Enter the Matrix: Unsupervised feature learning with matrix decomposition to discover hidden knowledge in high dimensional data

Aedin Culhane PhD

aedin@jimmy.harvard.edu


@AedinCulhane


Harvard T.H. Chan School of Public Health

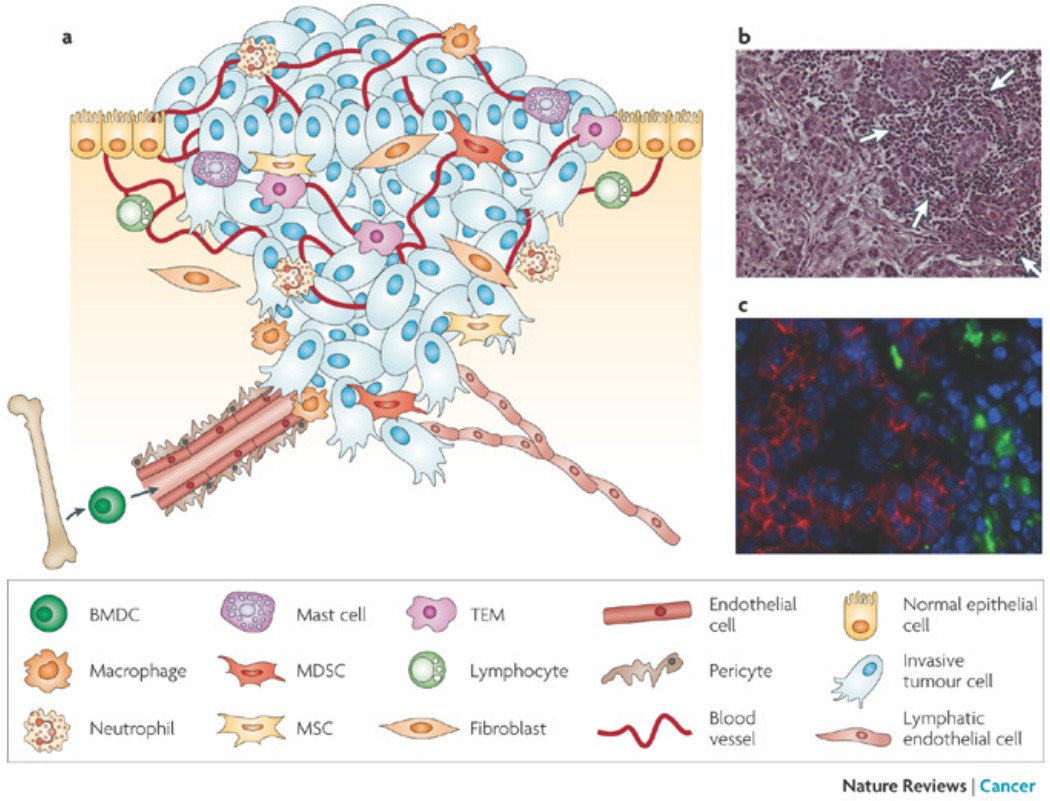

DANA-FARBER CANCER INSTITUTE


Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Overview of Talk

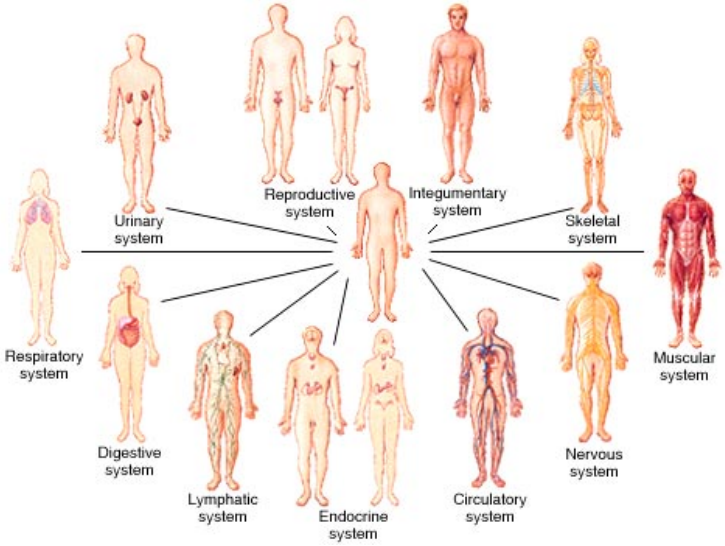
- Lab motivation –Cancer Progression
- Supervised learning - powerful tools in data science but it requires training data.
- Unsupervised dimension reduction, matrix factorization approaches including PCA, CA, ICA, NMF
- Applied to multiple data sets.
- Vignettes of applications to genomics
  - Meng & Zeleznik et al., 2016 (Briefings in Bioinformatics, 17:628, https://doi.org/10.1093/bib/bbv108)
  - Stein-O'Brien et al., 2017 (bioRxiv 196915; https://doi.org/10.1101/196915).

Cancer Microenvironment, immune cells influence tumor progression, drug response



Many cell types

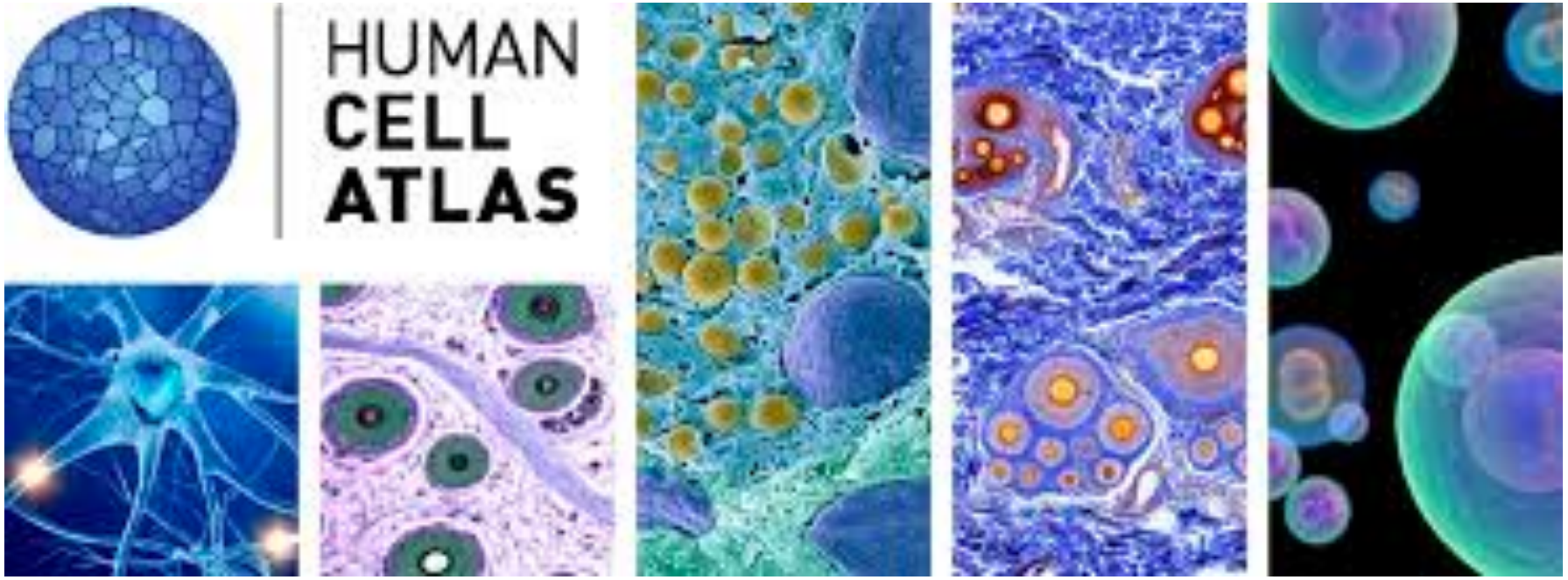
Homeostasis maintained through cell network dynamic, active, cross-talk, between local, regional and centrally regulated body systems



We are all made up of trillions of cells. Each cell in your body has its own job.

How many cells?

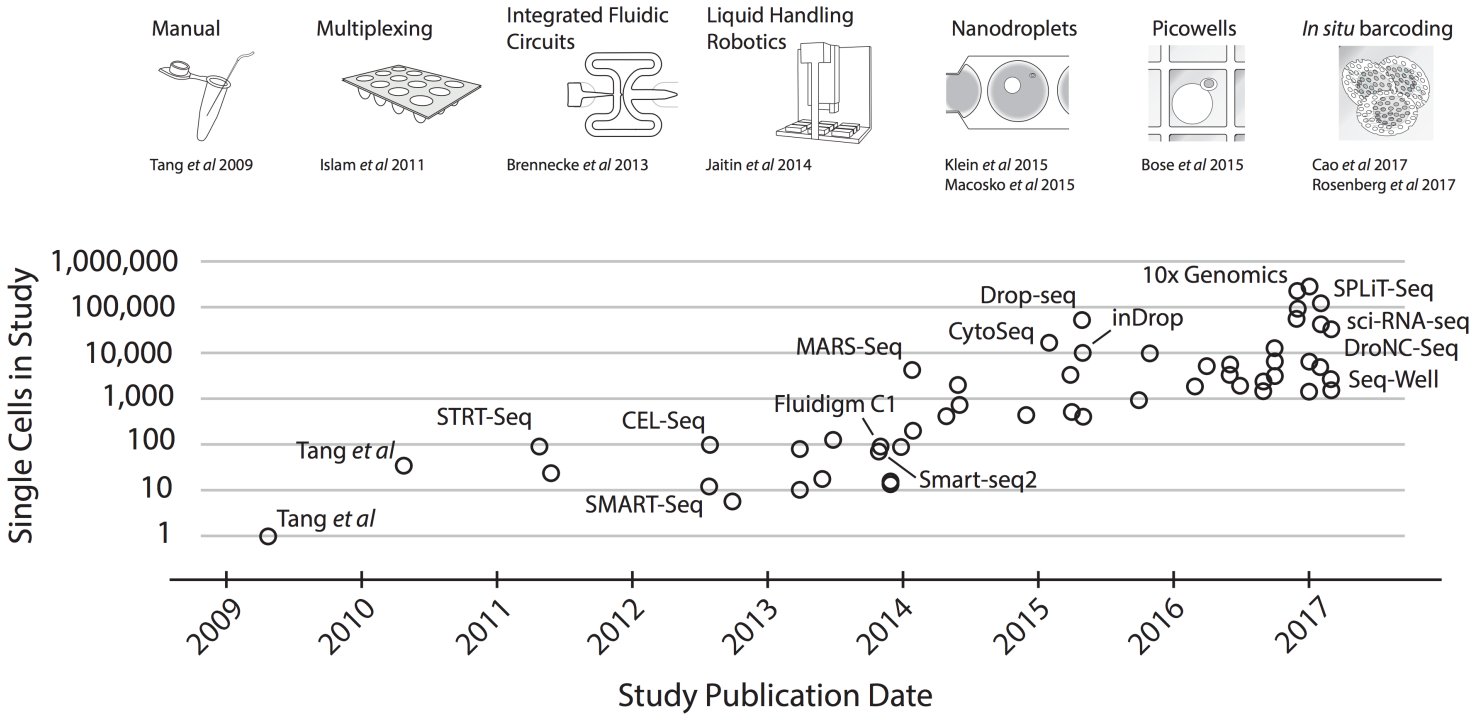
- Trillions in body (with many more Trillions more microbiota)
- 2.5 billion cells in one hand
- If every cell in your hand was the size of a grain of sand, your hand would be the size of a school bus!



