Rx Pharma

Friday 23rd August 2019

Multi-modal data integration

Aedín Culhane PhD

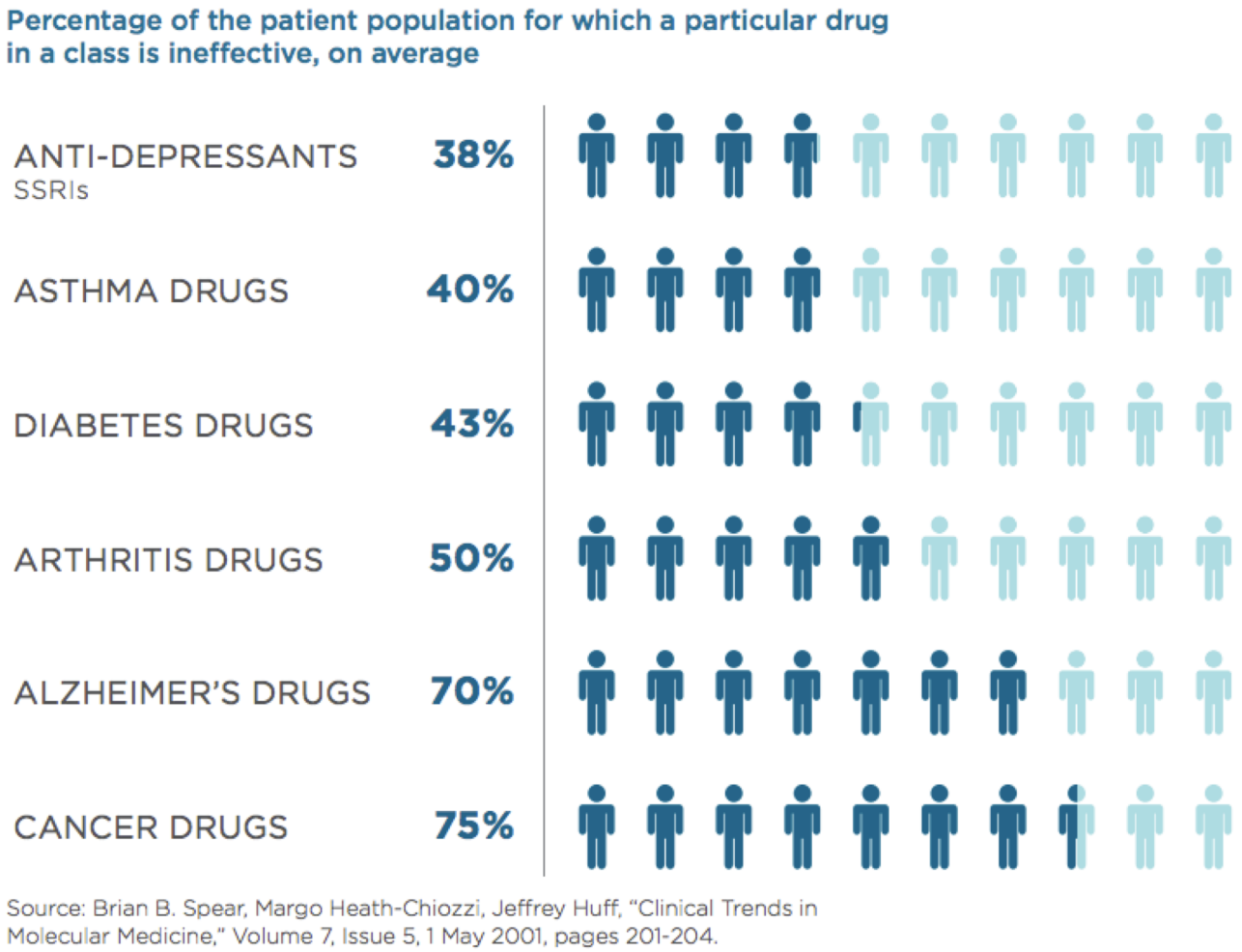
@AedinCulhane

Harvard T.H. Chan School of Public Health

DANA-FARBER CANCER INSTITUTE

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

One “model” does not fit all



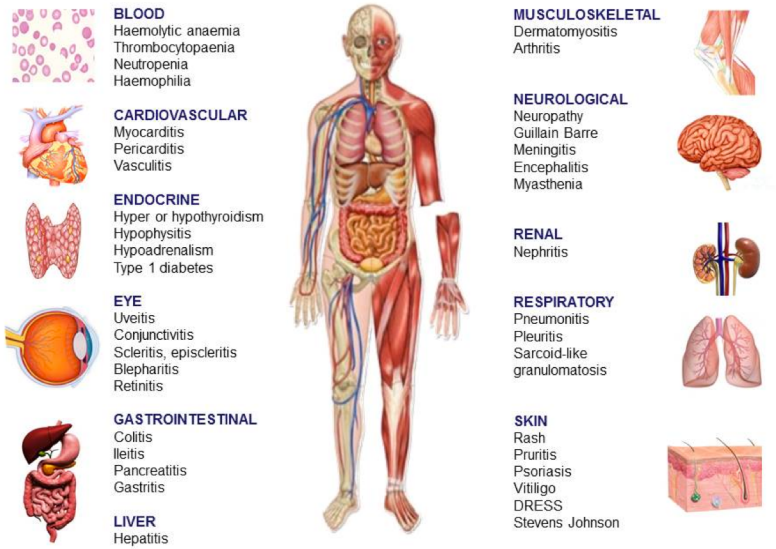
Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, “Clinical Trends in Molecular Medicine,” Volume 7, Issue 5, 1 May 2001, pages 201-204.

Prediction of drug response is complex - efficacy v side effects

Cancer Immunotherapy

2013 2014

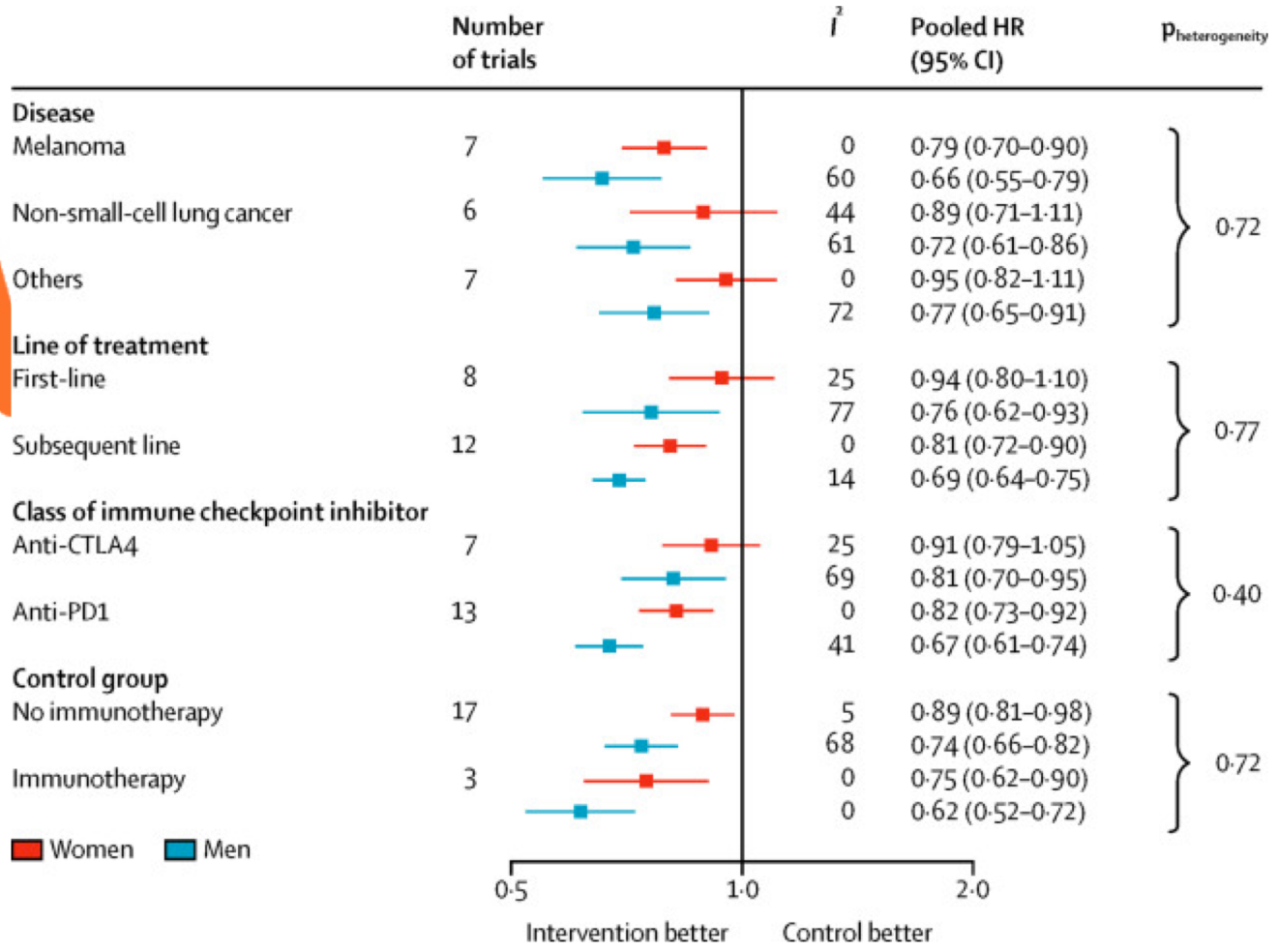
Toxicities



Immune Checkpoint Inhibitor-Related Toxicities
Brown & Mislang 2018 Cancer immunotherapy: at a new immune frontier Immunotherapy 42:1

Cancer immunotherapy efficacy and patients' sex: a systematic review and meta-analysis

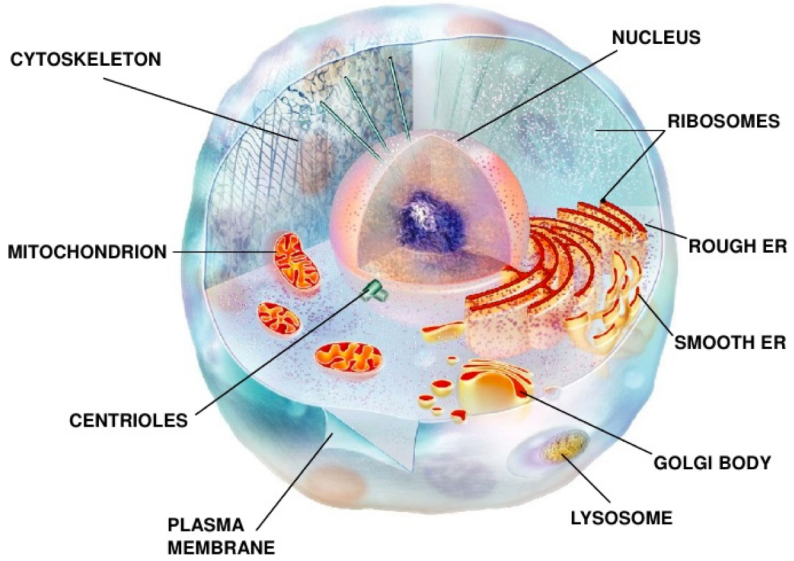
Fabio Conforti, Laura Pala, Vincenzo Bagnardi, Tommaso De Pas, Marco Martinetti, Giuseppe Viale, Richard D Gelber, Aron Goldhirsch



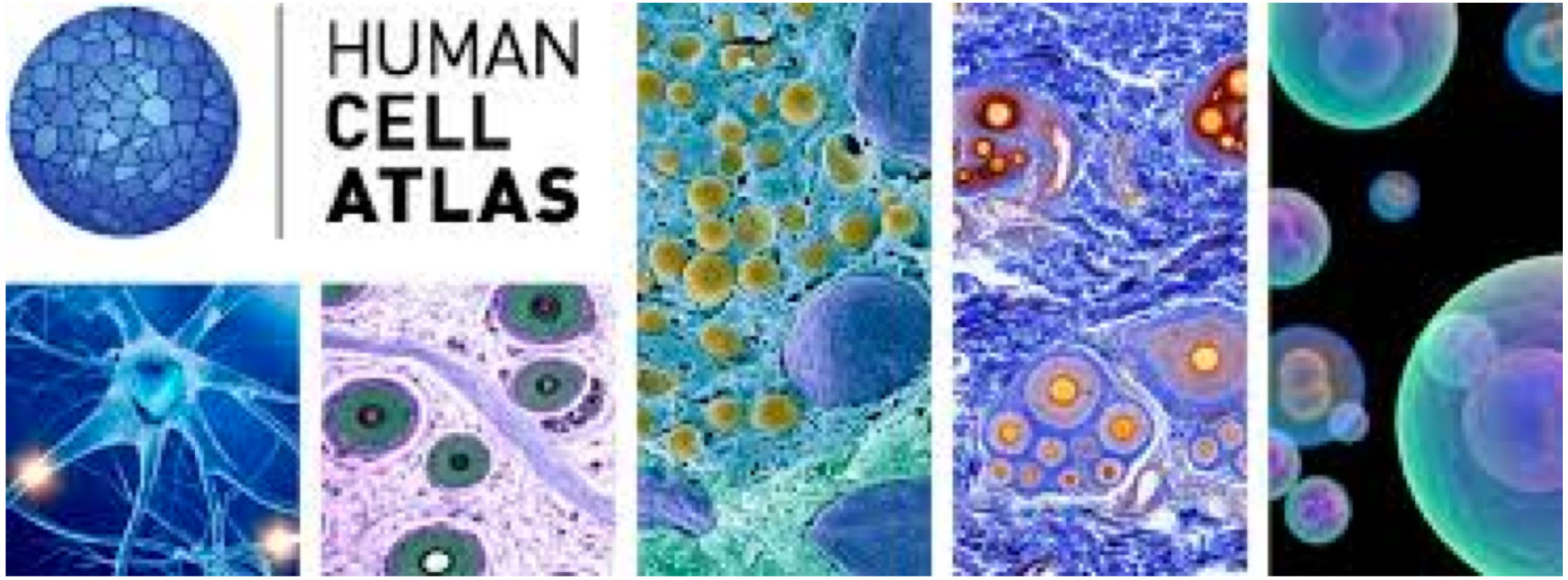
Between individuals. Study concluded patients' gender included in risk versus benefit assessment

