), 2016, 628–641

Simple Example

200 judges performed a blind tasting of 5 red wines from Bordeaux



	excellent	good	mediocre	boring
Cru_Bourgeois	45	126	24	5
Grand_Cru_classe	87	93	19	1
Bordeaux_d_origine	36	68	74	22
Vin_de_marque	0	30	111	59
Vin_de_table	0	0	52	148

```r
library(dplyr)
library(ggpomological)
library(ggplot2)
library(ade4)
data(bordeaux)
df<-bordeaux %>%
  tibble::rownames_to_column(var="Wine") %>%
  reshape2::melt(.,variable.name="Judge_Score",
                 value.name="Number")

p<- ggplot(df,
           aes(Judge_Score, Number, fill=Judge_Score))+
    geom_bar(color="black",stat = "identity") +
    facet_wrap(~Wine, nrow = 2)+
    scale_fill_pomological()+
    theme_pomological("Homemade Apple", 12)+
    theme(axis.text.x=element_blank(),
          legend.position = "bottom",
          legend.key = element_rect(colour = "black"))

 paint_pomological(p,res = 110) %>%
   magick::image_write("barplot-painted.png")
```

PCA of Bordeaux

```
data(bordeaux)
bordeauxS= scale(bordeaux,center = TRUE, scale = TRUE)
s= svd(bordeauxS)
s$u %*% diag(s$d) %*% t(s$v) #  X = U D V'
```

Original data

	excellent	good	mediocre	boring
Grand_Cru_classe	87	93	19	1
Cru_Bourgeois	45	126	24	5
Bordeaux_d_origine	36	68	74	22
Vin_de_table	0	0	52	148
Vin_de_marque	0	30	111	59

Scaled & Centred (z-score)

	excellent	good	mediocre	boring
Grand_Cru_classe	1.500	0.590	-0.98	-0.76
Cru_Bourgeois	0.320	1.300	-0.84	-0.69
Bordeaux_d_origine	0.066	0.092	0.47	-0.41
Vin_de_table	-0.930	-1.300	-0.10	1.70
Vin_de_marque	-0.930	-0.670	1.40	0.20

Run SVD

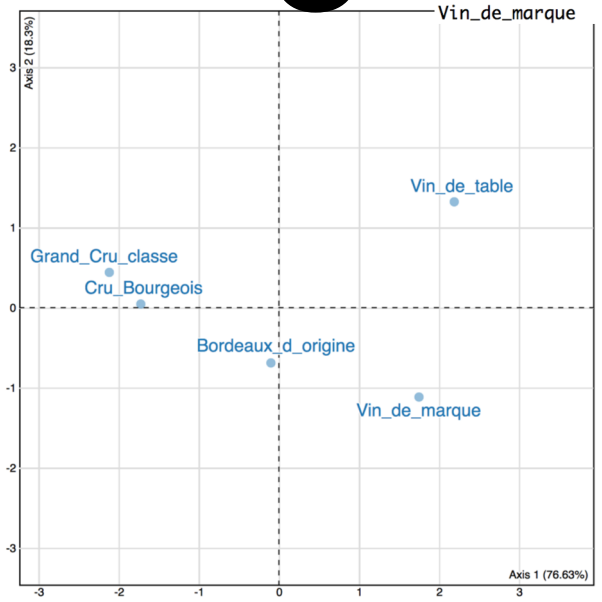
$$X = U\ D\ V'$$

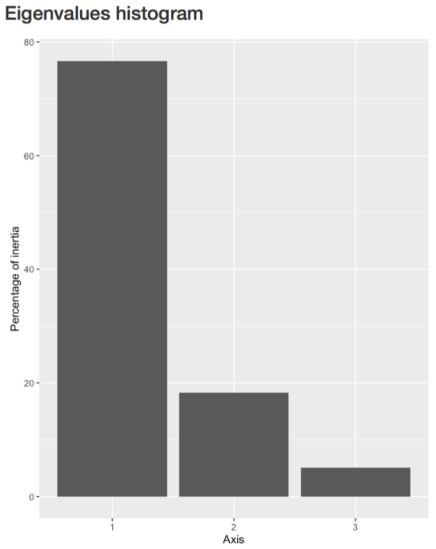
Visualize

```
library(ade4)
library(explor)
explor(dudi.pca(bordeaux, scan = FALSE))
```

U

	u1	u2
Cru_Bourgeois	-0.440	0.024
Grand_Cru_classe	-0.540	0.230
Vin_de_table	0.560	0.690
Bordeaux_d_origine	-0.025	-0.360
Vin_de_marque	0.450	-0.580





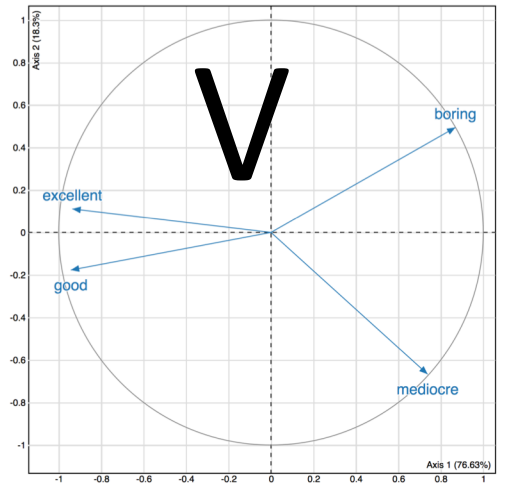
D

	PC1	PC2
PC1	3.5	0.0
PC2	0.0	1.7

Eigenvalues table

Axis	%	Cum. %
1	76.6	76.6
2	18.3	94.9
3	5.1	100.0

V

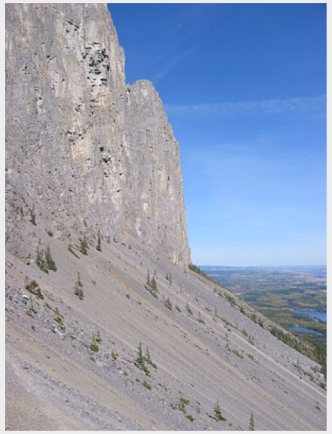


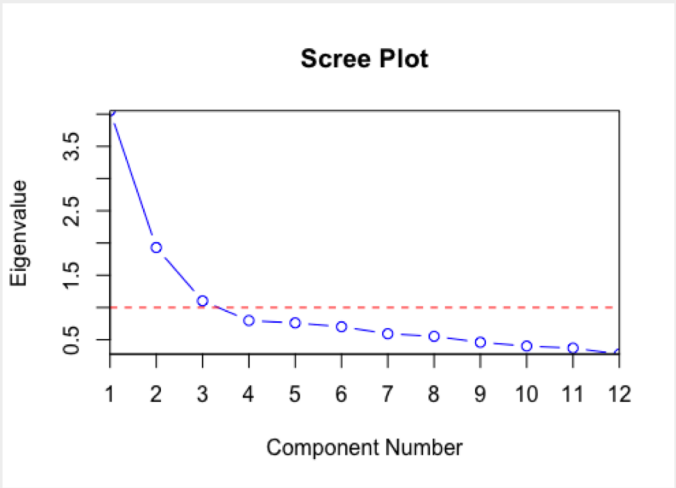
	excellent	good	mediocre	boring
CS1	-0.53	-0.54	0.42	0.50
CS2	0.13	-0.21	-0.78	0.58

Dimension reduction: PCA $\leftrightarrow$ SVD

- Widely used in statistical and computational science
- PCA can be computed by linear regression, eigen analysis, singular value decomposition (SVD), latent factor analysis
  - https://aedin.github.io/talks/PCA.html
- Vectors are called principal components, principal axes, latent vectors, eigen vector, etc and capture variance (information) in the data
- Number selected by **scree** plot or permutations.

Scree

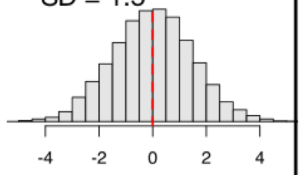
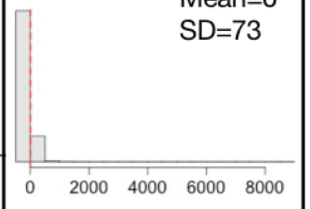
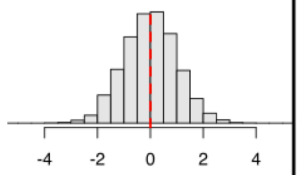
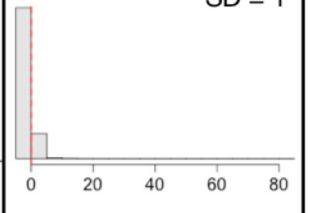




The "Kaiser rule" criteria is shown in red.

Covariance-based PCA

Correlation-based PCA

Center	$x_{i,j}{}^{*} = x_{i,j} - column\ mean$	Mean = 0 SD = 1.5 	Mean=0 SD=73 
Standardize	$x_{i,j}{}^{*} = \frac{x_{i,j} - column\ mean}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure. For example z-score subtracts means, divides by standard deviation	Mean = 0 SD = 1 	Mean = 0 SD = 1 

$X = U\ D\ V'$

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

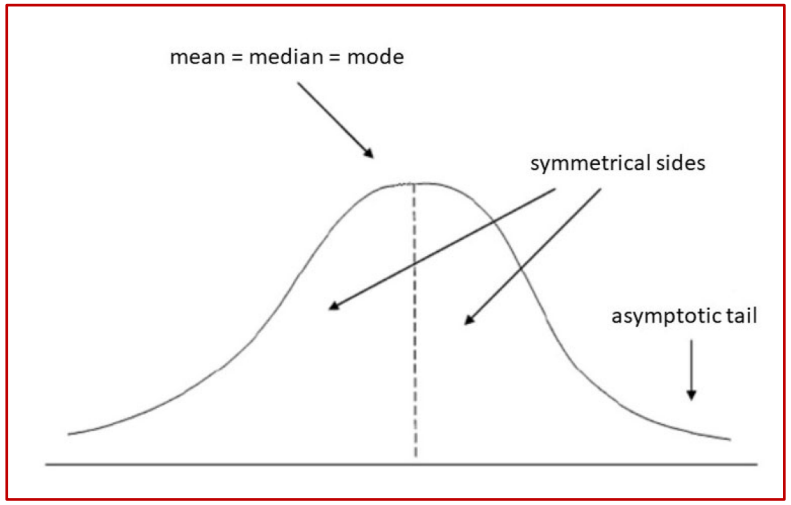
Covarianced normalized by Standard Deviation

Correlation between X and Y

Standard deviation of X

Standard deviation of Y

Each pre-processing step impacts the data



There are many other pre-processing steps
These are just some common ones

Method	Formula	Example raw datasets Graphical Examples Toy data Mean = 1.5 SD = 1.5	scMix, 10X counts Mean = 14.7 SD = 73.0
Scale	$x_{i,j}^{*} = \frac{x_{i,j}}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure	Mean = 0.7 SD = 0.7	Mean=0.2 SD = 1.0
Center	$x_{i,j}^{*} = x_{i,j} - column\ mean$	Mean = 0 SD = 1.5	Mean=0 SD=73
Standardize	$x_{i,j}^{*} = \frac{x_{i,j} - column\ mean}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure. For example z-score subtracts means, divides by standard deviation	Mean = 0 SD = 1	Mean = 0 SD = 1
Transform	$x_{i,j}^{*} = f(x_{i,j})$ where *f(x)* is the transformation function, for example logarithms are commonly used	Mean = 0.5 SD = 1.4 *Log_2 transformation*	Mean=2.0 SD = 1.9 *Log_2 transformation, pseudocount of 1*

