Matrix Factorization approaches for data integration. Towards better gene set and pathway analysis

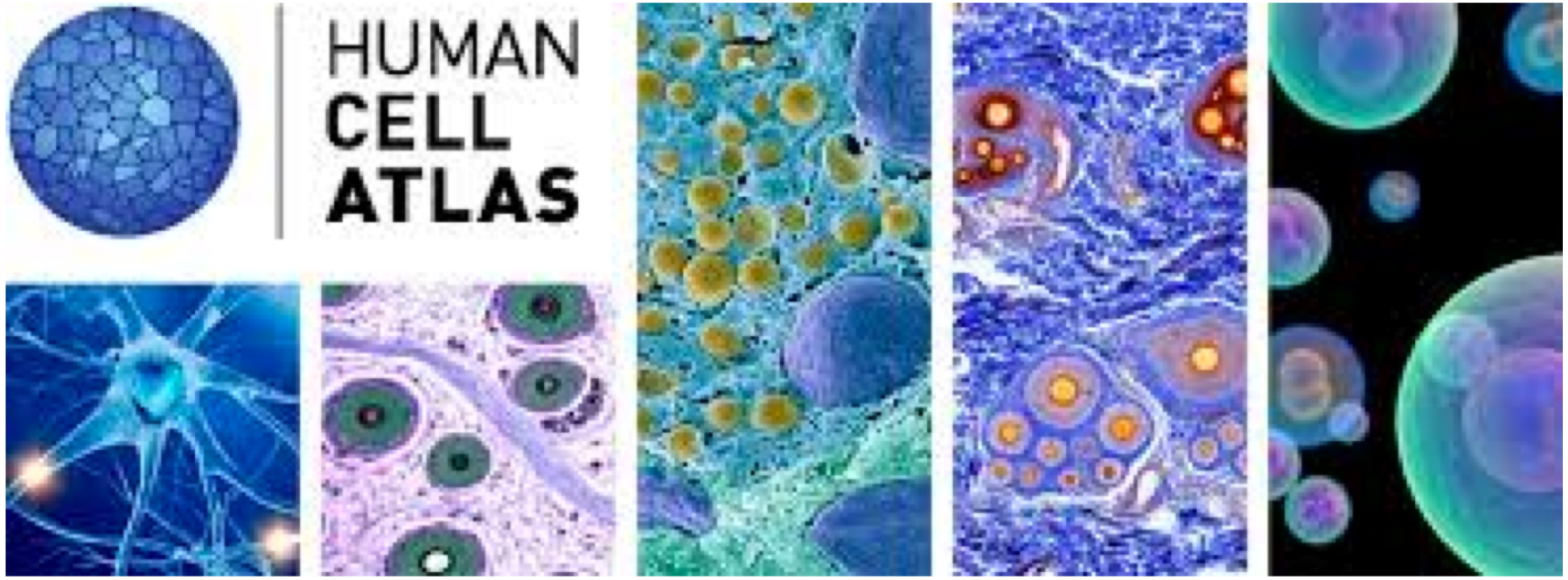
Aedín Culhane PhD

@AedinCulhane

Dana-Farber Cancer Institute

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Harvard T.H. Chan
School of Public Health

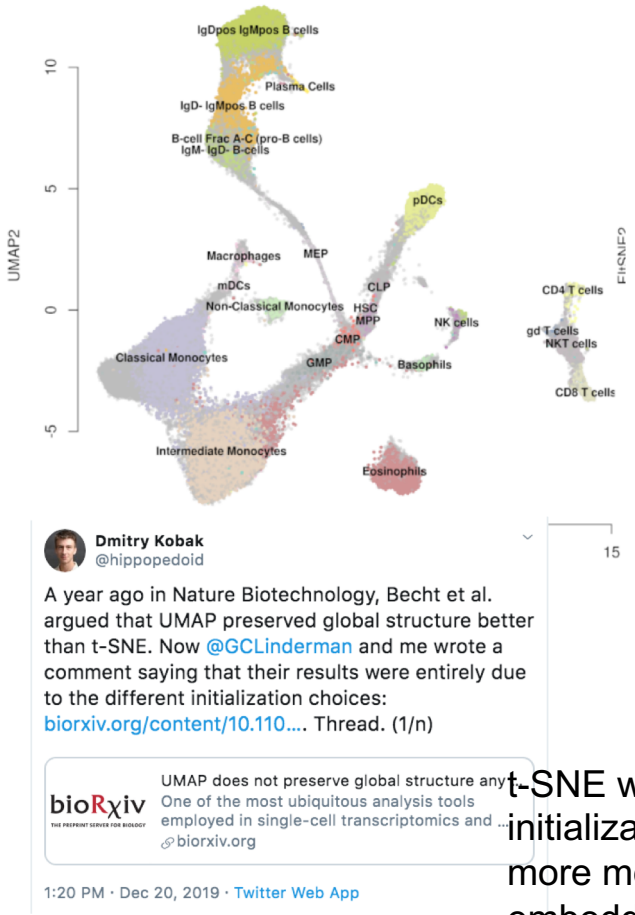


Goal: Create a Human Cell Atlas

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

Chan Zuckerberg Initiative

Dimension reduction is indispensable and is fundamental in almost all scRNAseq analysis



Dmitry Kobak
@hippopedoid

A year ago in Nature Biotechnology, Becht et al. argued that UMAP preserved global structure better than t-SNE. Now @GCLinderman and me wrote a comment saying that their results were entirely due to the different initialization choices: biorxiv.org/content/10.110…. Thread. (1/n)

bioRxiv
UMAP does not preserve global structure any… One of the most ubiquitous analysis tools employed in single-cell transcriptomics and …
biorxiv.org

1:20 PM · Dec 20, 2019 · Twitter Web App

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

Cell clustering, discovery

Input to dimension Reduction t-SNE, UMAP etc

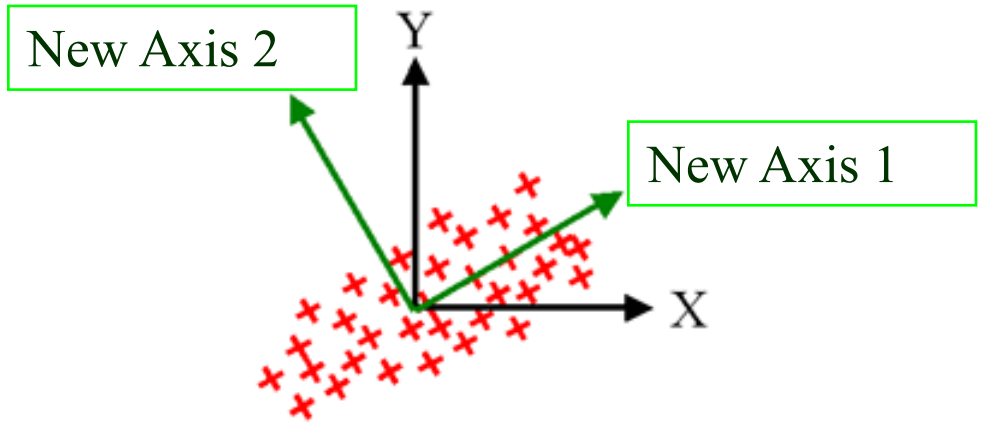
Dimension Reduction eg PCA

Lineage reconstruction

Exploratory Data analysis

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

PCA is most commonly used and is well suited to finding known & unknown (latent) patterns in data



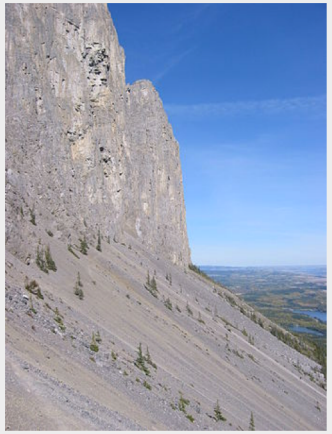
The first new axes will be projected through the data so as to explain the greatest proportion of the variance in the data.

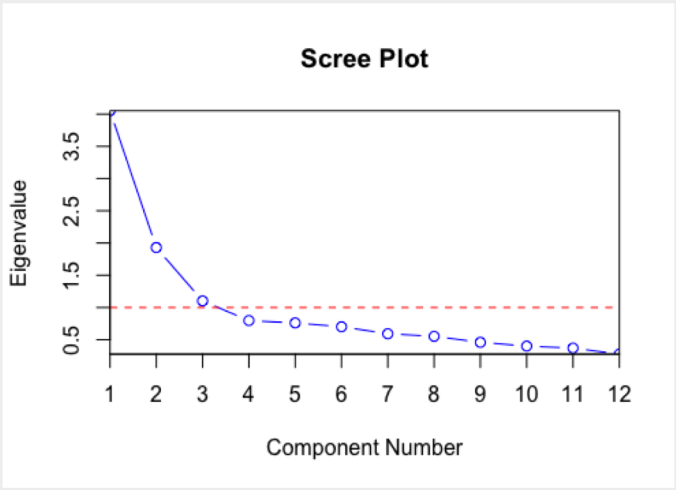
The second new axis will be orthogonal, and will explain the next largest amount of variance

Dimension reduction: PCA $\leftrightarrow$ SVD

- Widely used in statistical and computational science
- 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.

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').

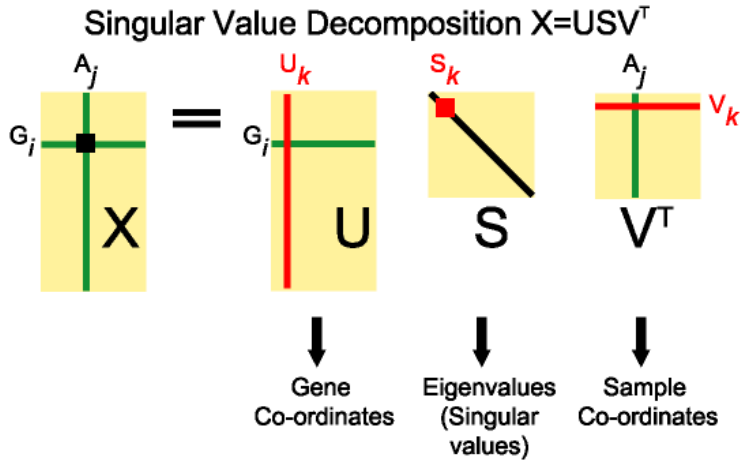
From our Briefings in Bioinformatics review. Meng et al., 2016 https://academic.oup.com/bib/article/17/4/628/2240645

Matrix of raw data

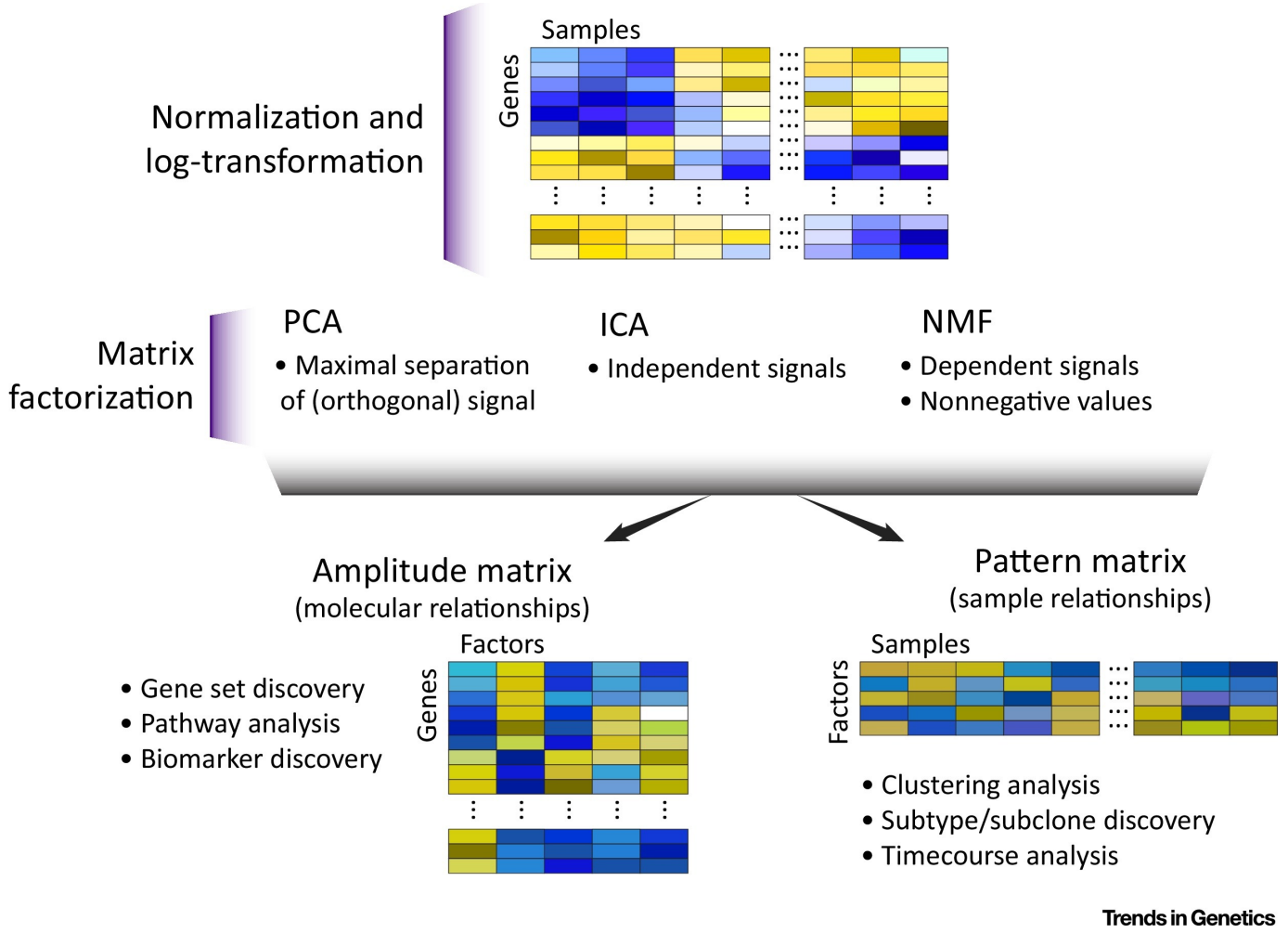
Data Processing

Method	Formula
Scale	$x_{i,j}{}^{*} = \frac{x_{i,j}}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure
Center	$x_{i,j}{}^{*} = x_{i,j} - column\ mean$
Standardize	$x_{i,j}{}^{*} = \frac{x_{i,j} - column\ mean}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure
Transform	$x_{i,j}{}^{*} = f(x_{i,j})$ where *f(x)* is the transformation function, for example logarithms are commonly used