‘omics data

> Everything you'll ever need to know is within you; the secrets of the universe are imprinted on the cells of your body.
>
> **Dan Millman**



Rapid changes in the scale and price of sequencing technology



Goal: Create a Human Cell Atlas

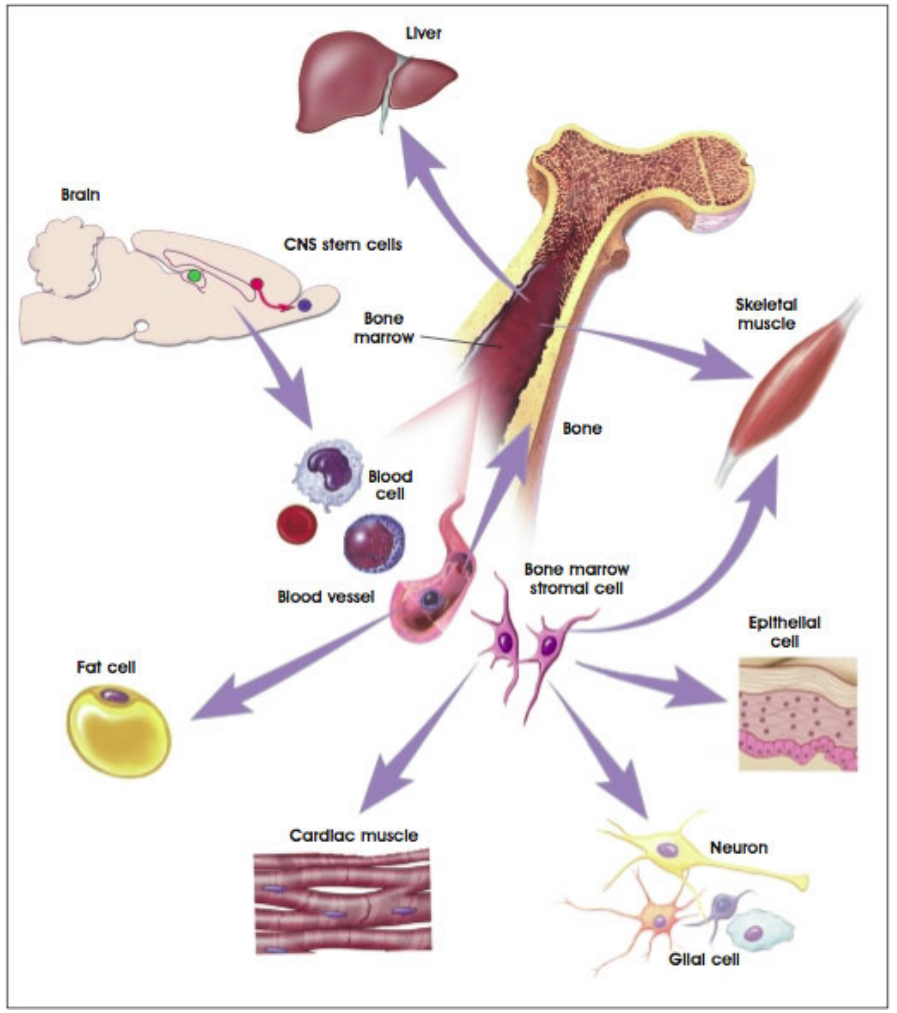
catalog and map of all cell types to the location within tissues and within the body; temporal, spatial, development, etc

If every cell in your hand was the size of a grain of sand,

how big would you hand be ?

A school bus - 2.5 billion cells in one hand

- Human disease including cancers are multicellular
- Can we discover inter-cellular pathways
- Molecular signatures of cross-talk between cells during disease
- How do we find primary and latent signals?



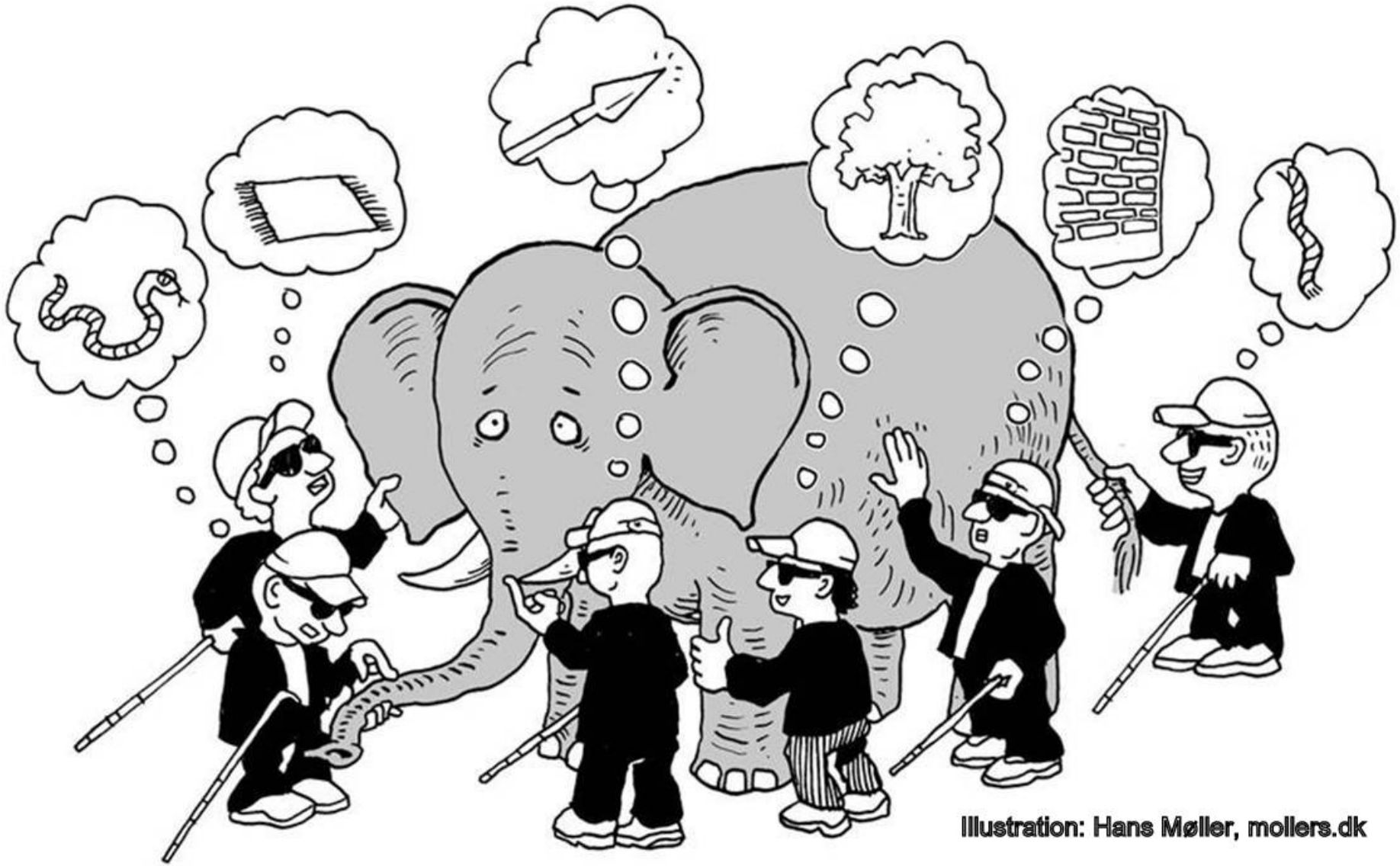
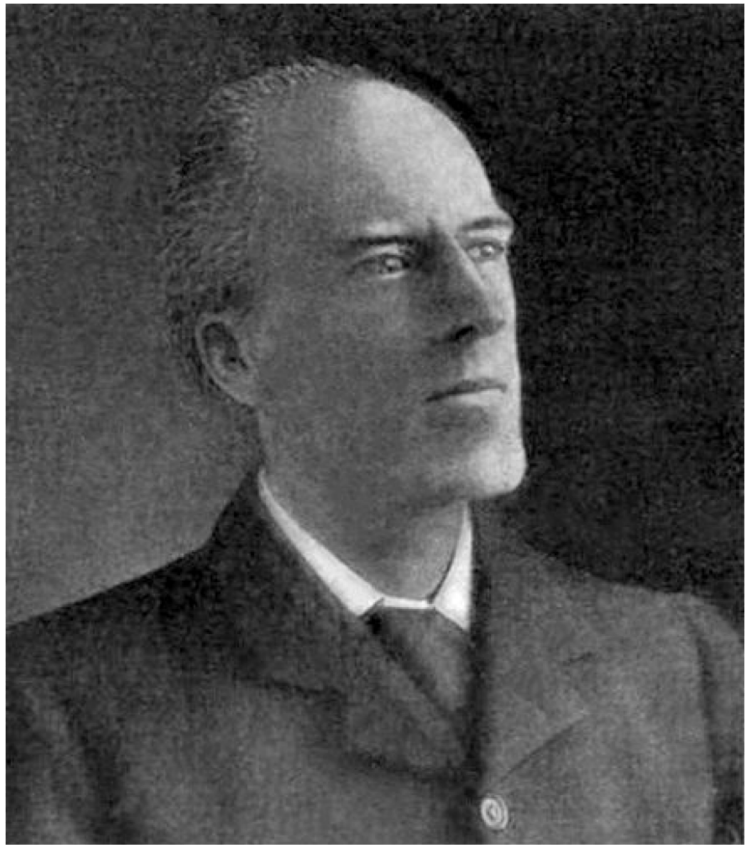


Illustration: Hans Møller, mollers.dk



LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* *By* KARL PEARSON, *F.R.S., University College, London* *.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$

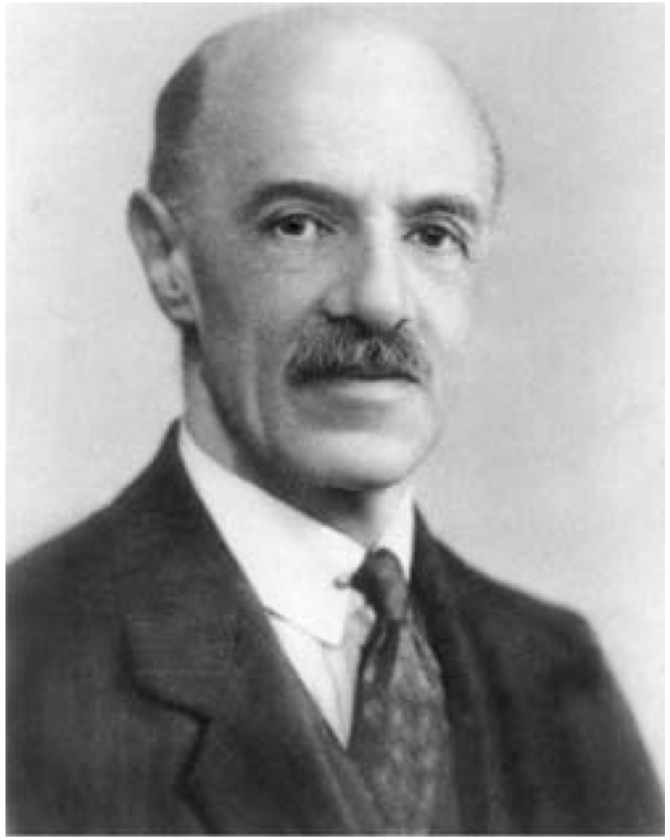
$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books of least squares, the variables on the right of our equations are treated as the independent, those on the left as the dependent variables. The result of this treatment is that we get one straight line or plane if we treat some one variable as independent, and a quite different one if we treat another variable as the independent variable. There is no paradox

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.