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.

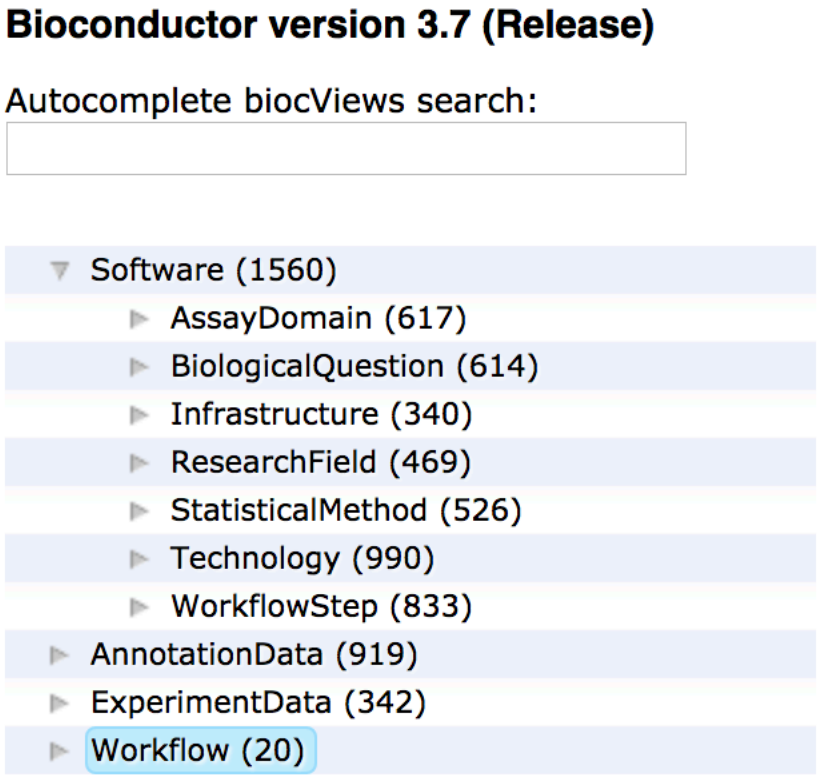
Study on single cells: Growth



Ideal Data pipeline: https://hemberg-lab.github.io/scRNA.seq.course/ideal-scrnaseq-pipeline-as-of-oct-2017.html
Data sets : https://www.10xgenomics.com/resources/datasets/, https://portals.broadinstitute.org/single_cell, http://jinglebells.bgu.ac.il/

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

- DelayedArray, restfulSE… out of memory processing
- SingleCellExperiment, Support for Loom HDF5 format
- New methods for sparsity etc, lots scRNAseq methods
- Interactive SummarizedExperiment Explorer iSEE



http://bioconductor.org

BOSTON R/Bioconductor for Genomics

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

- Meetup 4-6 times per year
- Most meetings (so far) in DFCI
- Next Meeting 18th May 2018

Bioconductor Annual Meeting
July 25-28th, Toronto, CA
http://bioc2018.bioconductor.org/

Conference Home
Registration
Scholarships
Call for Abstracts
Schedule (tentative)
Code of Conduct
Sponsor Opportunities

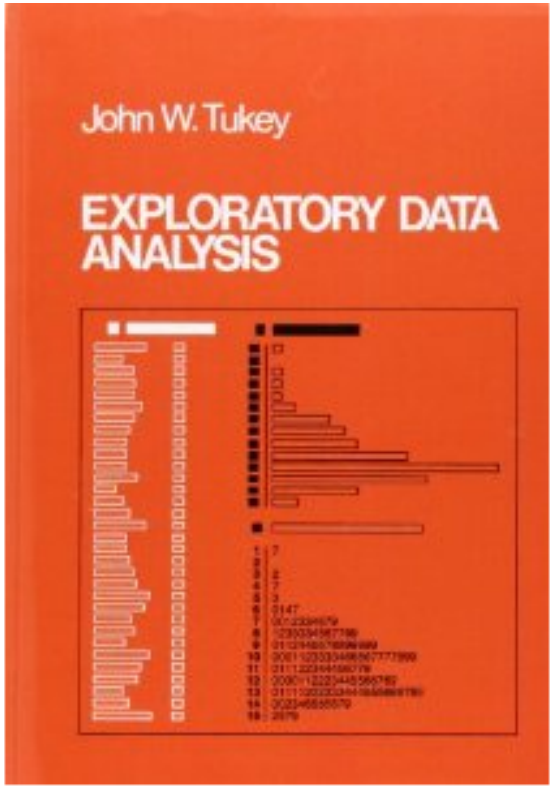
BioC 2018: Where Software and Biology Connect

When: July 25 (Developer Day), 26, and 27, 2018
Where: Victoria University, University of Toronto, Toronto, Canada

This conference highlights current developments within and beyond *Bioconductor*. Morning scientific talks and afternoon workshops provide conference participants with insights and tools required for the analysis and comprehension of high-throughput genomic data. 'Developer Day' precedes the main conference on July 25, providing developers and would-be developers an opportunity to gain insights into project direction and software development best practices.

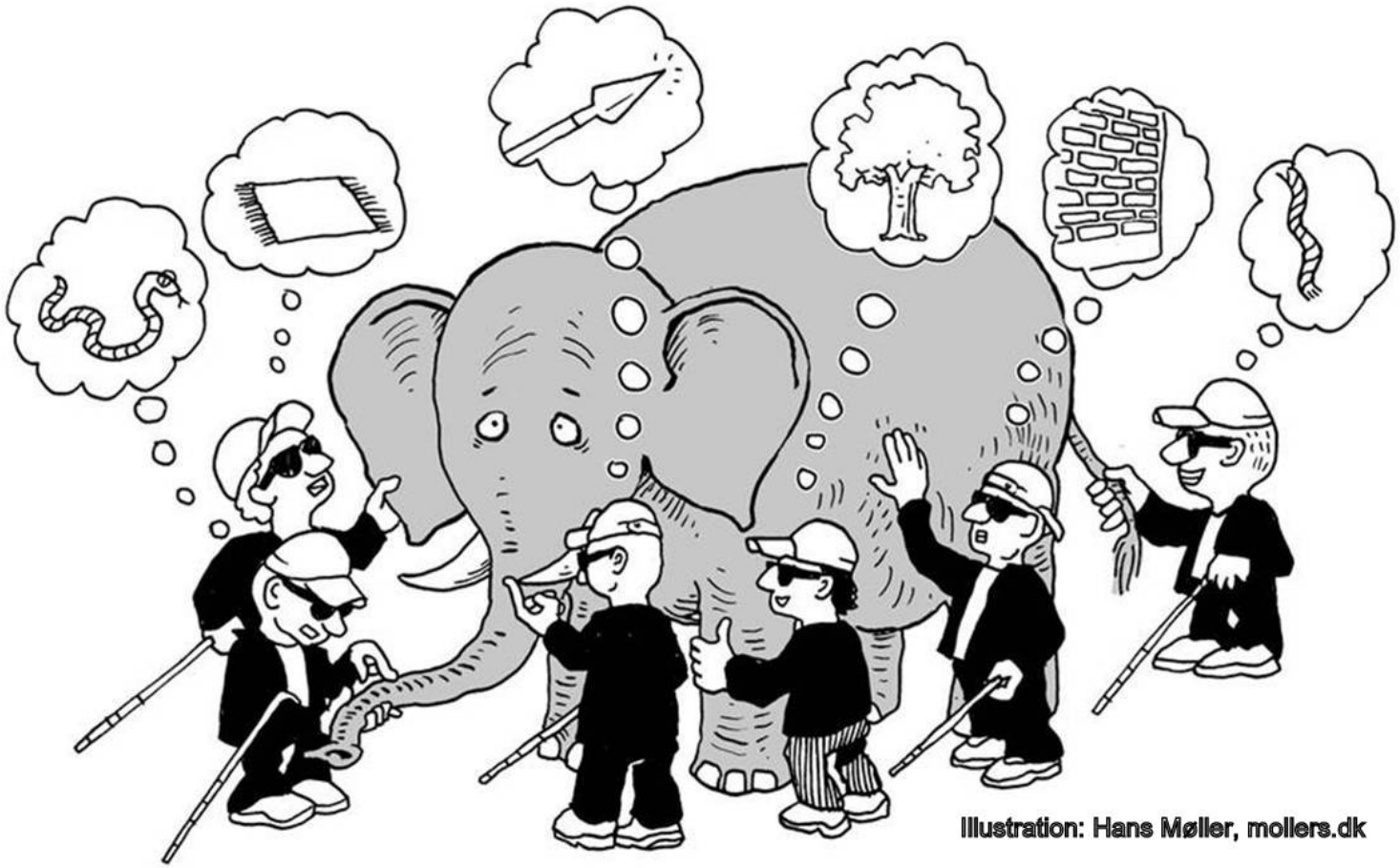
Exploratory data analysis (EDA)

> 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.

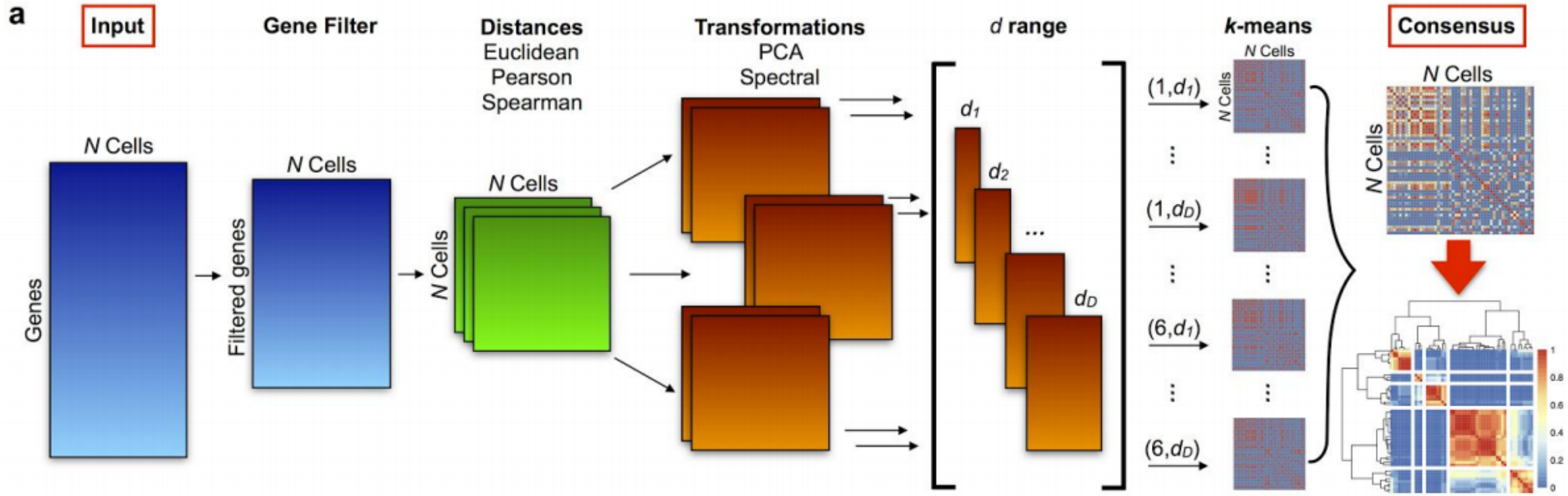


Exploratory Data Analysis [Paperback]
John W. Tukey (Author)
(8 customer reviews)

