BOSTON R/Bioconductor for Genomics

meetup

Matrix factorization & network inference approaches for multi-omics data fusion

Feb 28th 5:30-8:00 pm

Yawkey Building,
Room 306/307
Dana-Farber Cancer Institute
450 Brookline Ave, Boston

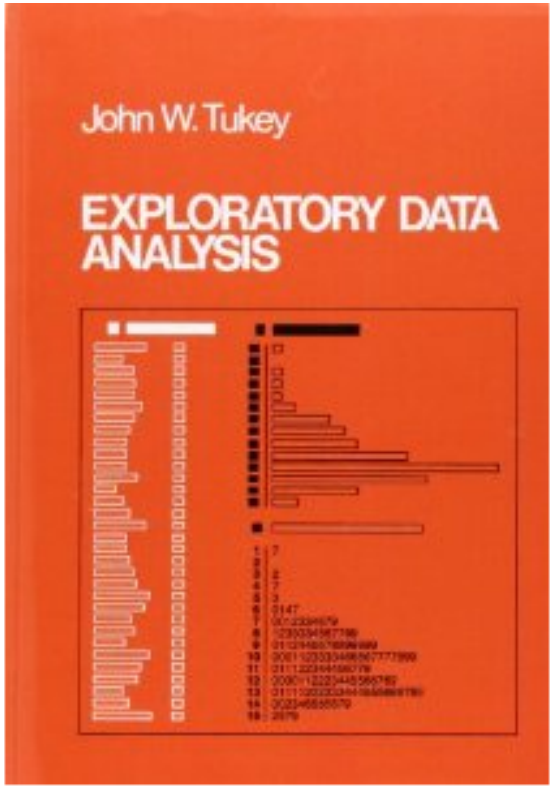
RSVP
https://www.meetup.com/Boston-R-Bioconductor-for-genomics/

Nathalie Pochet, Ph.D.
Harvard Medical School, Broad Institute of MIT and Harvard, Brigham and Women's Hospital
AMARETTO: a regulatory network inference tool for multi-omics data fusion across systems and diseases

Aedin Culhane, Ph.D.
Dana-Farber Cancer Institute,
Harvard TH Chan School of Public Health
moGSA, Integrative multi' omics single sample gene set analysis

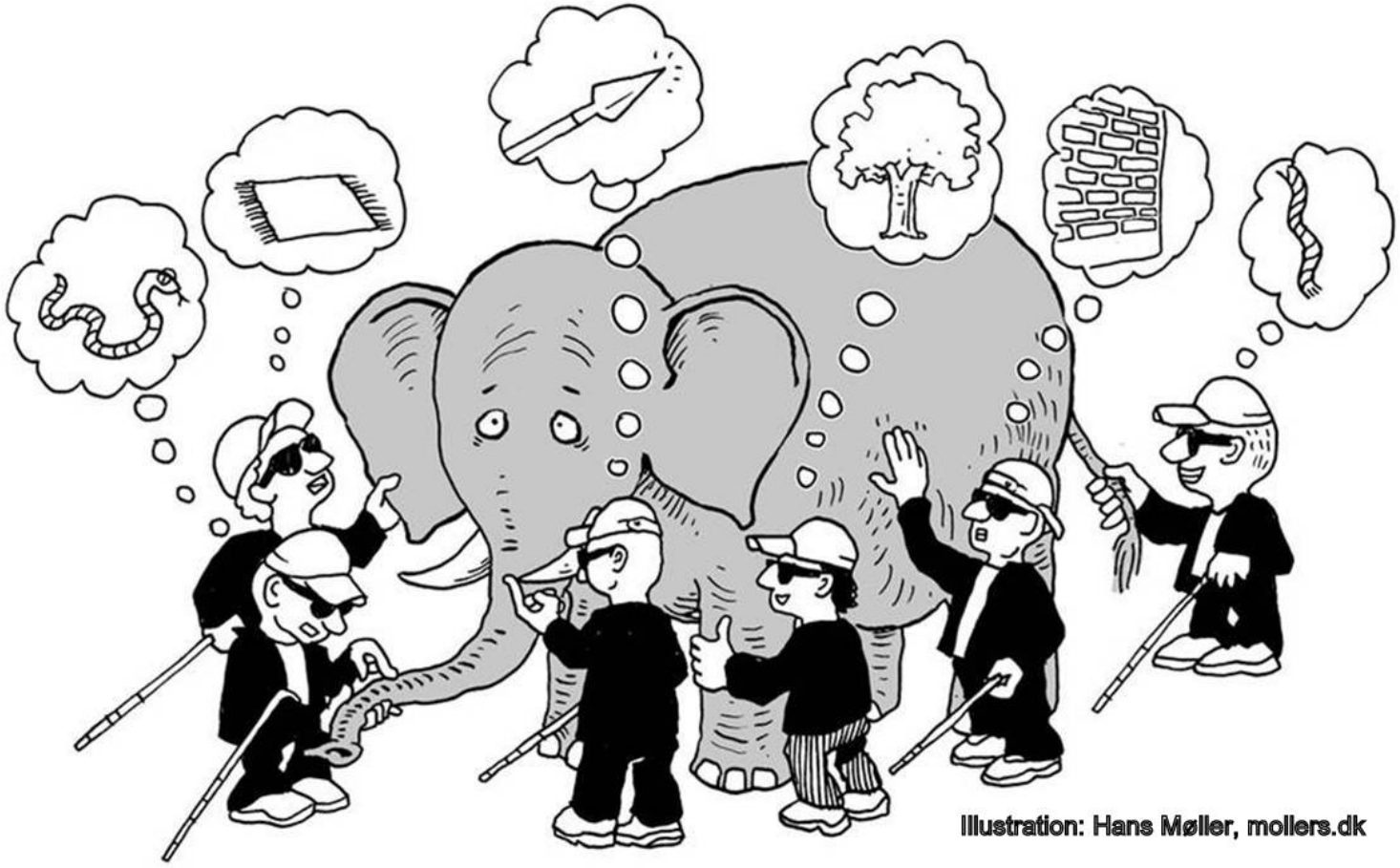
Exploratory data analysis (EDA)

> The greatest value of a picture is when it forces us to notice what we never expected to see.
>
> – John W. Tukey. Exploratory Data Analysis. 1977.



Exploratory Data Analysis [Paperback]
John W. Tukey (Author)
(8 customer reviews)

As data gets larger, EDA is important



Dimension Reduction, Latent or Matrix Factorization approaches

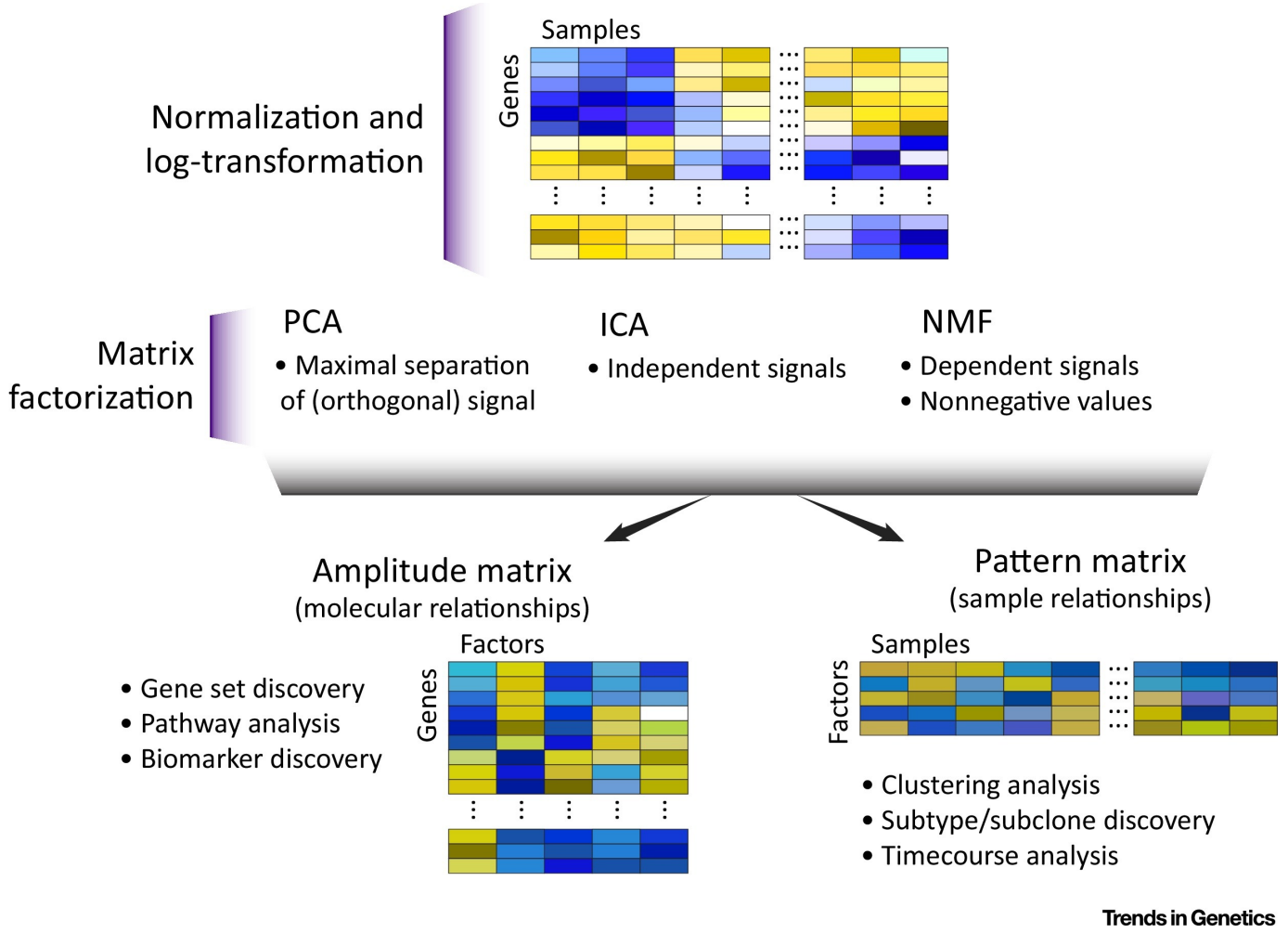
- Principal component analysis (PCA)
- Correspondence analysis (COA or CA)
- Nonmetric multidimensional scaling (NMDS, MDS)
- Principal co-ordinate analysis (PCoA)

latent

- present but not visible, apparent.

Do you see an old or young lady?