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

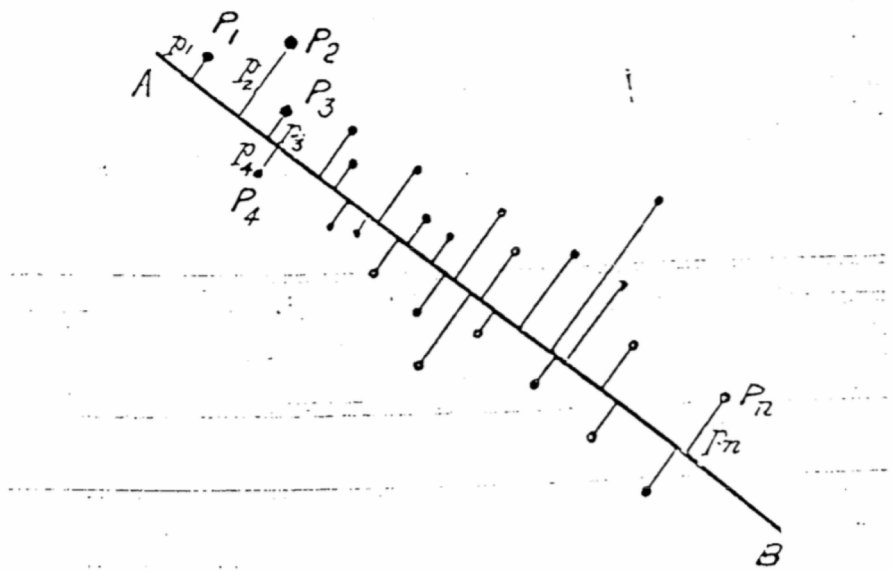
For example:—Let $P_1, P_2, \ldots P_n$ be the system of points with coordinates $x_1, y_1; x_2, y_2; \ldots x_n y_n$, and perpendicular distances $p_1, p_2, \ldots p_n$ from a line A B. Then we shall make

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

If y were the dependent variable, we should have made

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

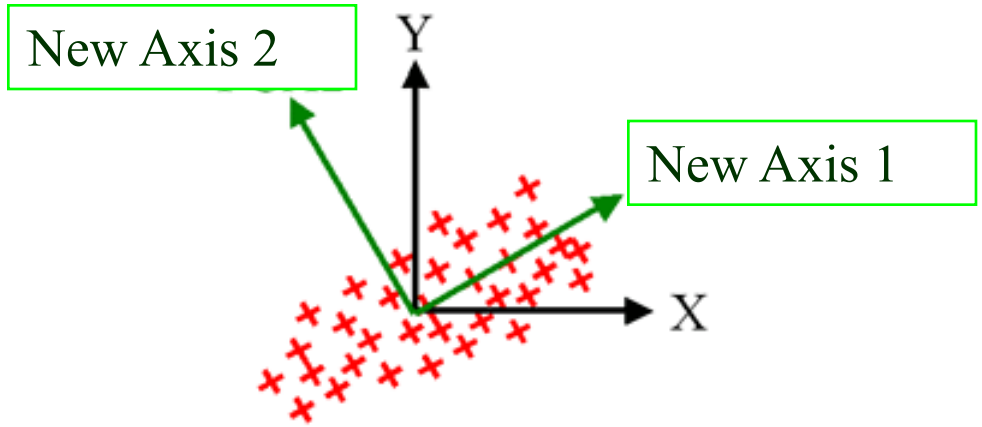
(y' being the ordinate of the theoretical line at the point x which corresponds to y), had we wanted to determine the best-fitting line in the usual manner.



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.

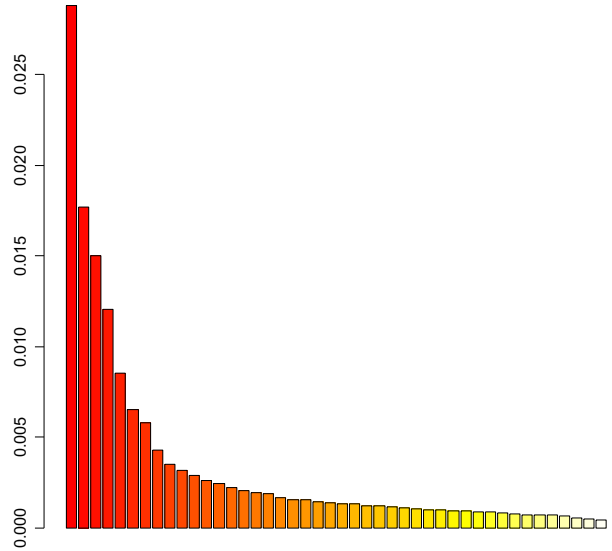
The second new axis will be orthogonal, and will explain the next largest amount of variance

Principal Axes

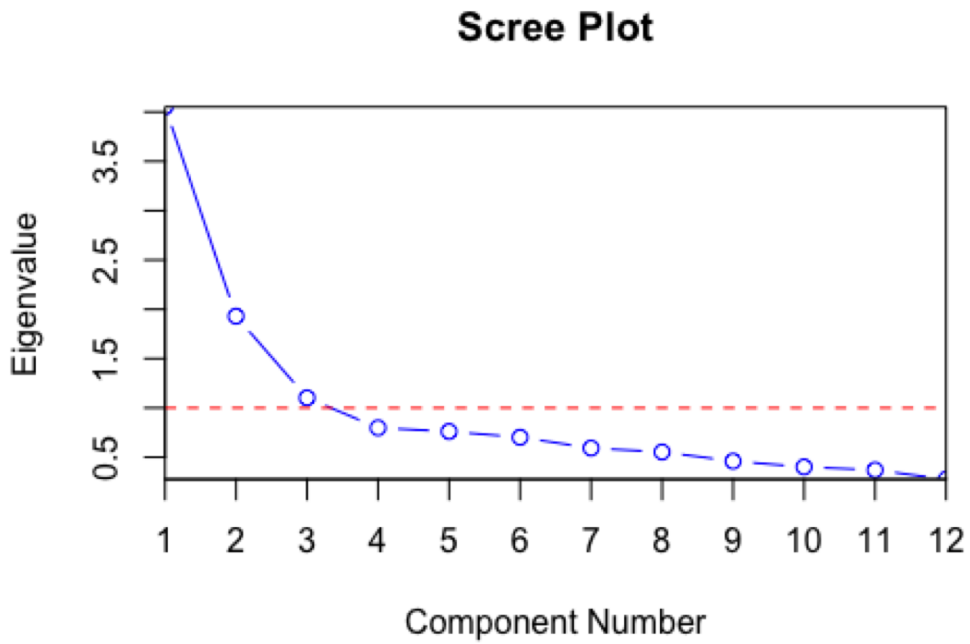
- Project new axes through data which capture variance. Each represents a different trend in the data.
- Orthogonal (decorrelated)
- Typically ranked: First axes most important
- Principal axis, Principal component, latent variable or eigenvector

Eigenvalues

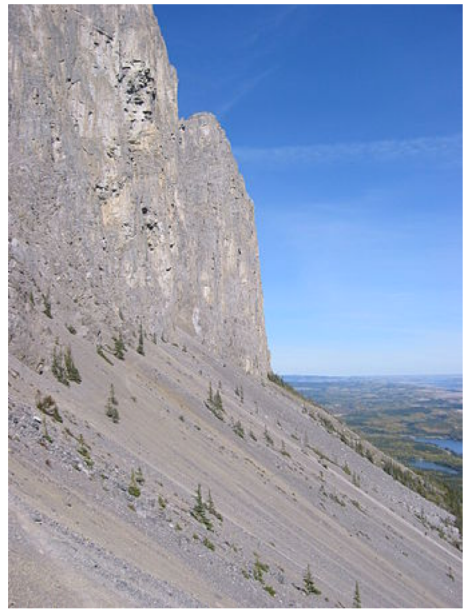
- Describe the amount of variance (information) captured by each eigenvector
- Ranked. First eigenvalue is the largest.
- Generally only examine 1st few components
  - scree plot



Selecting Components (based on eigenvalues)



The "Kaiser rule" criteria is shown in red.



Scree “elbow”

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]

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. Second axis - arched function of the first axis. Called horseshoe effect

Arch Effect

A horseshoe or arch structure in the points is often an indicator of a sequential latent ordering or gradient in the data (Diaconis, Goel, and Holmes 2007).

https://www.huber.embl.de/msmb/Chap-MultivaHetero.html

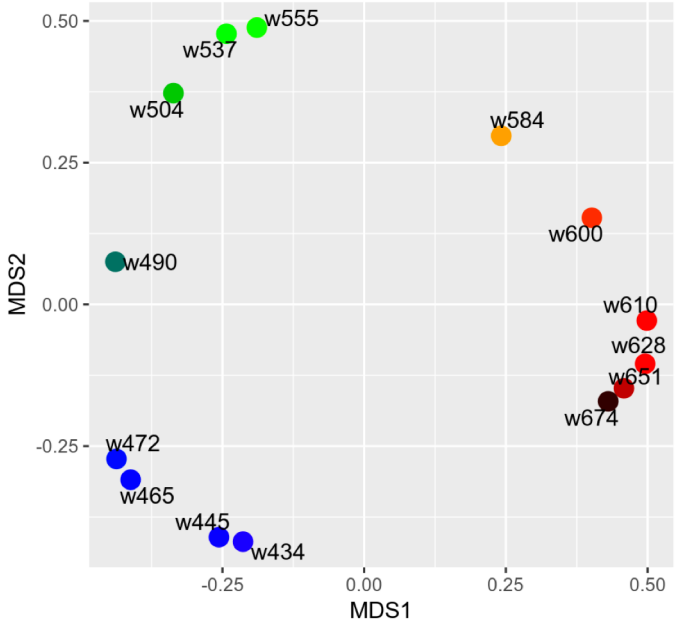
http://ordination.okstate.edu/PCA.htm

http://phylonetworks.blogspot.com/2012/12/distortions-and-artifacts-in-pca.html

https://statweb.stanford.edu/~susan/papers/horseshoes6.pdf

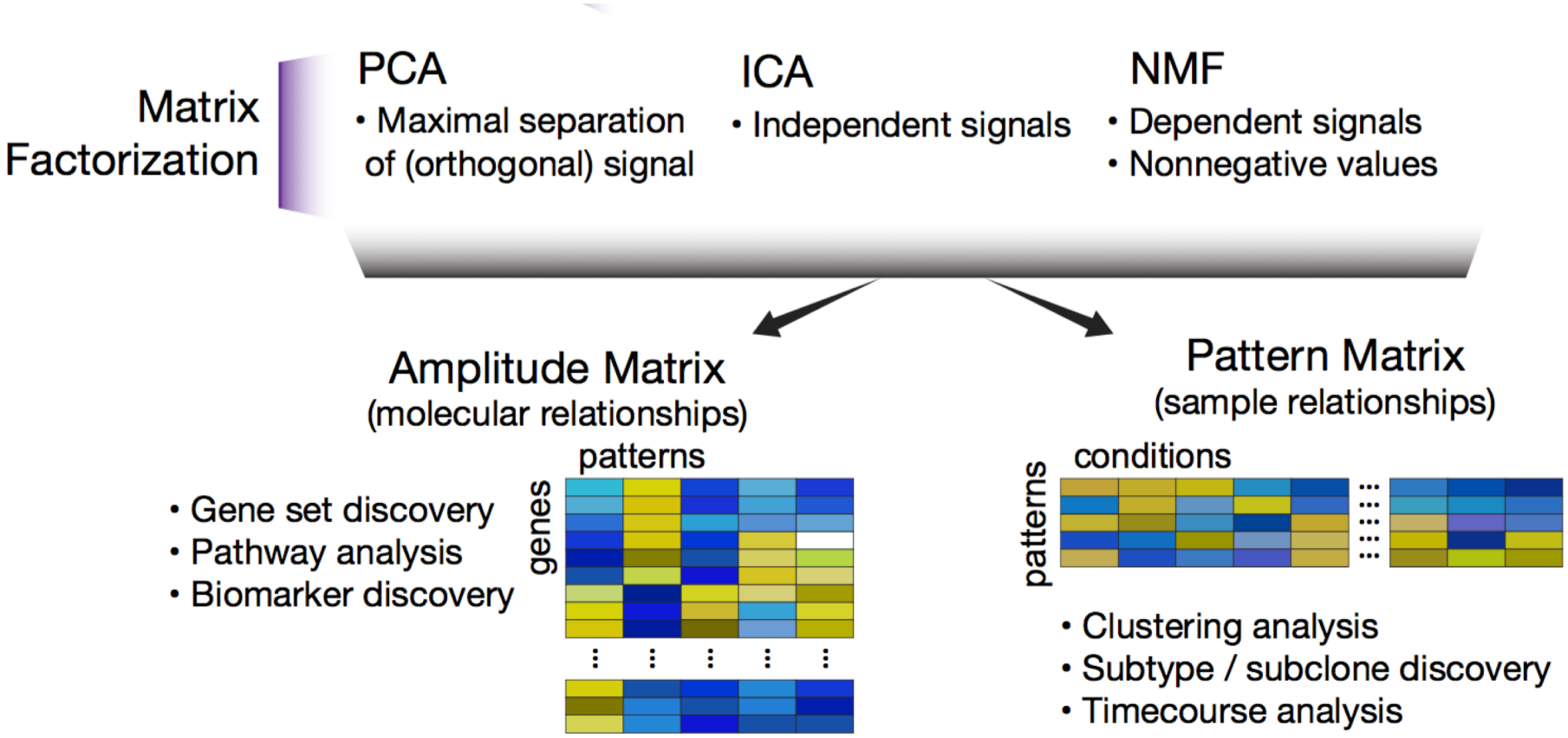
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320001/

https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006907



* Not presented during talk. Added afterwards because I got lots of questions about it