Matrix Decomposition

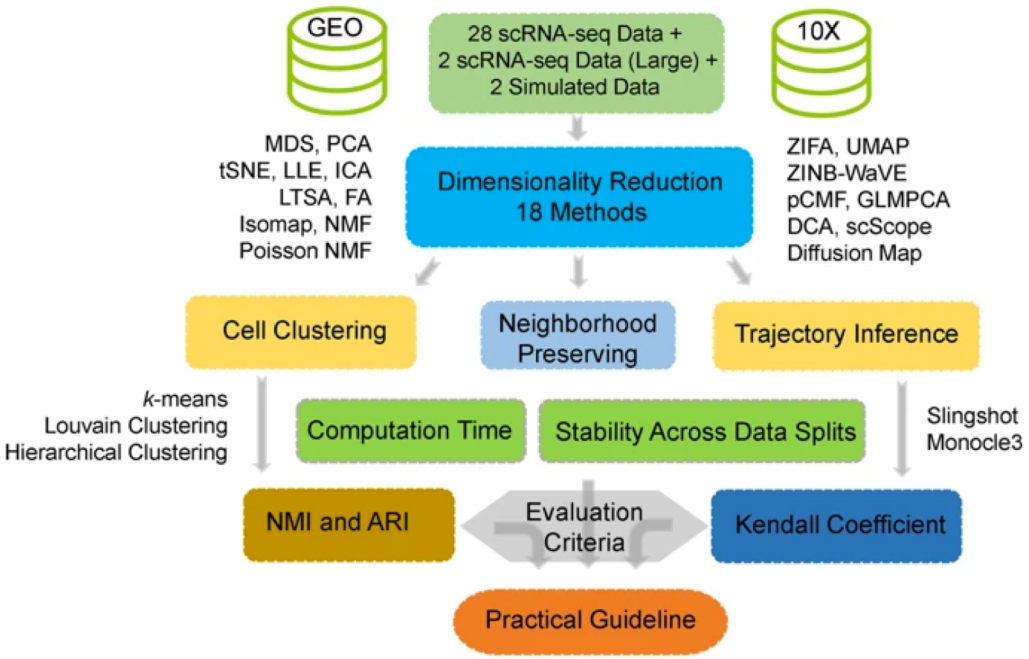


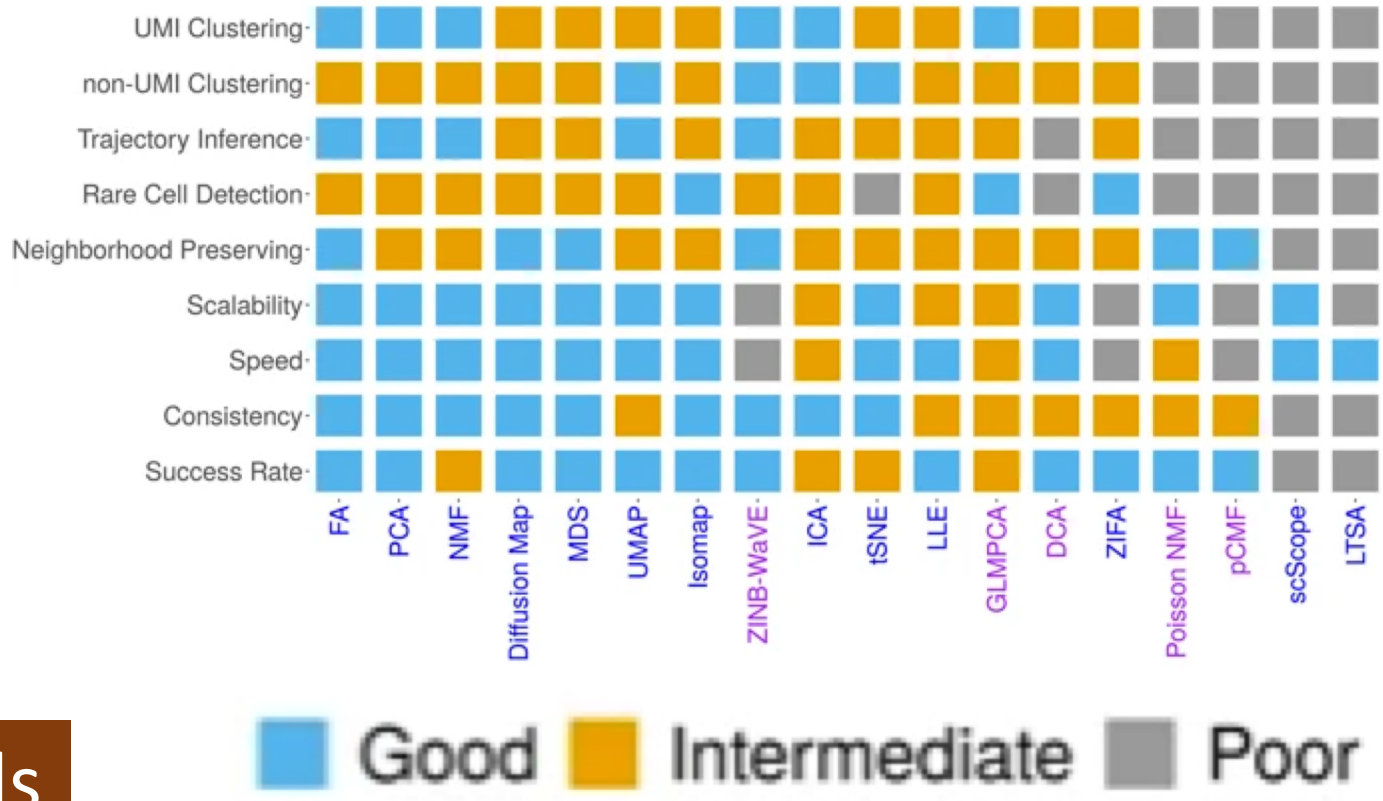
Matrix Decomposition: Global v Local



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

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)

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.

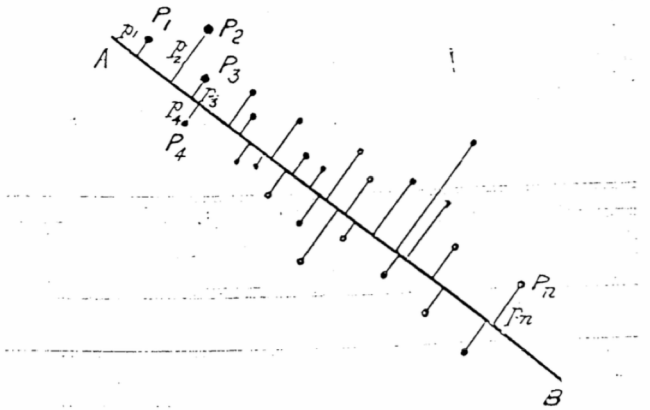
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

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

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.

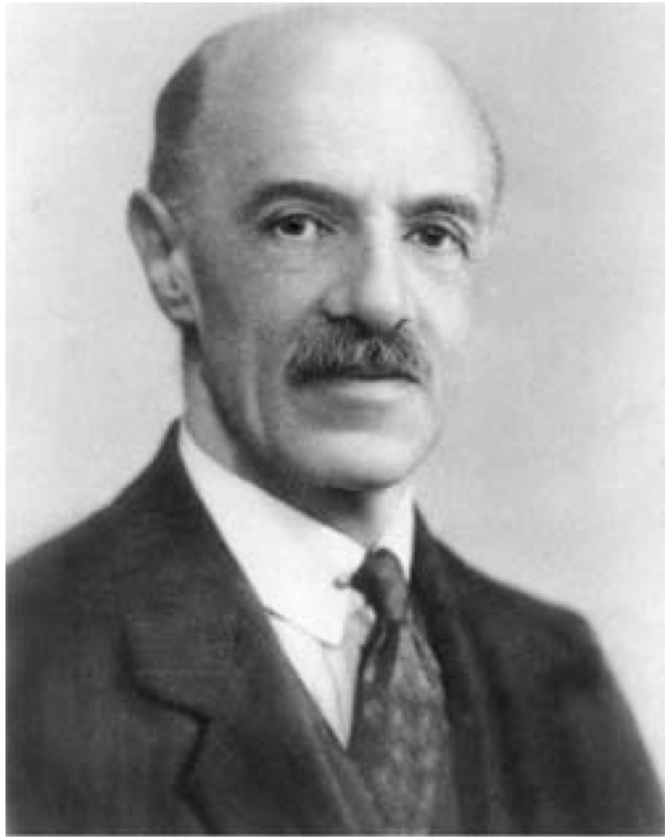


Now clearly $U = S(p^2)$ is the moment of momentum, the second moment of the system of points, supposed equally loaded, about the line A B. But the second moment of a system about a series of parallel lines is always least for the

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

Karl Pearson (1857 -1936)

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



Charles Spearman (1863- 1945)
1904- Factor Analysis

Psychometrician. Spearman was strongly influenced by the work of Francis Galton.

"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. (Family Measurements.)		MENTAL CHARACTERS. (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

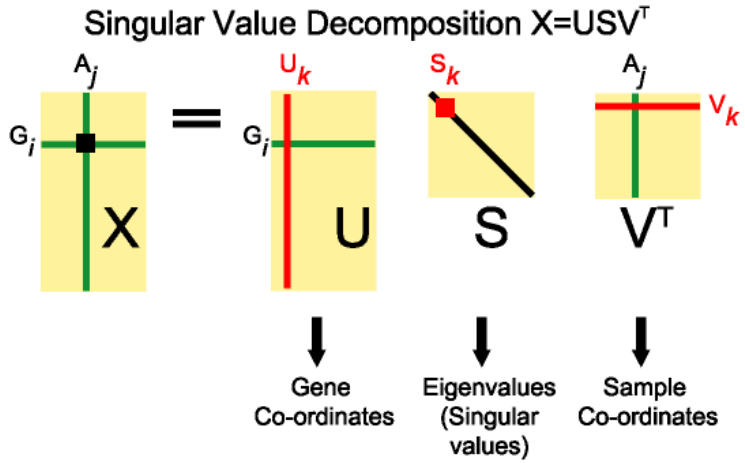
Dealing with the means for physical and mental characters, we are forced to the perfectly definite conclusion, *that the mental characters in man are inherited in precisely the same manner as the physical.*[1] Our mental and moral nature is, quite as much as our physical nature, the outcome of hereditary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

Matrix of raw data

Data Processing

Method	Formula
Scale	$x_{i,j}{}^{*} = \frac{x_{i,j}}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure
Center	$x_{i,j}{}^{*} = x_{i,j} - column\ mean$
Standardize	$x_{i,j}{}^{*} = \frac{x_{i,j} - column\ mean}{scaling\ factor}$ where the *scaling factor* can be a size or data dispersion measure
Transform	$x_{i,j}{}^{*} = f(x_{i,j})$ where *f(x)* is the transformation function, for example logarithms are commonly used