Matrix Decomposition: Find Latent Patterns

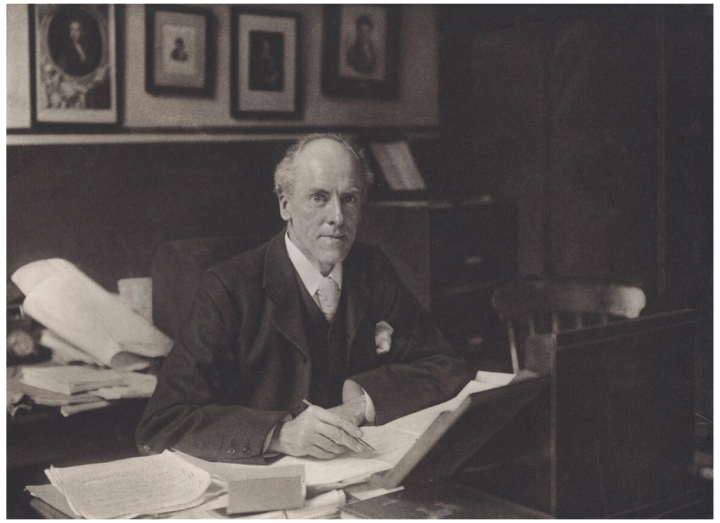


Stein-O'Brien GL, et al., Enter the Matrix: Factorization Uncovers Knowledge from Omics. 2018 *Trends in Genetics* DOI: (10.1016/j.tig.2018.07.003)

PCA: Pearson 1901

The best fit line passes through the centroid

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.



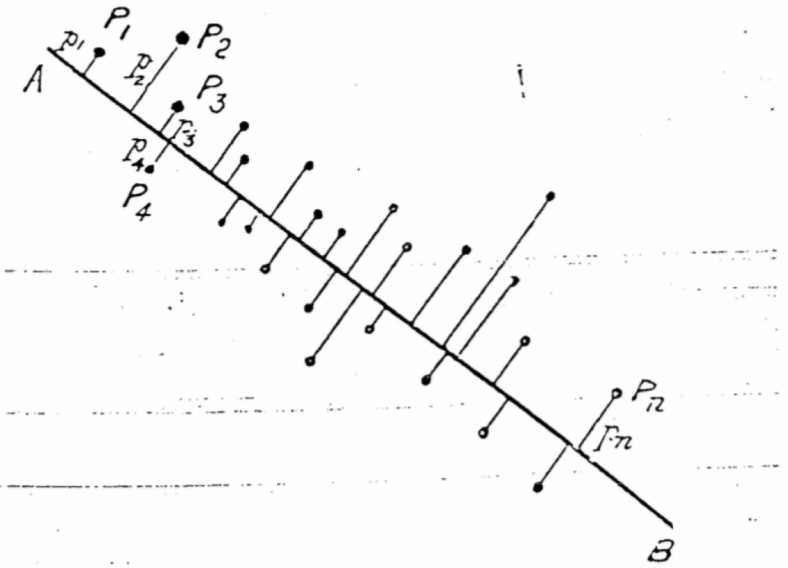
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

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

If y were the dependent variable, we should have made

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

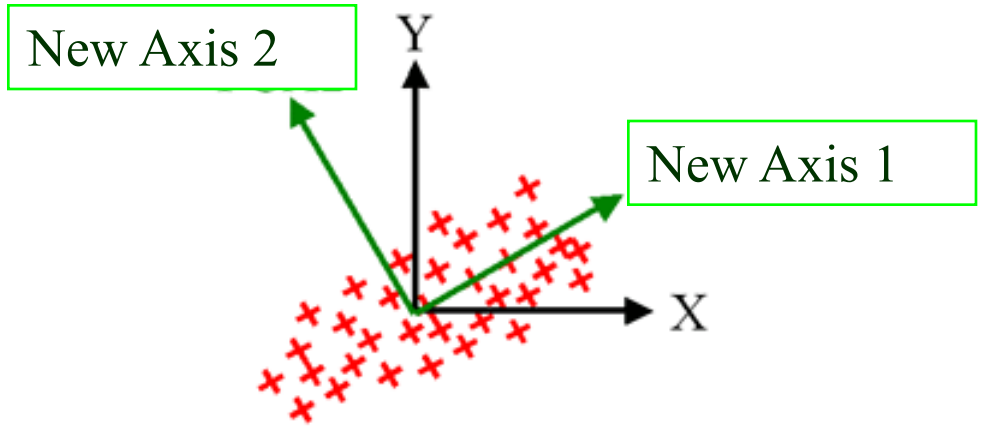
(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



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. *http://pbil.univ-lyon1.fr/R/pearson1901.pdf*

Matrix Decomposition is ideally suited to finding known & unknown (latent) patterns between datasets



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

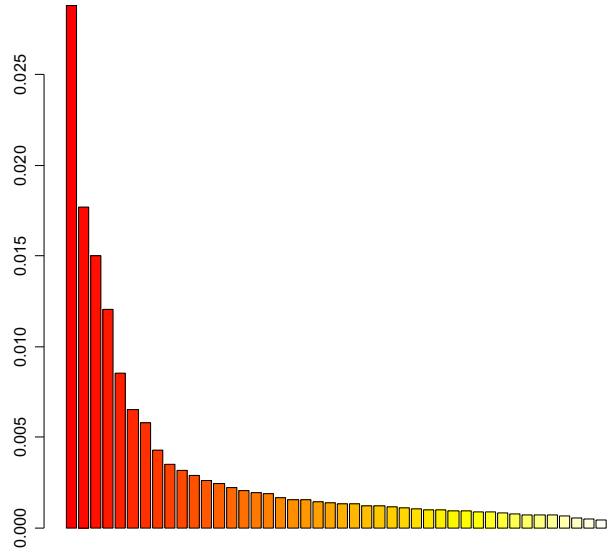
Review: Stein-O'Brien et al. *bioRxiv*. doi:10.1101/196915.

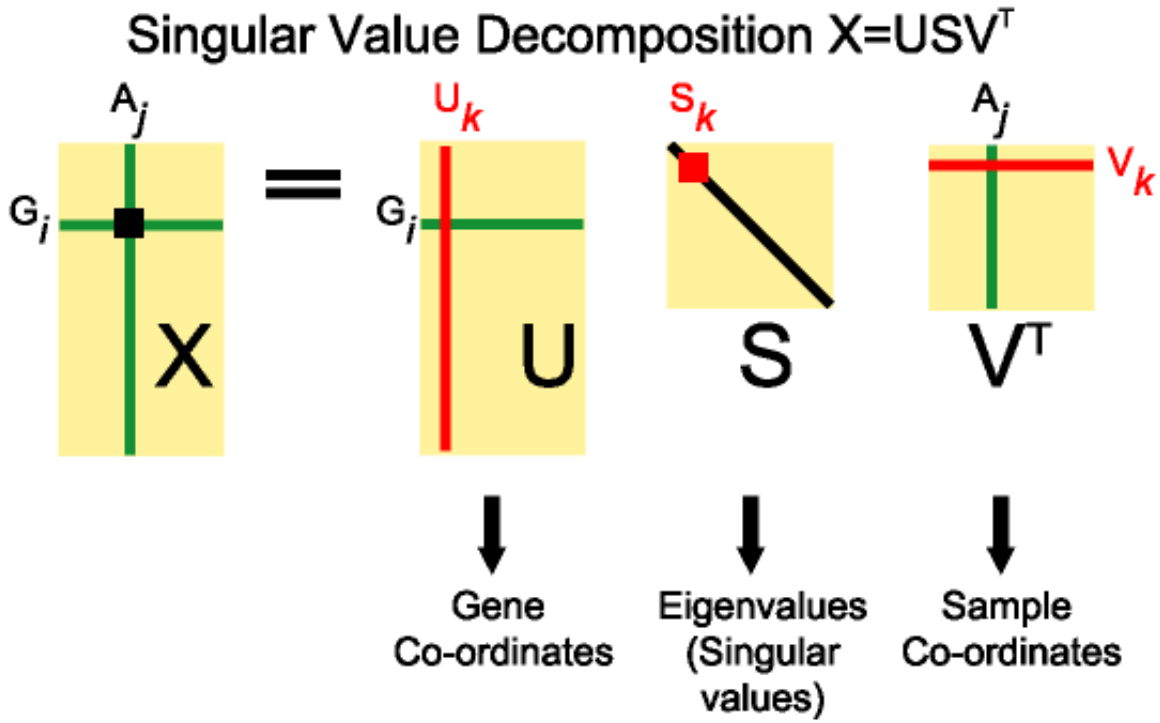
Principal Axes

- Project new axes through data which capture variance. Each represents a different trend in the data.
- Orthogonal (decorrelated)
- Typically ranked: First axes most important
- Principal axis, Principal component, latent variable or eigenvector

Eigenvalues

- Describe the amount of variance (information) captured by each eigenvector
- Ranked. First eigenvalue is the largest.
- Generally only examine 1st few components
  - scree plot