Global v Local methods



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
- Maybe try >1 approaches

Table 2. Dimension reduction methods for one data set

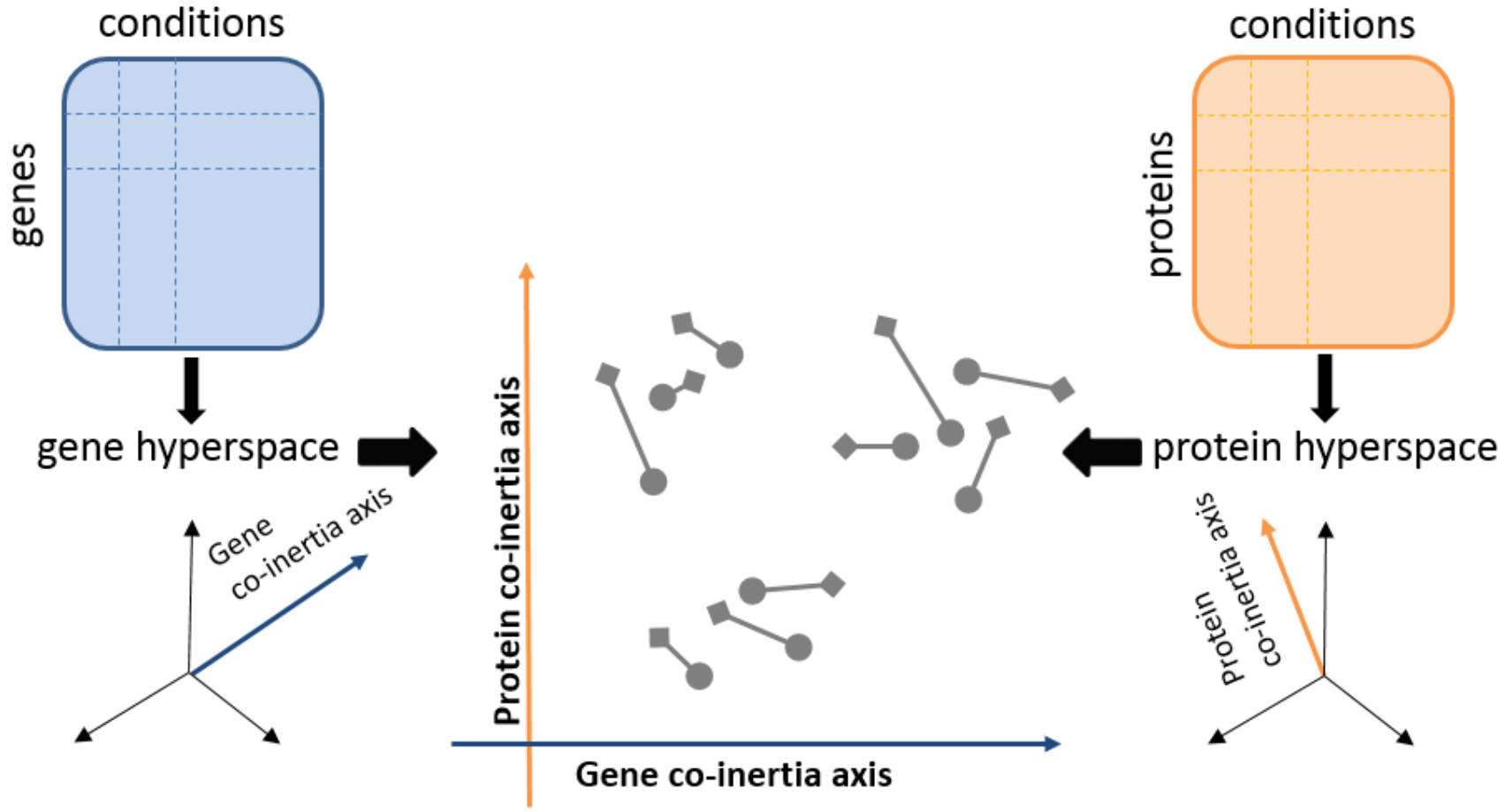
Method	Description	Name of R function {R package}
PCA	Principal component analysis	prcomp{stats}, princomp{stats}, dudi.pca{ade4}, pca{vegan}, PCA{FactoMineR}, principal{psych}
CA, COA	Correspondence analysis	ca{ca}, CA{FactoMineR}, dudi.coa{ade4}
NSC	Nonsymmetric correspondence analysis	dudi.nsc{ade4}
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


Please sir, I want some more data

>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.

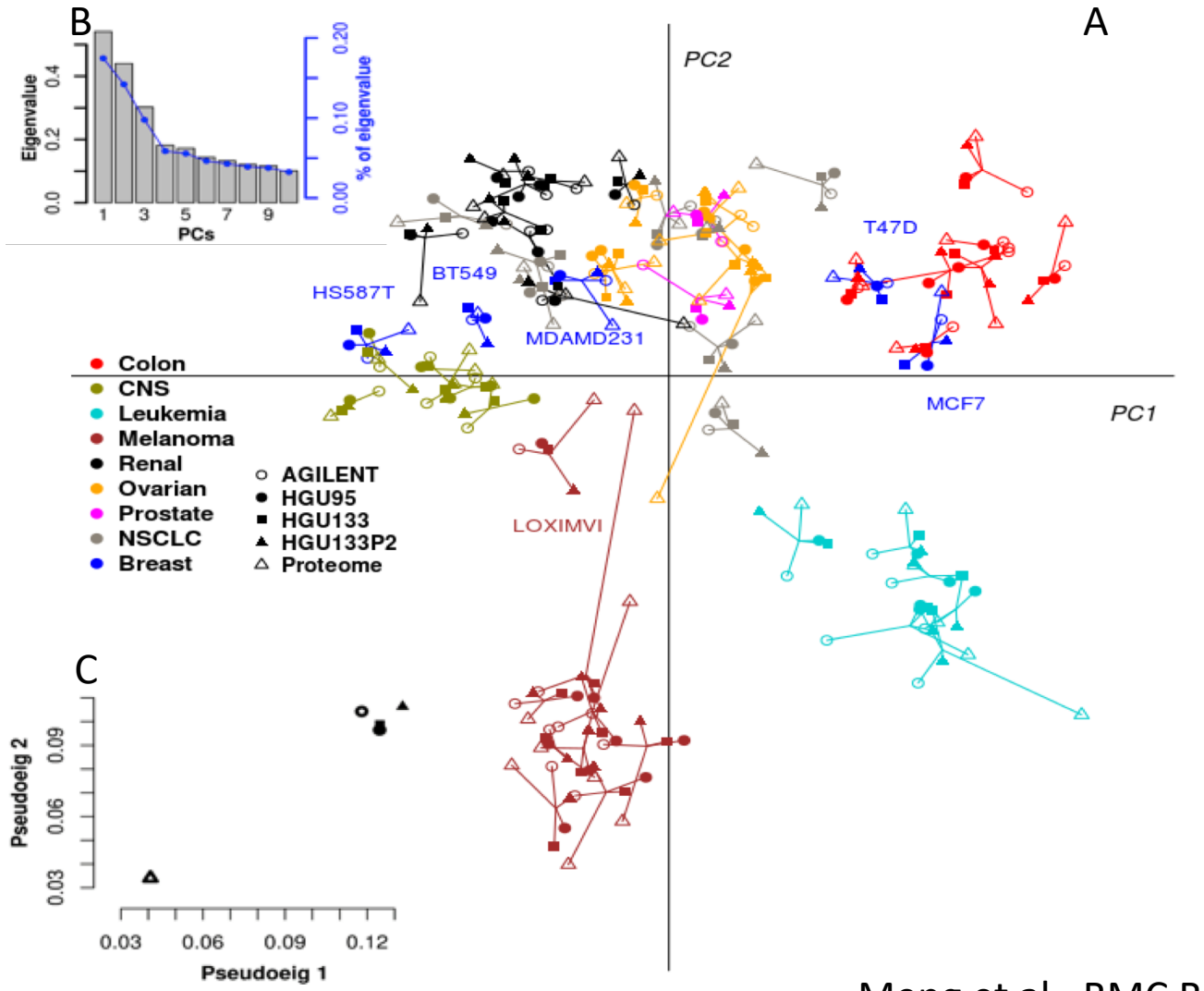
>2 datasets : Tensor data integration

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

Multiple Factor Analysis or Multiple Coinertia Analysis

- **MFA** (Abdi et al., 2013) the rows of each dataset are first centered and scaled, then each dataset is weighted by the reverse of its first eigenvalue (proc.row="center_ssq1", w.data="lambda1").

- **MCIA** the statis algorithm **statis=TRUE**, datasets are further weighted so those closer to the overall structure receive a higher weight.

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

MFA statis=FALSE (the default setting)..

Moa: weighting of datasets

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,

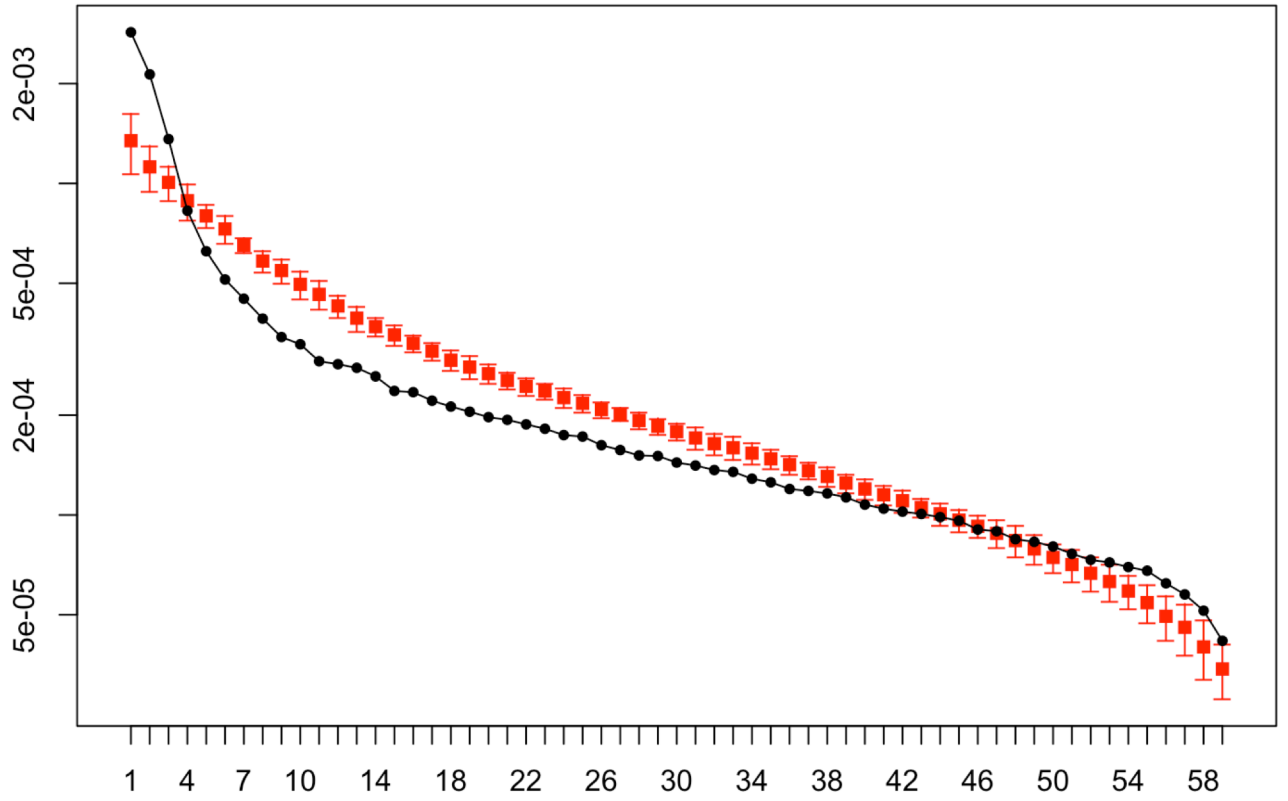
- uniform - no weighting
- lambda1 - weighted by the reverse of the first eigenvalue of each individual dataset
- inertia - weighted by the reverse of the total inertia.