As data gets larger, it important to sample and perform EDA



Single Cell Data Analysis Pipeline

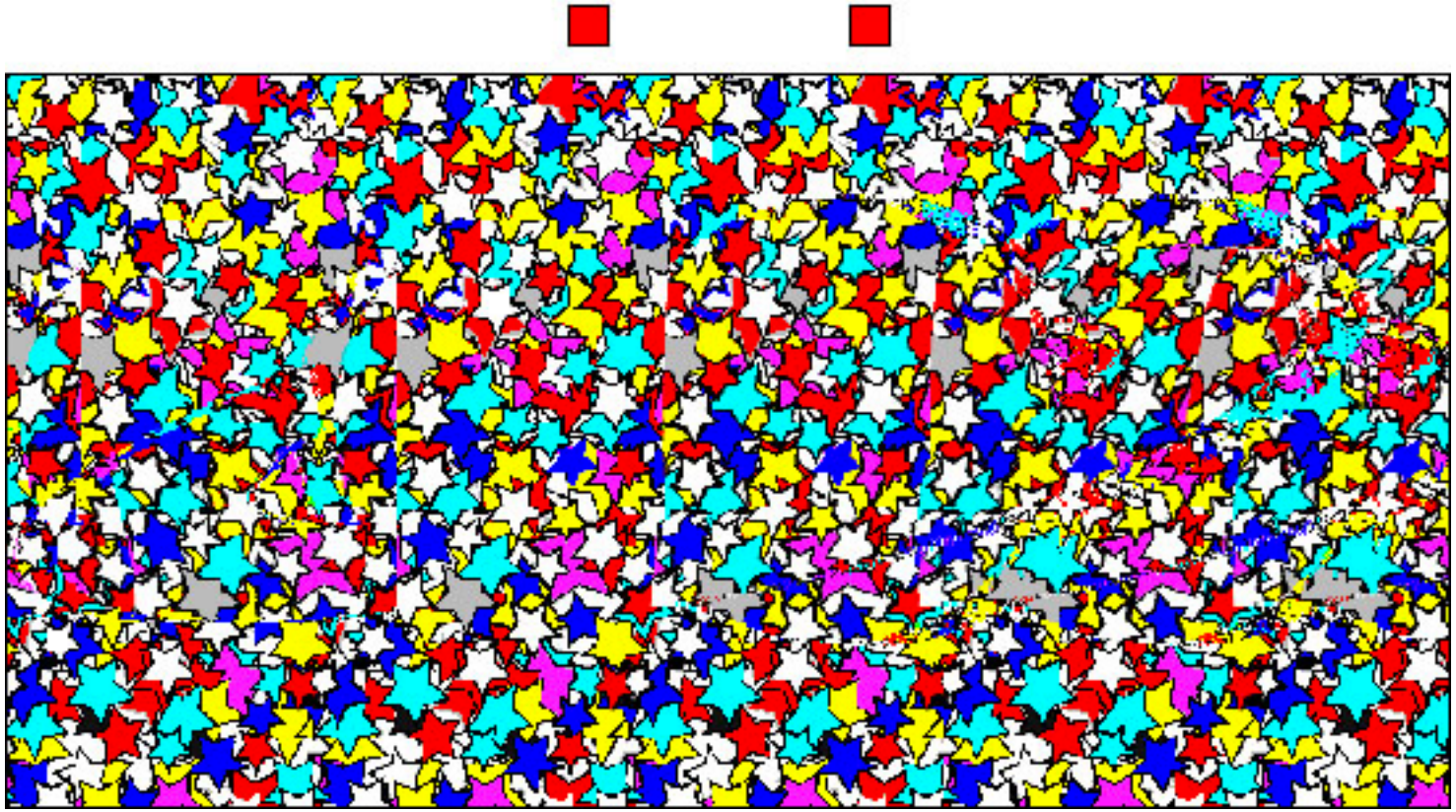


From Vladimir Kiselev (wikiselev), Tallulah Andrews, Jennifer Westoby (Jenni_Westoby), Davis McCarthy (davisjmcc), Maren Büttner (marenbuettner) and Martin Hemberg (m_hemberg)

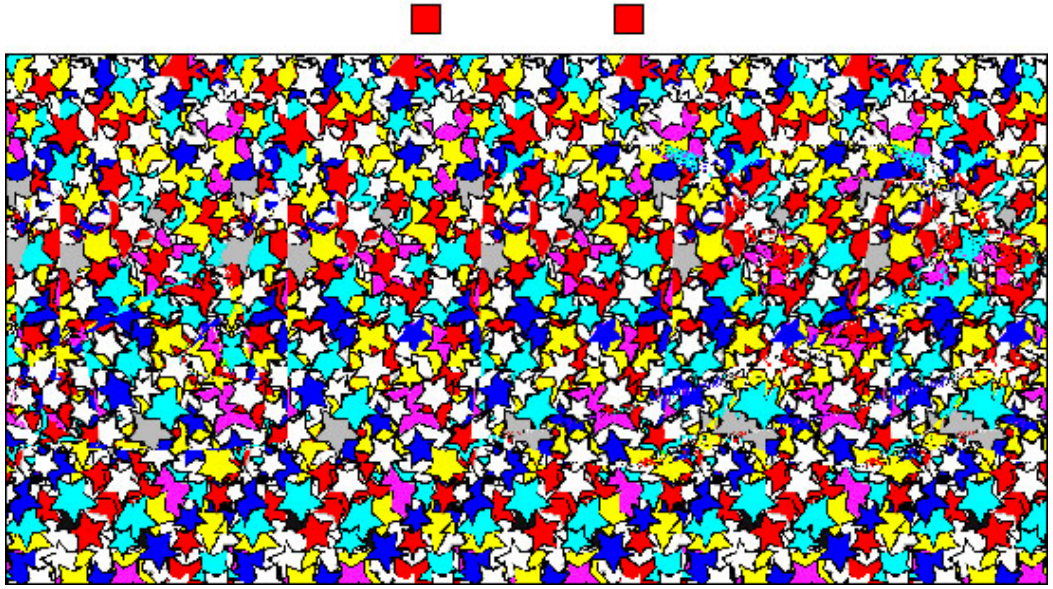
R Code, Tutorials: https://hemberg-lab.github.io/scRNA.seq.course/ideal-scrnaseq-pipeline-as-of-oct-2017.html

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

- *Latent: present but not visible, apparent.*



the answer



http://www.magiceye.com/faq_example.htm

Classical Dimension Reduction Matrix Factorization approaches

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

PCA

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.

In this case the 1-dimensional PCA subspace can be thought of as the *line* that best represents the average of the points

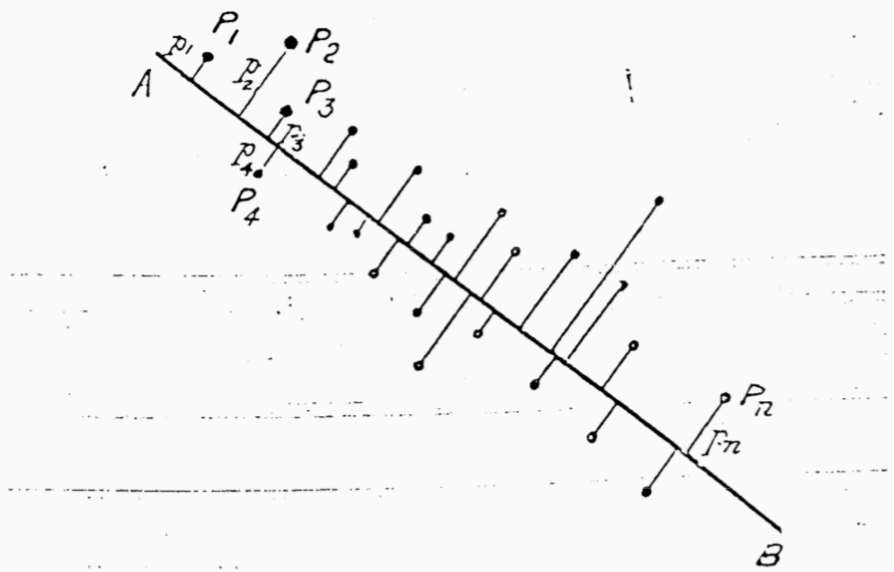
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

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

If y were the dependent variable, we should have made

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

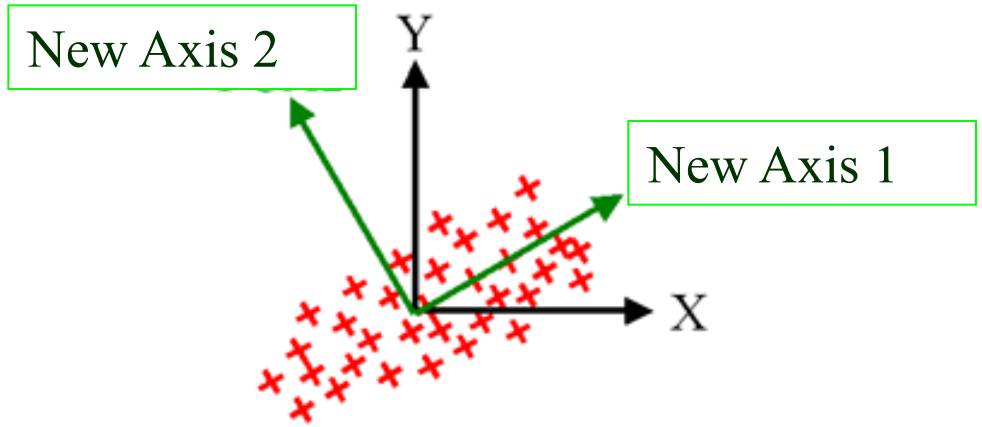
(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



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

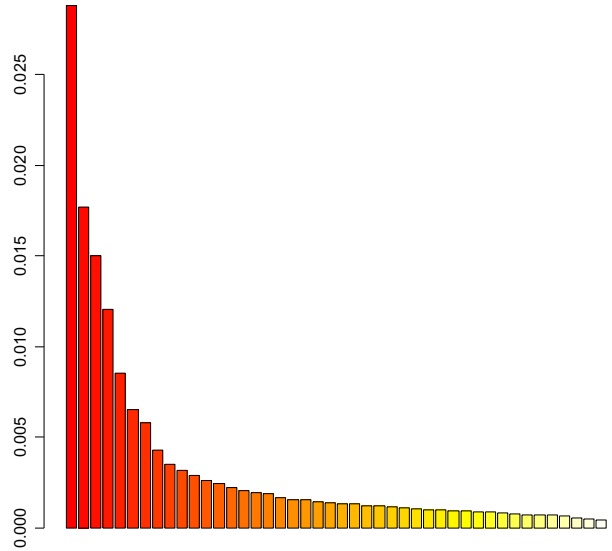
Review: Stein-O'Brien et al. *bioRxiv*. doi:10.1101/196915.