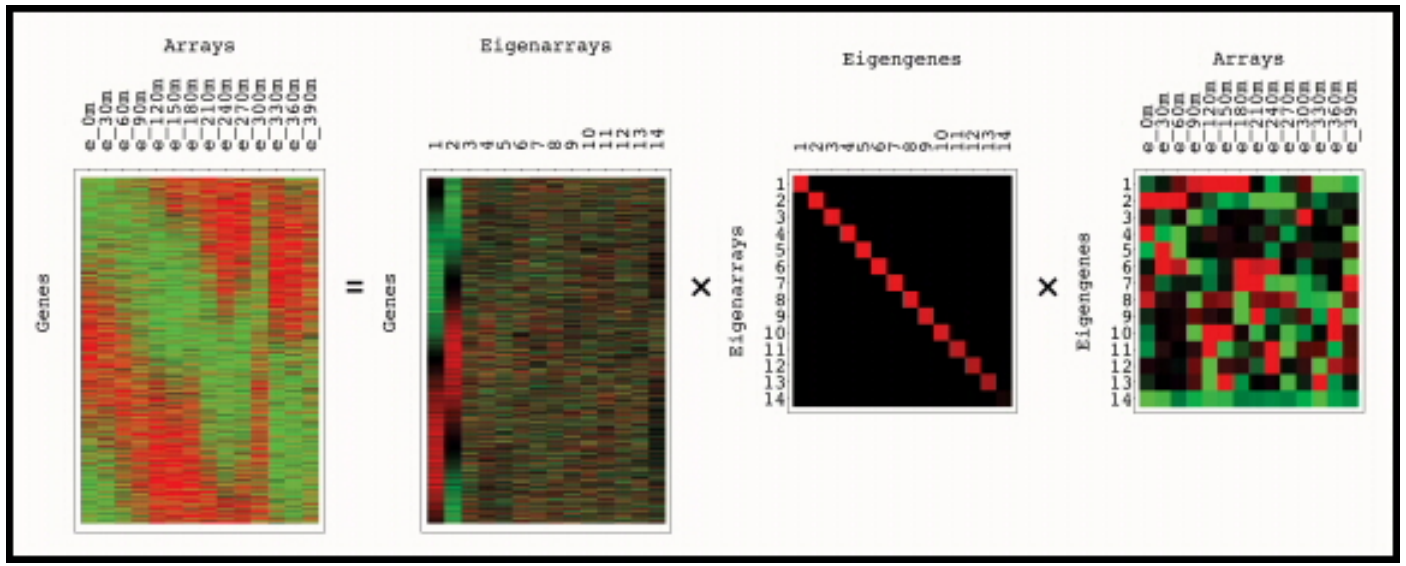


Image from

http://genome-www.stanford.edu/SVD/

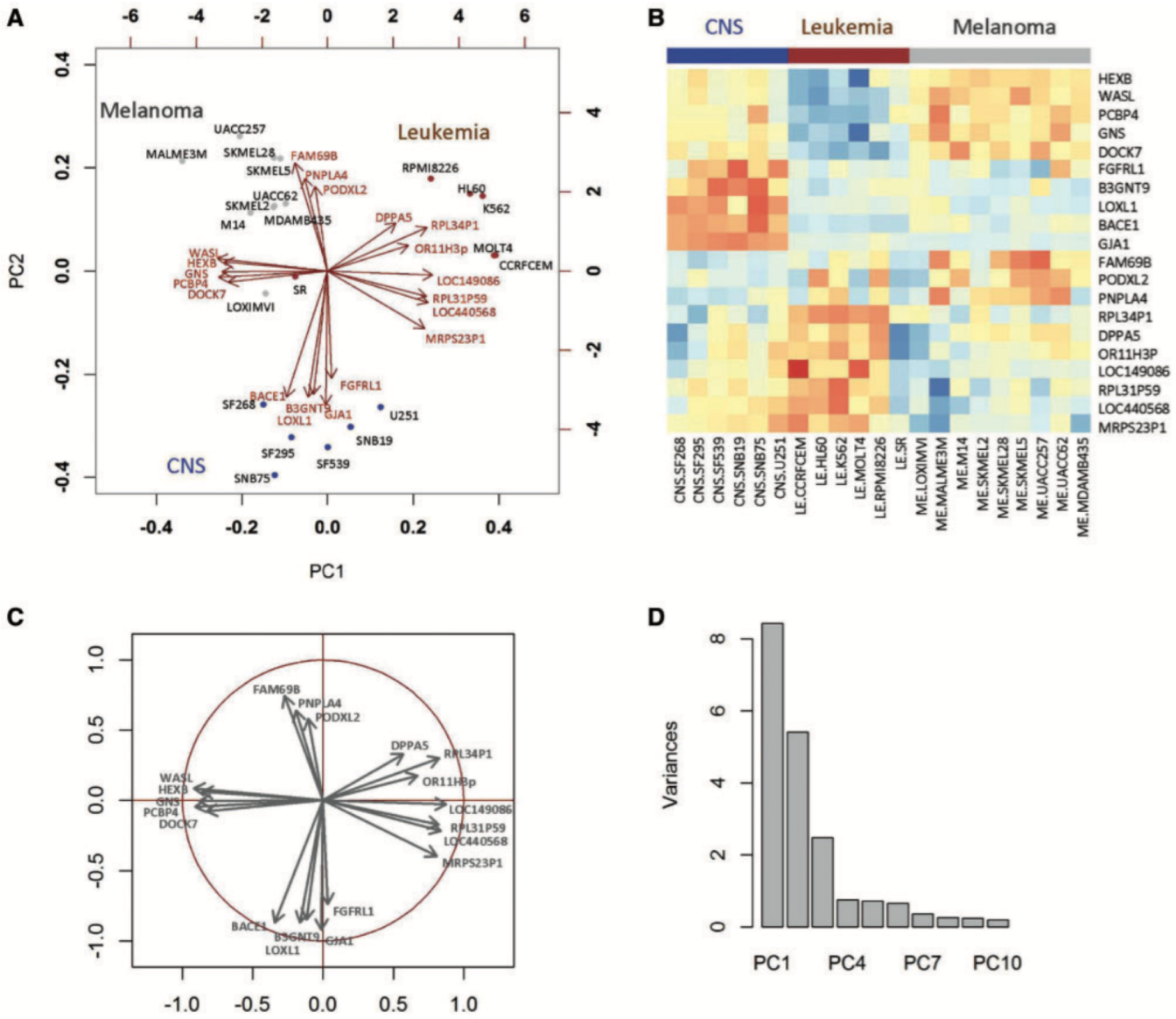


Figure 1. Results of a PCA analysis of mRNA gene expression data of melanoma (ME), leukemia (LE) and central nervous system (CNS) cell lines from the NCI-60 cell line panel. All variables were centered and scaled. Results show (**A**) a biplot where observations (cell lines) are points and gene expression profiles are arrows; (**B**) a heatmap showing the gene expression of the same 20 genes in the cell lines; red to blue scale represent high to low gene expression (light to dark gray represent high to low gene expression on the black and white figure); (**C**) correlation circle; (**D**) variance barplot of the first ten PCs. To improve the readability of the biplot, some labels of the variables (genes) in (**A**) have been moved slightly. A colour version of this figure is available online at BIB online: http://bib.oxfordjournals.org.

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

Correspondence Analysis

- **COA** (or CA) is an eigenanalysis of a Chi-square distance matrix.
- Measures the "strength" of association between an up-regulated gene and an array sample.
- Developed by numerous authors, also known as reciprocal averaging/ordering, dual scaling etc.
- Initially designed for analysis of 2-way contingency tables (frequency counts). Thus assumes matrix counts positive integers or zeros.

Multidimensional scaling (MDS)

- Input distance matrix
- Classical MDS is identical to principal coordinates analysis (PCoA).

- NMDS. Iterative. isoMDS (MASS), sammon (MASS).

Other related methods

Independent Component Analysis

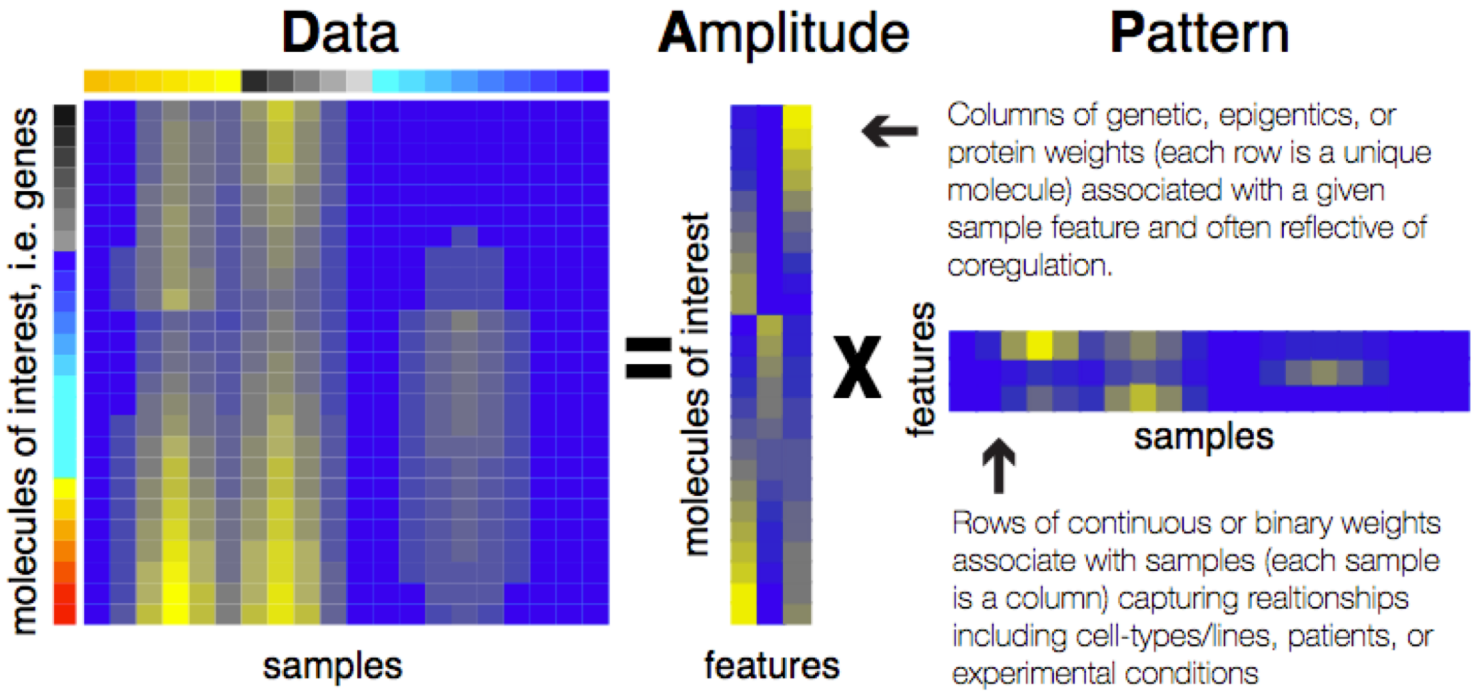
- does not constrain the axes to be orthogonal
- attempts to place them in the directions of statistical dependencies in the data.

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.

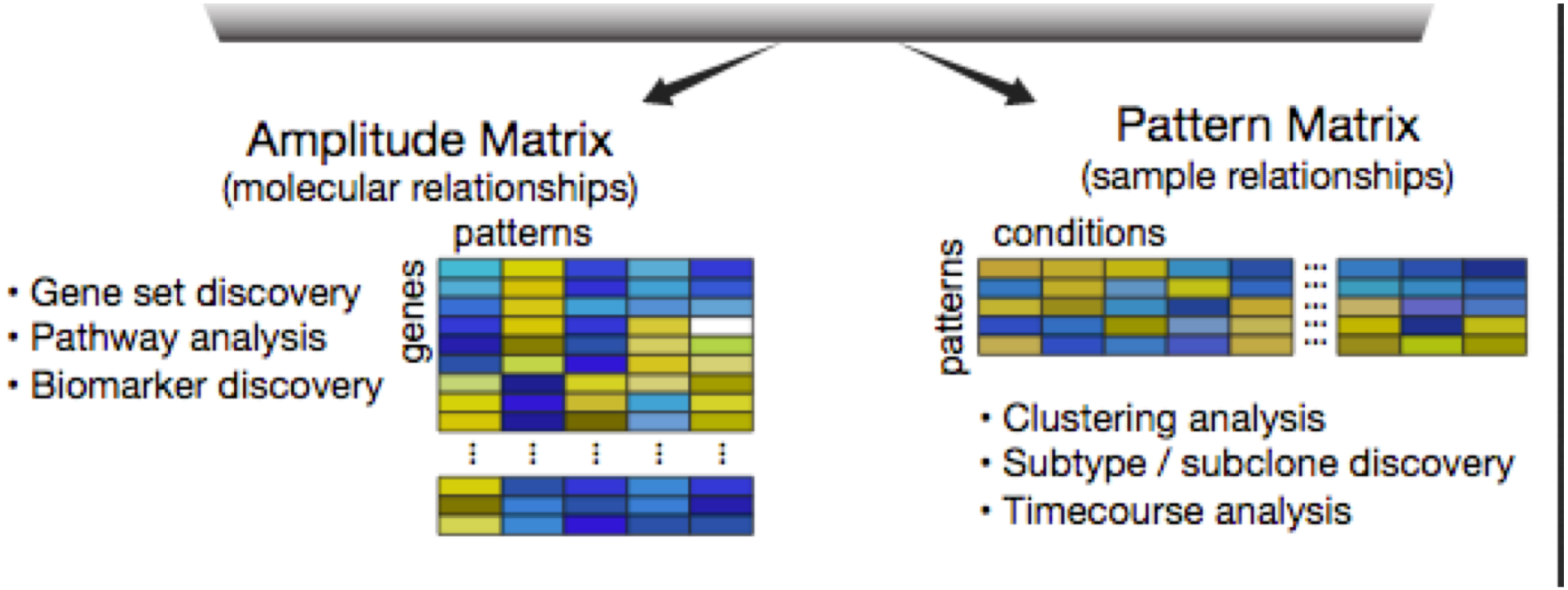
Global v local methods

- Non negative matrix factorization
  - Sparse, additive, local
  - Find by gradient descent