weight datasets closer to the overall structure

statis - FALSE

Determining 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)
```

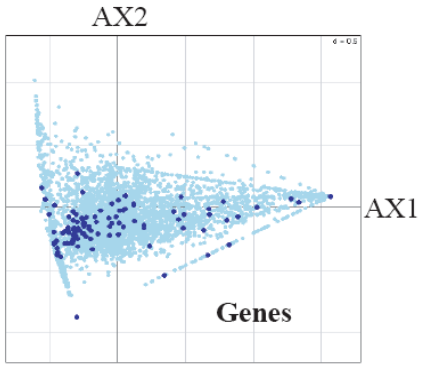


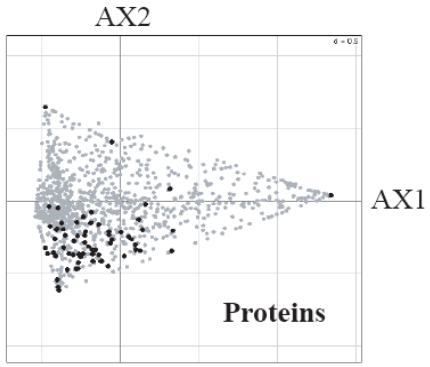
mogsa

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

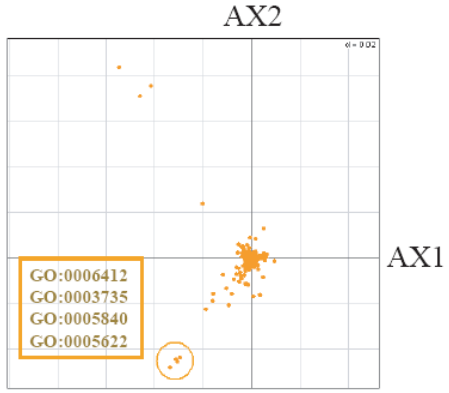

Please sir, I want some more data

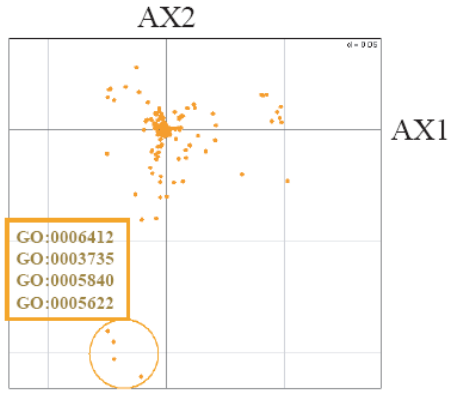
“Bucket Scores” ; Collecting features (Genes) across multiple datasets





Matrix decomposition of gene expression and proteomics onto same scale

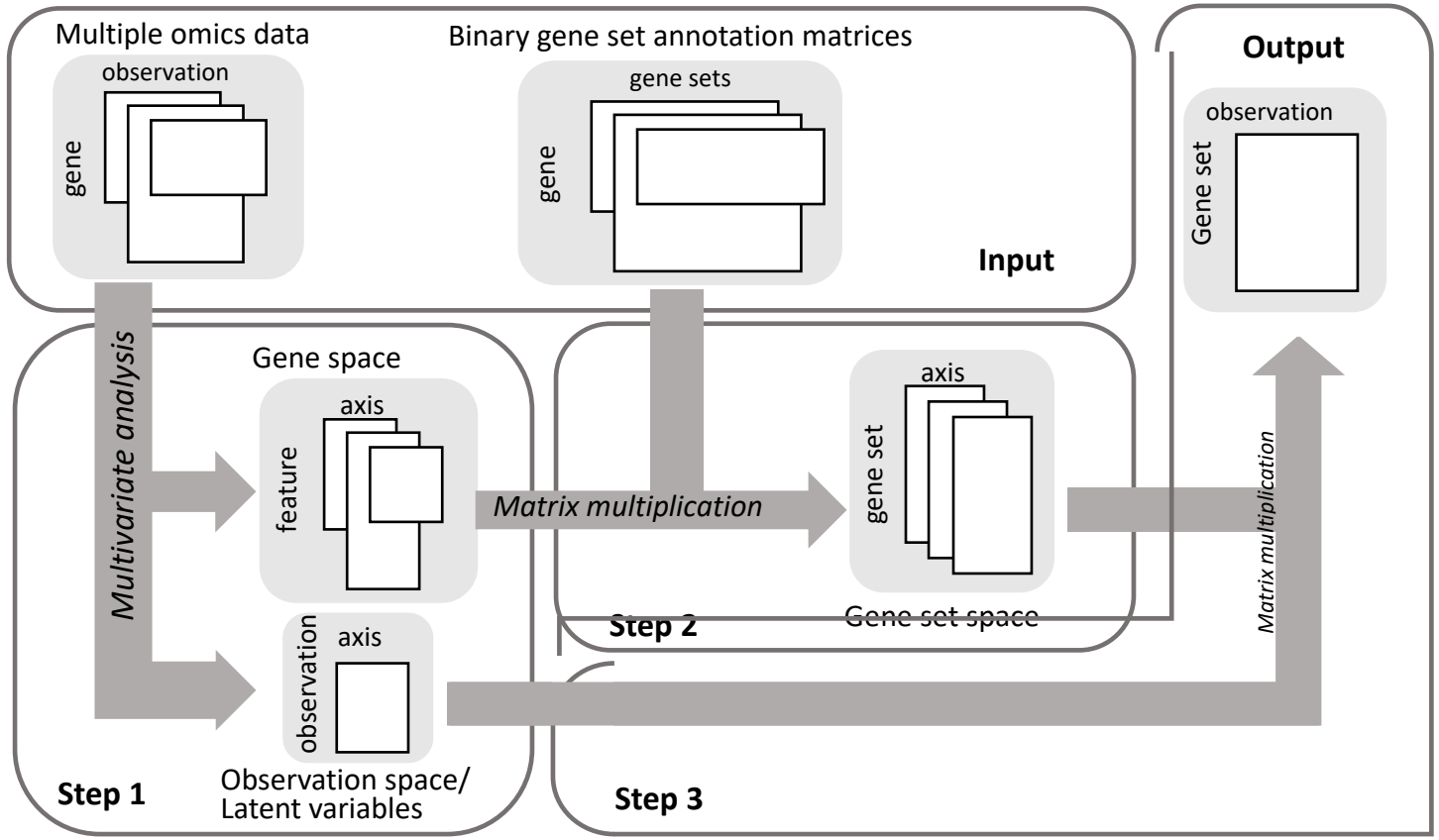




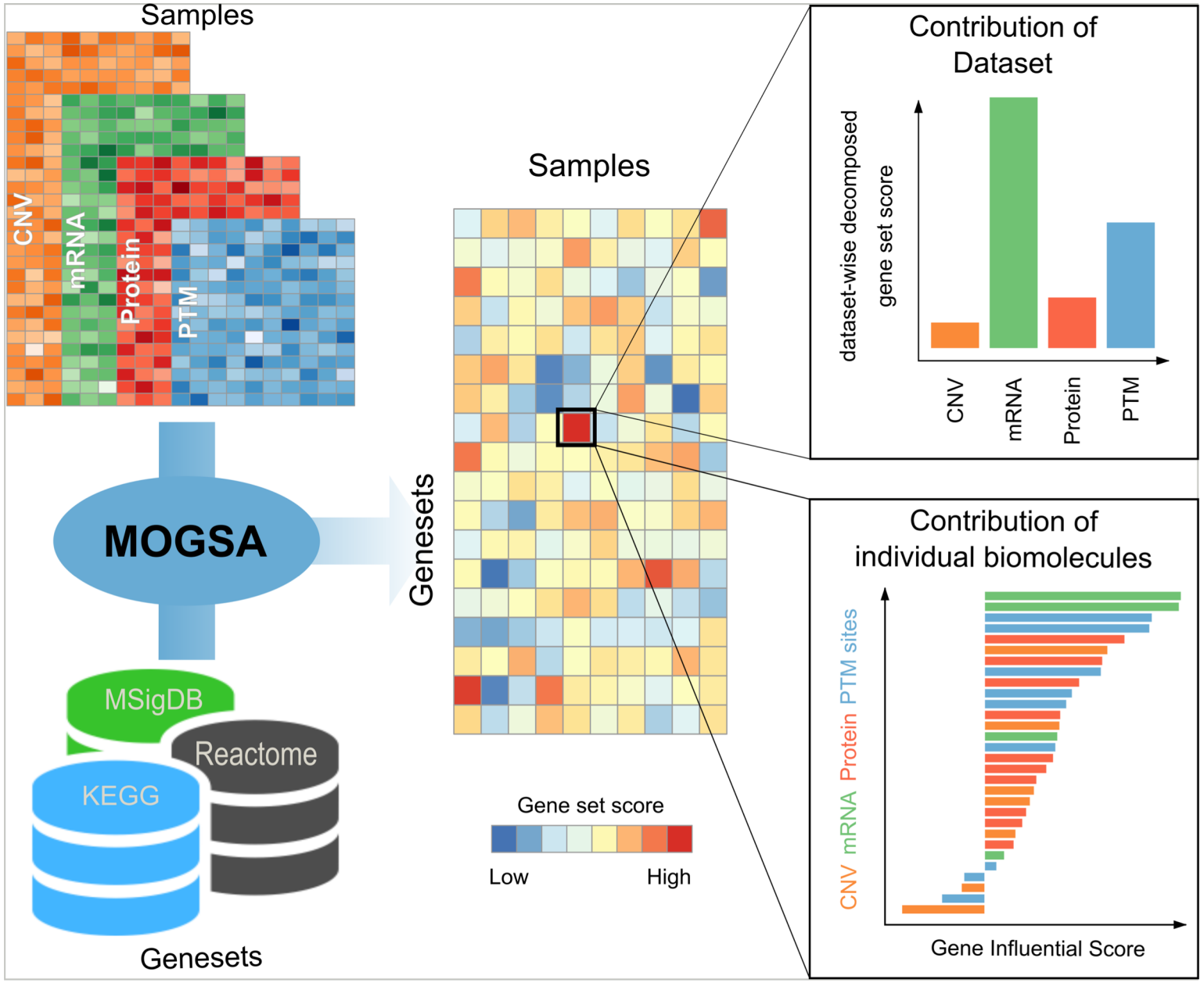
Project GO Terms (vector of gene) onto each to get a gene set “score” in each space

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. Mol Cell Proteomics. 2019 Aug 9;18(8 suppl 1):S153-S168.



Meng C, Basunia A, Kuster B , Peters B, Gholami AM, Culhane AC. moGSA: Integrative Single Sample Gene-Set Analysis of Multiple Omics Data. Mol Cell Proteomics. 2019 Aug 9;18(8 suppl 1):S153-S168.

mogsa::moa()

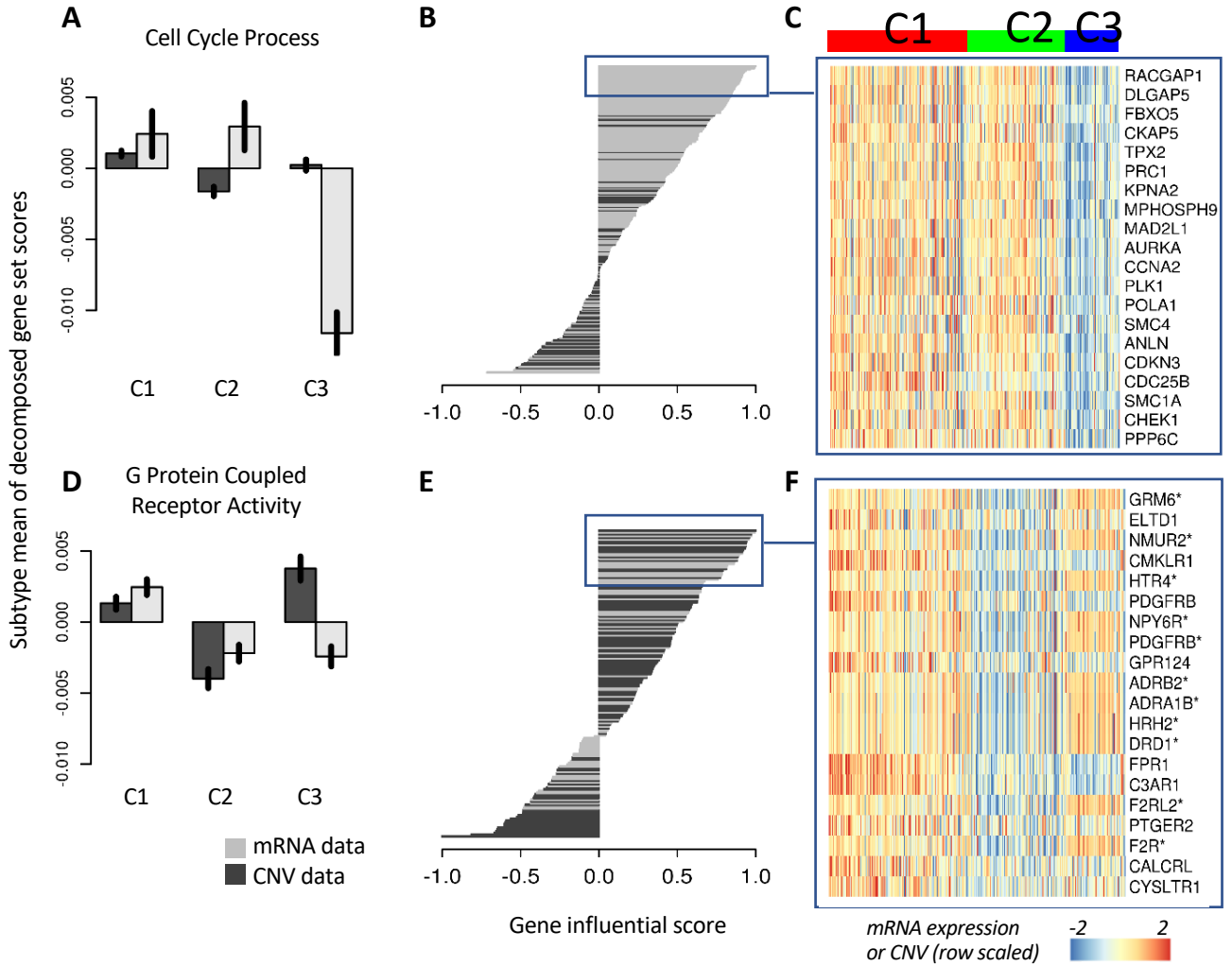
```
library(mogsa)
data(NCI60_4arrays)
ana<- moa(NCI60_4arrays, proc.row = "center_ssq1", w.data = "inertia", statis = TRUE)

plot(ana, value = "eig", type = 2) # plot eigen value
plot(ana, value = "tau", type = 2) # plot the normalized (percentage) eigen value

colcode <- as.factor(sapply(strsplit(colnames(NCI60_4arrays$agilent), split="\\."), "[", 1))
plot(ana, type = 1, value = "obs", col=colcode)
plot(ana, type = 2, value = "obs", col=colcode, data.pch=1:4)
plot(ana, value = "var", layout=matrix(1:4, 2, 2)) # plot variables/features in each data sets
plot(ana, value = "RV") # plot the RV coefficients for the data sets