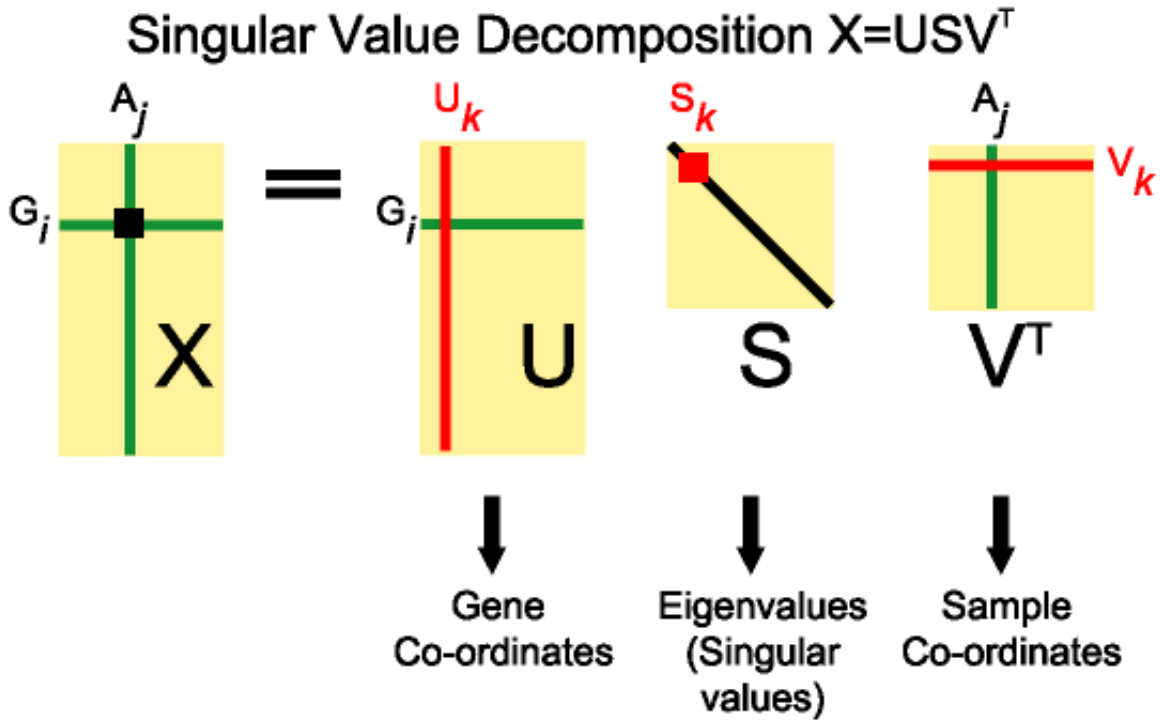
Principal Axes

- 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





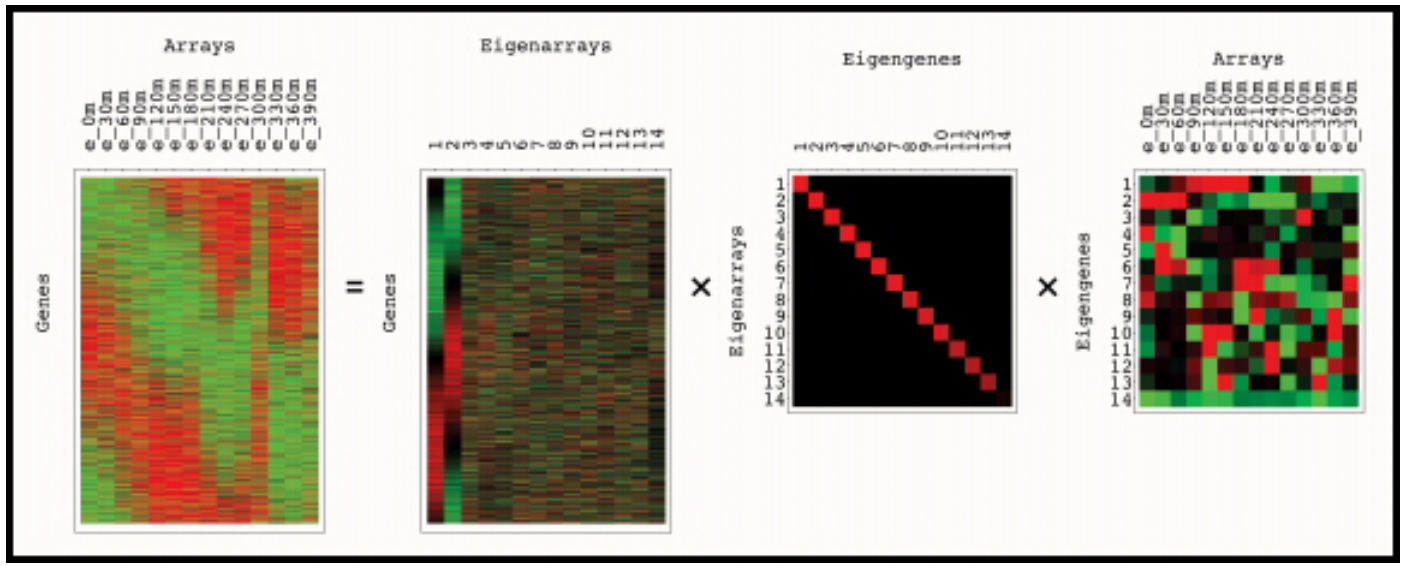


Image from

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

Considerations when applying PCA

- Distance – Euclidean
- Robust, but designed for analysis of multi-normal distributed data
- Row centre. Eliminate scale effect.
- Problems: if lots zero
- Problems: Unimodal or non-linear trends. Get distortion or artifact in plot, in which the second axis is an arched function of the first axis. Called horseshoe effect in PCA.

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.

Consideration when applying COA

- Data must be in same units (so they can be added)
- Data must be non-negative or made position by translation (scalar addition)
- In case of steep gradients (many zero) COA should produce better results than PCA
- Data is dual (column and row) scaled.
- Unimodal or non-linear trends may be represented as arch (2^{nd} axis). Less serious than PCA's horseshoe effect.

Multidimensional scaling (MDS)

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

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

Other related methods

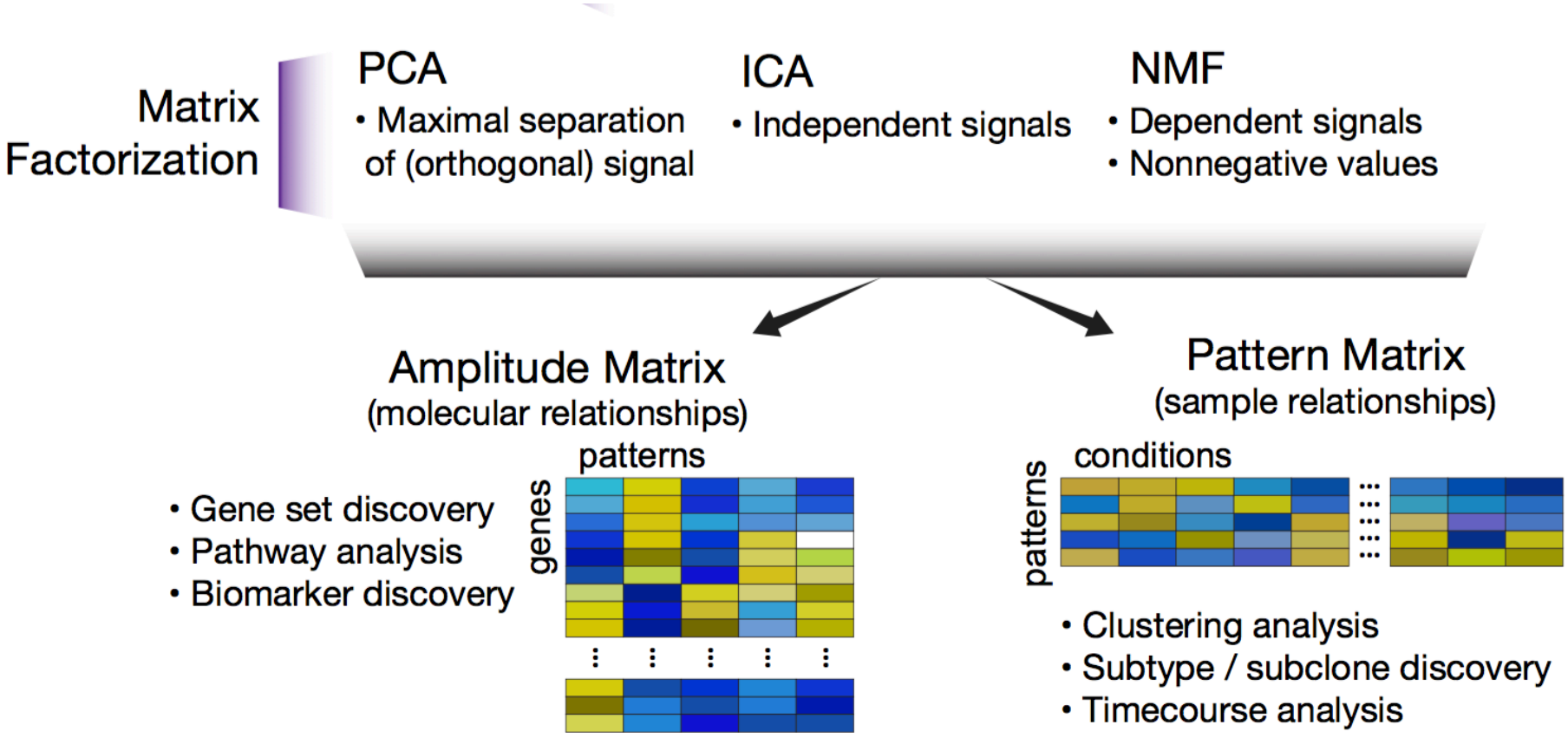
Independent Component Analysis

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

Spectral map analysis

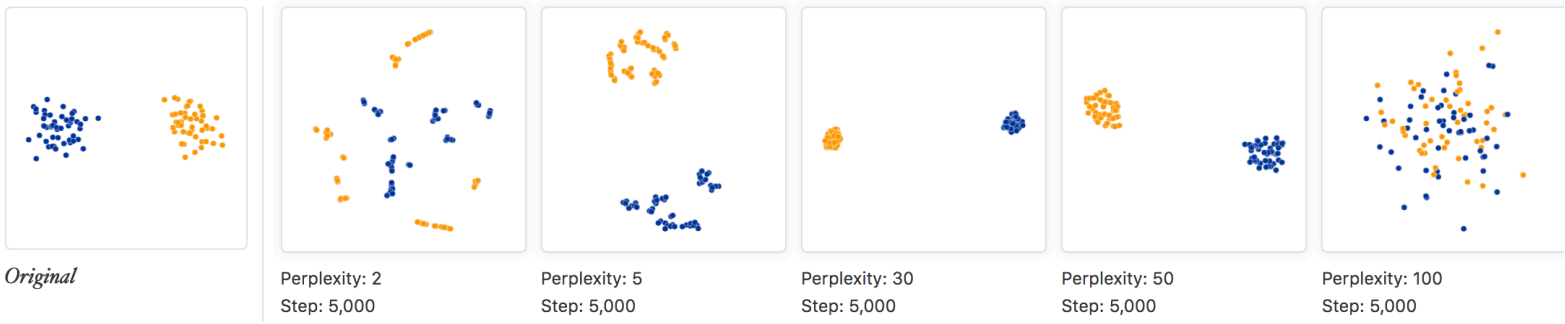
- related to COA (dual scaling of both rows + columns)
- not limited to contingency tables and cross-tabulations. possibility to use other weighting factors
- Wouters et al., 2003 showed SMA outperformed PCA, comparable to COA.