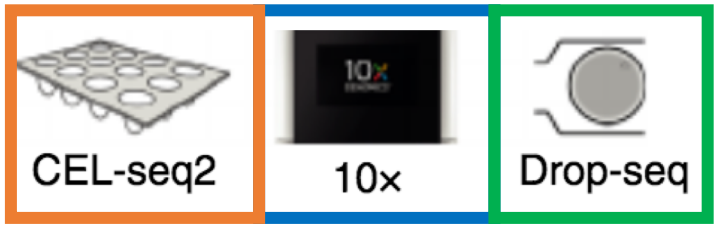
Single Cell RNAseq

- Stripped R code down to basics to examine impact of each processing step on results
- Implemented faster SVD (currently using IRBLA)
- Sparse matrix representation
- Benchmark data include scMix, Sun et al. data, etc
  - scMix Three cell lines, three sequencing platforms Tian et al,. Nat Methods. 2019

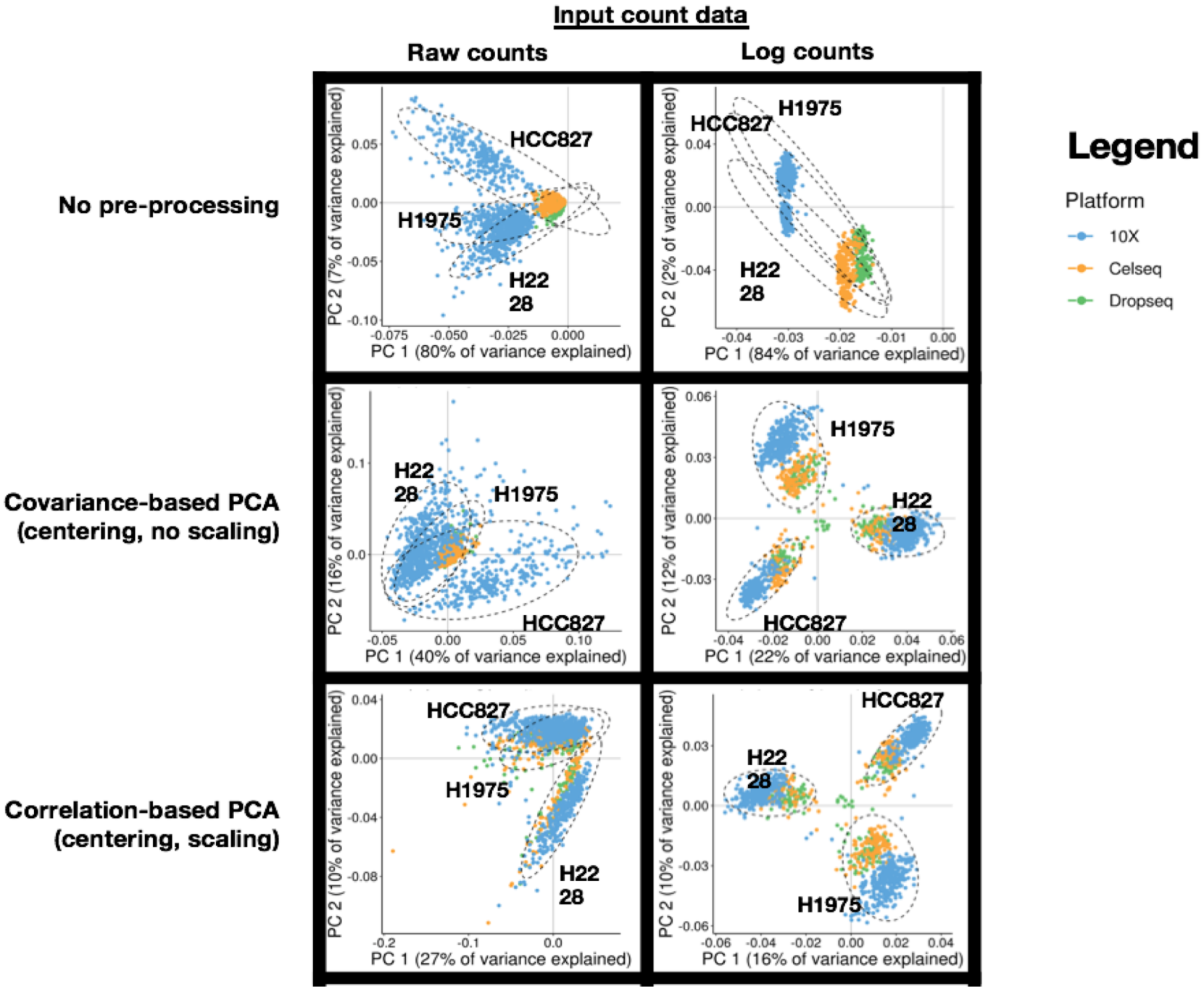
Work by Harvard TH Chan master's student Lauren Hsu

Processing steps impact results

human lung adenocarcinoma cell lines
HCC827
H1975
H2228



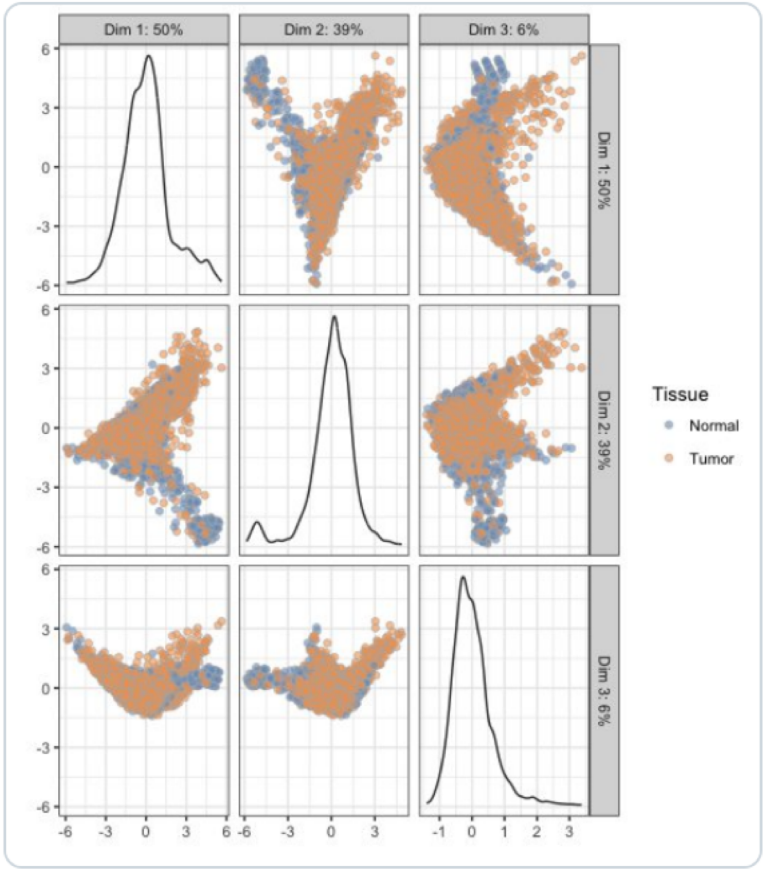
Submitted Frontiers in Oncolcogy



Watch out for Horseshoes

Aedin Culhane
@AedinCulhane

Pretty fun to find a "classic" PCA horseshoe artifact when analyzing some scRNAseq data. For a nice explanation of this see ordination.okstate.edu/PCA.htm. PCA likes data measured (or normalized to have) similar units. It expects linear relationships. It hates lots of zeros.



5:59 AM · Jun 14, 2018 · Twitter Web Client

Simina M. Boca @siminaboca · Jun 14, 2018
Replying to @AedinCulhane
Thanks for sharing! I've seen even clearer horseshoes with metabolomics data with many zeros - similar to the one in the link you sent - but had no idea they had this name!

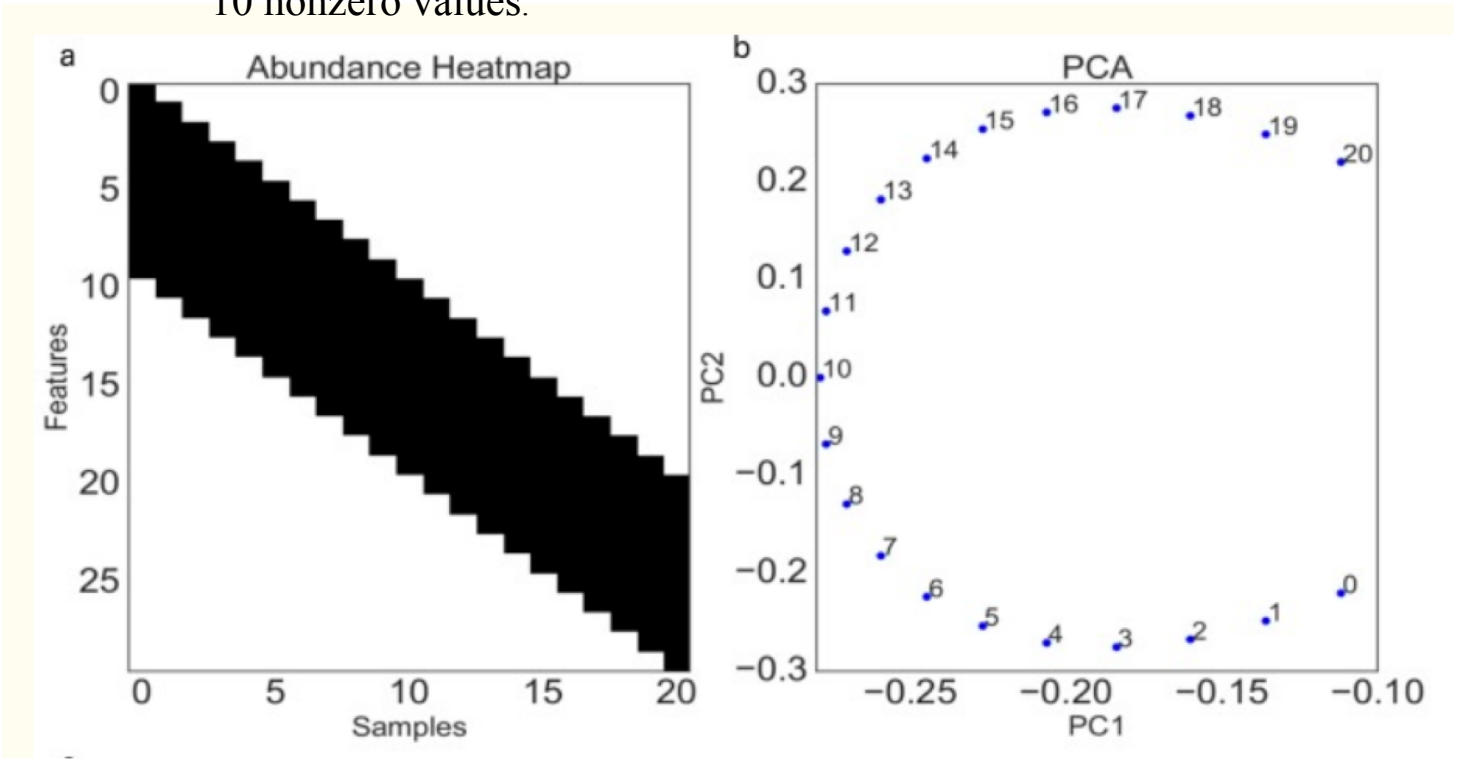
2

Mick Watson @BioMickWatson · Jun 14, 2018
Replying to @AedinCulhane
This is often the shape of PCA on e.g. human genetic variation data e.g. goo.gl/images/eZgpte so it's not always bad

Jamie Morton @jamietmorton · Jun 14, 2018
Replying to @AedinCulhane
Its possible that you have quite a few features unique to the tissue type - see our paper here that builds off of @SherlockpHolmes work:

Uncovering the Horseshoe Effect in Microbial Analyses
The horseshoe effect is a phenomenon that has long intrigued ecologists. The effect was commonly thoug...
msystems.asm.org

10 nonzero values.