#MOGSA
mgsa3 <- mogsa(x = ana, sup=smoa)
```

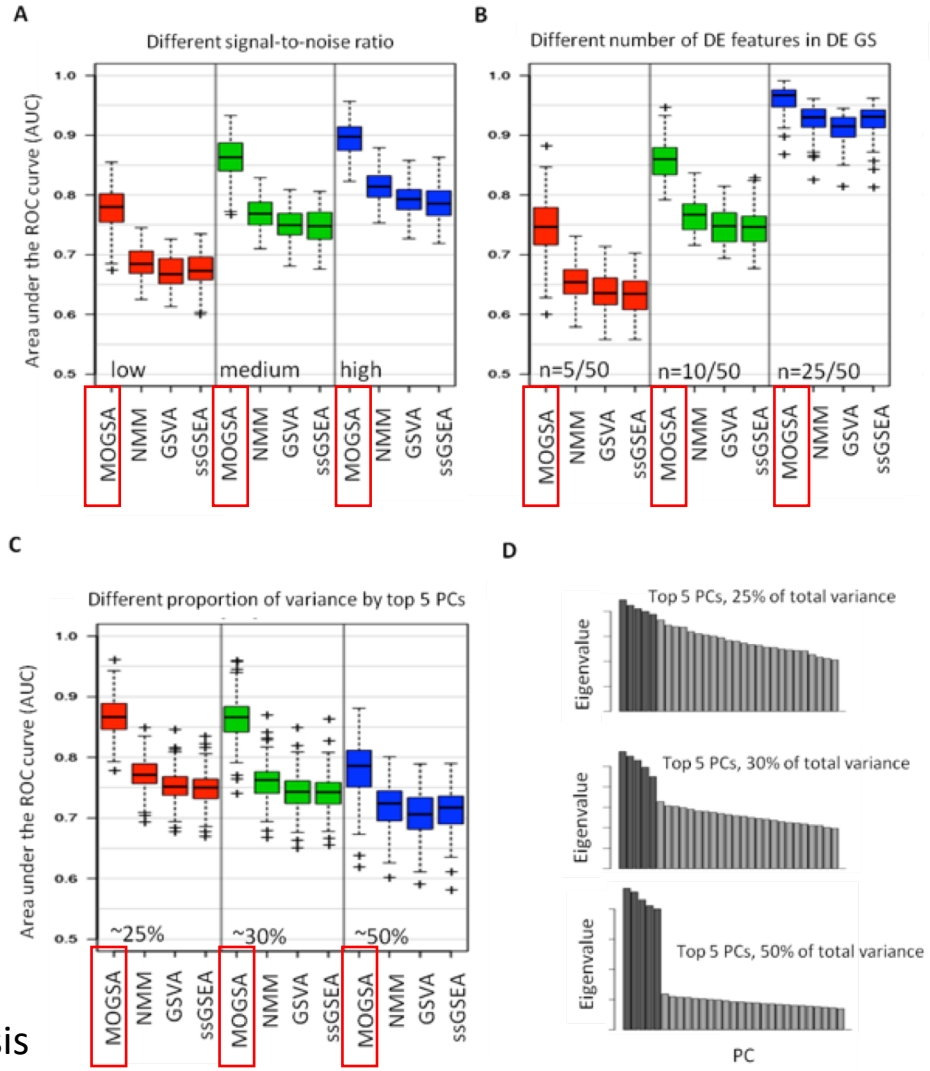
moGSA gene-set scores of CNV and mRNA data



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

moGSA single sample Gene Set analysis

MOGSA compares or outperforms other ssGSA approaches on Synthetic data



Meng C et al.,. moGSA: Integrative Single Sample Gene-Set Analysis of Multiple Omics Data Mol Cell Proteomics. 2019 Aug 9;18

Matrix factorization single sample Gene Set Analysis

1. Unsupervised..
2. Integrates multiple omics data
3. Multi-modal data has greater sensitivity to detect pathways
4. Scalable to large data
5. On simulated data outperforms popular existing methods (GSVA, ssGSEA)
6. Exacts ssGSA scores associated with each or composite of components each reflecting different “layers” of complex data
7. Can exclude components (batch effects) or select components of interest
8. AVAILABLE IN BIOCONDUCTOR. R PACKAGE moGSA

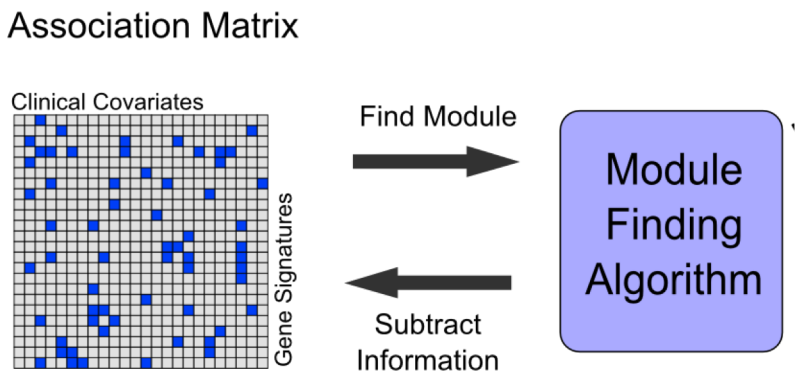
Meng et al., **MOGSA: integrative single sample gene-set analysis of multiple omics data.** Mol Cell Proteomics. 2019 Aug 9;18(8 suppl 1):S153

Iterative Binary Biclustering

(of Gene Sets)

iBBiG

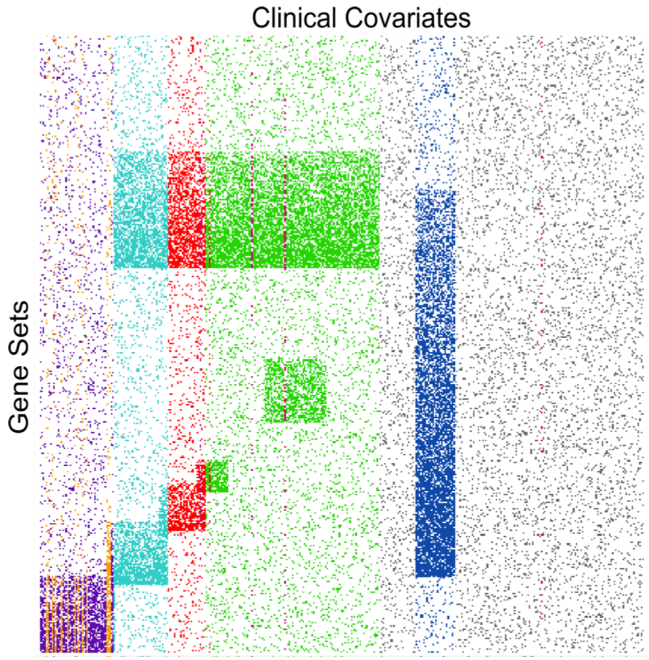
- **Iterative**
  - approach which iteratively extract the strongest signals in order to find weaker but more interesting signals.
- **Robust**
  - Data is intrinsically sparse and noisy.
  - Assymetric - Only associations important
- **Fuzzy**:
  - Allowing membership of >1 cluster, both covariates and gene sets



Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

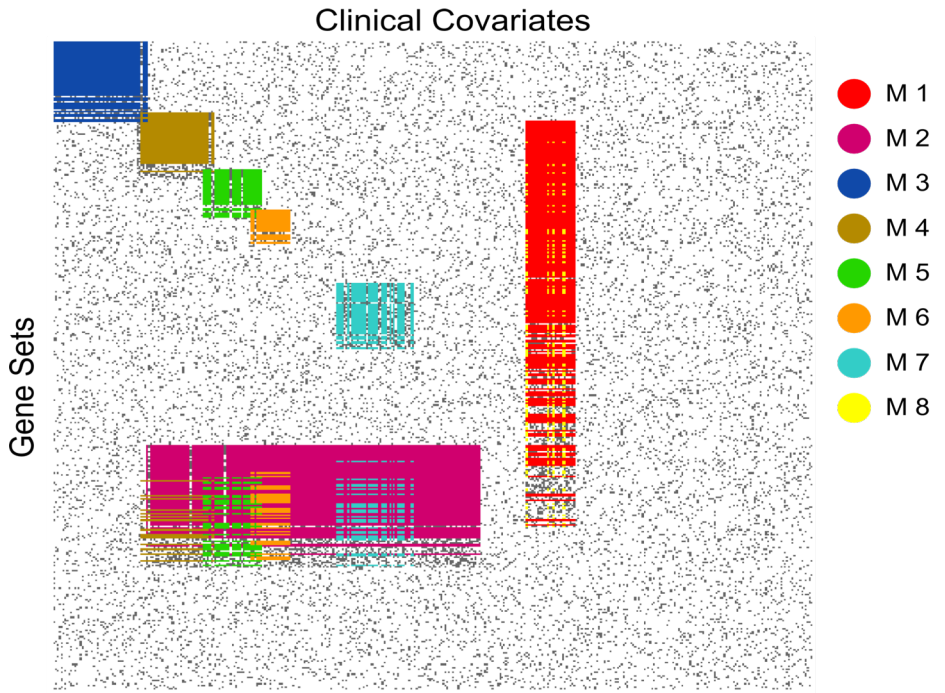
Bioconductor Package: iBBiG

K-Means



Used K=7

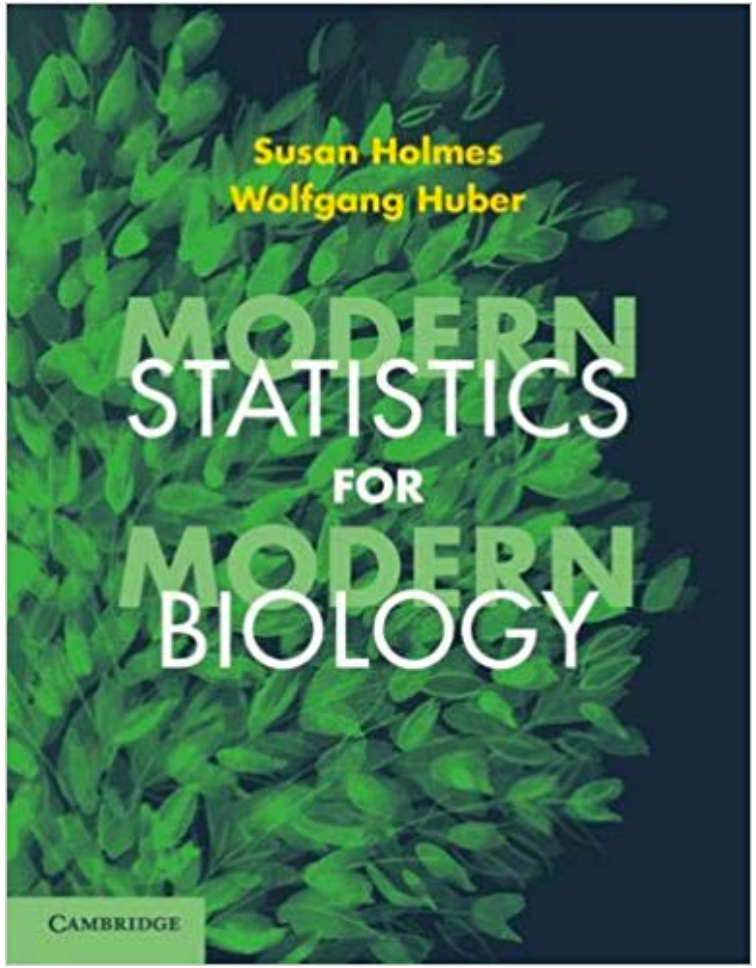
iBBiG



iBBiG allows for overlapping membership

Gusenleitner D, Howe EA, Bentink S, Quackenbush J, **Culhane AC.** iBBiG: Iterative Binary Bi-clustering of Gene Sets. *Bioinformatics.* 2012. 28(19):2484-92.

Resources



https://www.huber.embl.de/msmb/index.html

Modern Statistics for Modern Biology

Susan Holmes, Wolfgang Huber

Chapters

9 Multivariate methods for heterogeneous data

In Chapter 7, we saw how to summarize rectangular matrices whose columns were continuous variables. The maps we made used unsupervised dimensionality reduction techniques such as principal component analysis aimed at isolating the most important *signal* component in a matrix X when all the columns have meaningful variances.

Here we extend these ideas to more complex heterogeneous data where continuous and categorical data are combined or even to data where individual variables are not available. Indeed, sometimes our observations cannot be easily described by features – but it is possible to determine distances or (dis)similarities between them, or to put them into a graph or a tree. Examples include species in a species tree or biological sequences. Outside of biology, there are text documents or sound files, where we may have a reasonable method to determine (dis)similarity between objects, but no absolute 'coordinate system' of features.

This chapter contains more advanced techniques for which we have omitted many technical details. We hope that by giving the reader some hands-on experience with exampl

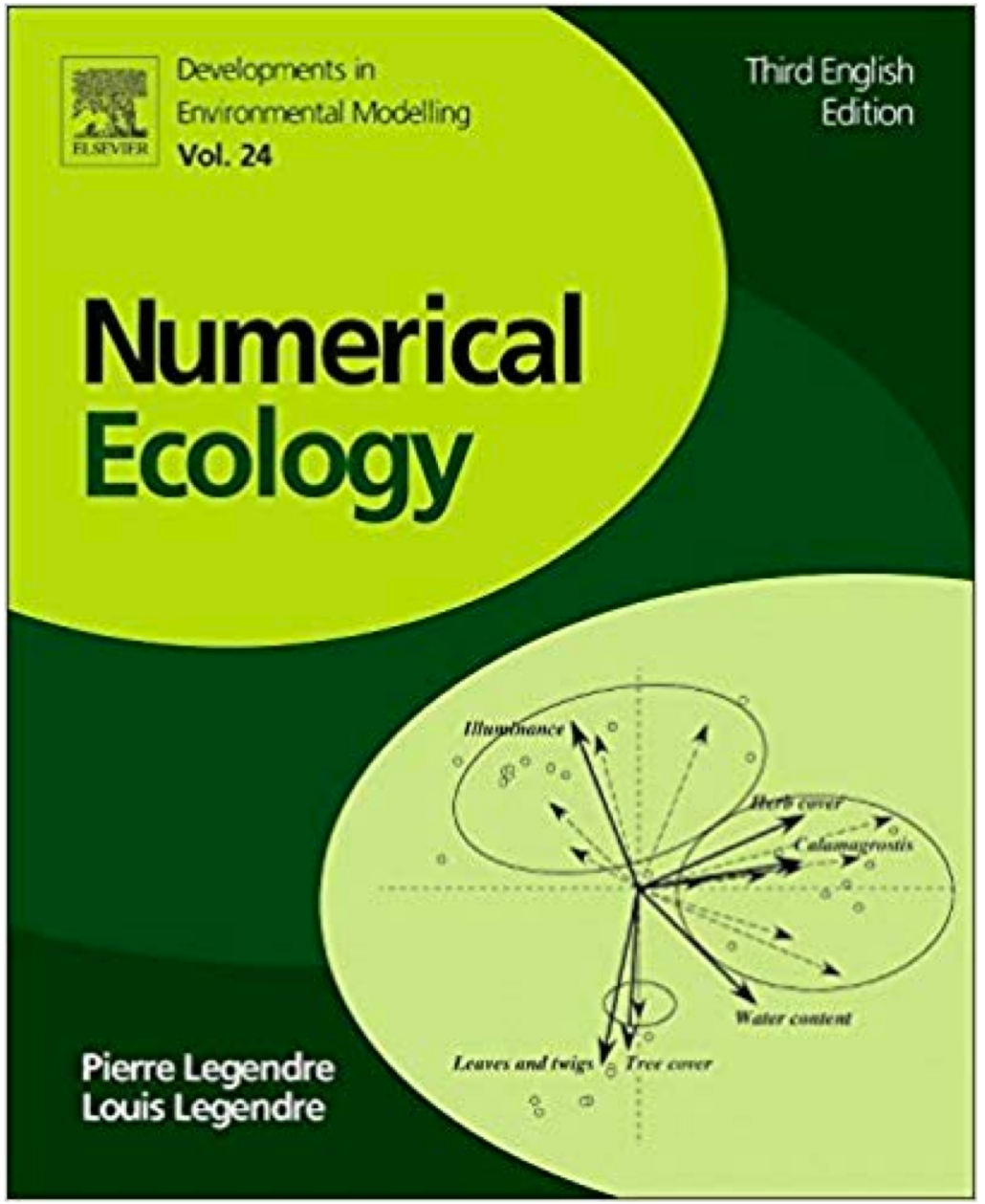
Susan Holmes

Wolfgang Huber

Resources

Louis Legendre

Pierre Legendre



* Following slides not presented during talk- made available for reference

Bioconductor Overview

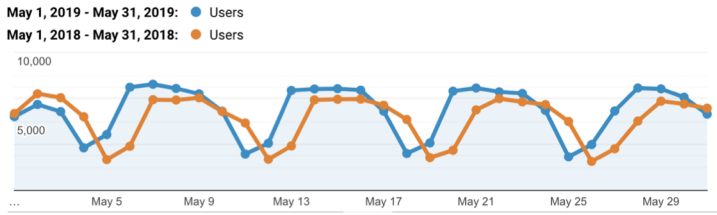
- Written by Robert Gentleman, 2003-

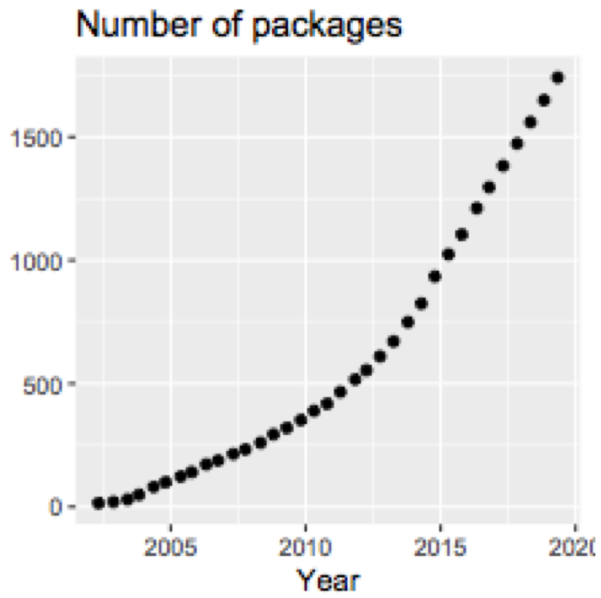
- 1751 software packages
- More than ½ million unique IP downloads last year
- More than 1200 maintainers
- More than 30,000 PMC full-text citations

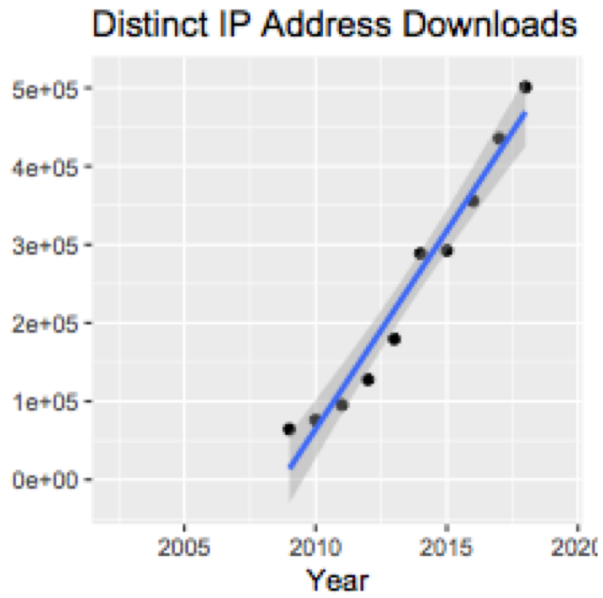
How?

- https://bioconductor.org
- https://support.bioconductor.org
- https://community-bioc.slack.org

Slack: join at https://bioc-community.herokuapp.com/







Martin Morgan, presented at Developer Day Bioc2019 http://bioc2019.bioconductor.org/schedule-developer-day

Installing Bioconductor

- https://cran.r-project.org/web/packages/BiocManager/vignettes/BiocManager.html