Matrix Decomposition



Sun et al., 2019 provide fully reproducible code on github

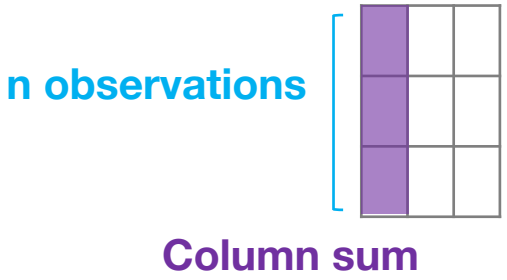
PCA

```
tryCatch({
        ct <- system.time({
        res_pca <- prcomp(norm_counts, center = TRUE, scale. = FALSE)
        res_pca <- res_pca$rotation[, seq_len(num_pc), drop = FALSE]
        })
        # count time
        ct <- c(user.self = ct[["user.self"]], sys.self = ct[["sys.self"]],
    user.child = ct[["user.child"]], sys.child = ct[["sys.child"]],
    elapsed = ct[["elapsed"]])
        list(res = res_pca, ctimes = ct)
},
```

https://github.com/xzhoulab/DRComparison/blob/master/algorithms/call_PCA.R

Covariance-based PCA

$$x^* = x_{original} - \frac{\text{Column sum}}{n}$$



Centering+ SVD

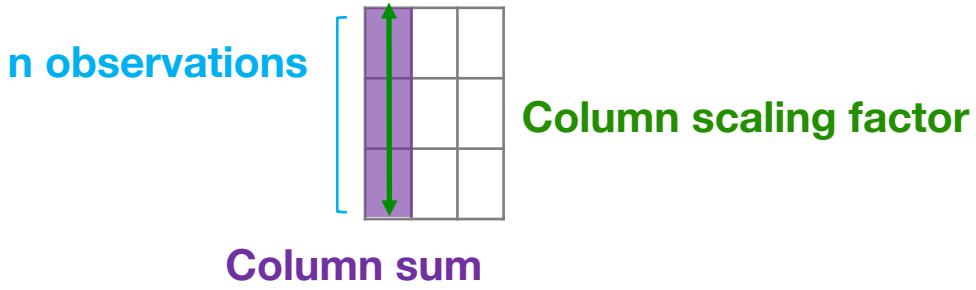
Subtract column mean from each value

+

SVD

Correlation-based PCA

$$x^* = \frac{x_{original} - \frac{\text{Column sum}}{n}}{\text{Column scaling factor}}$$



Centering and scaling +SVD

*Subtract column mean from each value and divide by the column scaling factor. If the scaling factor is the column standard deviation, this is a **Z-score standardization***

+

SVD

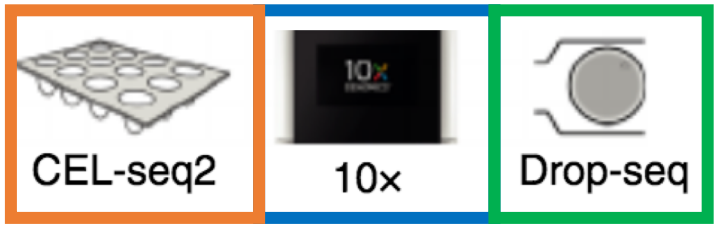
Single Cell RNAseq

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

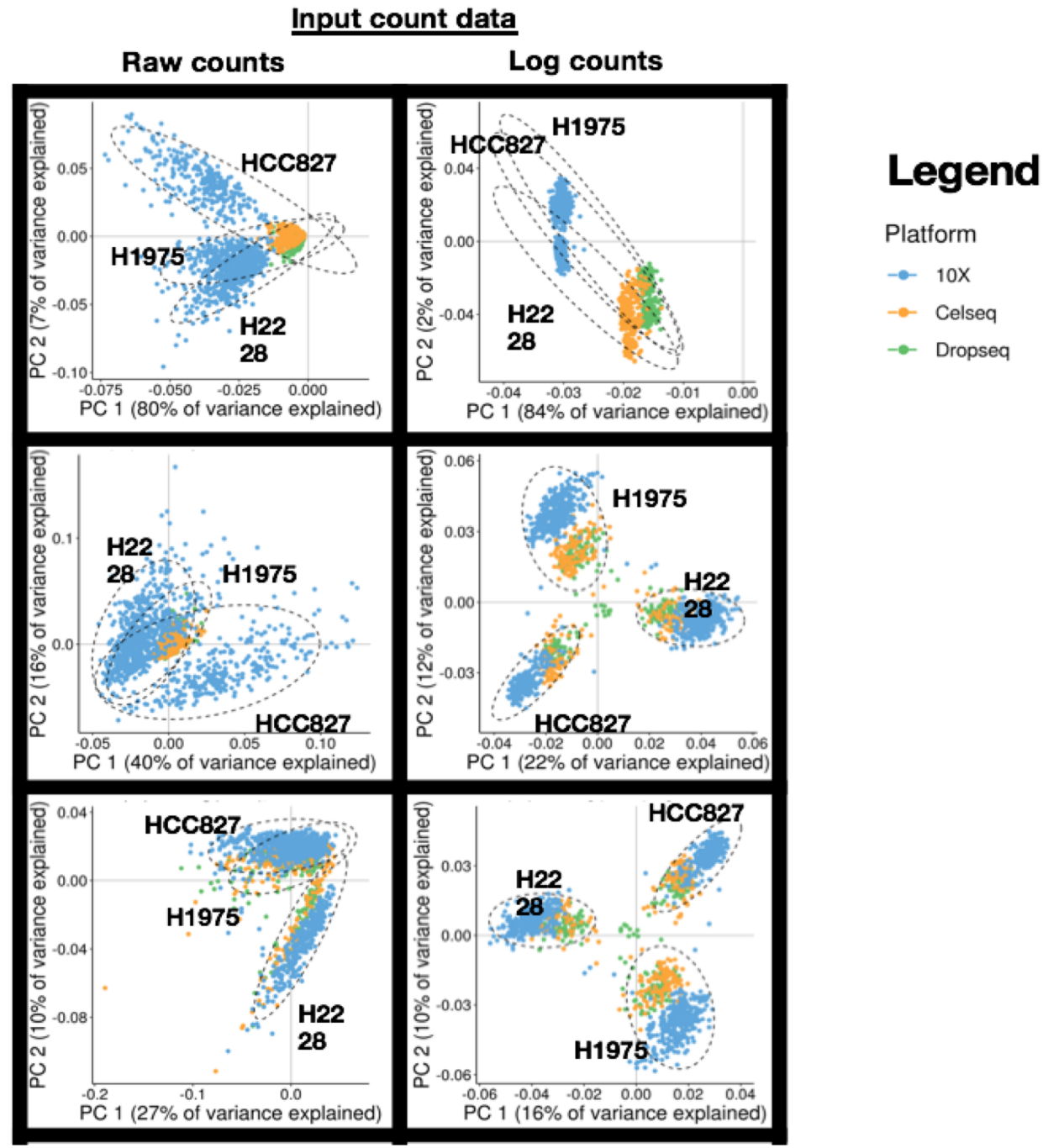
Work by Harvard TH Chan master's student Lauren Hsu

Processing steps impact results

human lung adenocarcinoma cell lines
HCC827
H1975
H2228



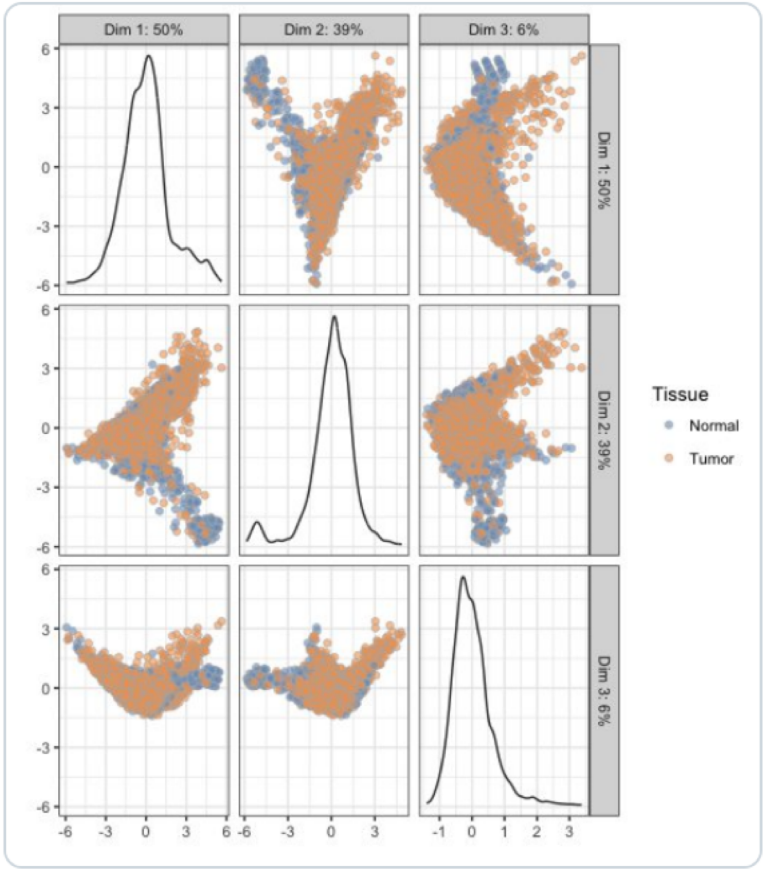
Unpublished



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!

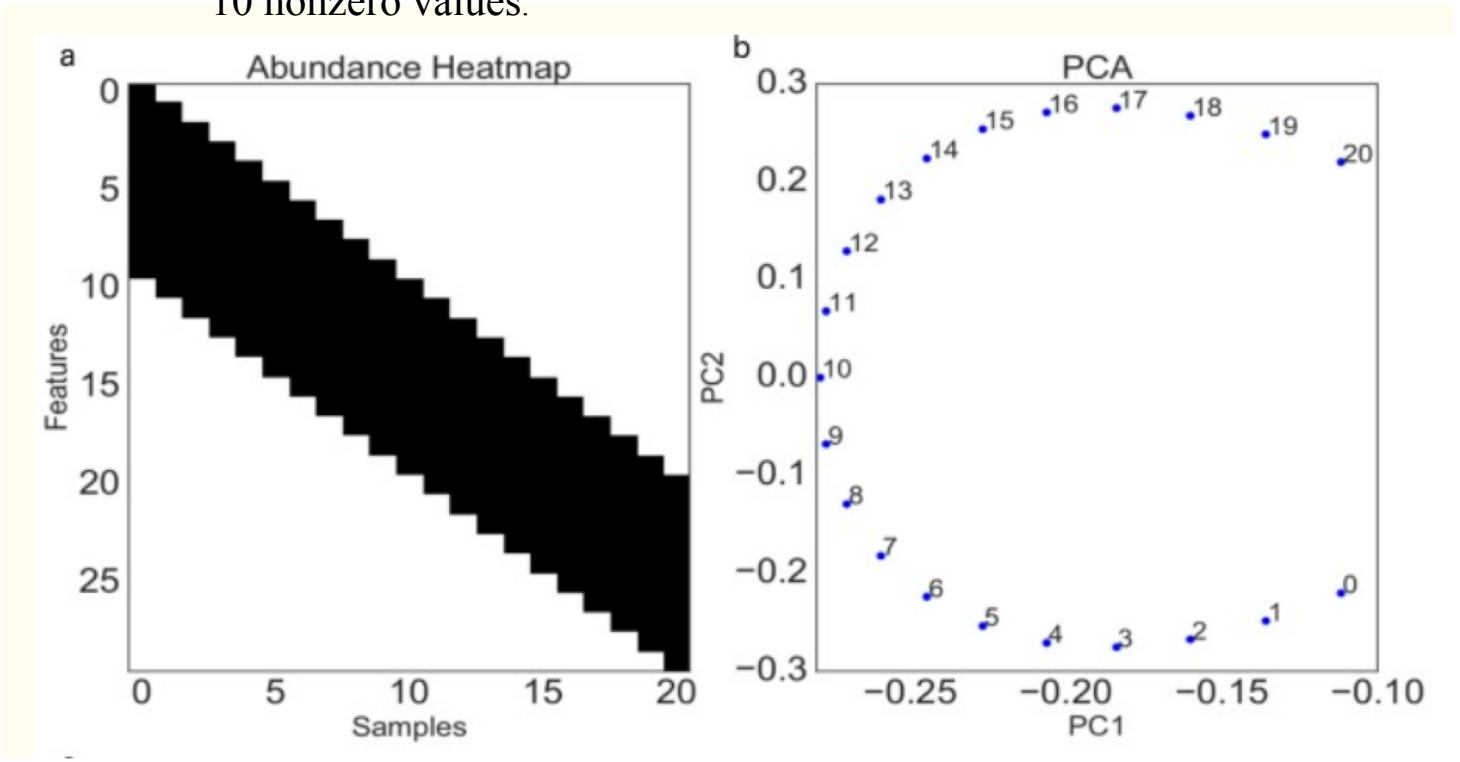
2

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.