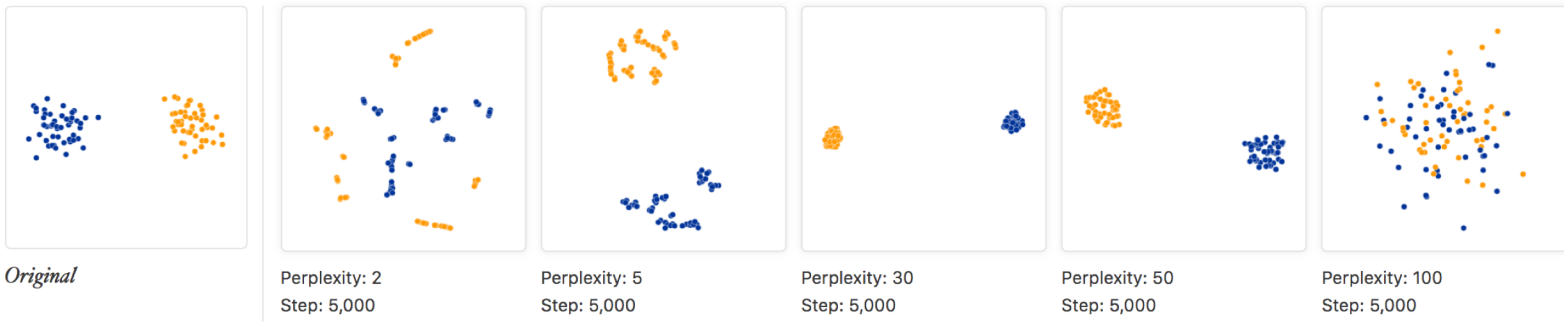
Stein-O'Brien GL, et al., Enter the Matrix: Factorization Uncovers Knowledge from Omics. 2018 *Trends in Genetics* DOI: (10.1016/j.tig.2018.07.003)

Extract weights of features and samples in new space



t-SNE t-distributed stochastic neighbor embedding

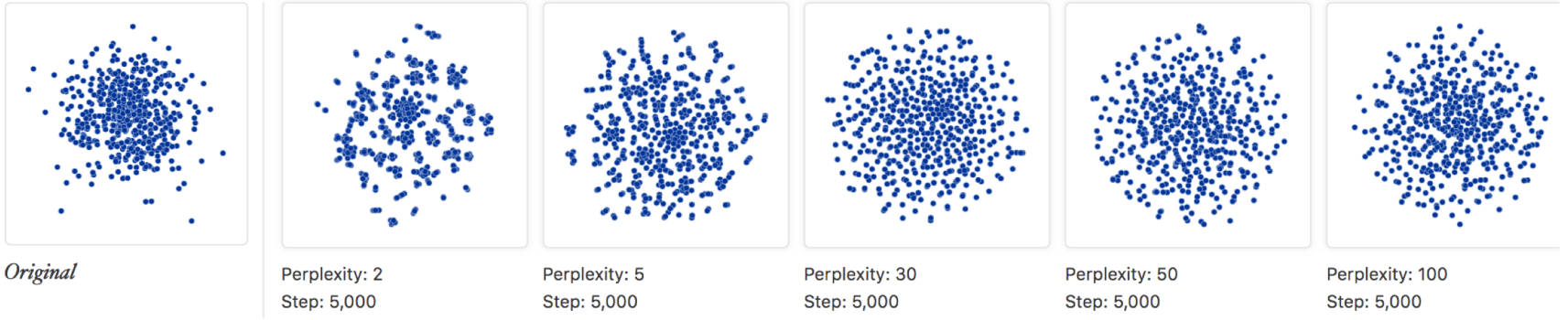
- Introduced van der Maaten and Hinton in 2008.
- non-convex objective function. objective function minimized using a gradient descent optimization. Possible that different runs give you different solutions.
- Good when large number (scale) objects as it avoids overlapping close points.
- t-SNE can "snap" groups apart farther than what represents the data generating mechanism. Parameters need to be optimized . “perplexity,” can balance trade-off between local and global components in data.
- Can be challenging to interpret t-SNE coordinates.



Be cautious inferring clusters with t-SNE

4. Random noise doesn't always look random.

A classic pitfall is thinking you see patterns in what is really just random data. Recognizing noise when you see it is a critical skill, but it takes time to build up the right intuitions. A tricky thing about t-SNE is that it throws a lot of existing intuition out the window. The next diagrams show genuinely random data, 500 points drawn from a unit Gaussian distribution in 100 dimensions. The left image is a projection onto the first two coordinates.

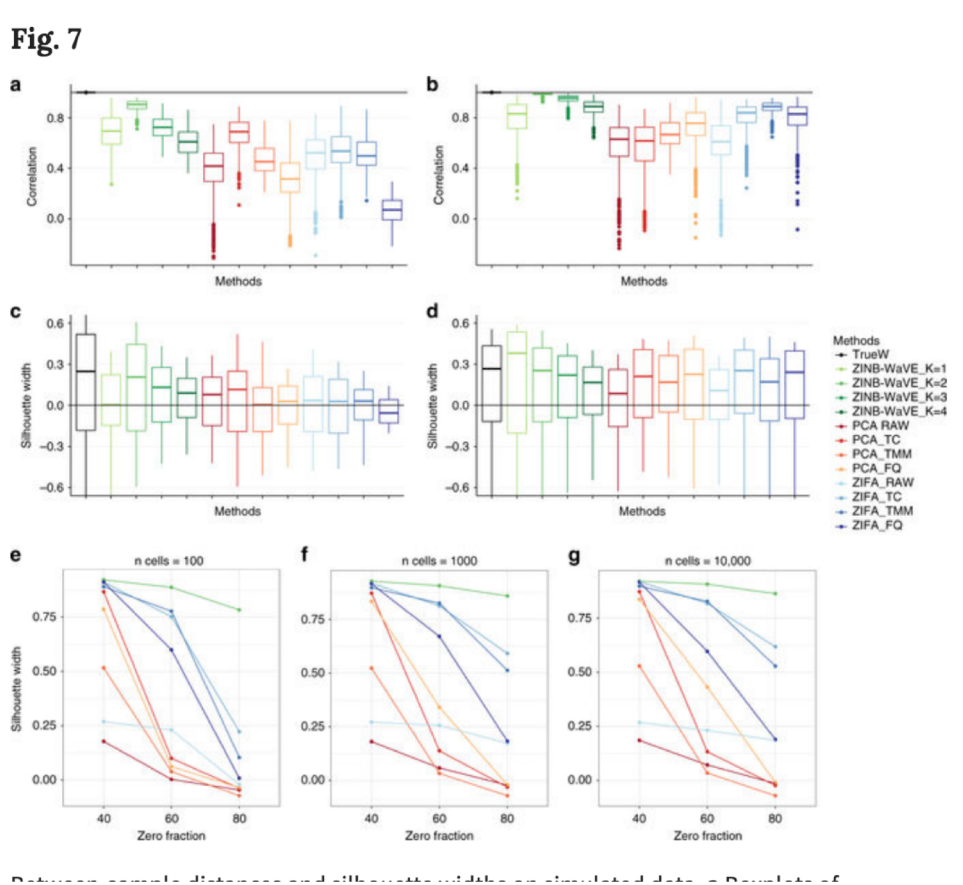


https://support.bioconductor.org/p/97594/#97598

How to use t-SNE effectively https://distill.pub/2016/misread-tsne/

Zero-inflated negative binomial model (ZINB-WaVE),

- Factor analysis- Clear interpretation of the reduced space
- Generates low-dimensional representations of the data that account for zero inflation (dropouts), over-dispersion. Applied to the count data.
- Assumes that the "true" signal is intrinsically low-dimensional

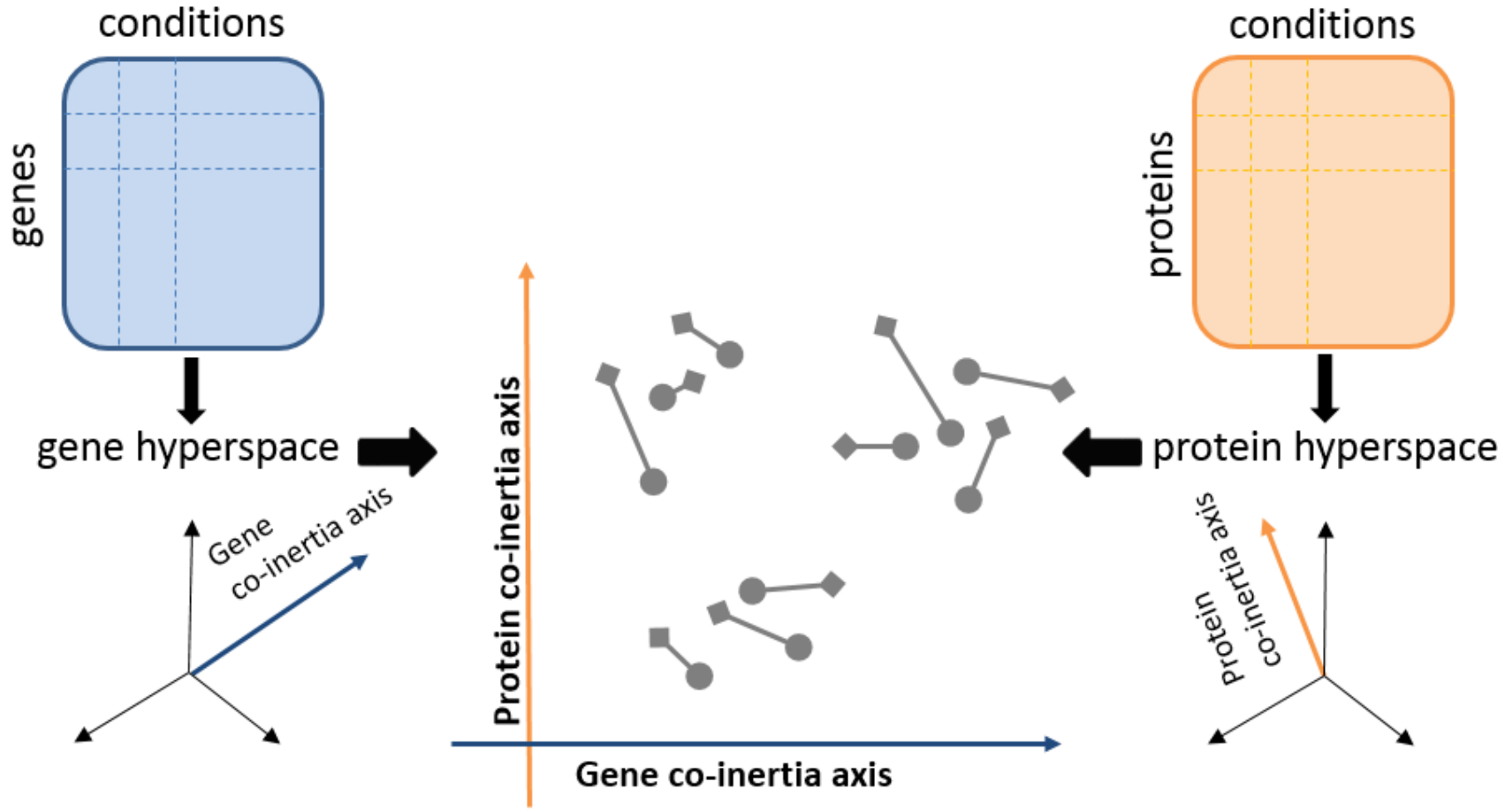


Risso et al., 2017 https://www.nature.com/articles/s41467-017-02554-5

Summary (single dataset methods)