Global v Local methods



t-SNE t-distributed stochastic neighbor embedding

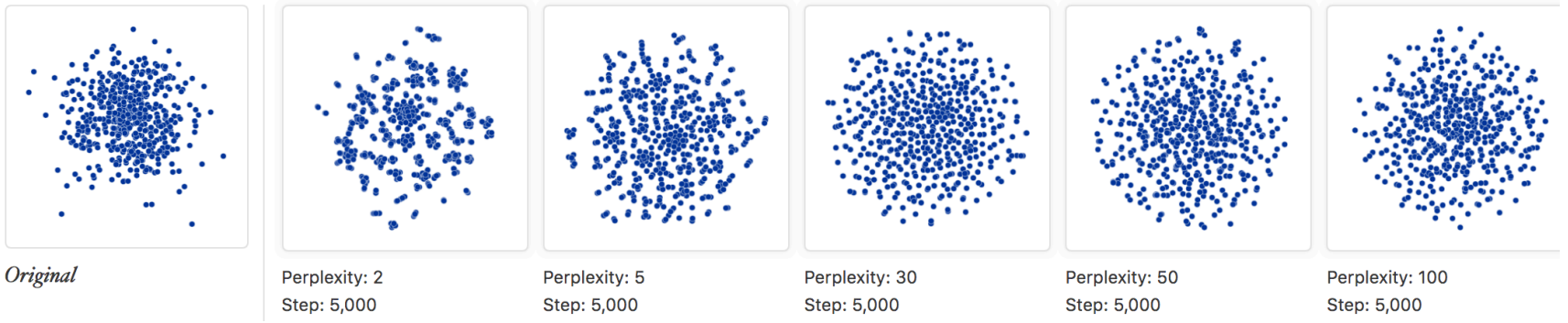
- Introduced van der Maaten and Hinton in 2008.
- non-convex objective function. objective function minimized using a gradient descent optimization. Possible that different runs give you different solutions.
- Good when large number (scale) objects as it avoids overlapping close points.
- t-SNE can "snap" 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.



Original

Perplexity: 2
Step: 5,000

Perplexity: 5
Step: 5,000

Perplexity: 30
Step: 5,000

Perplexity: 50
Step: 5,000

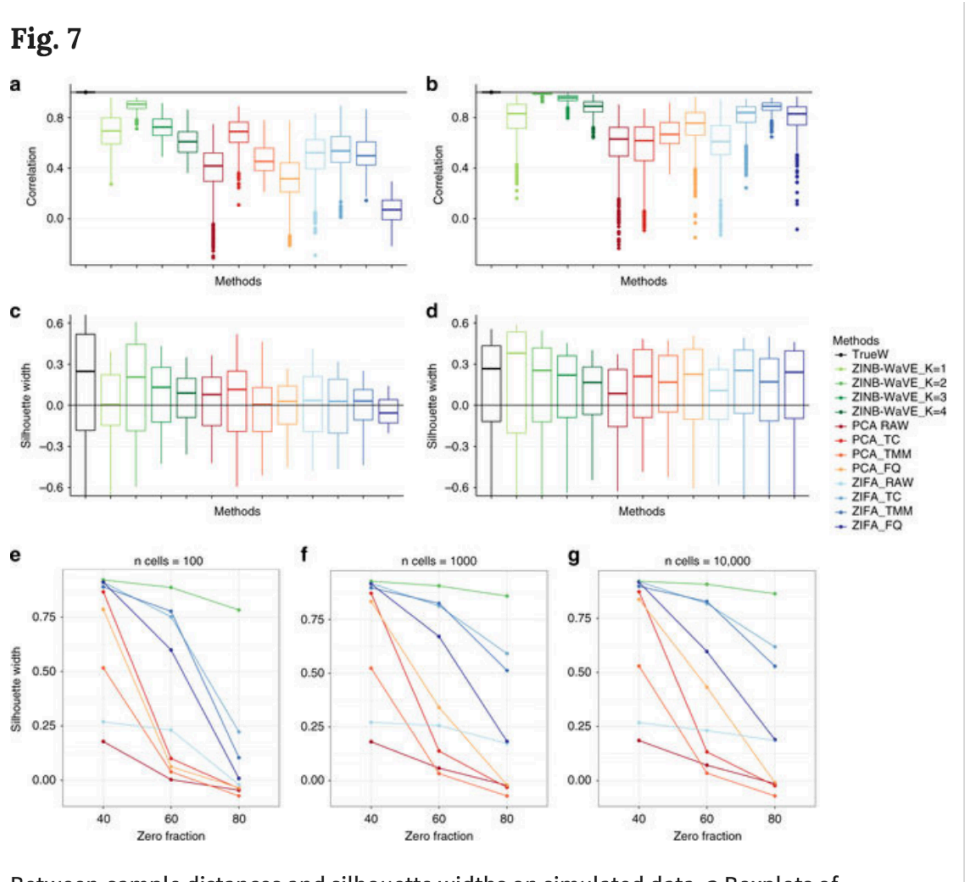
Perplexity: 100
Step: 5,000

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

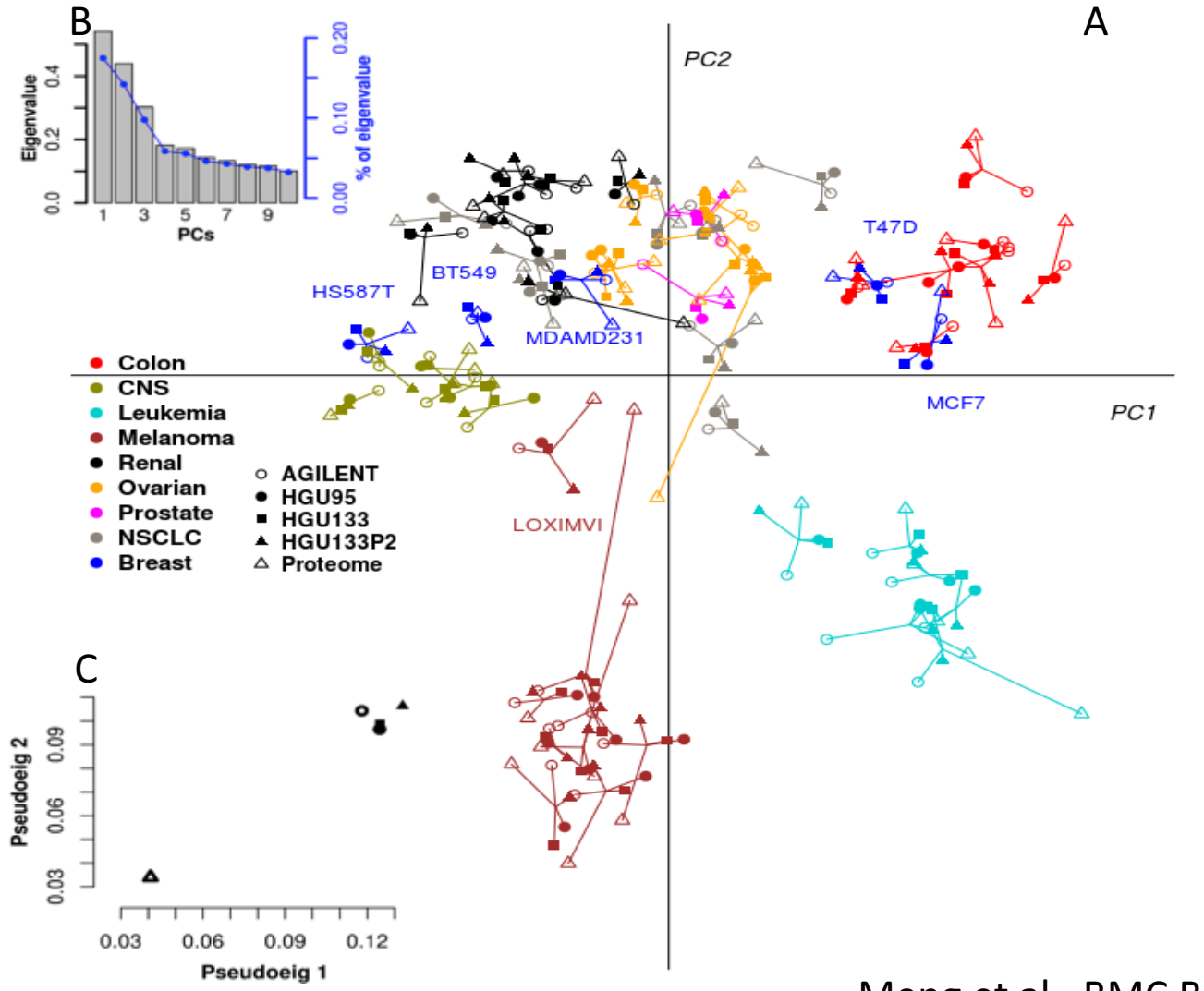
>2 datasets : Tensor data integration

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}

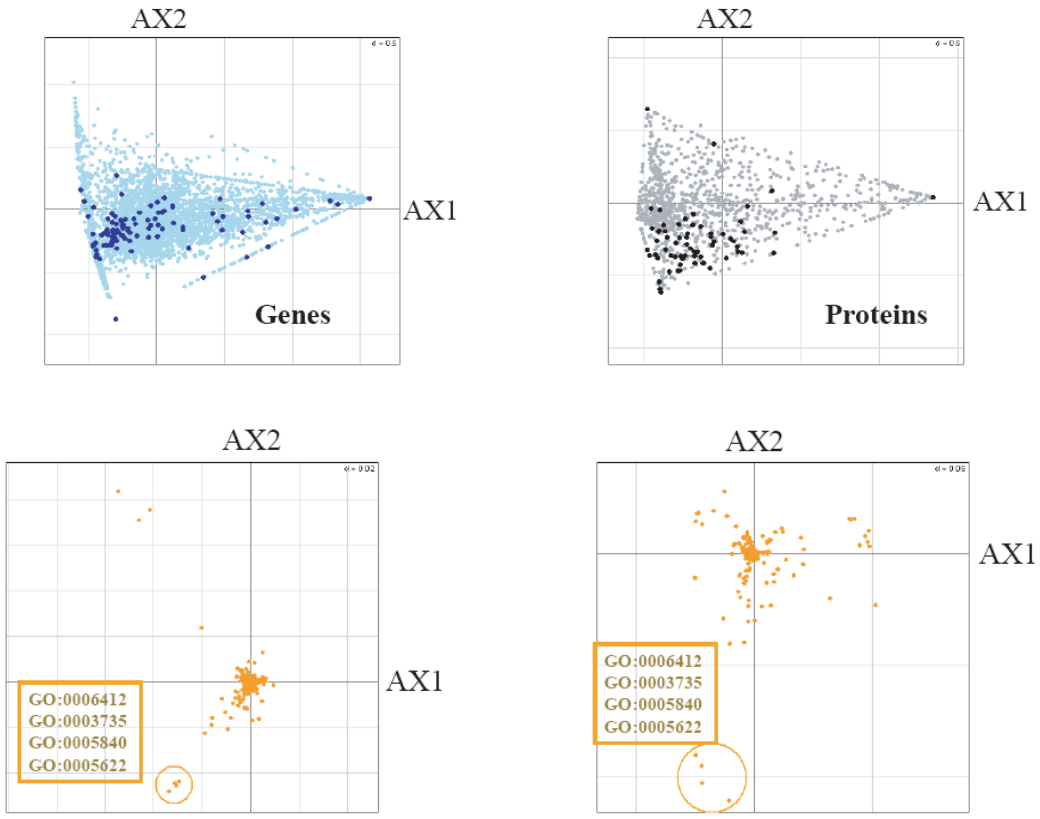
Meng & Zeleznik et al., (2016) Dimension reduction techniques for the integrative analysis of multi-omics data. Briefings in Bioinformatics, 17(4), 2016, 628–641