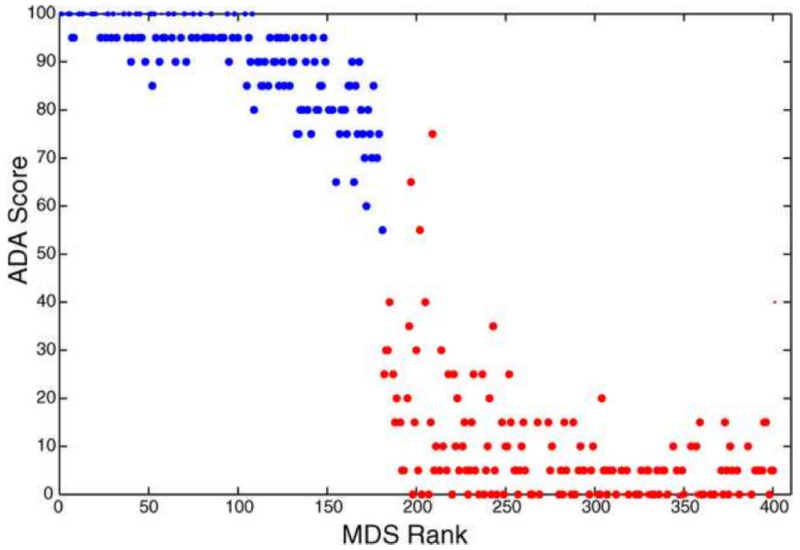
PCA Horseshoes

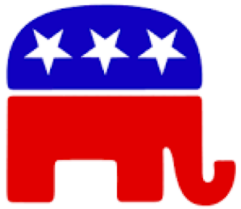
- Arch, horseshoe effect or Guttman effect
- Arise as a consequence of distance metrics that saturate.
- Morton et al., 2017 mSystems. 2(1): e00166-16

Congress PCA Horseshoes

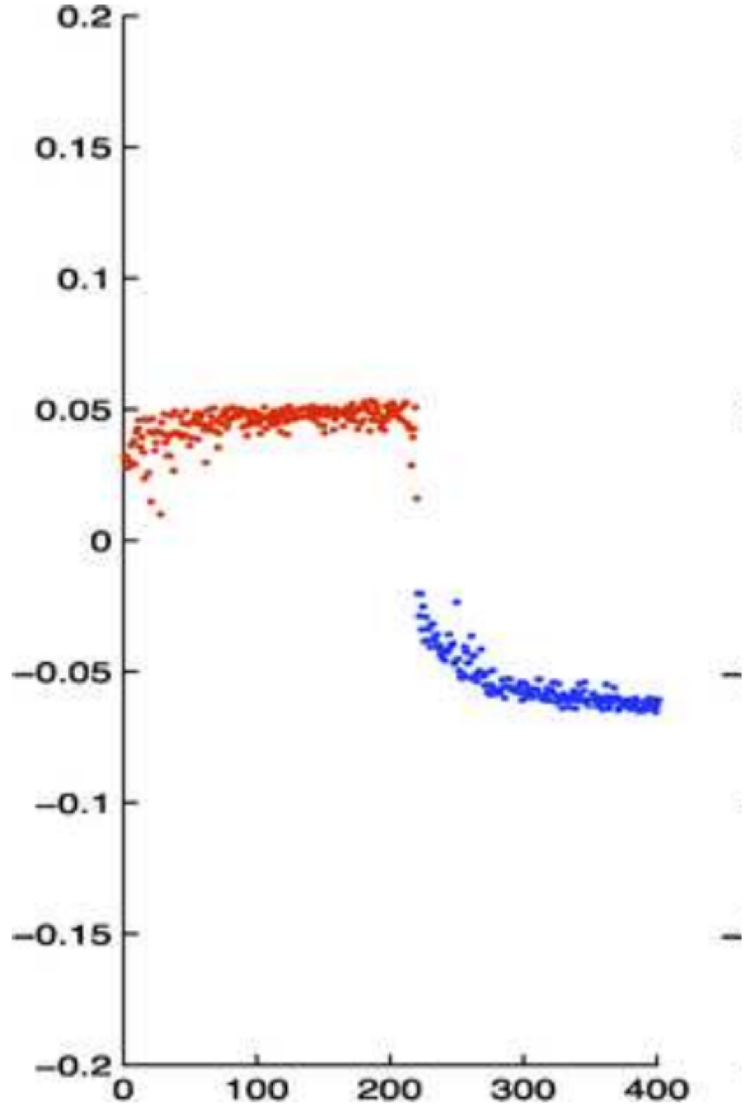
- no natural underlying Euclidean space
- horseshoe or arch indicator of a sequential latent ordering or gradient in the data

Rank of Americans for Democratic Action Liberal Quotient





PCA of “polarized” voting records



Diaconis, Goel, and Holmes 2007

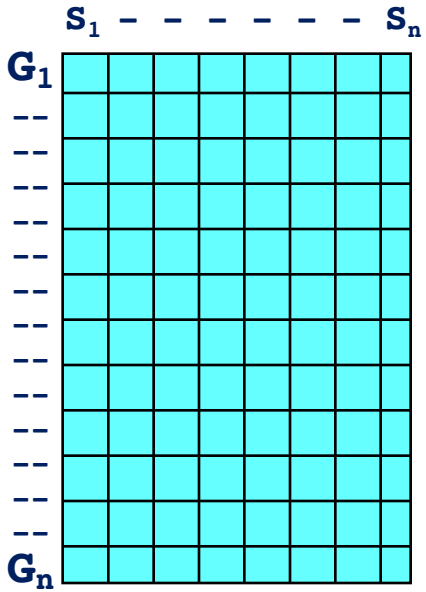
PCA, scRNAseq

- Dimension reduction is fundamental to scRNAseq data analysis
- PCA is the most popular approaches and is included in many algorithm
- Better approaches may exist and may increase performance of analysis

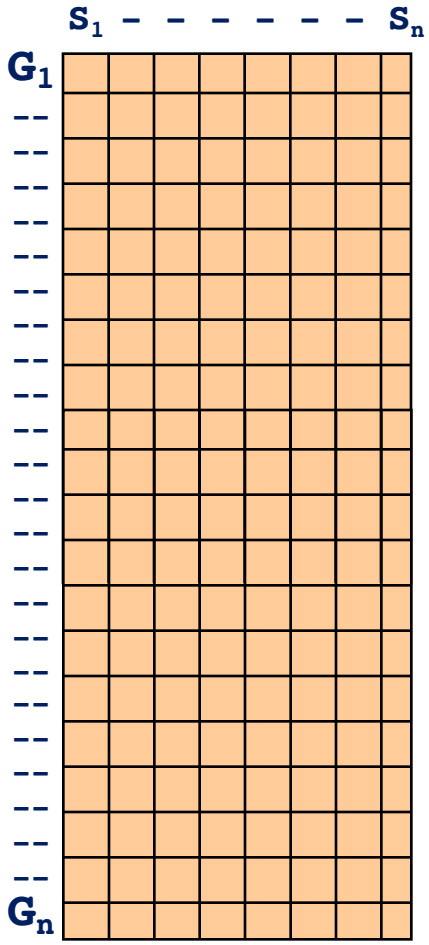
- Where are we going??? Our goals?

Aligning two datasets

Dataset 1

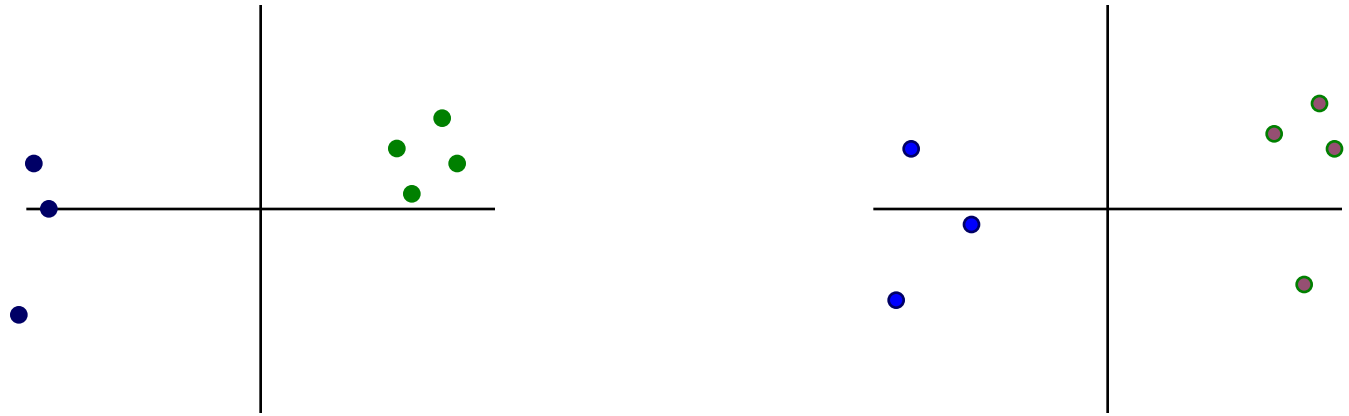


Dataset 2



Datasets have “matching” columns or rows

“Aligning” or integrating 2 datasets



Given A and B datasets.
Find 2 matrix decompositions most similar.

Find successive axes (a_i and b_i) such that
CORRELATION (a_i, b_i) is maximum (CCA)
COVARIANCE (a_i, b_i) is maximum (CIA)

Culhane AC, Perriére G, Higgins DG. (2003) *BMC Bioinformatics*. 4(1):59

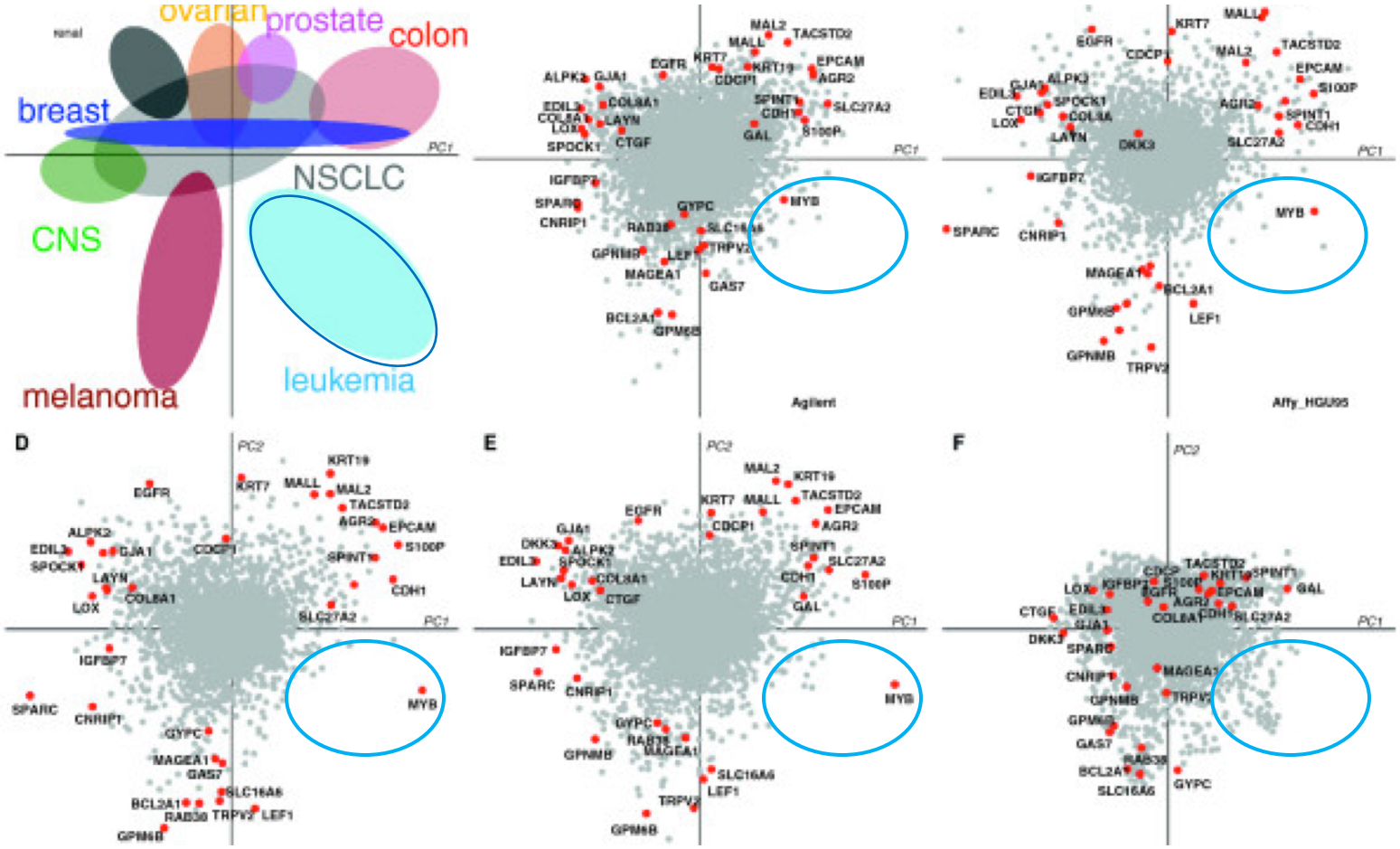
>2 datasets : Tensor data integration

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 Across 5 transcriptomics datasets using MCIA tensor Integration