- Classical methods (PCA, CA, MDS) global methods. Limitations on “big” data
- Local methods NMFs. slower not determined (gradient descent)
- t-SNE powerful but need to watch parameters
- ZINB-WaVE outperforms PCA. Possibly more “robust” that t-SNE in some cases
- In reality try >1 approaches

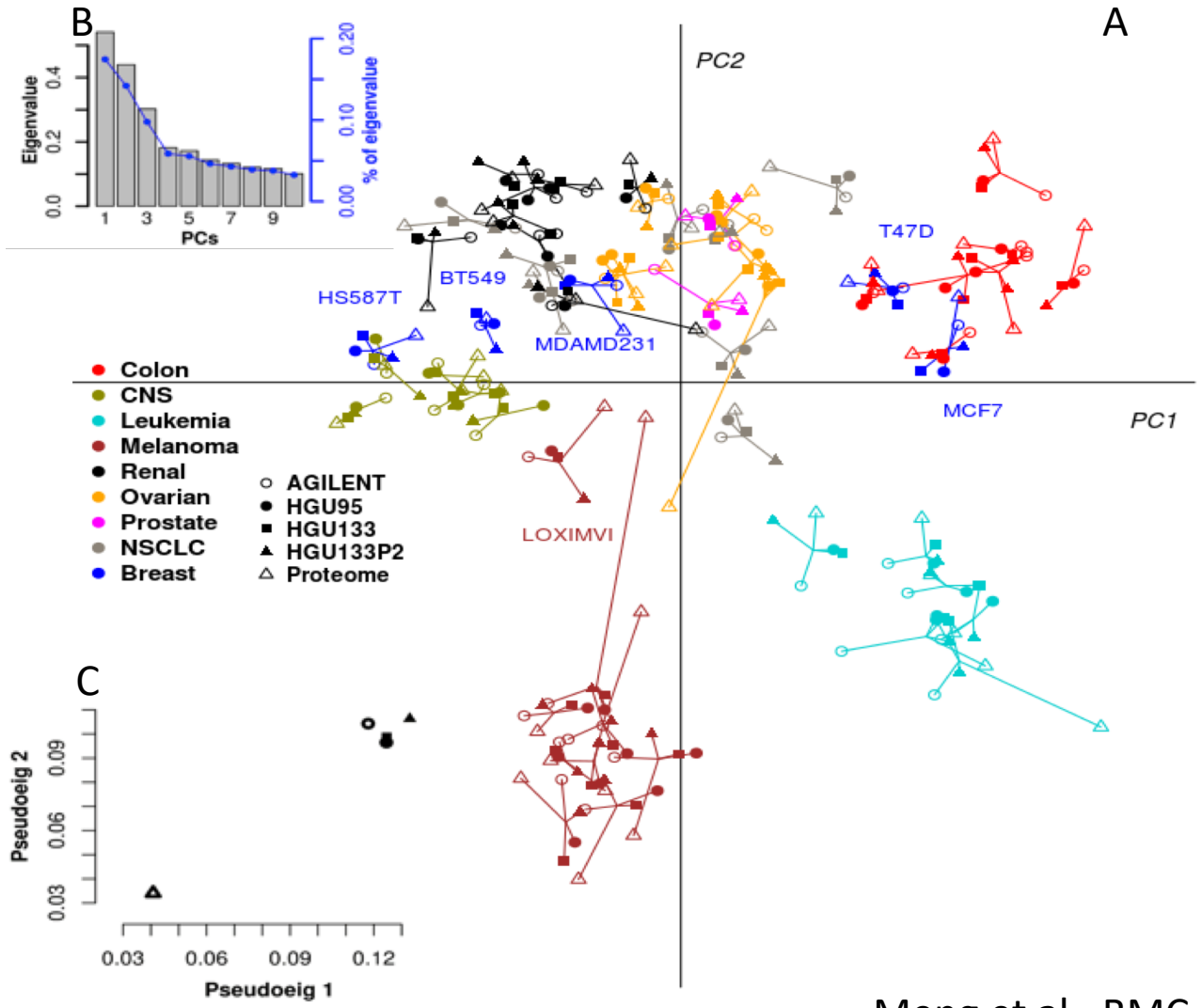
>1 Dataset.



Doledec S *et al.*, Freshwater Biology 1994, 31:277-294
Culhane AC *et al.*, BMC Bioinformatics 2003, 4:59-74

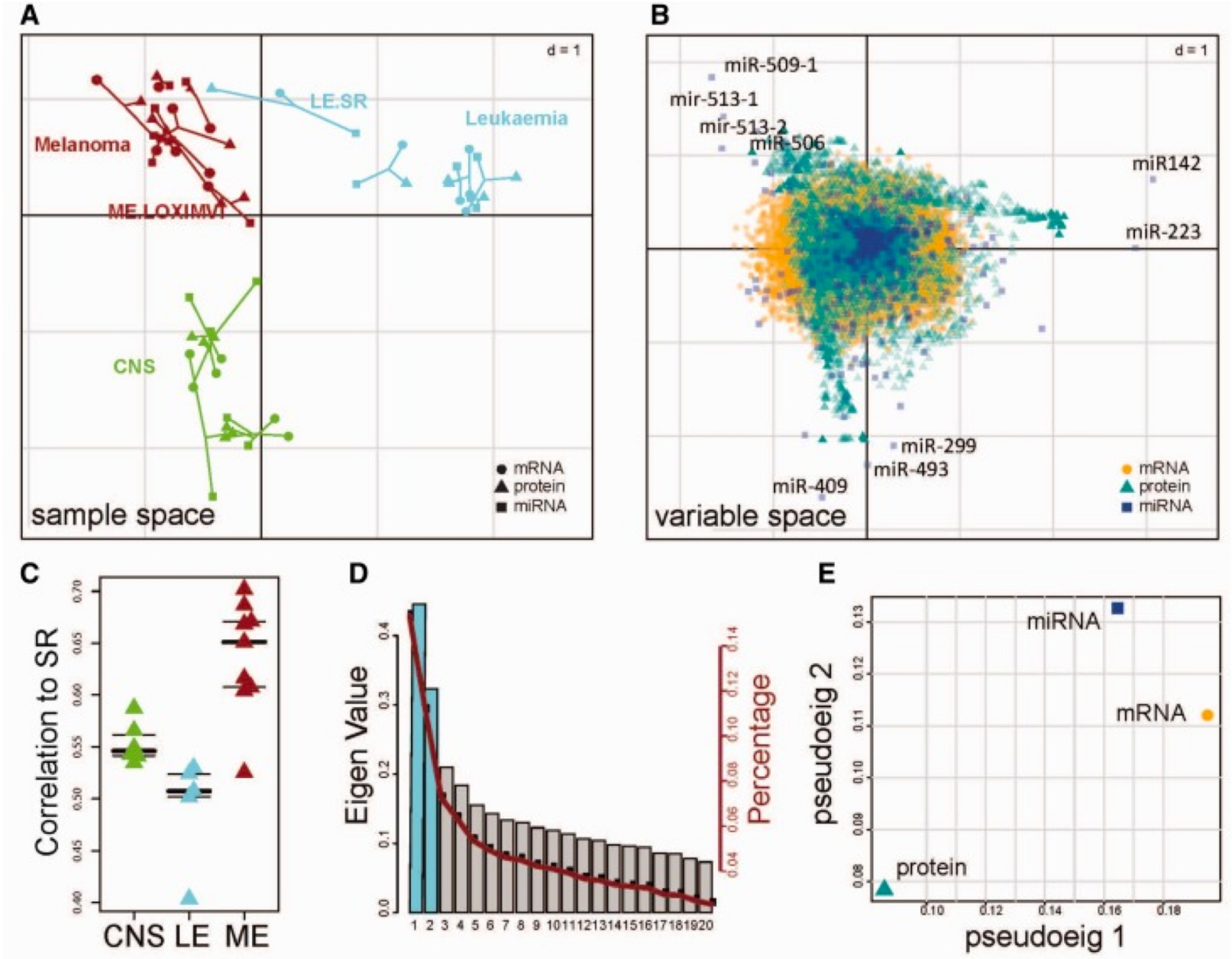
Meng et al., BMC Bioinformatics 2014, 15:162
Meng et al., Brief Bioinform. 2016 Jul; 17(4): 628–641.

Tensor Integration of 5 mRNA studies using MCIA



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

MCIA to analyze mRNA, miRNA and proteomics expression profiles of melanoma, leukemia and CNS cells lines from the NCI-60 panel.



RV coefficients from 0.78 to 0.84.