Tensor Integration of 5 data sets (NCI60) using multi-CIA



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

Group features.. Project extra information onto axis

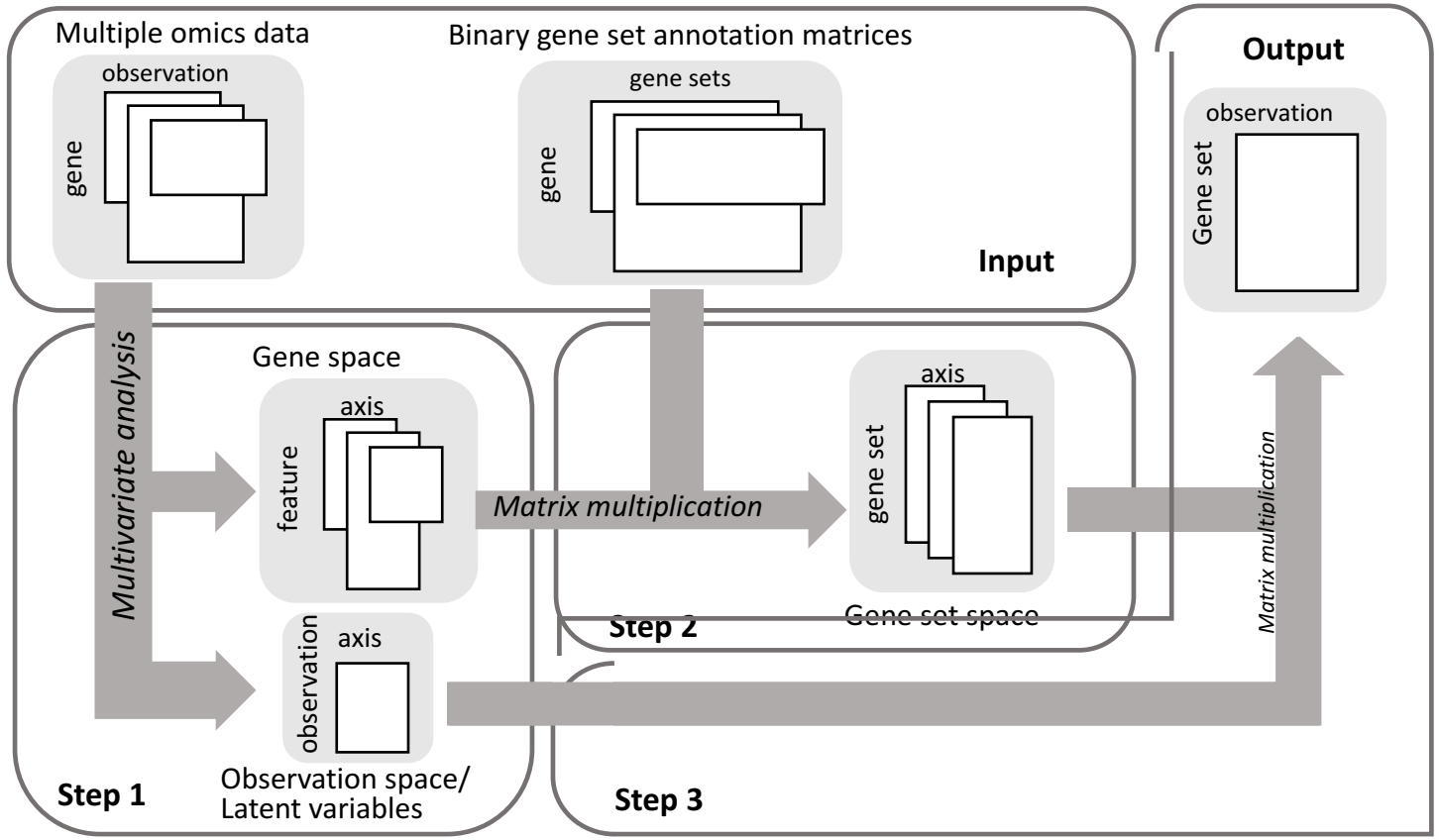


Matrix decomposition of gene expression and proteomics onto same scale

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.

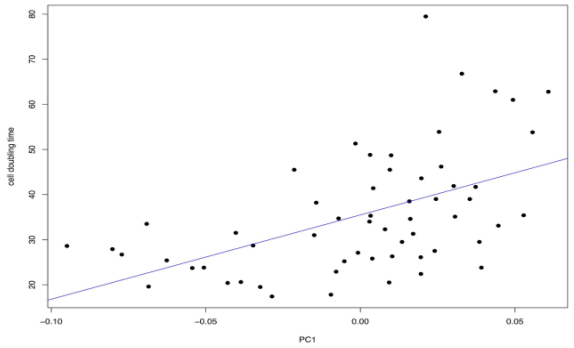
moGSA integrating complex data

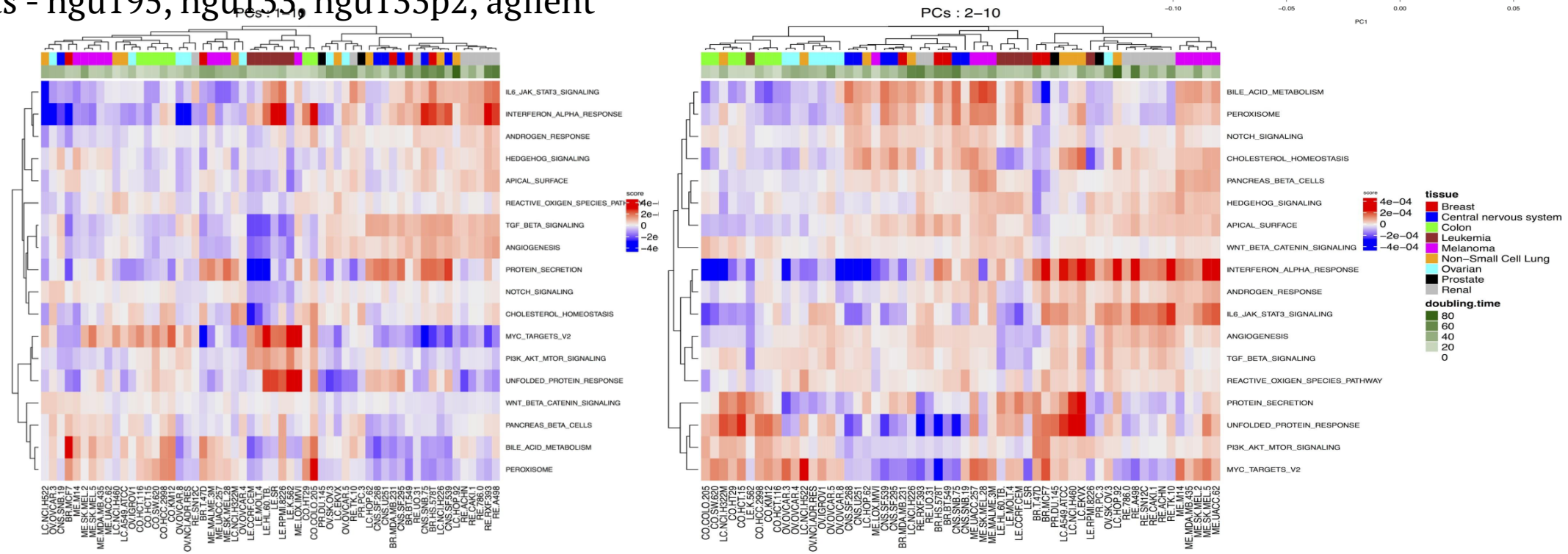
Retrieve ssGSA scores on each component.

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

58 NCI60 samples and 18 Hallmark Genesets

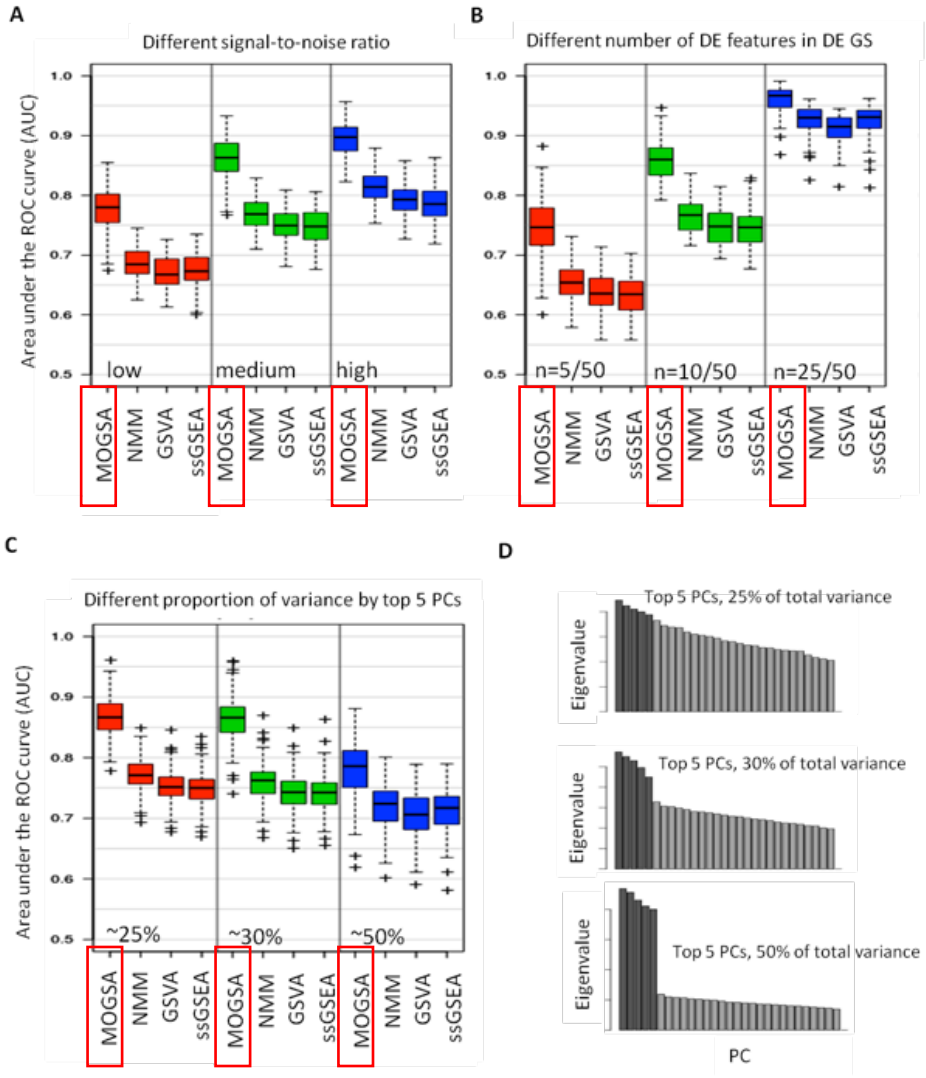
4 datasets - hgu195, hgu133, hgu133p2, agilent