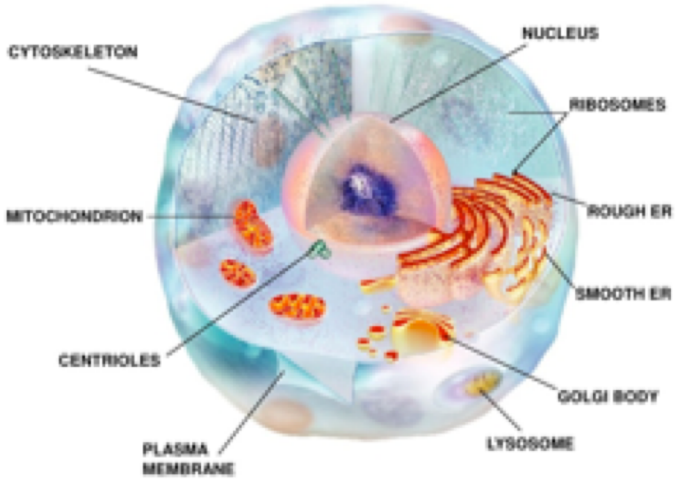


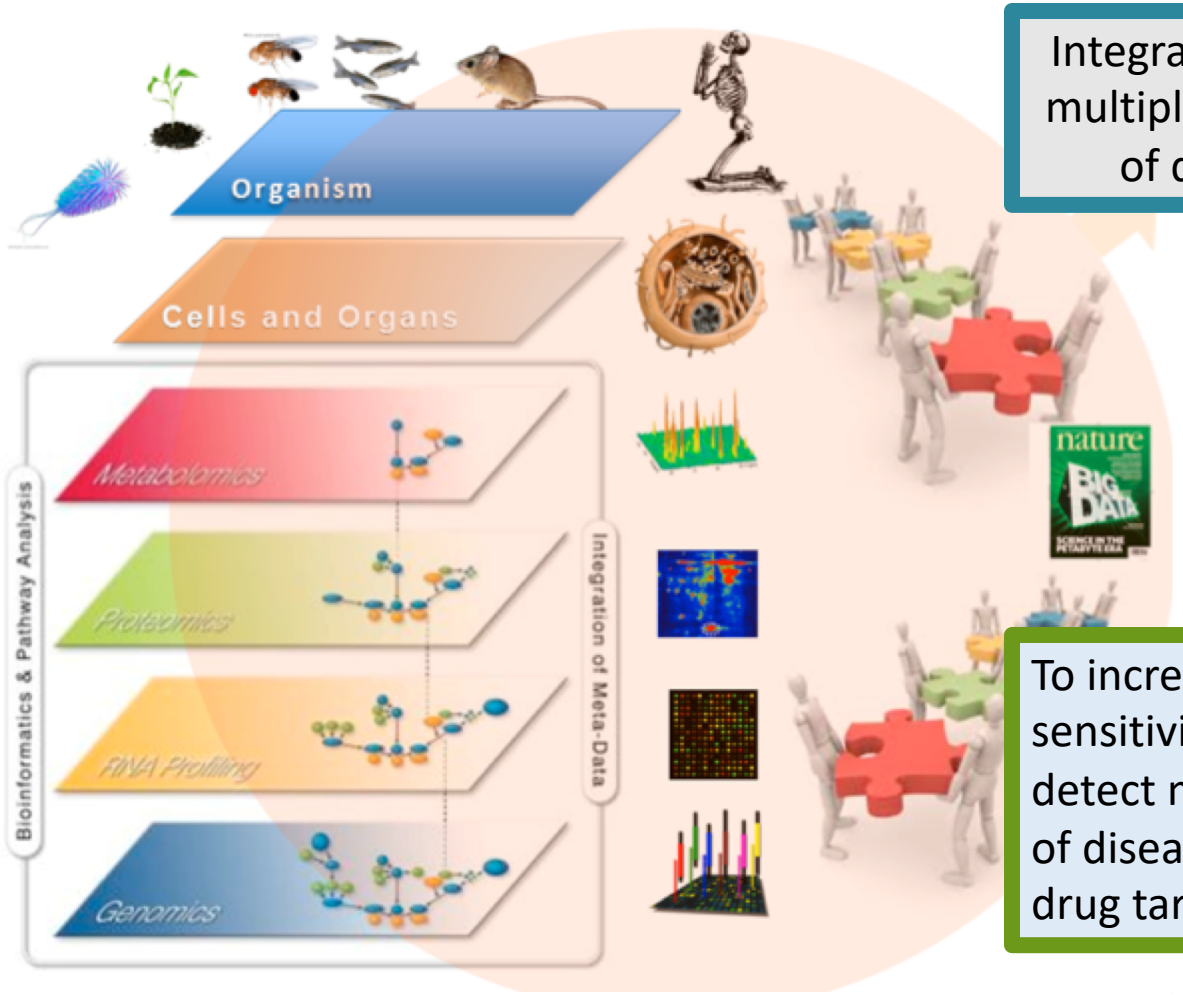
Meng et al., BMC Bioinformatics 2014, 15:162

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Package:mogsa

Finding Pathways in Multi-Omics Data

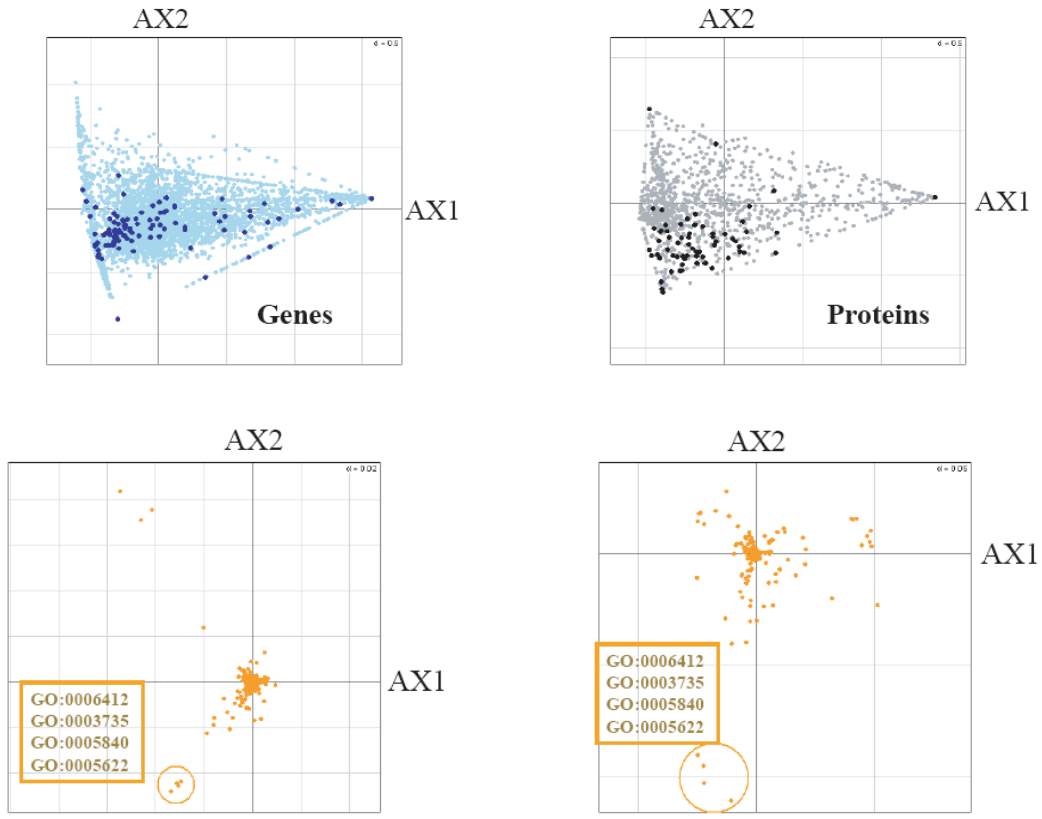




Integrate over multiple layers of data

To increase sensitivity to detect markers of disease and drug targets

Simple approach to generate a gene set scores



Matrix decomposition of gene expression and proteomics onto same scale

Project GO 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.

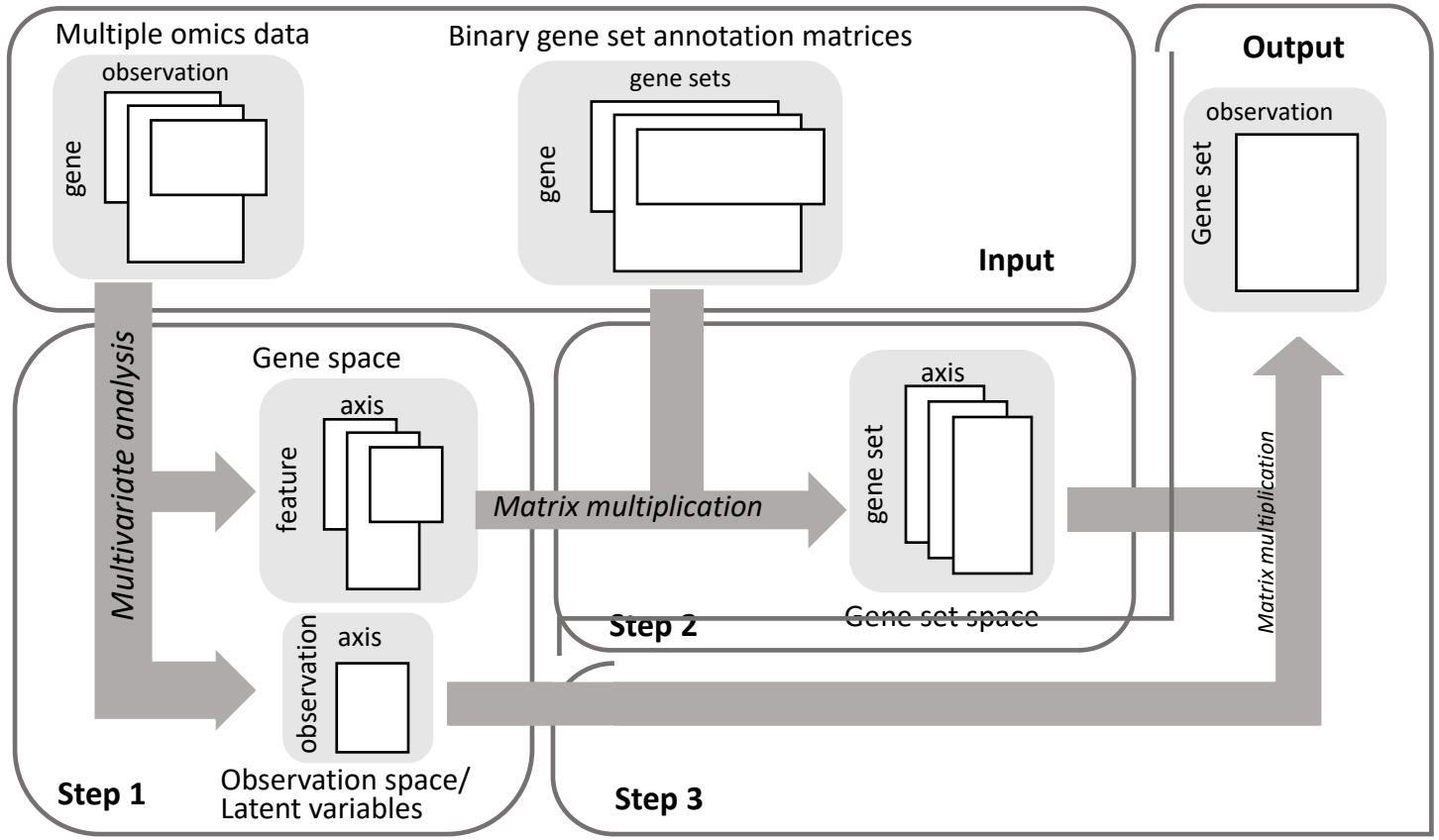
How to integrate & extract pathway scores from multi-omics data?