PCA Horseshoes

- Arch, horseshoe effect or Guttman effect
- Arise as a consequence of distance metrics that saturate.
- Morton et al., 2017 mSystems. 2(1): e00166-16

Beyond PCA- Other matrix factorization approaches

Table 2. Dimension reduction methods for one data set

Method	Description	Name of R function {R package}
PCA	Principal component analysis	prcomp{stats}, princomp{stats}, dudi.pca{ade4}, pca{vegan}, PCA{FactoMineR}, principal{psych}
CA, COA	Correspondence analysis	ca{ca}, CA{FactoMineR}, dudi.coa{ade4}
NSC	Nonsymmetric correspondence analysis	dudi.nsc{ade4}
PCoA, MDS	Principal co-ordinate analysis/multiple dimensional scaling	cmdscale{stats} dudi.pco{ade4} pcoa{ape}
NMF	Nonnegative matrix factorization	nmf{nmf}
nmMDS	Nonmetric multidimensional scaling	metaMDS{vegan}
sPCA, nsPCA, pPCA	Sparse PCA, nonnegative sparse PCA, penalized PCA. (PCA with feature selection)	SPC{PMA}, spca{mixOmics}, nsprcomp{nsprcomp}, PMD{PMA}
NIPALS PCA	Nonlinear iterative partial least squares analysis (PCA on data with missing values)	nipals{ade4} pca{pcaMethods}[a] nipals{mixOmics}
pPCA, bPCA	Probabilistic PCA, Bayesian PCA	pca{pcaMethods}[a]
MCA	Multiple correspondence analysis	dudi.acm{ade4}, mca{MASS}
ICA	Independent component analysis	fastICA{FastICA}
sIPCA	Sparse independent PCA (combines sPCA and ICA)	sipca{mixOmics} ipca{mixOmics}
plots	Graphical resources	R packages including scatterplot3d, ggord[b], ggbiplot[c], plotly[d], explor

[a]Available in Bioconductor.
[b]On github: devtools::install_github ('fawda123/ggord').
[c]On github: devtools::install_github ('ggbiplot', 'vqv').
[d]On github: devtools::install_github ('ropensci/plotly').

From our Briefings in Bioinformatics review. Meng et al., 2016 https://academic.oup.com/bib/article/17/4/628/2240645

Other related methods

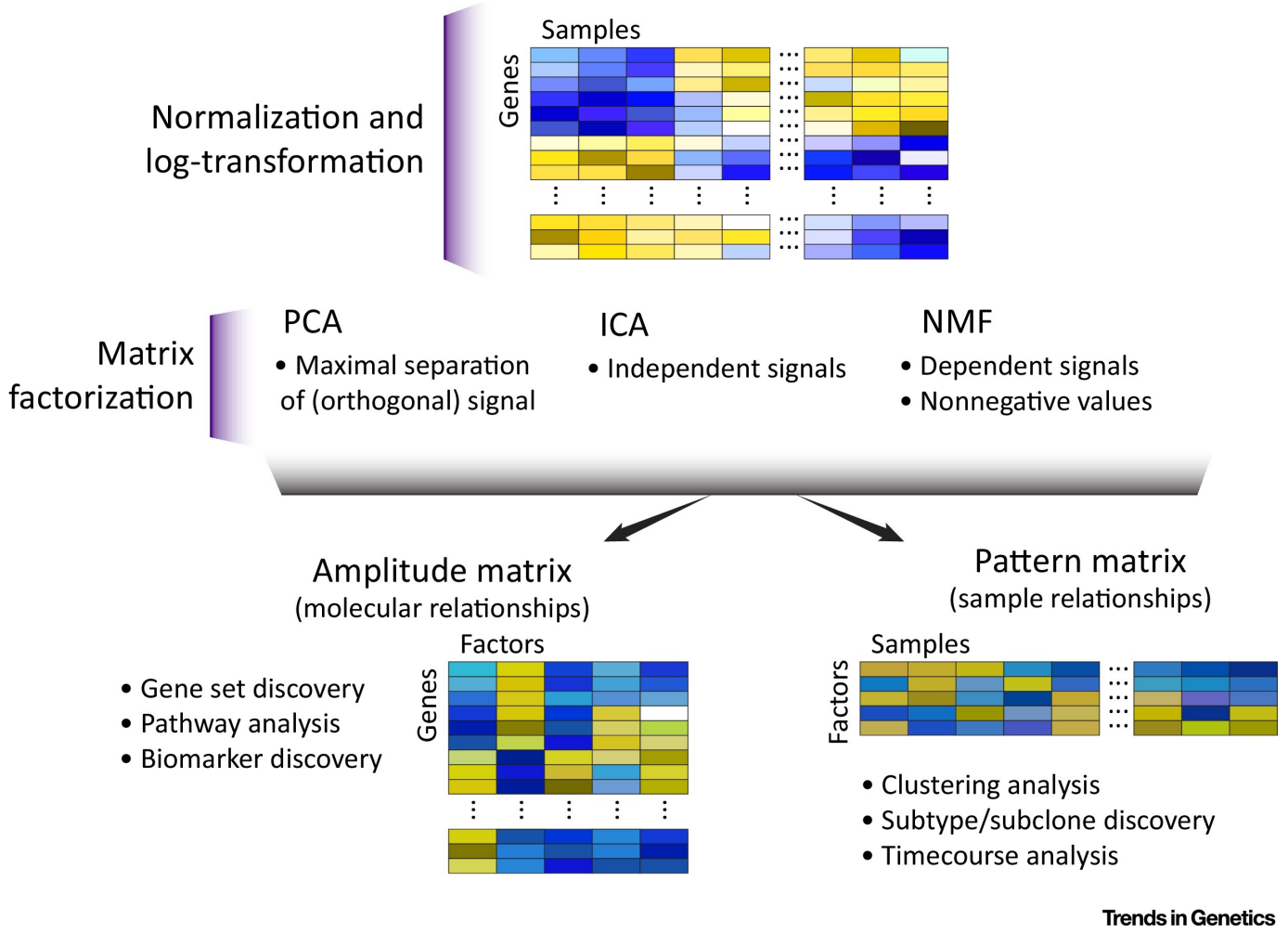
Independent Component Analysis

- does not constrain the axes to be orthogonal
- attempts to place them in the directions of statistical dependencies in the data.

Spectral map analysis

- related to COA (dual scaling of both rows + columns)
- not limited to contingency tables and cross-tabulations. possibility to use other weighting factors
- Wouters et al., 2003 showed SMA outperformed PCA, comparable to COA.

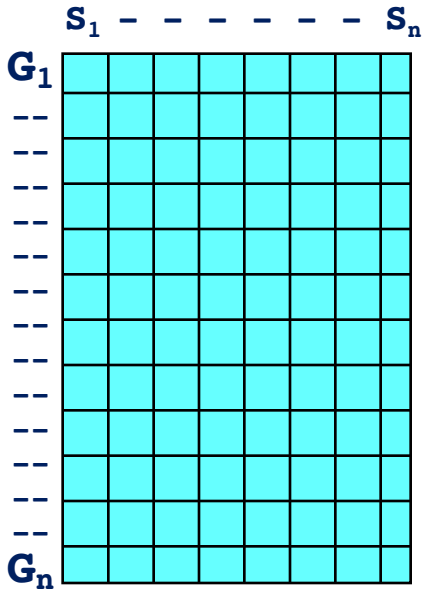
Matrix Decomposition: Global v Local



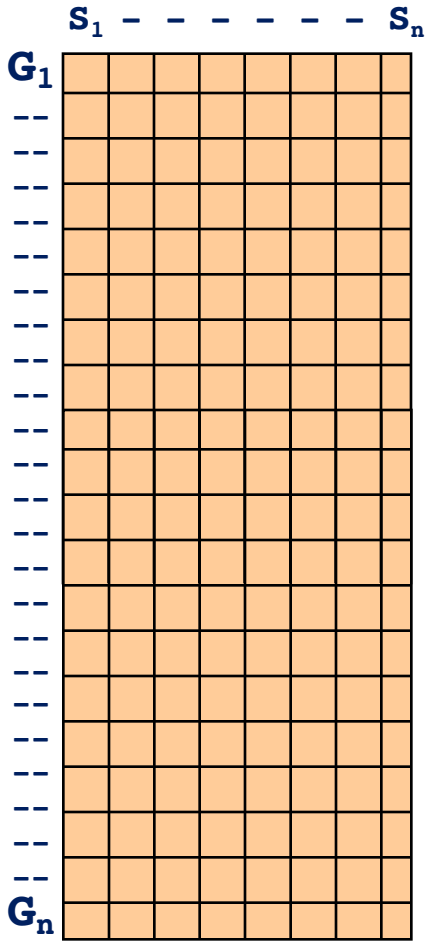
From our review Stein-O'Brien GL, et al., Enter the Matrix: Factorization Uncovers Knowledge from Omics. 2018 *Trends in Genetics*