```
install.packages("BiocManager")

BiocManager::install() # Install current release of Bioconductor

BiocManager::version() # Check Bioconductor version

BiocManager::install(version="3.7")  # Install older release

BiocManager::install(version = "devel")  #Devel

BiocManager::valid() # checks for out of date packages
```

Community:

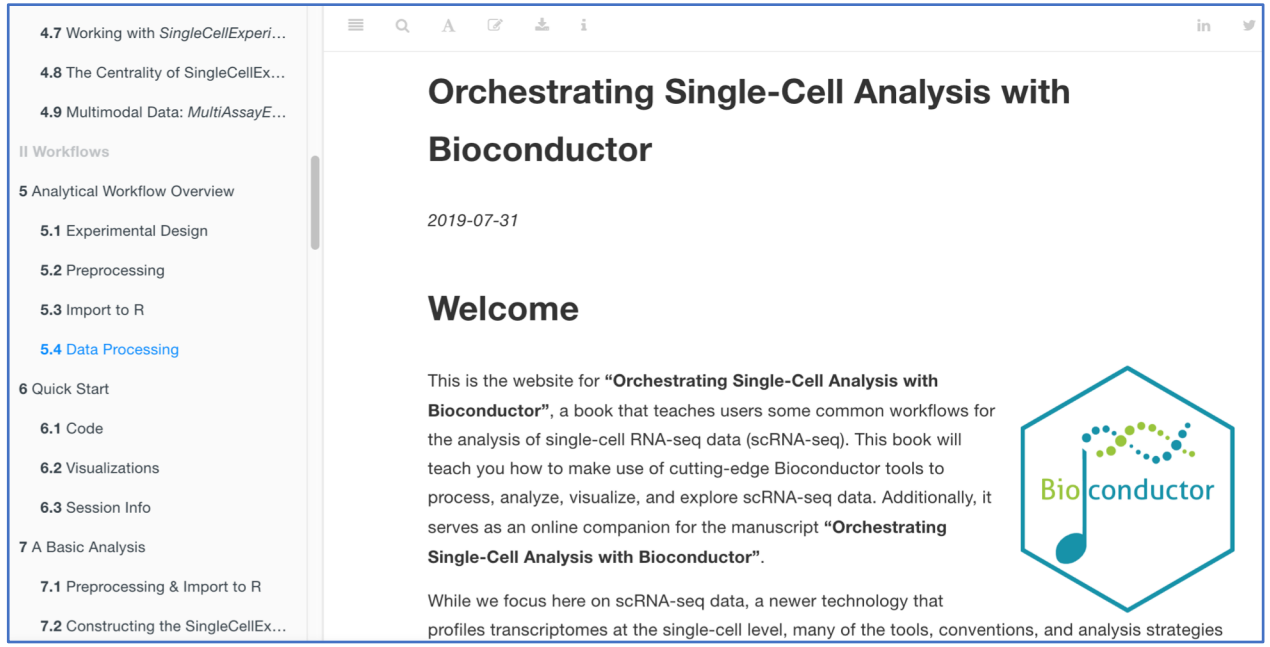
Online Support Site, Slack Channel,
Bioconductor Meetings (2020 Boston, Europe, Asia),
Courses, Local Meetup groups (Boston/New York)

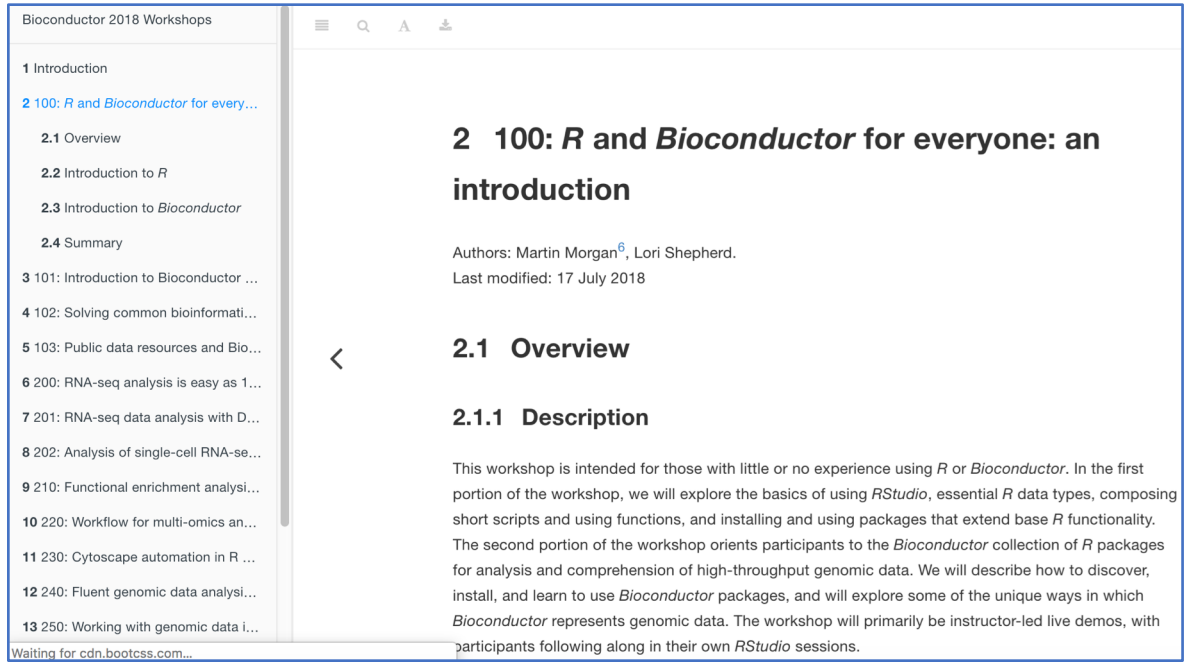
Where to Start?

https://bioconductor.org/help/course-materials/ (Course materials)
http://bioconductor.org/packages/devel/BiocViews.html#___Workflow (Workflows)
https://f1000research.com/gateways/bioconductor (F1000 channel)

Online Cookbooks of Bioconductor Workflows

- https://osca.bioconductor.org/ OSCA
- https://bioconductor.github.io/BiocWorkshops/ Meeting Workshop Bookdown





Classes, Vignettes, Documentation

- http://bioconductor.org/checkResults/ (nightly build reports)

Importing

- GTF, GFF, BED, BigWig, etc., – rtracklayer::import()
- VCF – VariantAnnotation::readVcf()
- SAM / BAM – Rsamtools::scanBam(), GenomicAlignments::readGAlignment*()
- FASTA – Biostrings::readDNAStringSet()
- FASTQ – ShortRead::readFastq()
- MS data (XML-based and mgf formats) – MSnbase::readMSData(), MSnbase::readMgfData()

Common Classes

- Rectangular feature x sample data – SummarizedExperiment::SummarizedExperiment() (RNAseq count matrix, microarray, ...)
- Genomic coordinates – GenomicRanges::GRanges() (1-based, closed interval)
- DNA / RNA / AA sequences – Biostrings::*StringSet()
- Gene sets – GSEABase::GeneSet() GSEABase::GeneSetCollection()
- Multi-omics data – MultiAssayExperiment::MultiAssayExperiment()
- Single cell data – SingleCellExperiment::SingleCellExperiment()
- Mass spec data – MSnbase::MSnExp()

Bioconductor and S4 classes

dangerpeel 10 points · 6 months ago

"use R effectively" is a too broad of a topic to be completely encompassed by tidyverse operations. For example, I regularly work with giant matrices and with S4 objects within Bioconductor packages. Don't get me wrong, I love the tidyverse. I am annoyed at the comparative clunkiness of the code for the matrices and S4 objects and wish that I could use tidyverse-style code. However, as it is, I need to know base matrix operations and S4-class code to use R effectively for these kinds of data.

jasonpbecker 7 points · 6 months ago

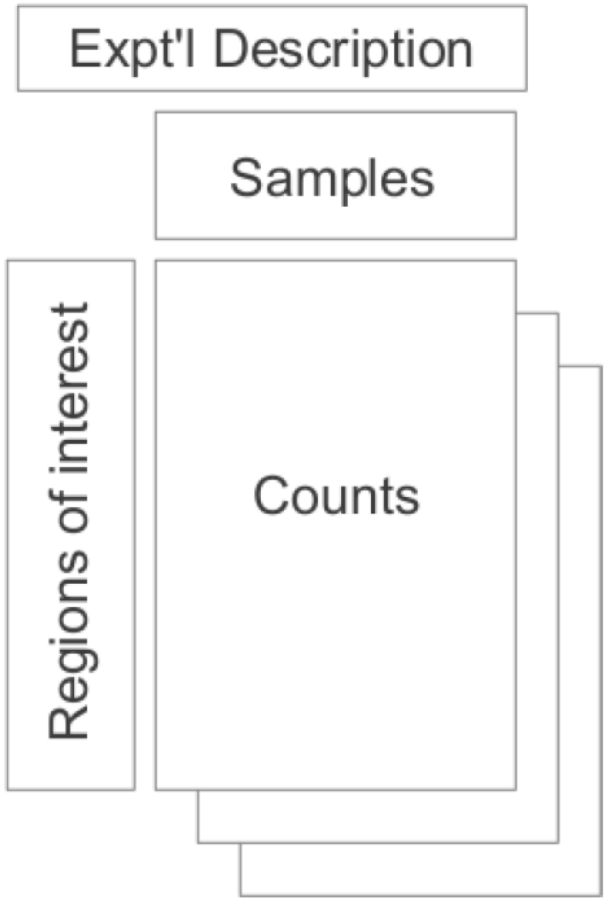
BioConductor world is so much it's own thing, that a lot of R users who are quite advanced have no idea what's going on over there.

Continue this thread →

https://kasperdanielhansen.github.io/genbioconductor/html/R_S4.html

https://www.bioconductor.org/help/course-materials/2017/Zurich/S4-classes-and-methods.html

RangedSummarizedExperiment



Regions of interest × samples

- `assay()` – `matrix`, e.g., counts of reads overlapping regions of interest.
- `rowData()` – regions of interest as `GRanges` or `GRangesList`
- `colData()` – `DataFrame` describing samples.