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

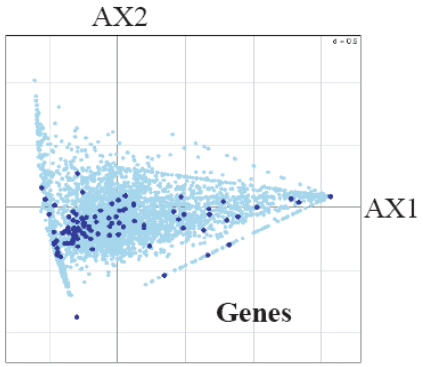
>2 datasets : Tensor data integration

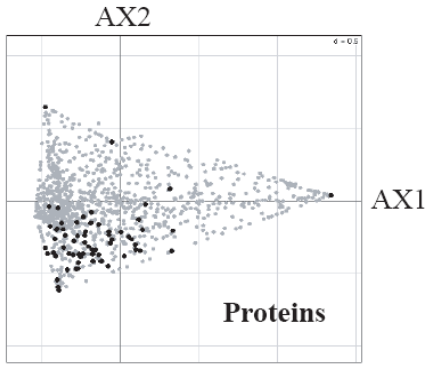
Table 4. Dimension reduction methods for multiple (more than two) data sets

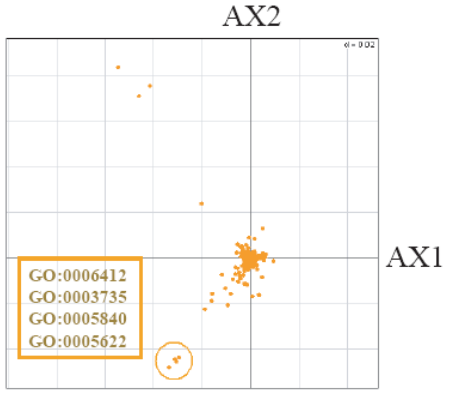
Method	Description	Feature selection	Matched cases	R Function {package}
MCIA	Multiple coinertia analysis	No	No	mcia{omicade4}, mcoa{ade4}
gCCA	Generalized CCA	No	No	regCCA{dmt}
rGCCA	Regularized generalized CCA	No	No	regCCA{dmt} rgcca{rgcca} wrapper.rgcca{mixOmics}
sGCCA	Sparse generalized canonical 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}

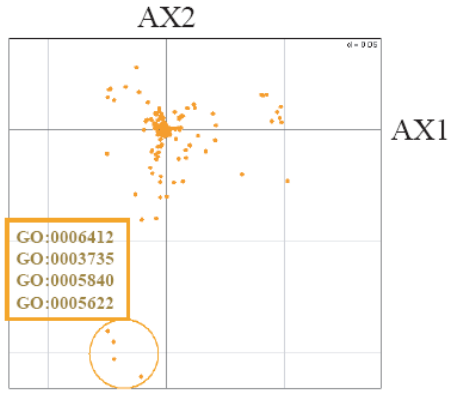
Meng & Zeleznik et al., (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics, 17(4), 2016, 628–641

Can generate a gene set scores over union of all features in new space







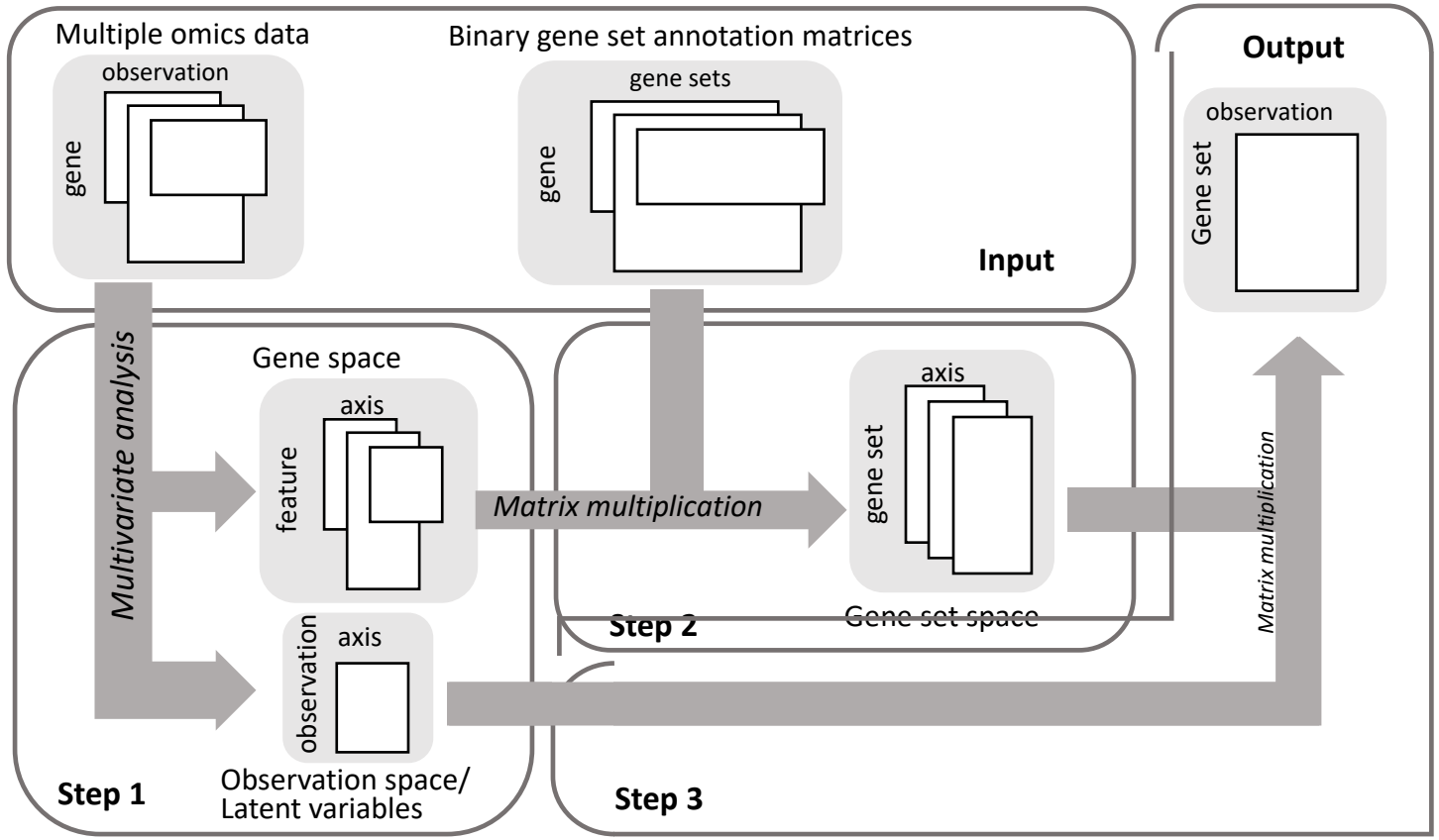


Matrix decomposition of gene expression and proteomics onto same scale

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.

Reduce features to “groups of genes” to score get groups feature level single per case (moGSA)



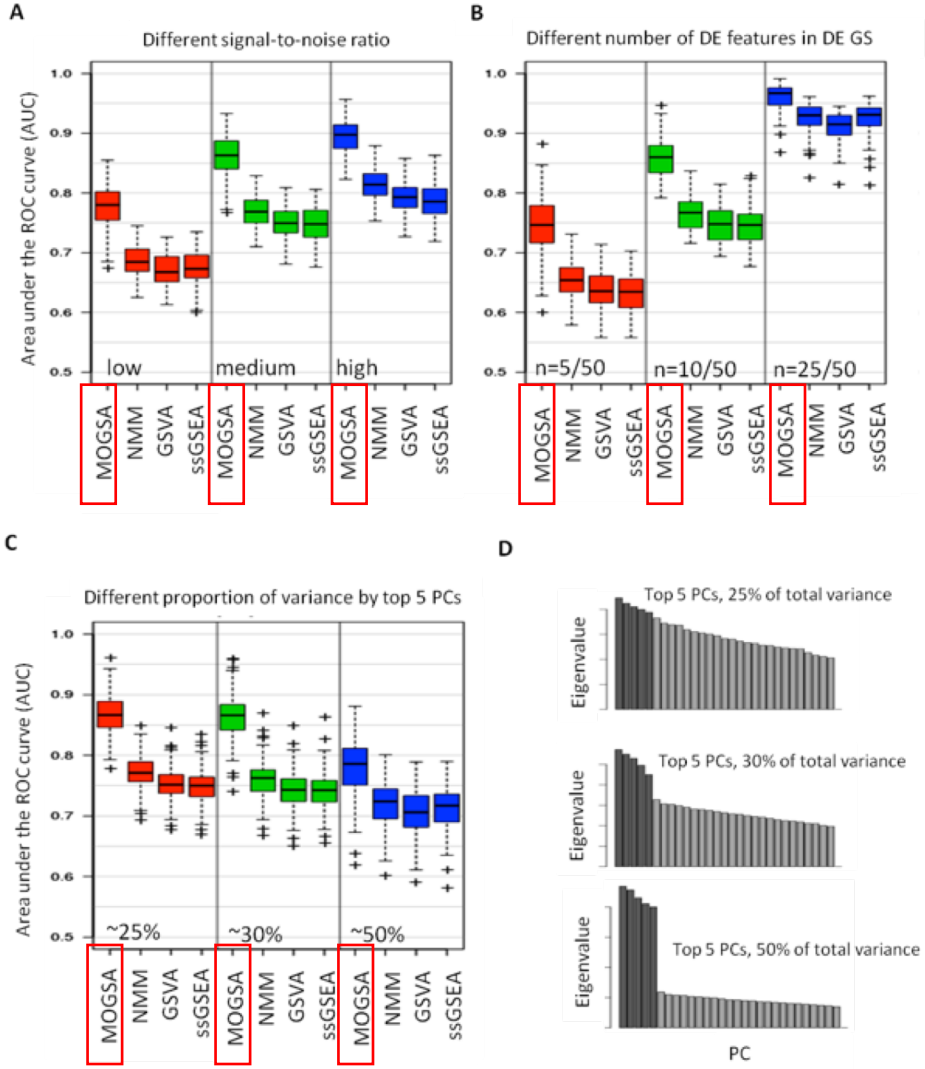
Meng C, Basunia A, Kuster B , Peters B, Gholami AM, Culhane AC. moGSA: Integrative Single Sample Gene-Set Analysis of Multiple Omics Data. *bioRxiv*, 046904.

moGSA single sample Gene Set analysis

MOGSA outperforms other ssGSA approaches when applied to Synthetic data



Meng C et al.,. moGSA: Integrative Single Sample Gene-Set Analysis of Multiple Omics Data. *bioRxiv*, 046904.