Integrating >1 datasets

Dataset 1

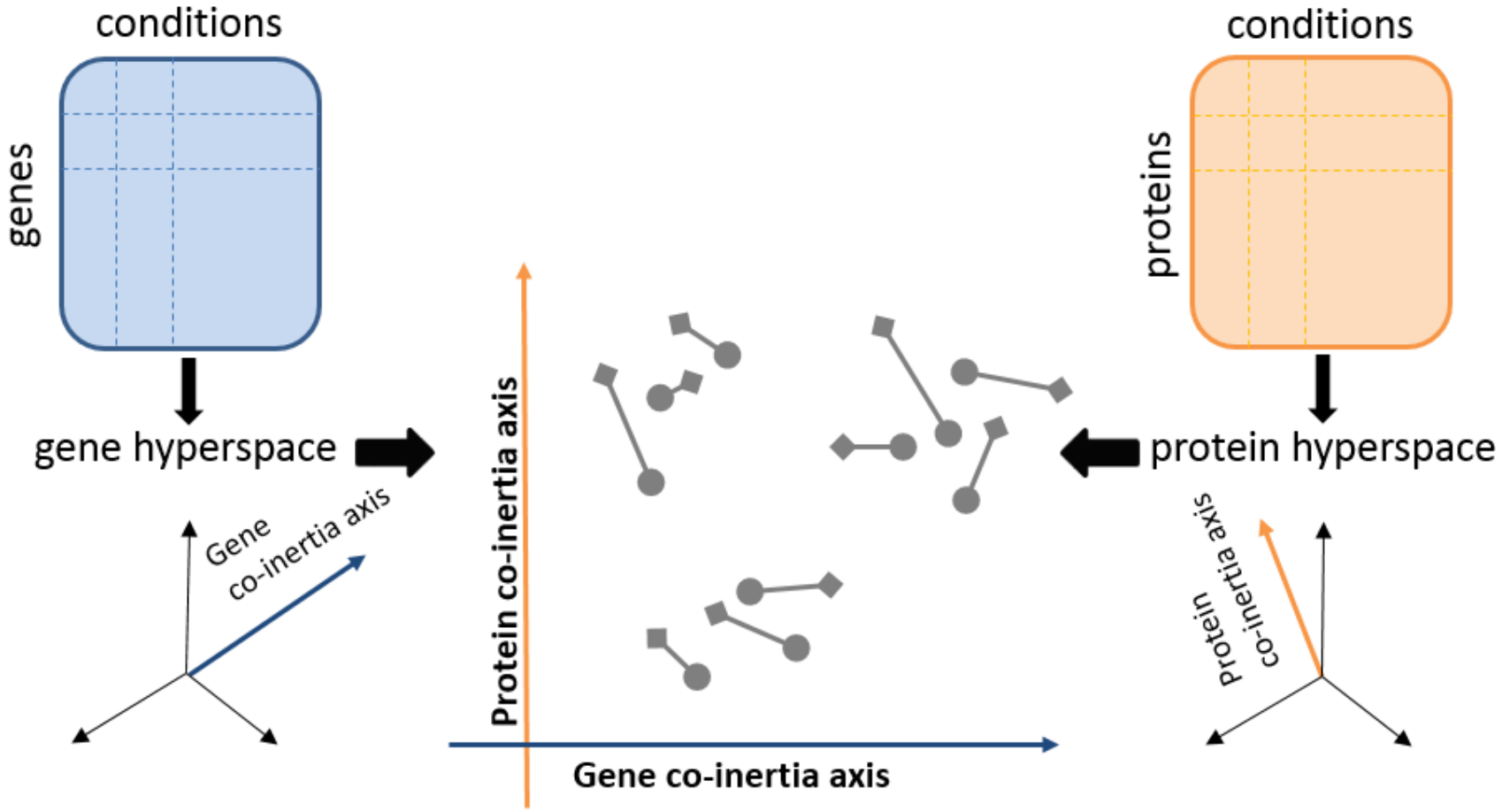


Dataset 2



Datasets have “matching” columns or rows

Integration of features from multiple datasets



Doledec S *et al.*, Freshwater Biology 1994, 31:277-294

Culhane AC *et al.*, BMC Bioinformatics 2003, 4:59-74

Meng et al., BMC Bioinformatics 2014, 15:162

Meng et al., Brief Bioinform. 2016 Jul; 17(4): 628–641.

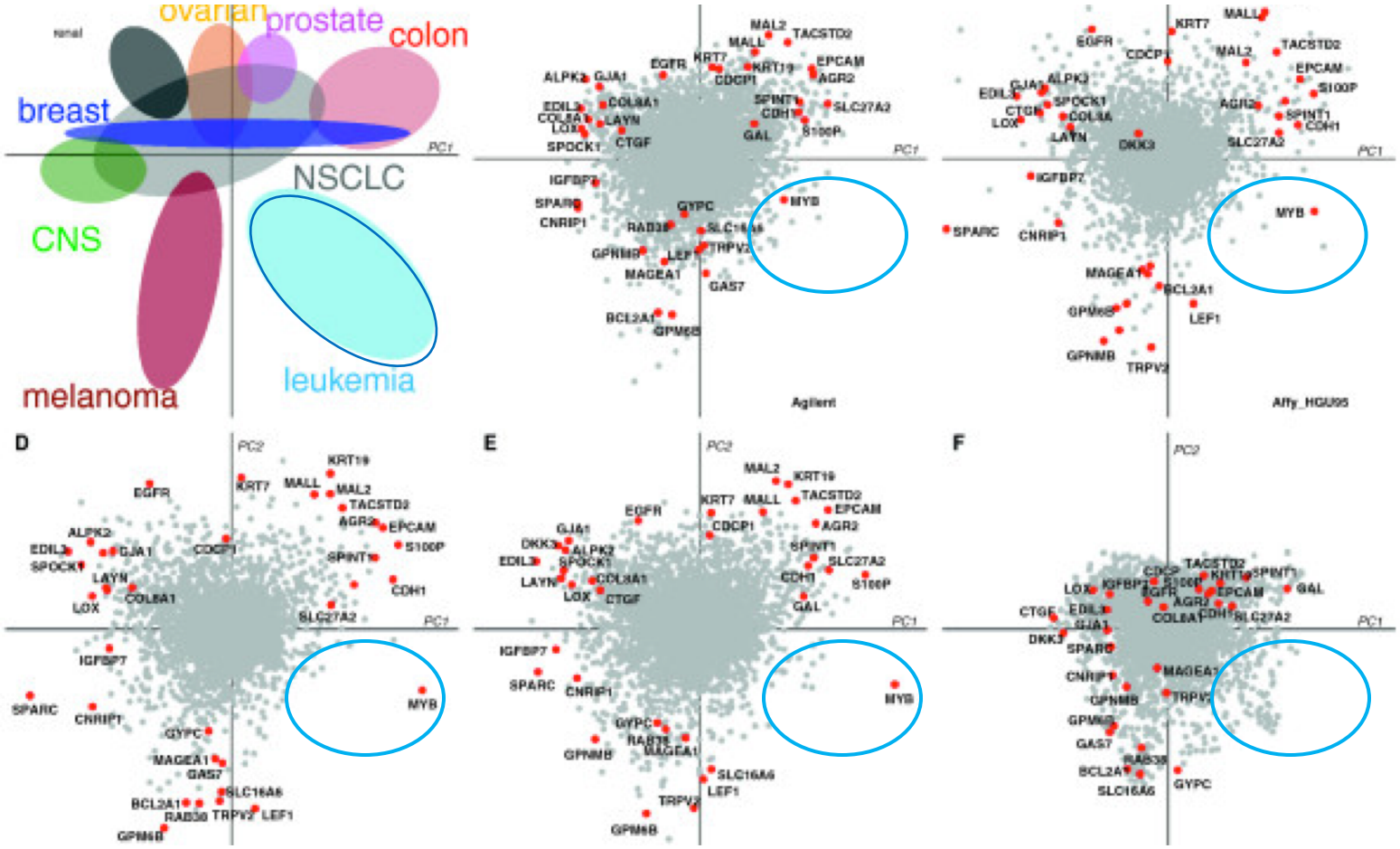
>2 datasets : Tensor data integration

Table 4. Dimension reduction methods for multiple (more than two) data sets

Method	Description	Feature selection	Matched cases	R Function {package}
MCIA	Multiple coinertia analysis	No	No	mcia{omicade4}, mcoa{ade4}
gCCA	Generalized CCA	No	No	regCCA{dmt}
rGCCA	Regularized generalized CCA	No	No	regCCA{dmt} rgcca{rgcca} wrapper.rgcca{mixOmics}
sGCCA	Sparse generalized canonical correlation analysis	Yes	No	sgcca{rgcca} wrapper.sgcca{mixOmics}
STATIS	Structuration des Tableaux á Trois Indices de la Statistique (STATIS). Family of methods which include X-statis	No	No	statis{ade4}
CANDECOMP/ PARAFAC / Tucker3	Higher order generalizations of SVD and PCA. Require matched variables and cases.	No	Yes	CP{ThreeWay}, T3{ThreeWay}, PCAn{PTaK}, CANDPARA{PTaK}
PTA	Partial triadic analysis	No	Yes	pta{ade4},
statico	Statis and CIA (find structure between two pairs of K-tables)	No	No	statico{ade4}

Meng & Zeleznik et al., (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics, 17(4), 2016, 628–641

Find correlated structure in features across Datasets



Chen Meng

Meng et al., BMC Bioinformatics 2014, 15:162

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Package:mogsa

Multiple Factor Analysis, Multiple Coinertia Analysis, Consensus PCA

```
moa(lapply(se, exprs), proc.row = "center_ssq1",
w.data = "inertia", statis = TRUE)  #MCIA
```

MFA statis=FALSE (the default setting)..

Scaling, weighting of datasets

Implements STATIS, MFA, MCIA, CCA for K-table or multi block integration

Preprocessing of rows of datasets;

- none - no preprocessing,
- center - center only,
- center_ssq1 - center and scale (sum of squares values equals 1),
- center_ssqN - center and scale (sum of squares values equals the number of columns),
- center_ssqNm1 - center and scale (sum of squares values equals the number of columns - 1)

weights of each separate dataset,

- uniform - no weighting
- lambda1 - weighted by the reverse of the first eigenvalue of each individual dataset
- inertia - weighted by the reverse of the total inertia.

weight datasets closer to the overall structure

statis – FALSE

Parallel (permutation) based selection of components

- Horn's Parallel Analysis for factor retention
  - https://www.r-bloggers.com/determining-the-number-of-factors-with-parallel-analysis-in-r/

    library(paran)