1. Analyze each dataset, exact list of differentially expressed features, perform Gene Set Analysis, combine P-values (eg R-topper).
    - High specificity, good for “low hanging fruit”
    - May miss subtle signals
    - Requires known covariates (for supervised differential analysis)
2. Map features to a gene identifier (eg Entrez ID), & perform single sample gene set analysis/pathway enrichment
    - Loose information from data with little feature annotation (eg lipids, glyco)
    - Mapping to common identifier
3. Consensus clustering + Gene Set Pathway Analysis of new clusters
4. Network Analysis.
5. Matrix Decomposition

Interpreting integrated datasets: Gene Set/Pathway Analysis

- To allows investigators understand key programs involved in a experimental phenomena
- Increase signal to noise by averaging over systems of individuals genes
- Borrows information from more well annotated genes

Single Sample Gene Set Analysis using MOGSA



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

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

mogsa

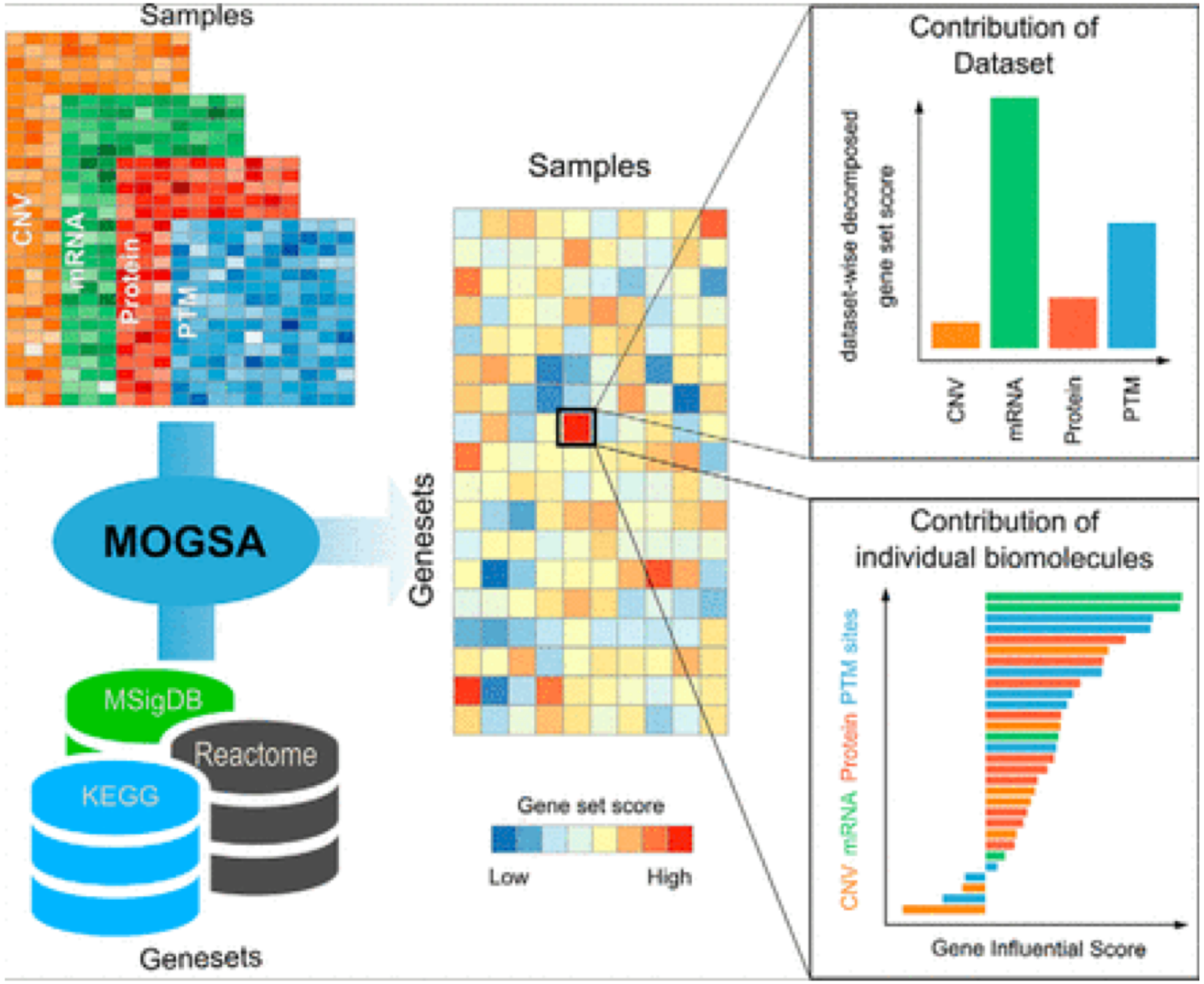
MOGSA

Multi-Omics Gene Set Analysis

Fast: Integrates different ‘omics data using matrix factorization

Flexible: Gene/Protein Annotation Projected onto space.

Metrics: GeneSet x Sample, Gene, Data or Gene Contribution



MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data
Meng C, Basunia A, Peters B, Gholami AM, Kuster B, **Culhane AC**
Molecular & Cellular Proteomics **DOI:** 10.1074/mcp.TIR118.001251

mogsa

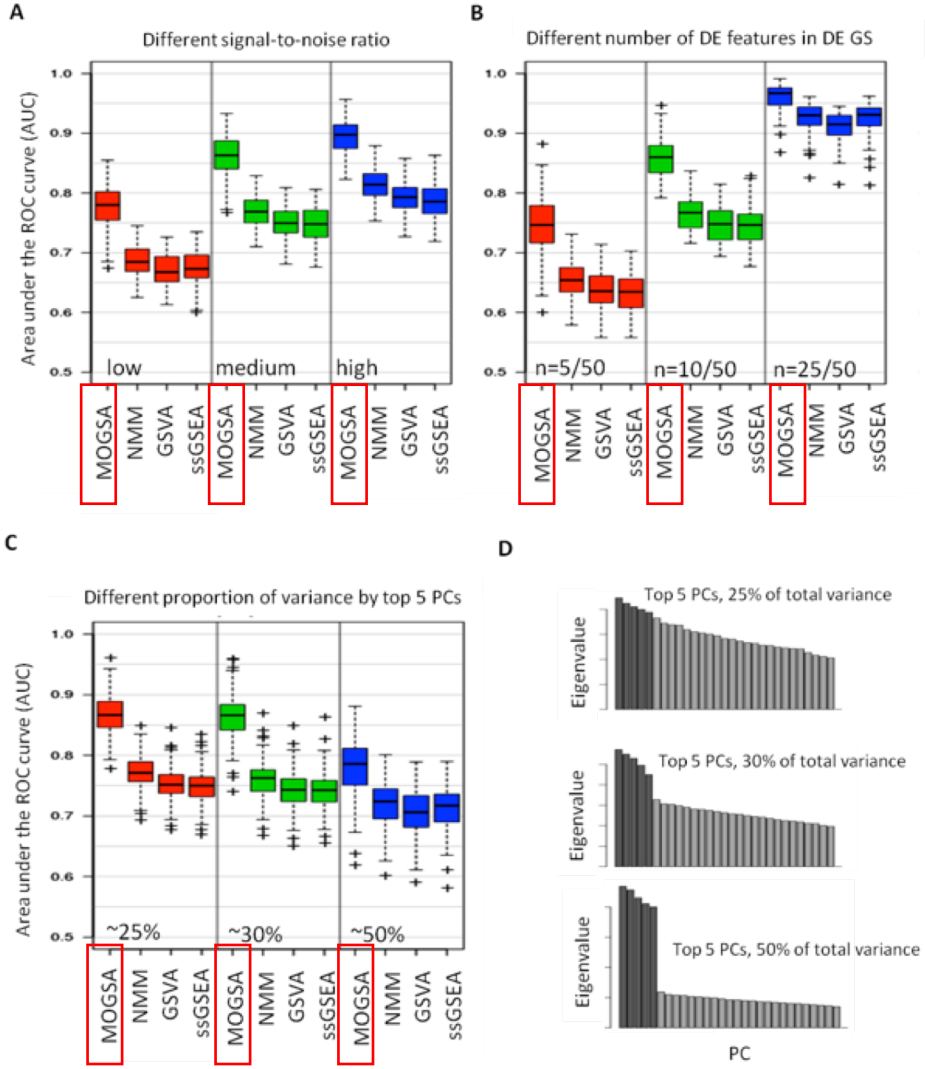
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

moGSA single sample Gene Set analysis

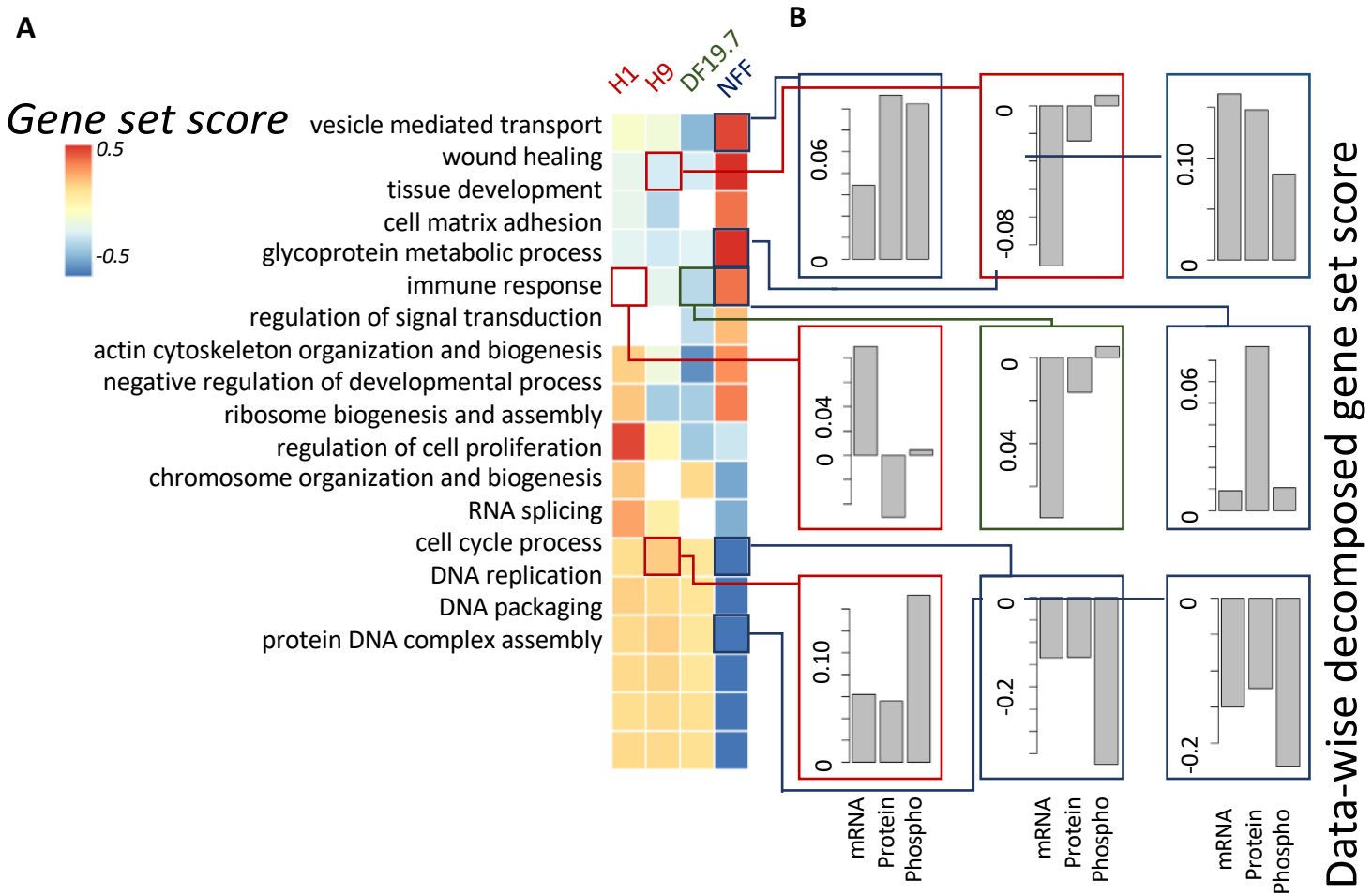
MOGSA outperforms other ssGSA approaches when applied to Synthetic data