moGSA single sample Gene Set analysis

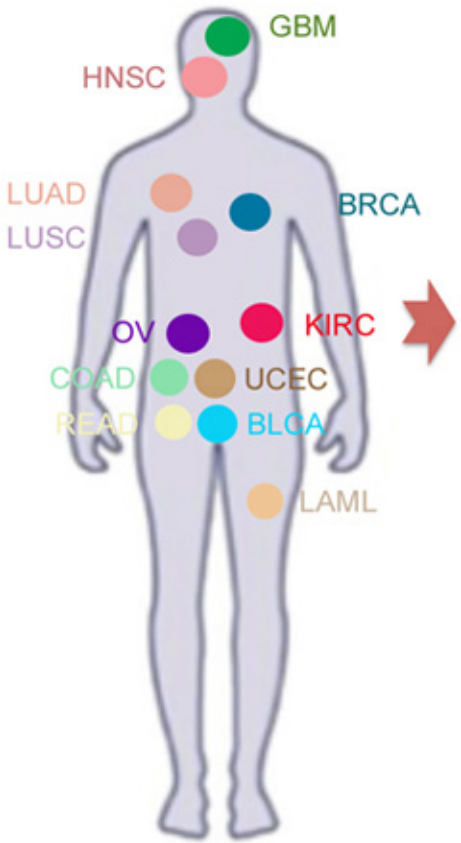
MOGSA outperforms other ssGSA approaches when applied to Synthetic data

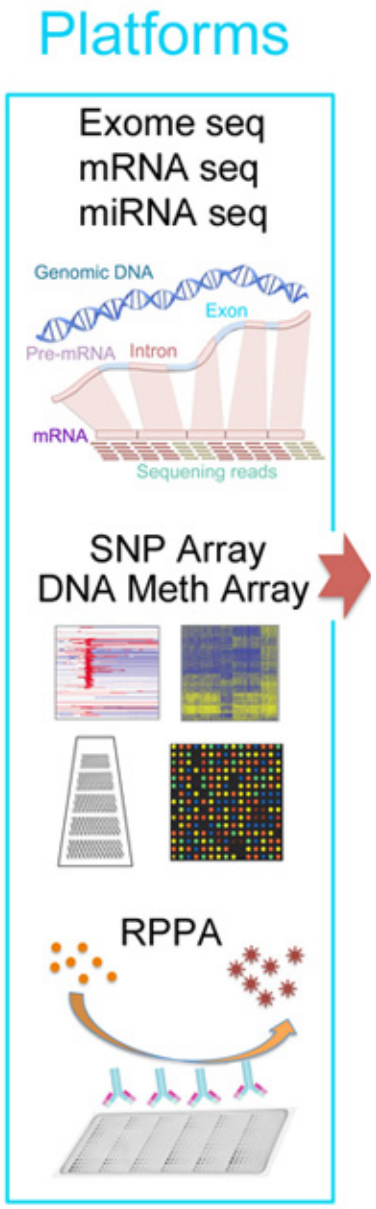


Meng C et al.,. moGSA: Integrative Single Sample Gene-Set Analysis of Multiple Omics Data. *bioRxiv*, 046904.

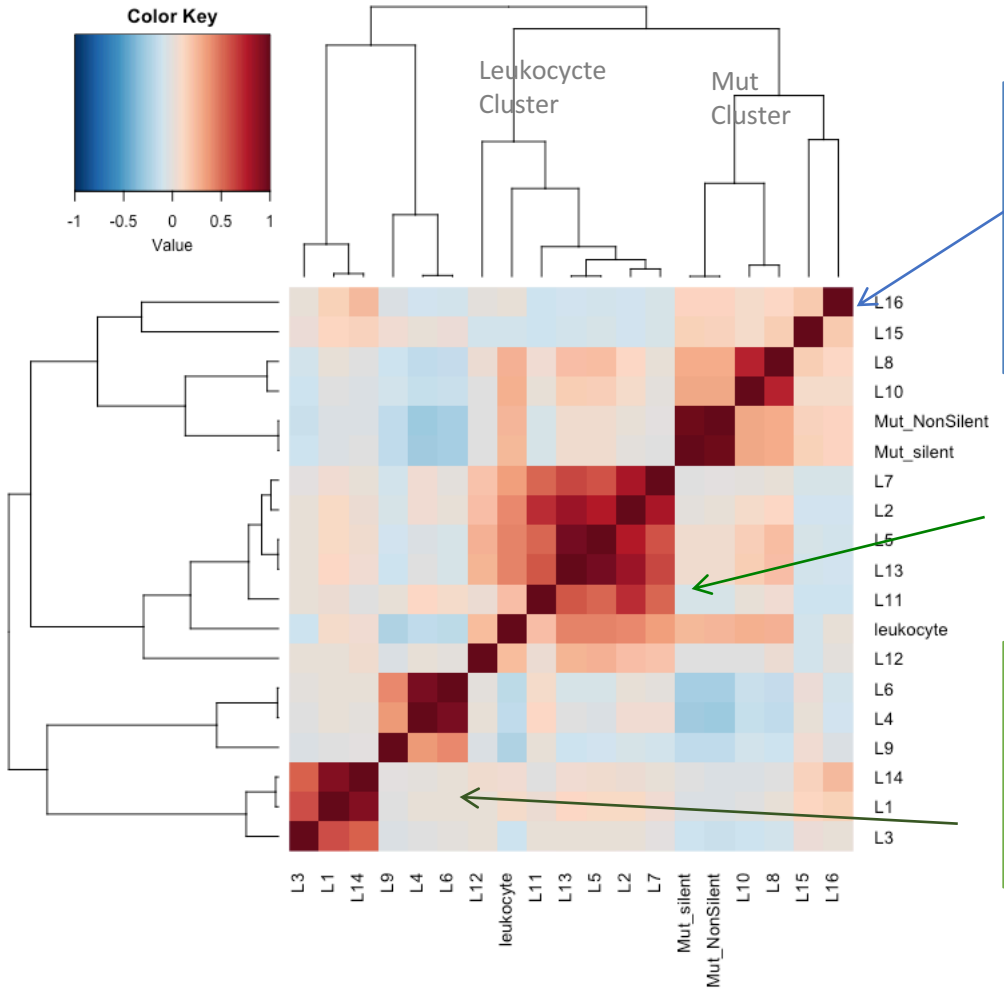
Application of moGSA to finding PanCancer Immune subtypes

- 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





Correlation between 16 Clusters, leucocyte fraction and mutation load



L8 and L10 correlated with Mutation load and leukocytes Fraction

L15, L16 – Mutation but not leukocytes

Clusters correlated to leukocyte fraction but weak to mutation load

2 clusters negatively correlated with mutation load. Very weak correlation to leukocyte fraction

Summary: multiple dataset integration

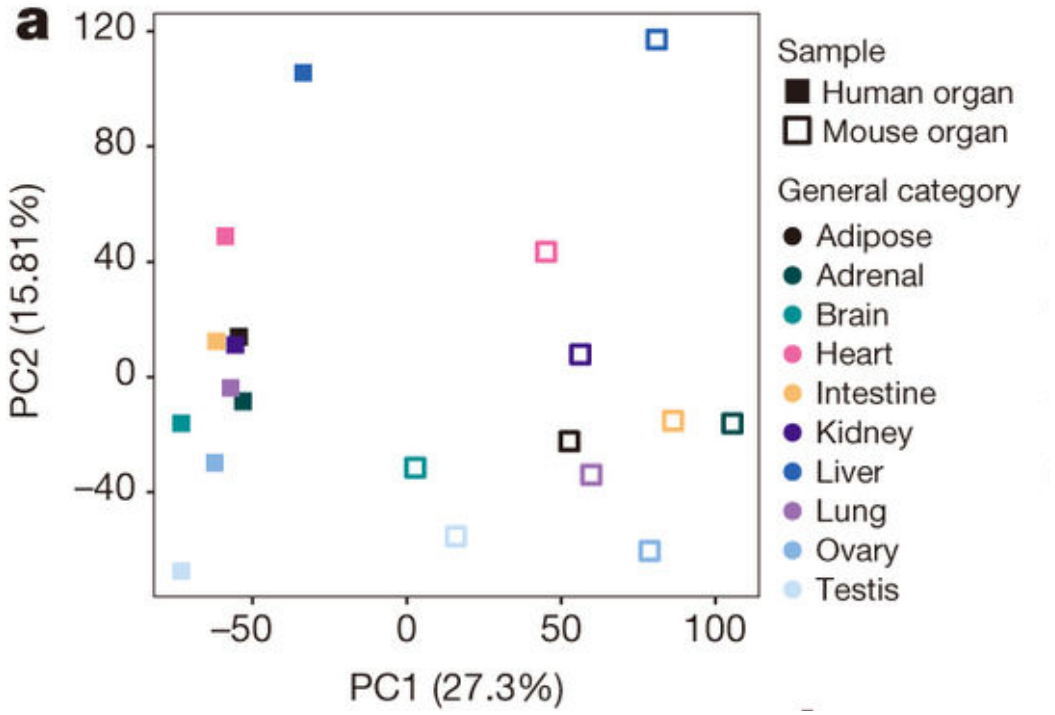
1. Unsupervised..
2. Can extract feature scores in data with unknown or complex phenotypes.
3. Integrates multiple dataset summarizing each case (tumor sample) by groups of features.
4. Scalable to large data
5. Outperforms popular existing methods (GSVA, ssGSEA)
6. Among components, one can exclude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

Final Take Away : Expect the unexpected

"Exploratory data analysis' is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there." (p. 806, Tukey)

ENCODE

Nature **volume515**, pages355–364 (20 November 2014)