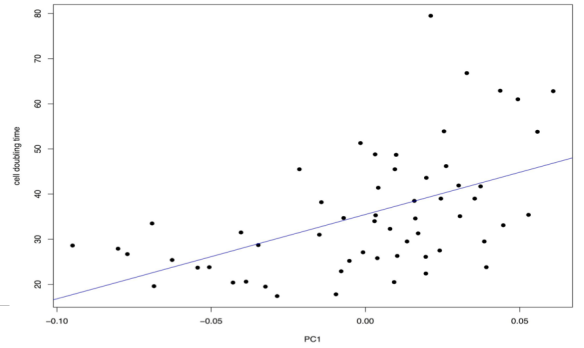
moGSA is ideally suited to ssGSEA of complex data

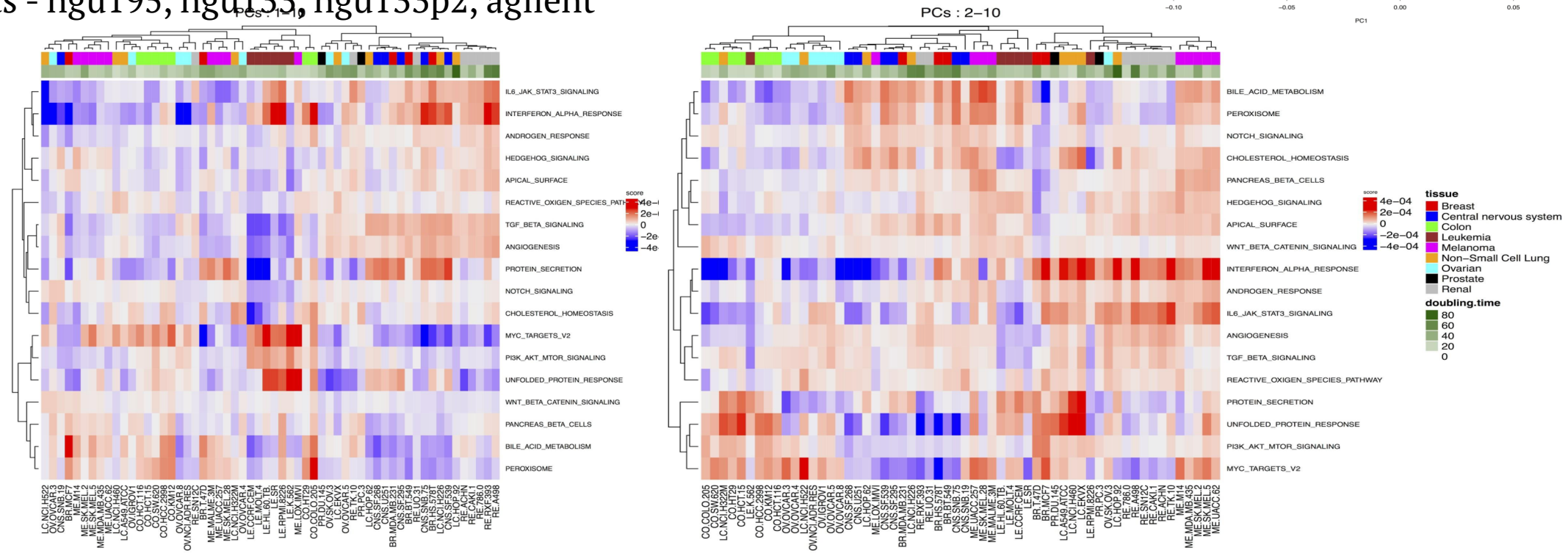
Retrieve ssGSA scores on each component.

Can select which component to analyze or **exclude**

58 NCI60 samples and 18 Hallmark Genesets

4 datasets - hgu195, hgu133, hgu133p2, agilent





After removing PC1, and recomputing moGSA. Cell cycle “signal” is lost.

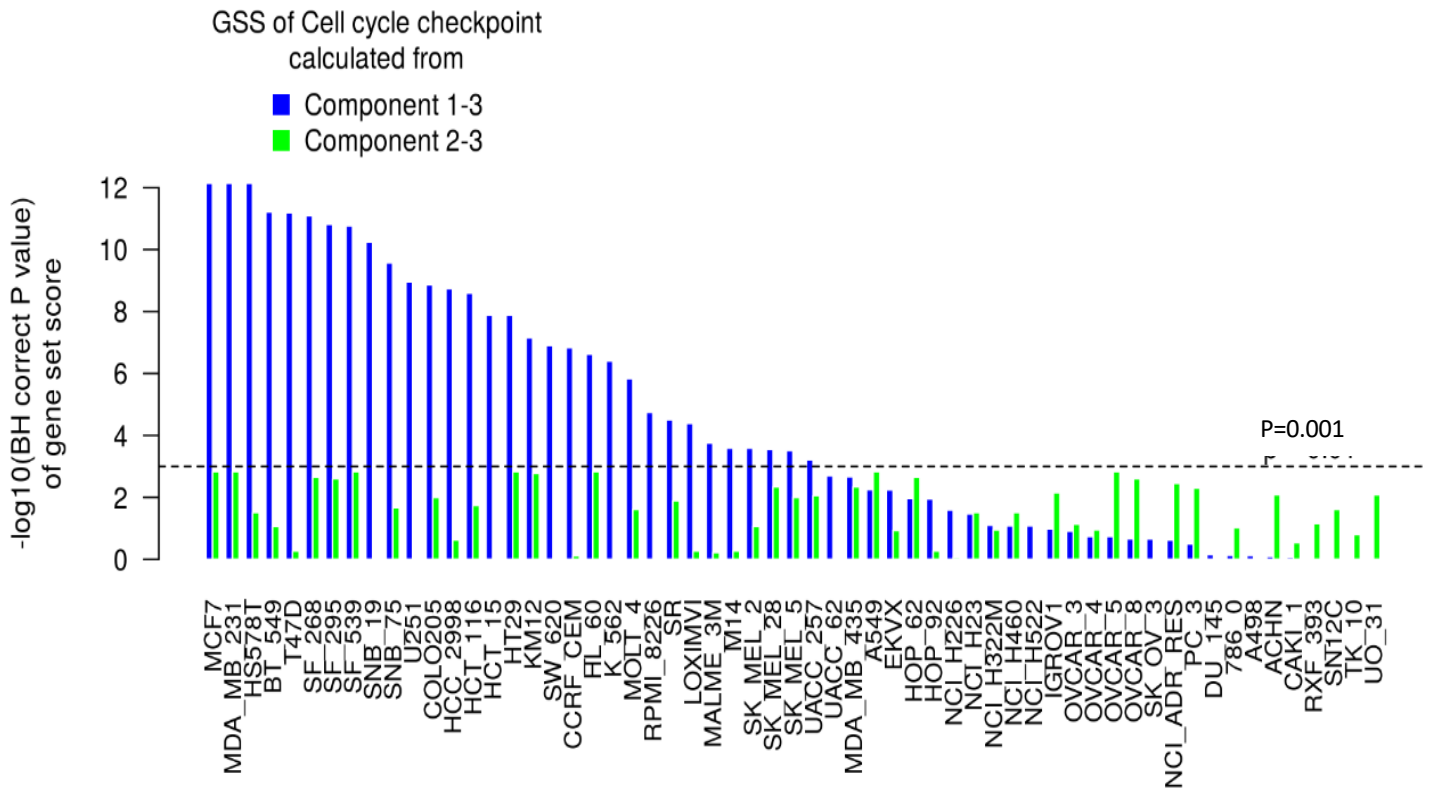
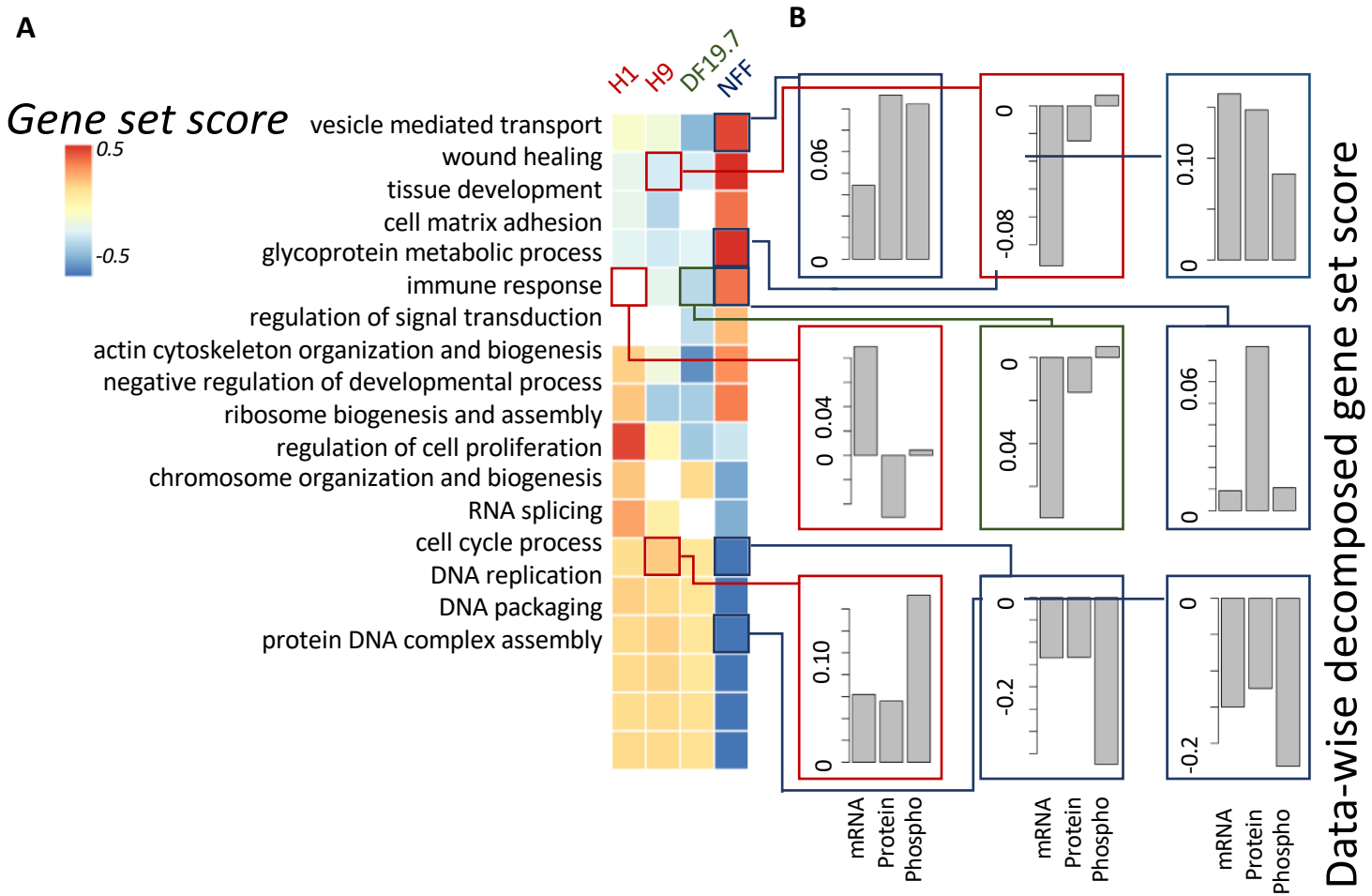