- **Edgar Dobriban**
  - https://github.com/dobriban/DPA

J. R. Statist. Soc. B (2019)
81, *Part* 1, *pp.* 163–183

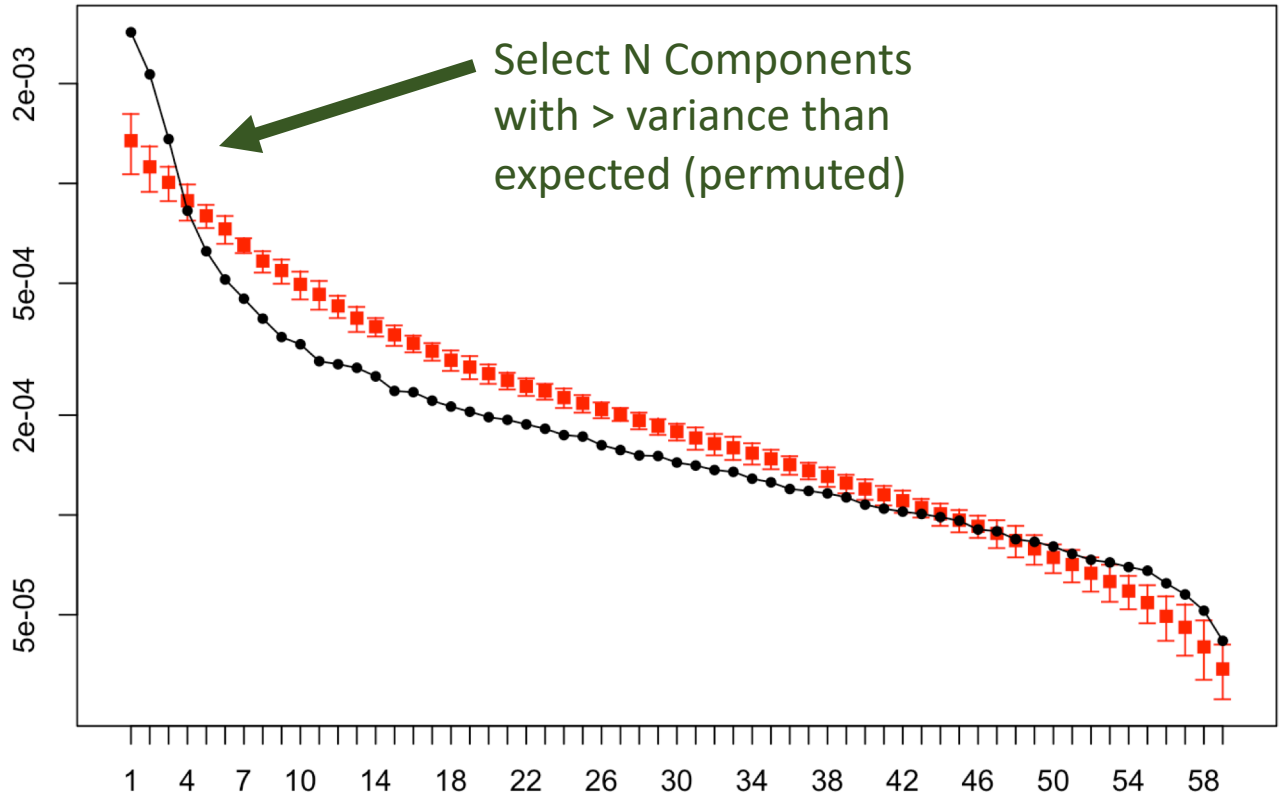
Deterministic parallel analysis: an improved method for selecting factors and principal components

Edgar Dobriban
University of Pennsylvania, Philadelphia, USA
and Art B. Owen
Stanford University, USA

[Received November 2017. Final revision October 2018]

Determine Number of Components (by permutation) representing concordant structure between datasets

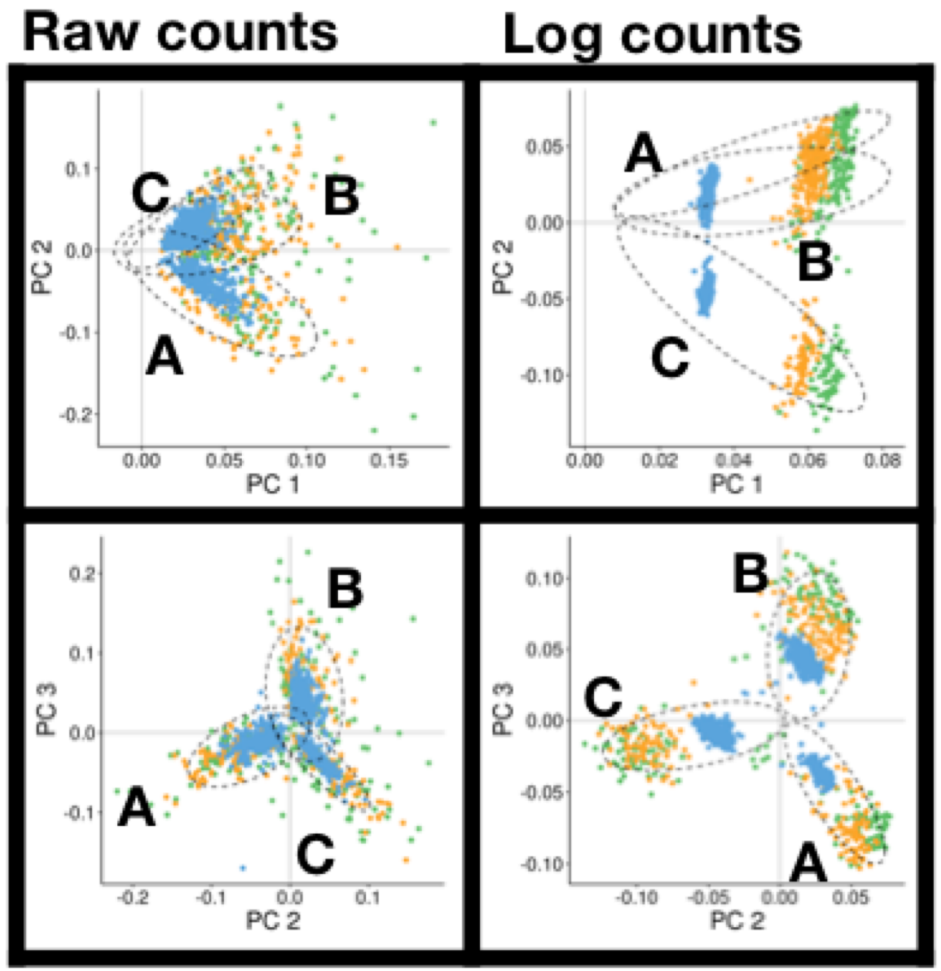
```
bootMoa(
moa = ana,
proc.row = "center_ssq1",
w.data = "inertia",
statis = TRUE,
B = 20,
plot=TRUE)
```



mogsa

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

C. Canonical correlation analysis



Platform
- 10X
- Celseq
- Dropseq

Cell lines

A = HCC827

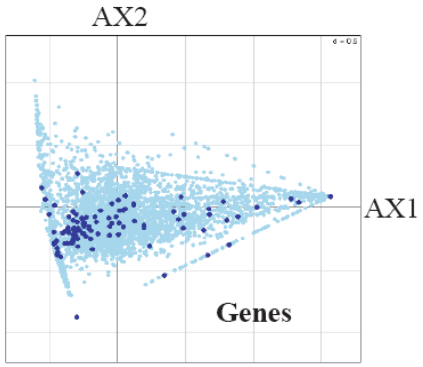
B = H1975

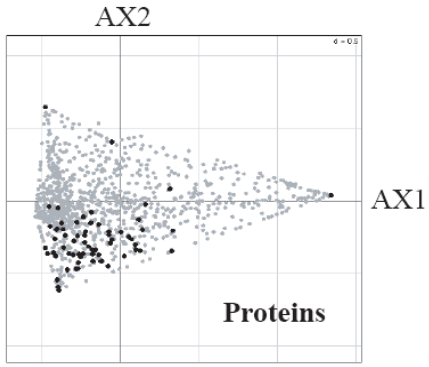
C = H2228

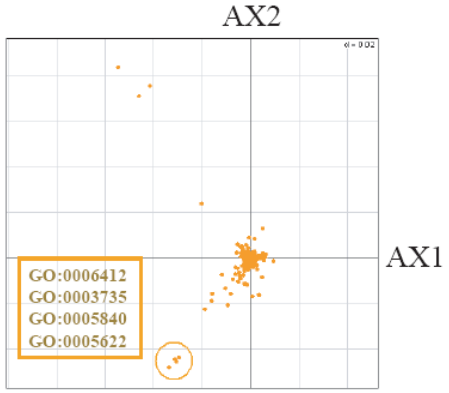
Chan Zuckerberg Initiative

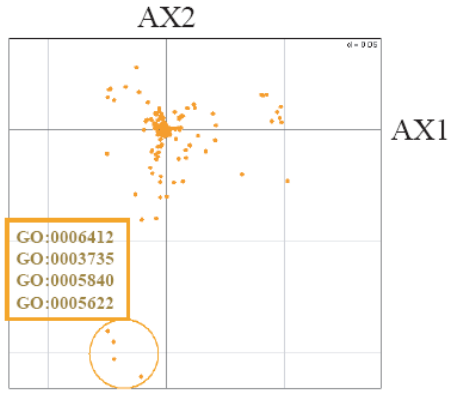
HUMAN CELL ATLAS

Layering Annotation









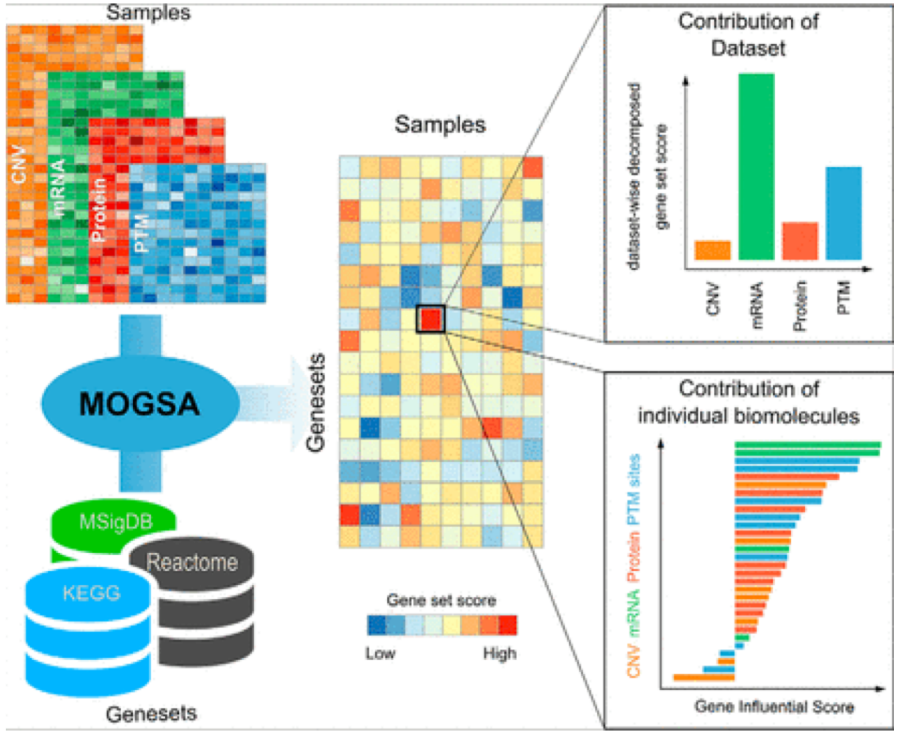
Matrix decomposition of gene expression and proteomics onto same scale

Project Gene Ontology Terms (vector of gene) onto each to get a gene set “score” in each space