A comparative encyclopedia of DNA elements in the mouse genome

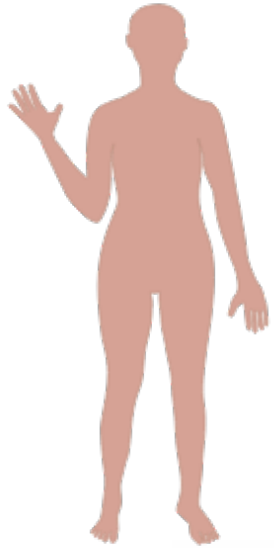
Divergent and conserved gene expression patterns

Previous studies have revealed remarkable examples of species-specific gene expression patterns that underlie phenotypic changes during evolution[38,39,40,41,42]. In these cases changes in expression of a single gene between closely related species led to adaptive changes. However, it is not clear how extensive the changes in expression patterns are between more distantly related species, such as mouse and human, with some studies emphasizing similarities in transcriptome patterns of orthologous tissues[43,44,45] and others emphasizing substantial interspecies differences[46]. Our initial analyses revealed that gene expression patterns tended to cluster more by species rather than by tissue (Fig. 2a). To resolve the sets of genes contributing to different components in the clustering, we employed variance decomposition

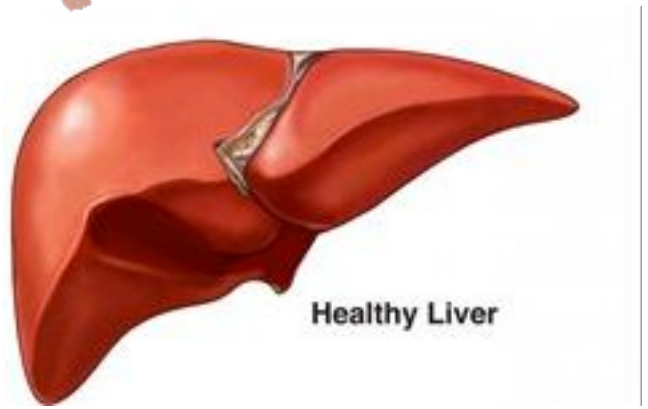
https://rmflight.github.io/post/mouse-human-transcriptomics-and-batch-effects/

Tissues or Species (mouse, human)

species

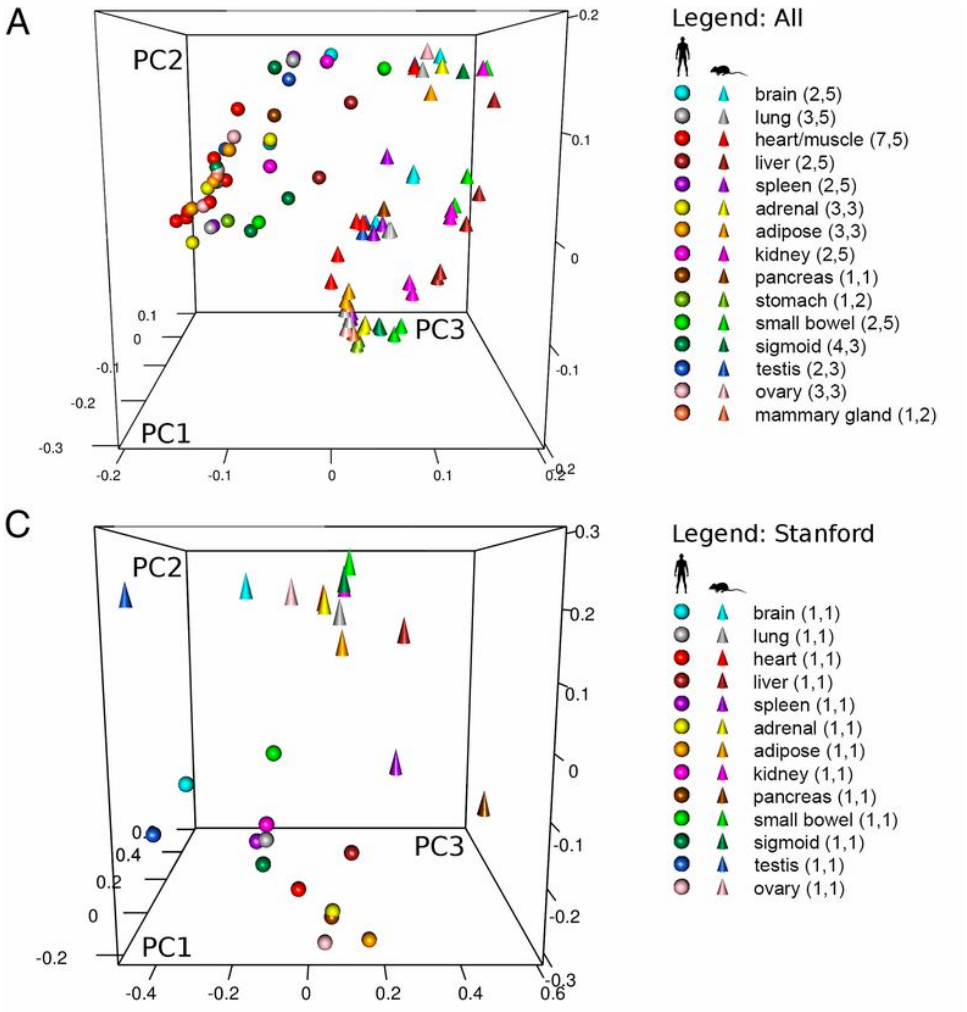


liver



ENCODE

Lin et al., PNAS December 2, 2014. 111 (48) 17224-17229 https://doi.org/10.1073/pnas.1413624111

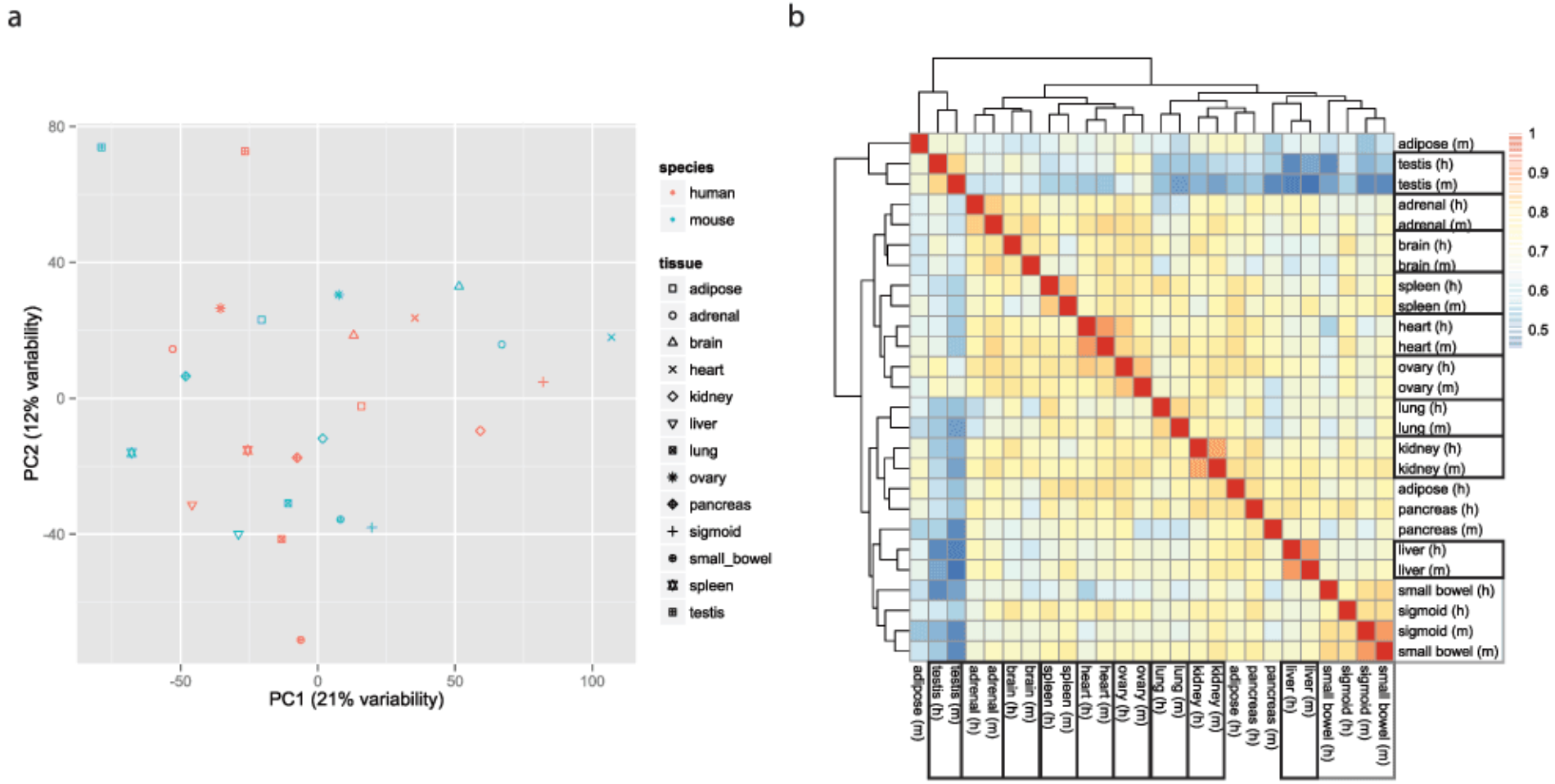


Methods). In contrast to what was reported previously (**1**, **2**, **5**), surprisingly, we found that the mouse and human samples cluster by species when the data are projected onto the first three principal components (**Fig. 1*A***). Because the same tissues of the same species produced by different laboratories did not cluster together, the possibility of methodologic differences among laboratories confounding our results was considered. To address this issue, analysis of only the 13 paired samples processed under one experimental protocol yielded the same species-specific clustering (**Fig. 1*C***). The same species-specific clustering was observed when other combinations of 10 or more tissues were examined, indicating that the clustering is not due to the particular 13–15 tissues selected. Finally, different normalization methods (e.g., quantile normalization) applied to the data produced similar groupings.

But Yoav Gilad https://f1000research.com/articles/4-121/v1,

D87PMJN1 (run 253, flow cell D2GUAACXX, lane 7)	D87PMJN1 (run 253, flow cell D2GUAACXX , lane 8)	D4LHBFN1 (run 276, flow cell C2HKJACXX , lane 4)	MONK (run 312, flow cell C2GR3ACXX , lane 6)	HWI-ST373 (run 375, flow cell C3172ACXX , lane 7)
heart	adipose	adipose	heart	brain
kidney	adrenal	adrenal	kidney	pancreas
liver	sigmoid colon	sigmoid colon	liver	brain
small bowel	lung	lung	small bowel	spleen
spleen	ovary	ovary	testis	Human
testis		pancreas		Mouse

Clustering of data after correcting for batch effect



Gilad Y and Mizrahi-Man O. A reanalysis of mouse ENCODE comparative gene expression data [version 1]. F1000Research 2015, 4:121 (doi: 10.12688/f1000research.6536.1)

F1000Research

Final Take Away : EXPLORE your data

Acknowledgements

Azfar Basunia

Daniel Gusenleitner

Chen Meng

Matthew Schwede

Oana A. Zeleznik

- **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 (Univ Toronto)

David Livingston (DFCI)
David Harrington (DFCI)
John Quackenbush (DFCI)


With You May the Fourth Be!