Meng C et al.,. 2019



moGSA of small dataset (n=4) of mRNA, protein, phosphoproteins. ESP v IPS cells



Cluster discovery in BLCA TCGA data

Clustering analysis of samples weights

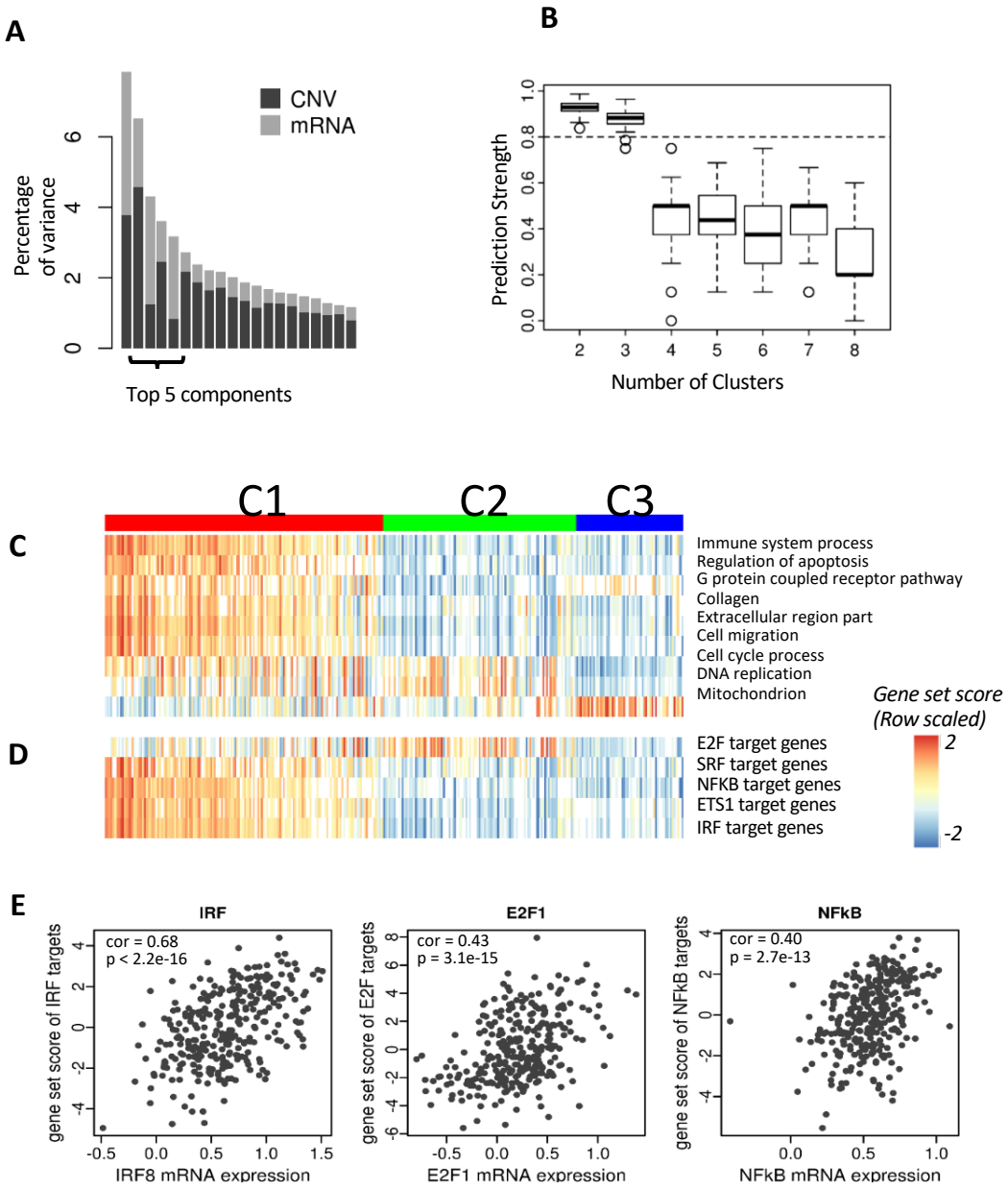


Figure 5

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,

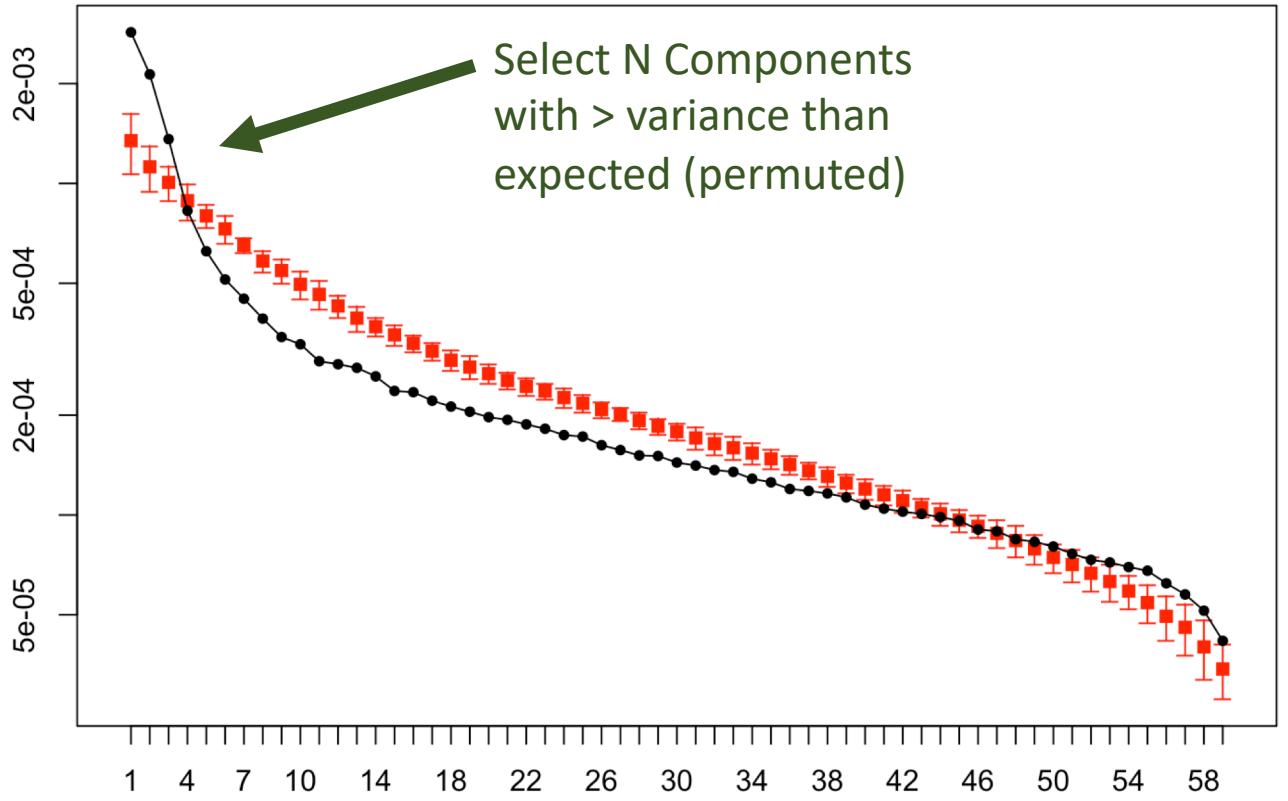
- uniform - no weighting
- lambda1 - weighted by the reverse of the first eigenvalue of each individual dataset
- inertia - weighted by the reverse of the total inertia.

weight datasets closer to the overall structure

statis – FALSE

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

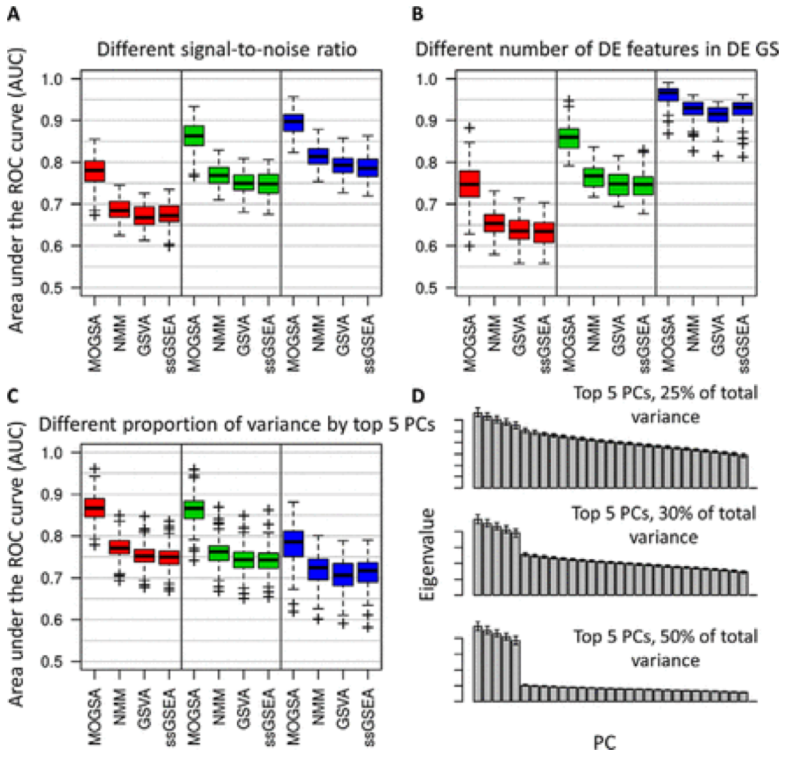
```
bootMoa(
moa = ana,
proc.row = "center_ssq1",
w.data = "inertia",
statis = TRUE,
B = 20,
plot=TRUE)
```