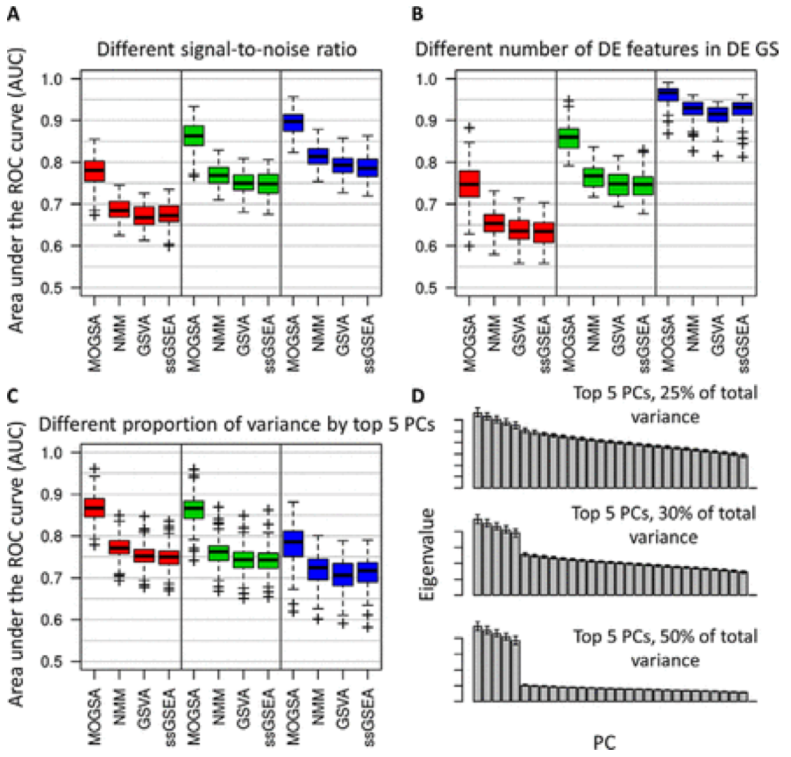
Fagan et al., (2007) A Multivariate Analysis approach to the Integration of Proteomic and Gene Expression Data. *Proteomics.* 7(13):2162-71.
Jeffery et al., (2007) Integrating transcription factor binding site information with gene expression datasets. Bioinformatics 23(3):298-305.

Multiple ‘Omics Gene Set Analysis

Single Sample/Cell Gene Set Scores

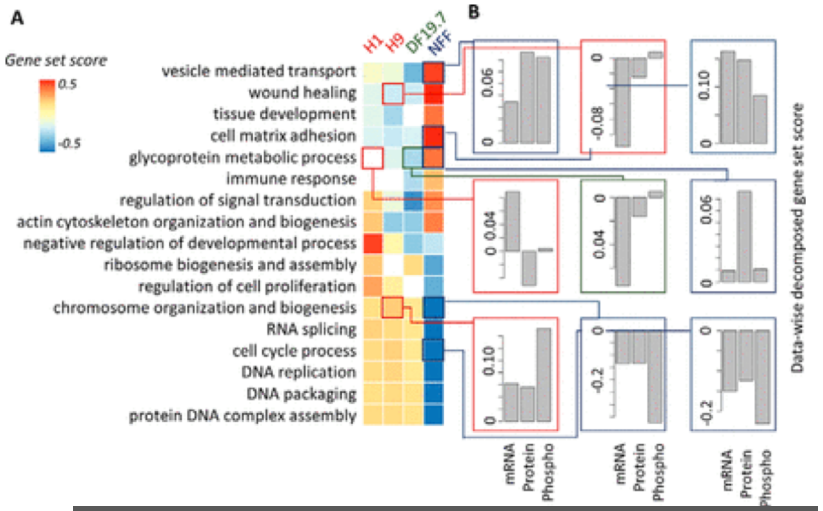


Fast & Performant

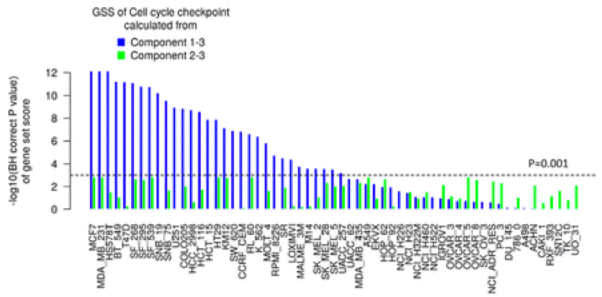


MOGSA Meng C, et al., 2019
MCP **DOI:** 10.1074/mcp.TIR118.001251

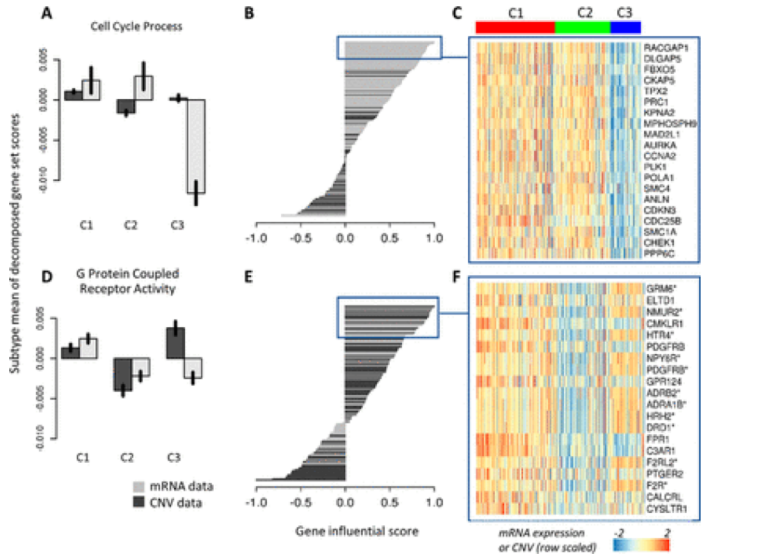
Data weight in Gene Set Scores



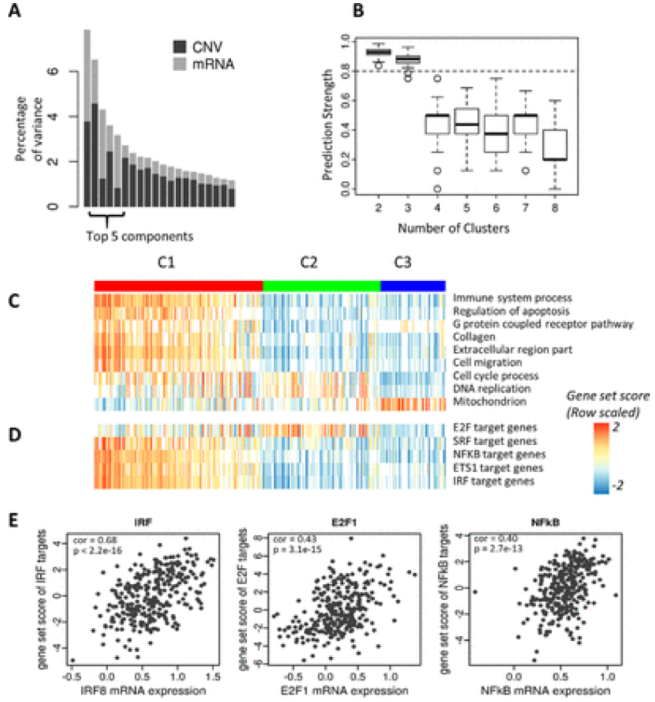
Removal of Batch Effects



Gene weight in Gene Set Scores



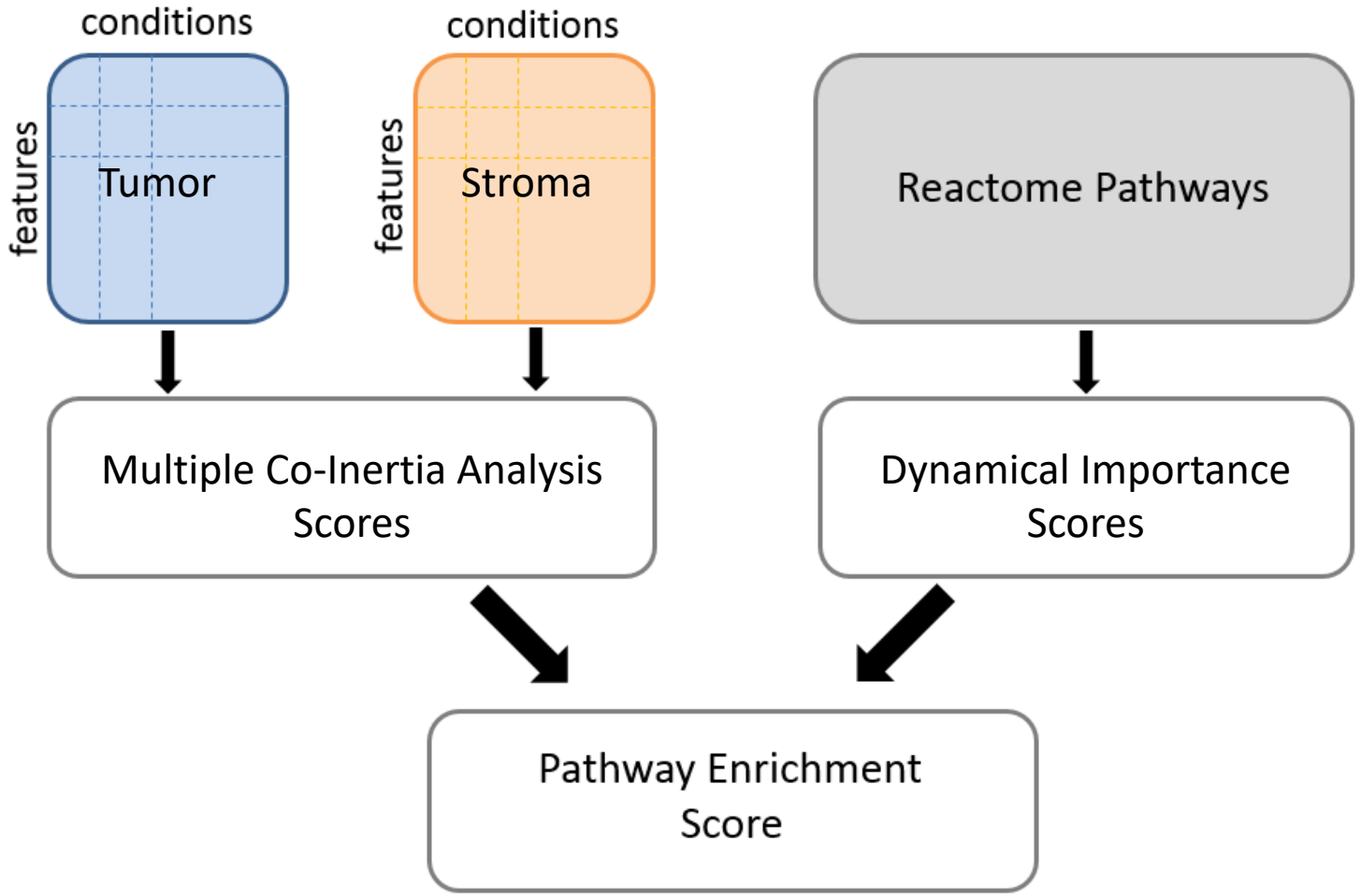
Cluster Discovery



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

MOGSA
Meng C, et al., 2019
MCP **DOI:** 10.1074/mcp.TIR118.001251

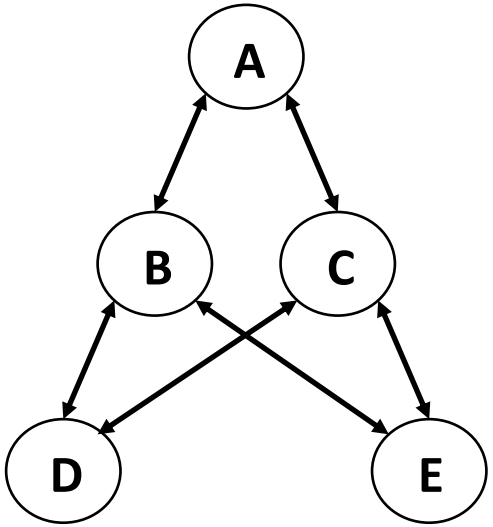
Integrative Pathway Enrichment Analysis (IPEA)



Zeleznik et al., (2017) Trends in Mathematics, vol 7.