Figure 3

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

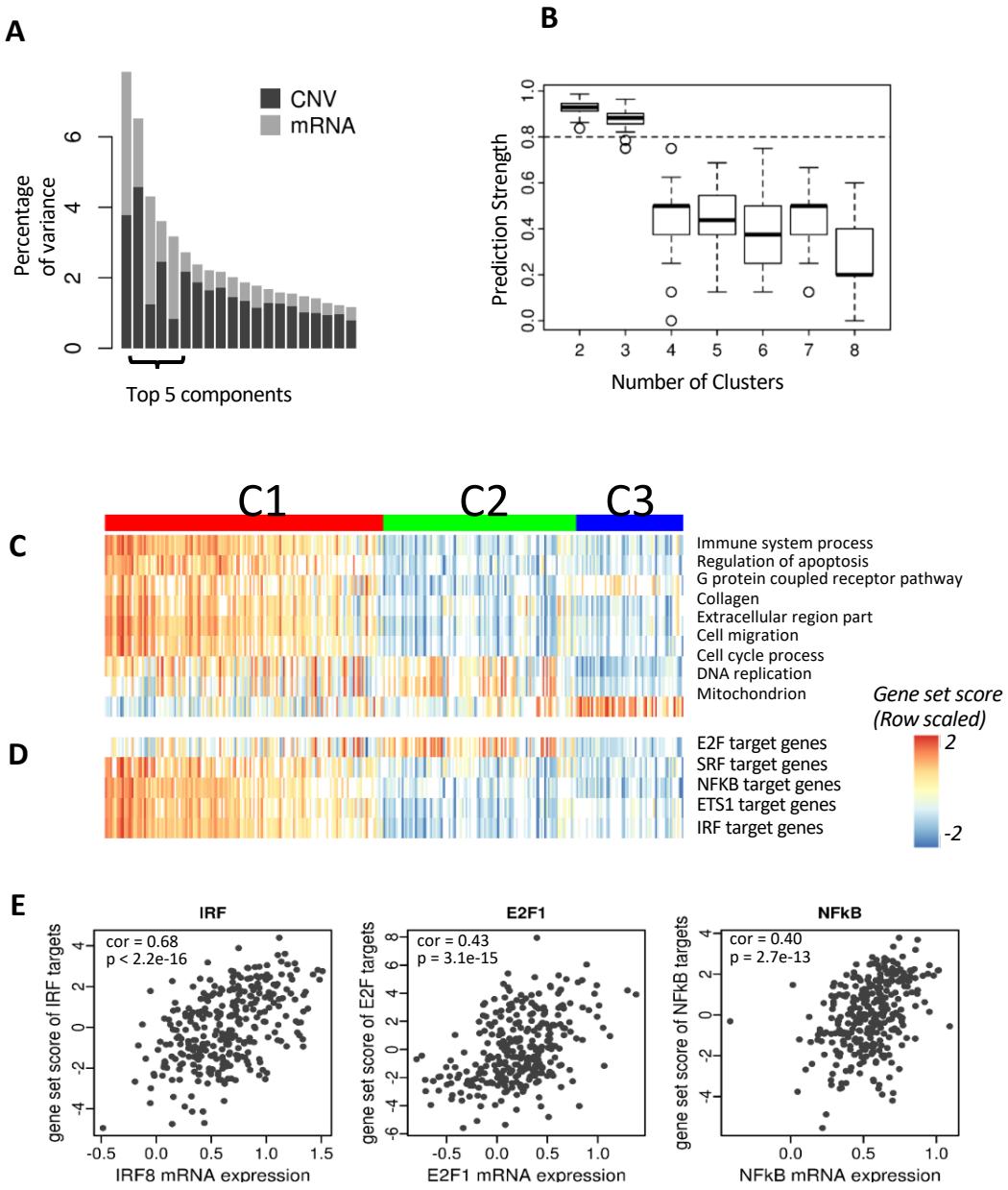


Figure 5

Cluster discovery in BLCA TCGA data

Extracting moGSA genesets associated with each cluster, with CNV or mRNA

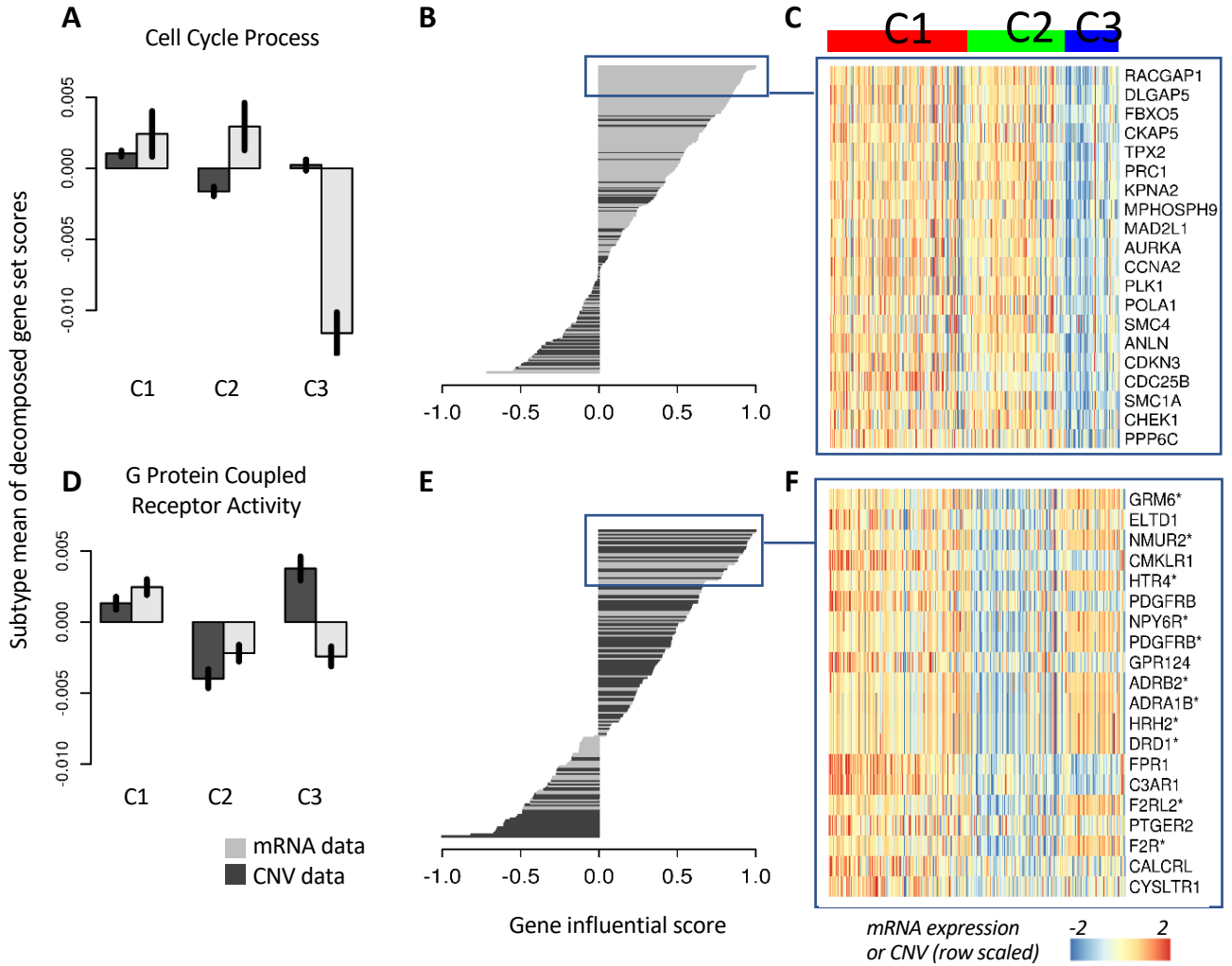


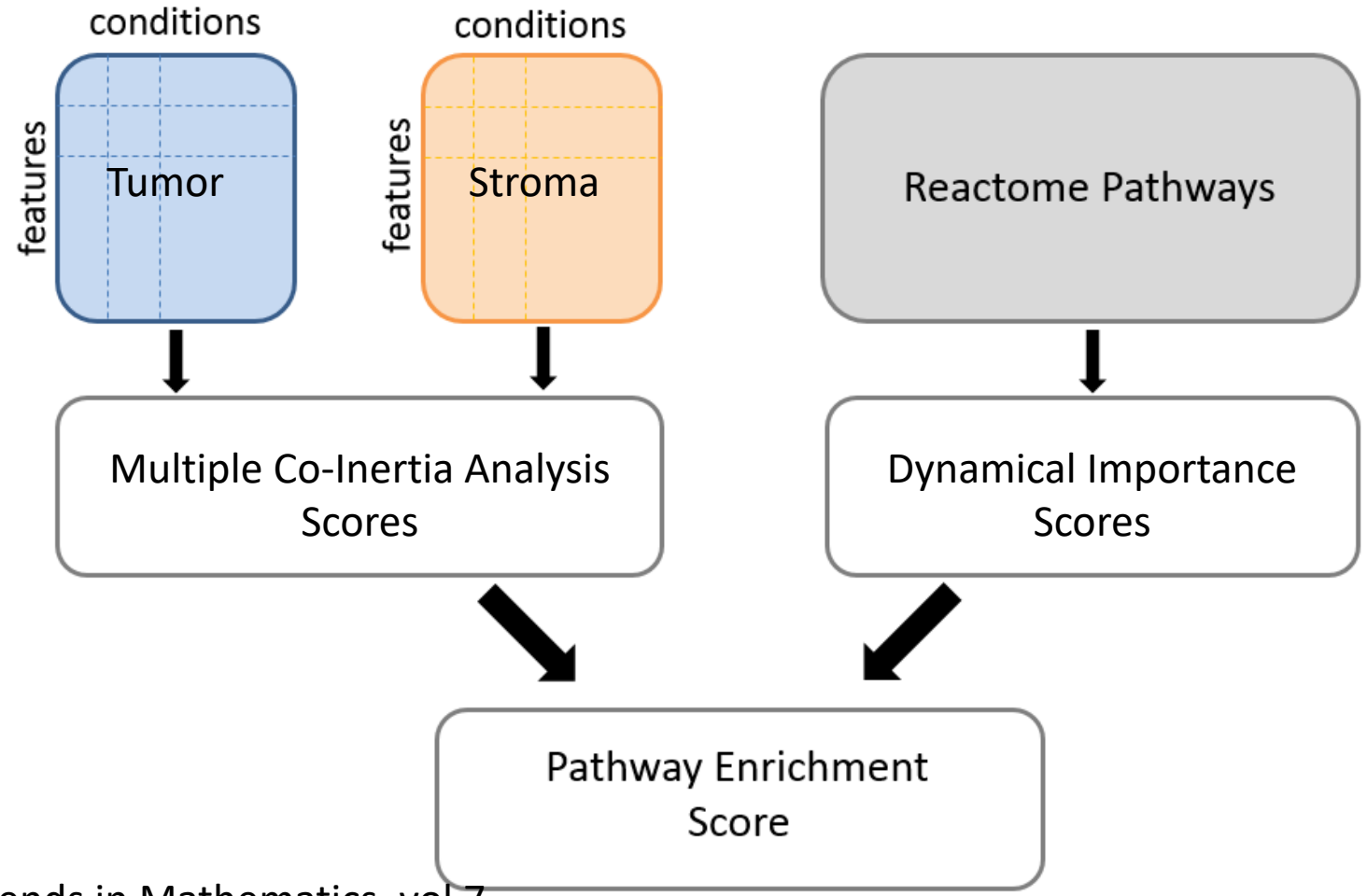
Figure 6

Summary: multiple dataset integration

1. Unsupervised..
2. Can extract feature scores in data with unknown or complex phenotypes.
3. Integrates multiple dataset summarizing each case (tumor sample) by groups of features.
4. Scalable to large data
5. Outperforms popular existing methods (GSVA, ssGSEA)
6. Among components, one can exclude (batch effects) or select components of interest

Integrative Pathway Enrichment Analysis (IPEA)

Oana Zeleznik



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

Code Demo:
Bioconductor moGSA

http://bioconductor.org/packages/release/bioc/vignettes/mogsa/

Acknowledgements

Chen Meng * (with Amin, Bernard)

Azfar Basunia

Daniel Gusenleitner

Matthew Schwede

Oana A. Zeleznik* (with Gerhard)

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 Hopkins)

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