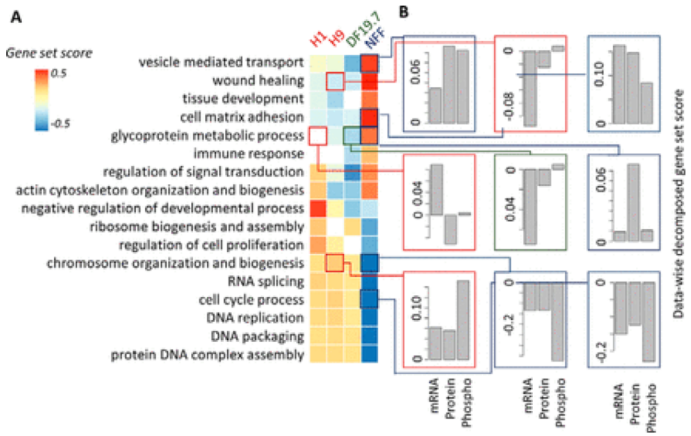
mogsa

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

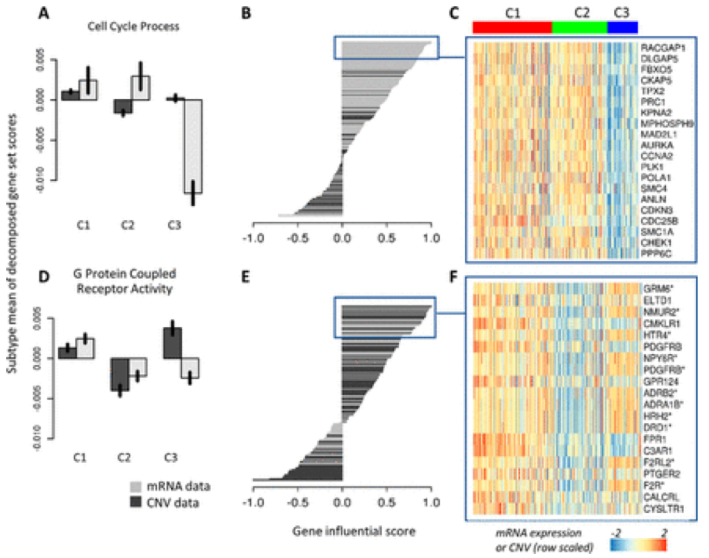
Performance



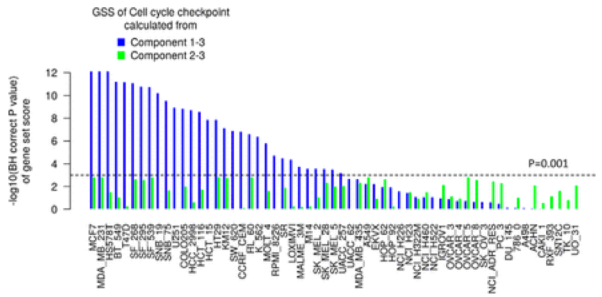
Data weight in Gene Set Scores



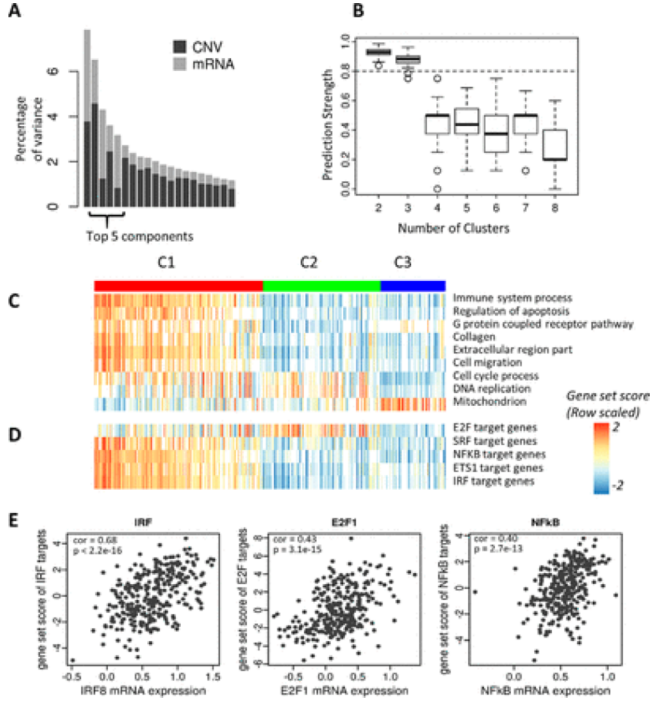
Gene weight in Gene Set Scores



Removal of Batch Effects



Cluster Discovery

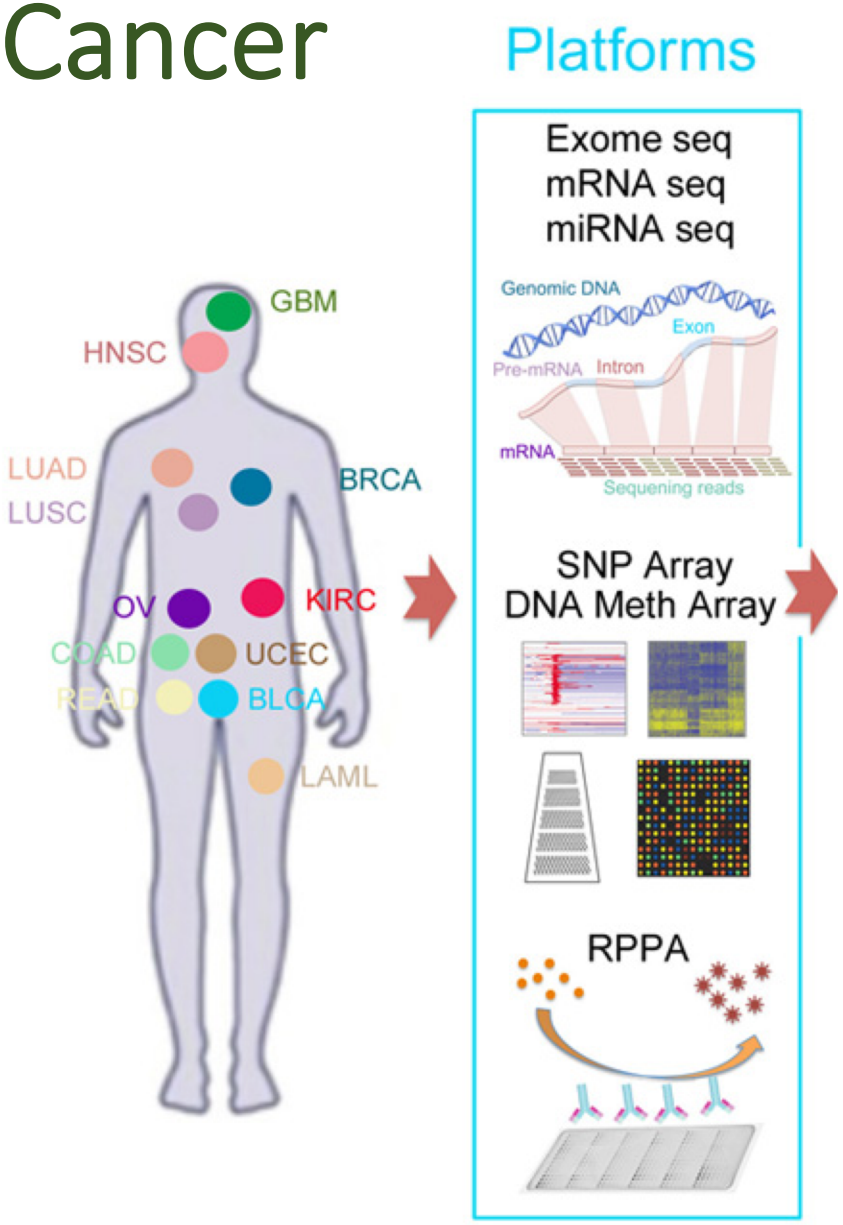


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

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

Using cell admixture to PanCancer Immune subtypes

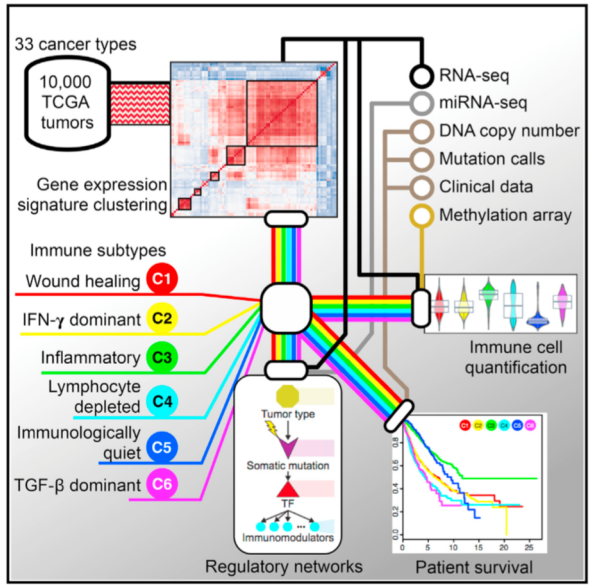
- PanCancer Project
- PanImmune Working Group
- 33 different tumor sources
- >10,000 tumors
- Discover immune subtypes in TCGA tumors
- Hypothesis- Immune subtypes (and infiltrating cells) span cancer types



Immunity

The Immune Landscape of Cancer

Graphical Abstract



Authors

Vésteinn Thorsson, David L. Gibbs, Scott D. Brown, ..., Mary L. Disis, Benjamin G. Vincent, Ilya Shmulevich

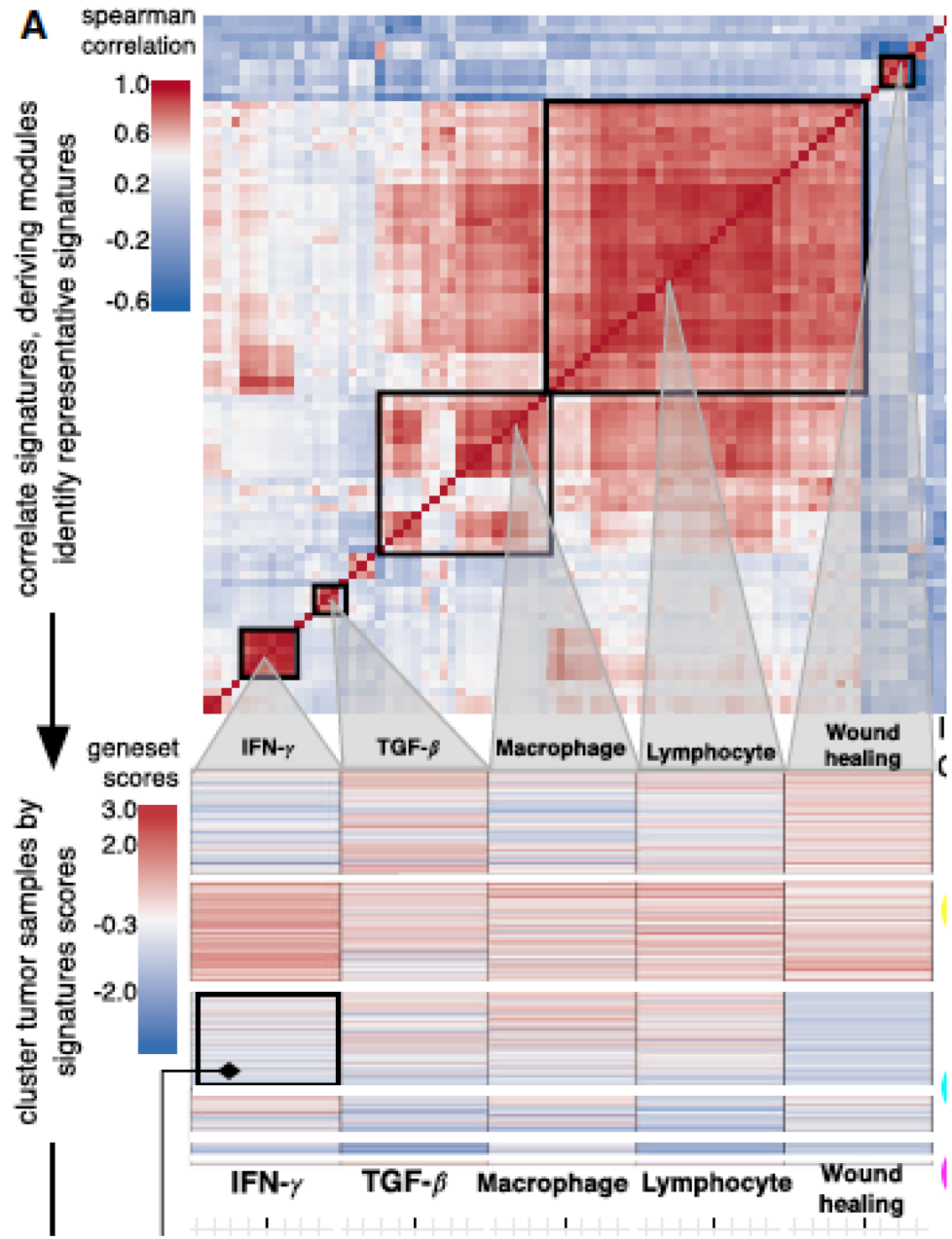
Correspondence

vesteinn.thorsson@systemsbiology.org (V.T.), benjamin.vincent@unchealth.unc.edu (B.G.V.), ilya.shmulevich@systemsbiology.org (I.S.)

In Brief

Thorsson et al. present immunogenomics analyses of more than 10,000 tumors, identifying six immune subtypes that encompass multiple cancer types and are hypothesized to define immune response patterns impacting prognosis. This work provides a resource for understanding tumor-immune interactions, with implications for identifying ways to

	Macrophage: lymphocyte	Th1:Th2	Proliferation	Intratumoral heterogeneity	Other
Wound healing	Balanced	Low	High	High	
IFN-γ dominant	Lowest	Lowest	High	Highest	Highest M1 and highest CD8 T cells
Inflammatory	Balanced	High	Low	Lowest	Highest Th17
Lymphocyte depleted	High	Minimal Th	Moderate	Moderate	
Immunologically quiet	Highest	Minimal Th	Low	Low	Highest M2
TGF-β dominant	High	Balanced	Moderate	Moderate	Highest TGF-β signature



#Bioc2020

(Co-Chair)

http://bioc2020.bioconductor.org/

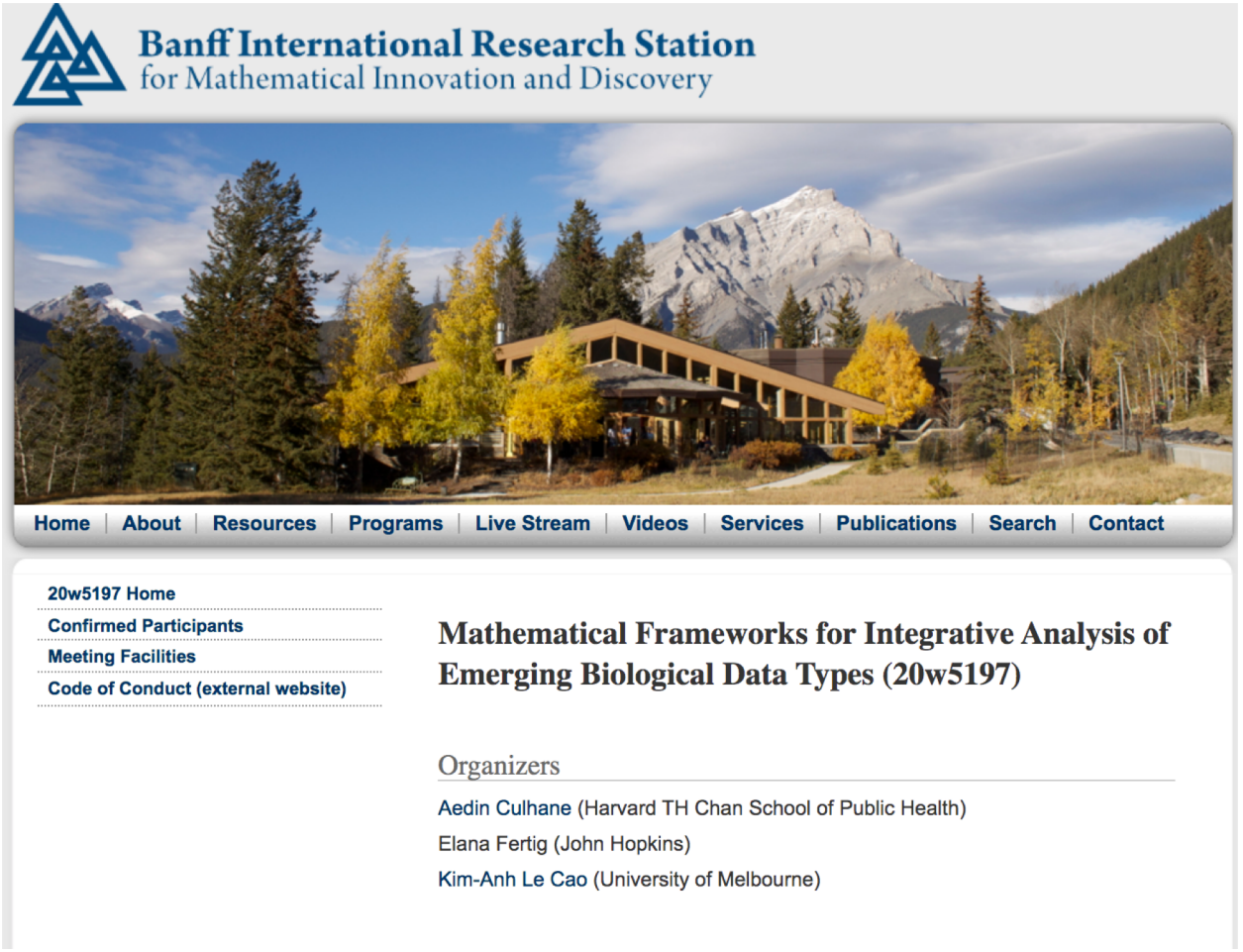
BioC 2020: Where Software and Biology Connect

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



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

Its all about chemistry…