Extract feature weights in Network using Dynamical Importance

- Input: adjacency matrix from network of interest
- DI of one node in a network: change of the highest eigenvalue of the adjacency matrix upon removal of that node

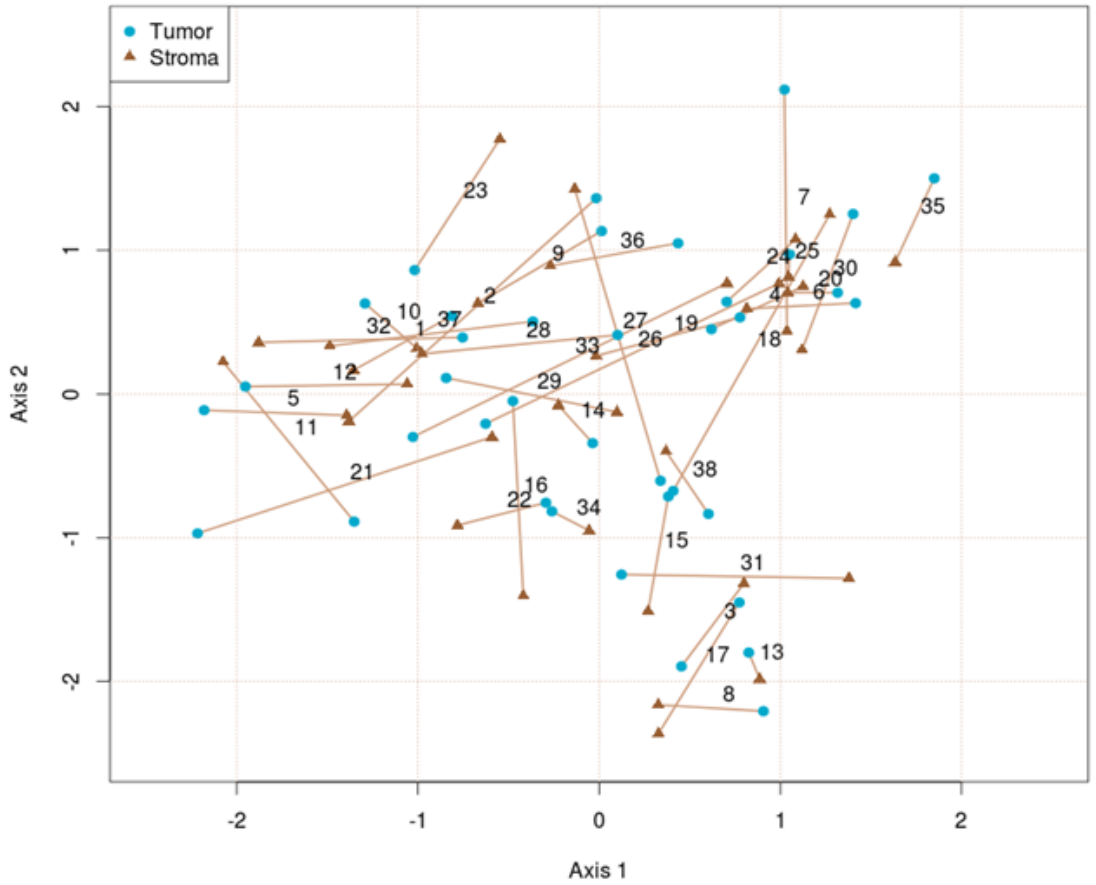


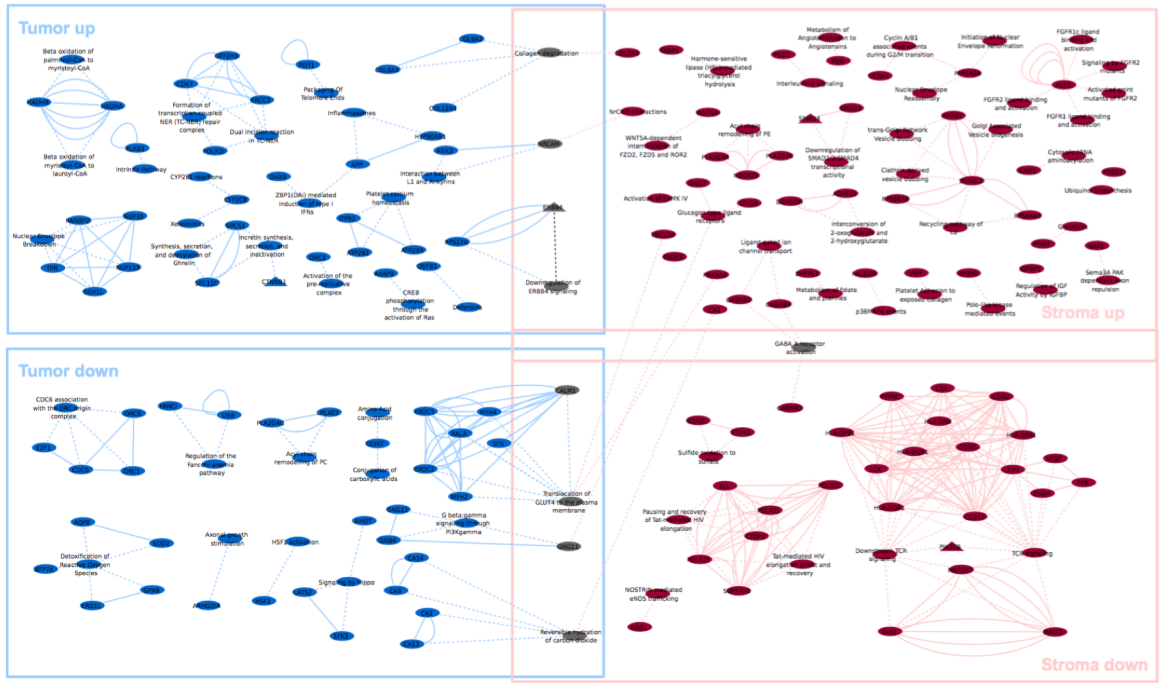
	A	B	C	D	E
A	0	1	1	0	0
B	1	0	0	1	1
C	1	0	0	1	1
D	0	1	1	0	0
E	0	1	1	0	0

	DI score
A	0.17
B	0.25
C	0.25
D	0.17
E	0.17

[F. Restrepo *et al.*, Physics Review Letters 2006, 97:094102]

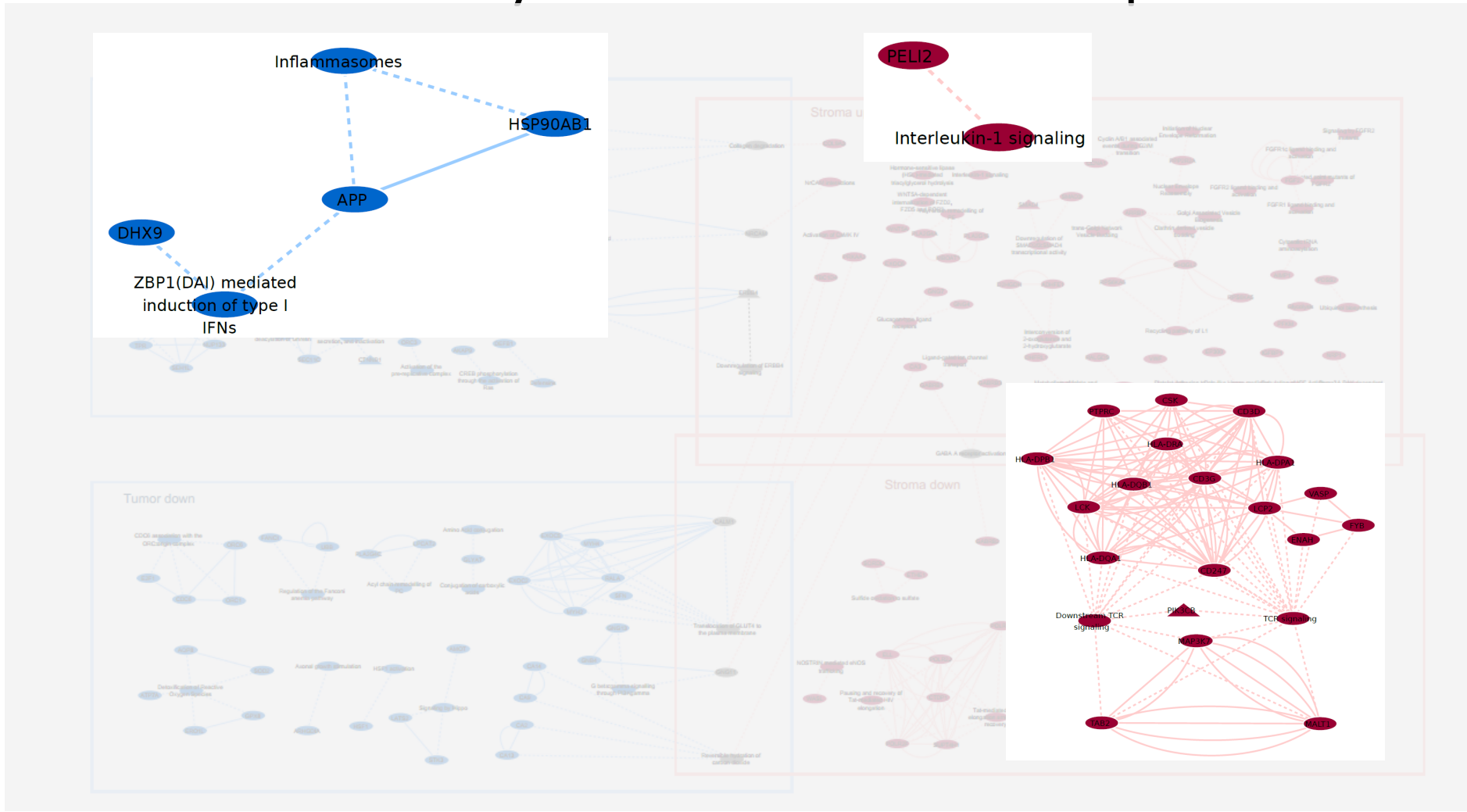
Discovering Cross talk between tumor and stroma





Zeleznik et al., (2017) Trends in Mathematics, vol 7.

Enriched Pathways – Immune Response



Summary

- Many forms of dimension reduction beyond PCA
- Pre-processing of data is critical, some approaches are more robust to zeros, count data, or gradients in data
- Many approaches beyond CCA for joint decomposition of data but consider how datasets are normalized and weighted in the joint decomposition
- Finally annotation or additional information can be projected onto any PCA or dimension reduction to group feature and get average joint score

Final Take Away : EXPLORE your data

#Bioc2020

(Co-Chair)

http://bioc2020.bioconductor.org/

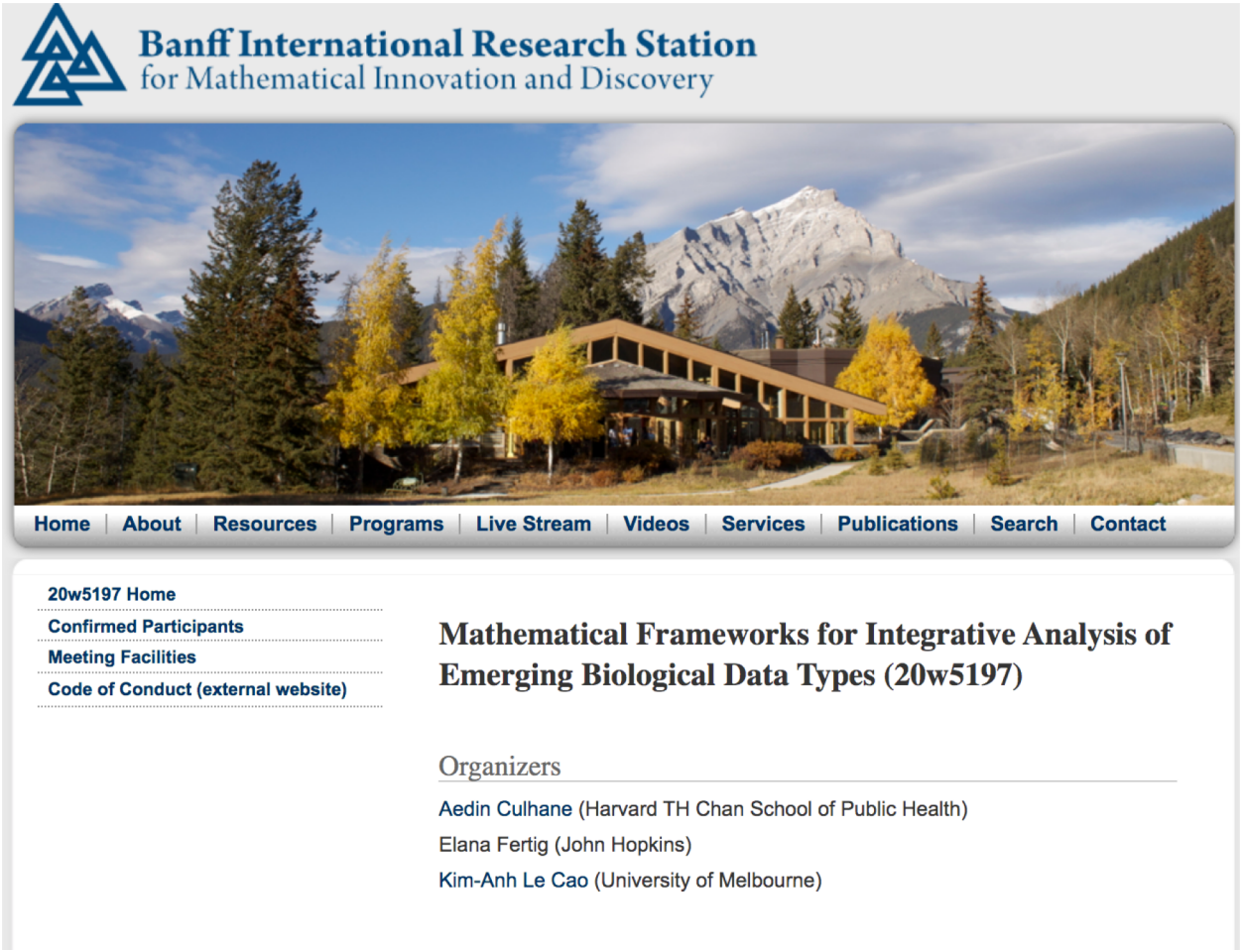
BioC 2020: Where Software and Biology Connect

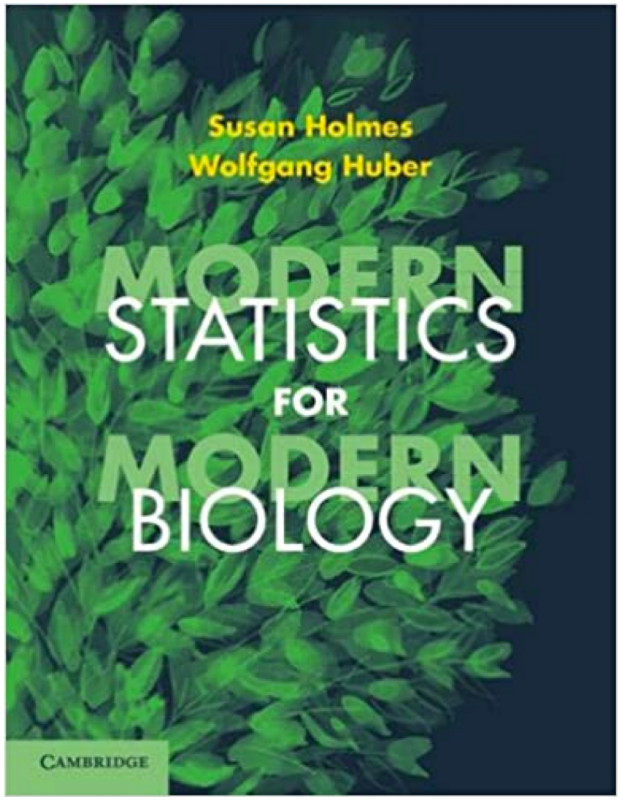
When: July 29 - 31, 2020
What: Community/Developer Day, Main Conference
Where: venue, Boston, USA
Slack: Bioconductor Team (#bioc2020 channel)
Twitter: #bioc2020

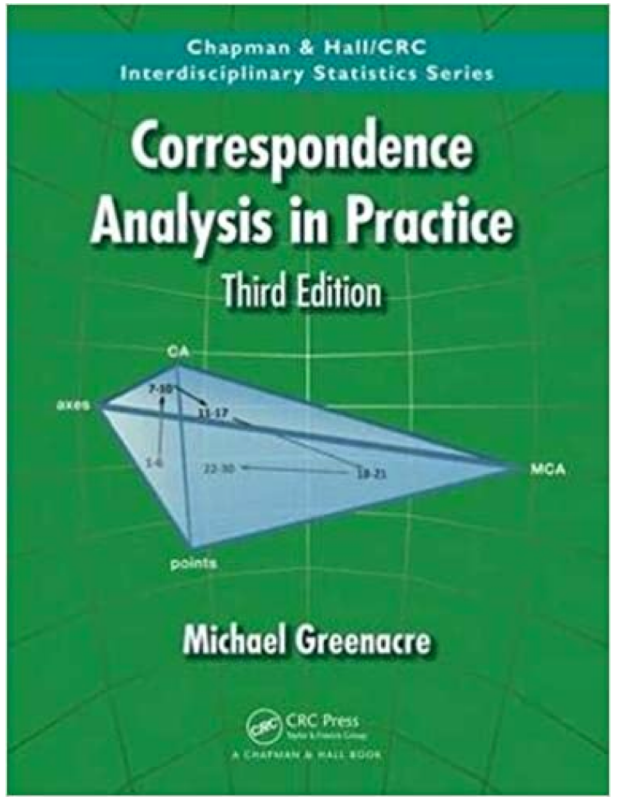
#BIRSBioIntegration

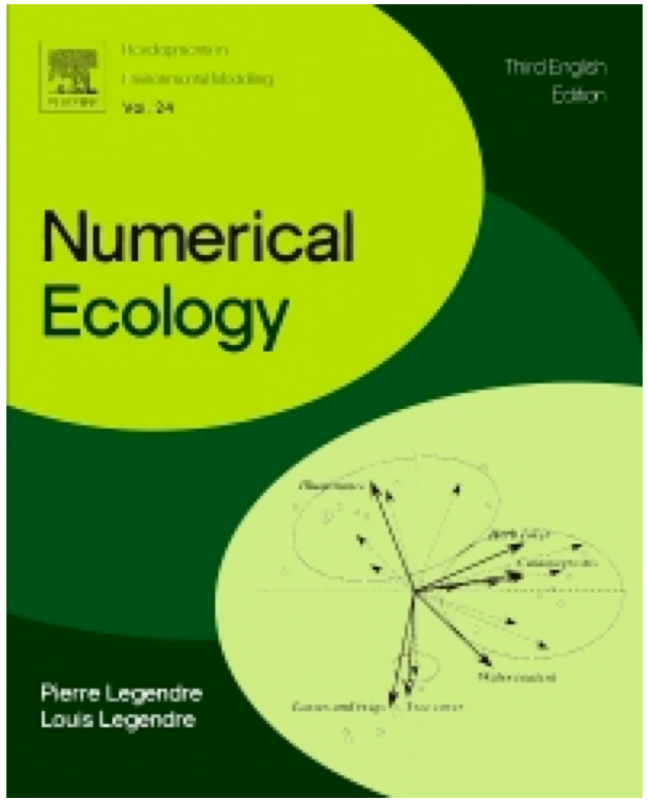
(Co Chair)

https://www.birs.ca/events/2020/5-day-workshops/20w5197









Open source (free) at

http://web.stanford.edu/class/bios221/book /

Useful Reference Books

Acknowledgements

Lauren Hsu

Azfar Basunia

Chen Meng (with Amin, Bernard)

Matthew Schwede

Oana A. Zeleznik

Technische Universitaet Muenchen, Germany

Amin Moghaddas Gholami, Bernard Kuster

Graz University of Technology, Graz, Austria

Gerhard G. Thallinger

TCGA PanCanAtlas Immune Response Working Group

Vésteinn Thorsson

Ilya Shmulevich

Benjamin Vincent

Thanks also to collaborators

Constanine Mitsiades (DFCI)

Levi Waldron (CUNY)

Vince Carey (Channing)

Toni Choueiri (DFCI)

Kathleen Mahoney (BIDMC)

Elana Fertig (John Hopins)

Rafa Irizarry (DFCI)

Benjamin Haibe Kains (Pharmacodb, Univ Toronto)

Mike Birrer (MGH)

David Livingston (DFCI)

David Harrington (DFCI)

John Quackenbush (DFCI)

Chan Zuckerberg Initiative

NIH NATIONAL CANCER INSTITUTE

Congressionally Directed Medical Research Programs CDMRP Department of Defense