```
> assay(se)[,se$Treatment == "Control"]  # Control counts
```

GRanges. GenomicRanges


```
> gr = exons(TxDb.Hsapiens.UCSC.hg19.knownGene); gr
GRanges with 289969 ranges and 1 metadata column:
           seqnames              ranges strand |   exon_id
             <Rle>           <IRanges>  <Rle> | <integer>
       [1]     chr1       [11874, 12227]      + |         1
       [2]     chr1       [12595, 12721]      + |         2
       [3]     chr1       [12613, 12721]      + |         3
       ...      ...                  ...    ... .       ...
  [289967]     chrY [59358329, 59359508]      - |    277748
  [289968]     chrY [59360007, 59360115]      - |    277749
  [289969]     chrY [59360501, 59360854]      - |    277750
  ---
  seqinfo: 93 sequences (1 circular) from hg19 genome
```

GRanges
```
length(gr); gr[1:5]
seqnames(gr)
start(gr)
end(gr)
width(gr)
strand(gr)
```

DataFrame
```
mcols(gr)
gr$exon_id
```

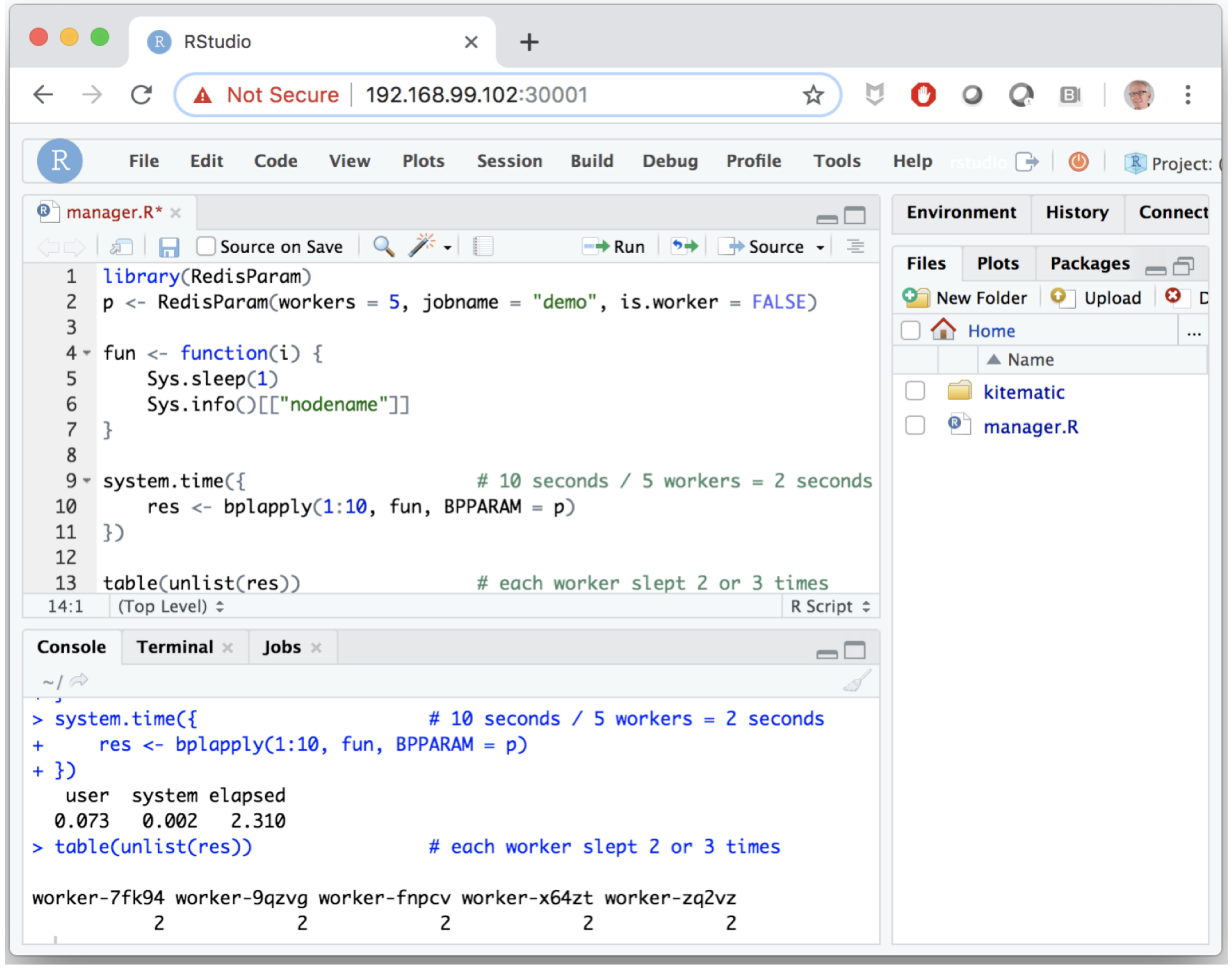
Seqinfo
```
seqlevels(gr)
seqlengths(gr)
genome(gr)
```

- Data: aligned reads, called peaks, SNP locations, CNVs, ...
- Annotation: gene models, variants, regulatory regions, ...
- `findOverlaps()`, `nearest()`, and many other useful range-based operations.

Variant Calling Workflow: https://www.bioconductor.org/help/course-materials/2014/BioC2014/Lawrence_Tutorial.pdf

Directions: Supporting bigger data

- Containers, Clouds & kubernetes
  - For more meetup Sep 25th in Boston
- ALTREP: C level ALTernative REPresentation of R objects
- Delayed implementations, LoomExperiment, support for HDF5



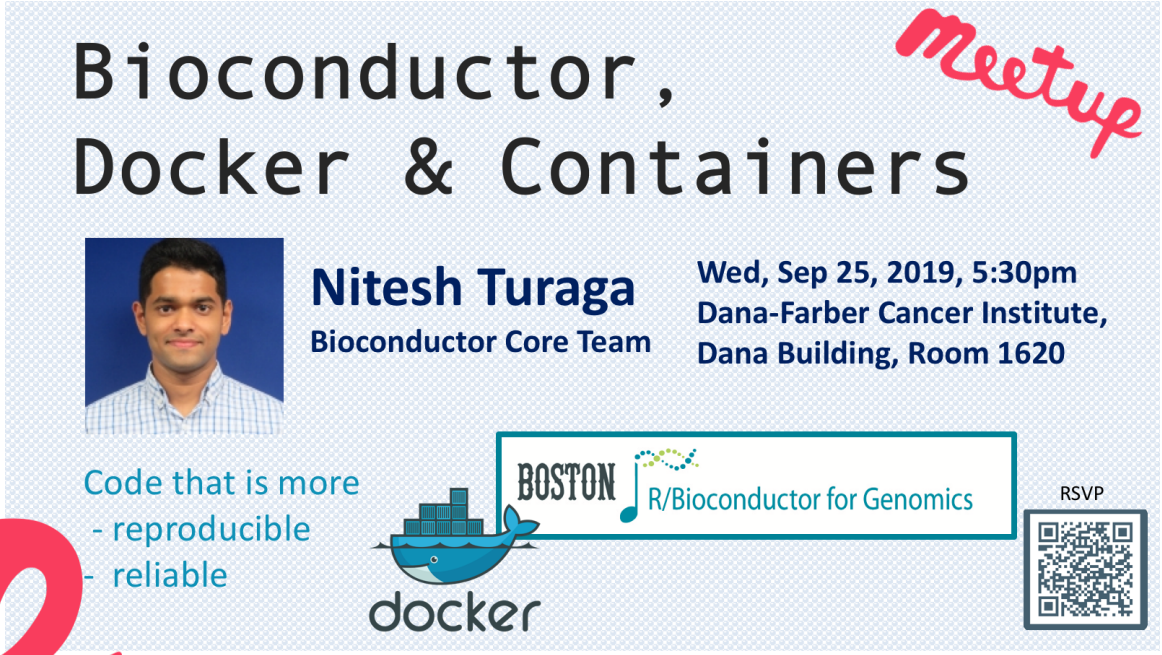
Martin Morgan, presented at Developer Day Bioc2019 http://bioc2019.bioconductor.org/schedule-developer-day

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 Sep 25th

meetup



Find Bioconductor meetups under the R Consortium account

Part of **R User Groups – 74 groups**

Boston R/Bioconductor for Genomics

Boston, MA
711 members · Public group
Organized by **Aedin Culhane** and **6 others**

Part of **R User Groups – 74 groups**

New York City R/Bioconductor for Genomics

New York, NY
780 members · Public group
Organized by **Levi Waldron** and **4 others**

Slack: join at https://bioc-community.herokuapp.com/

Acknowledgements

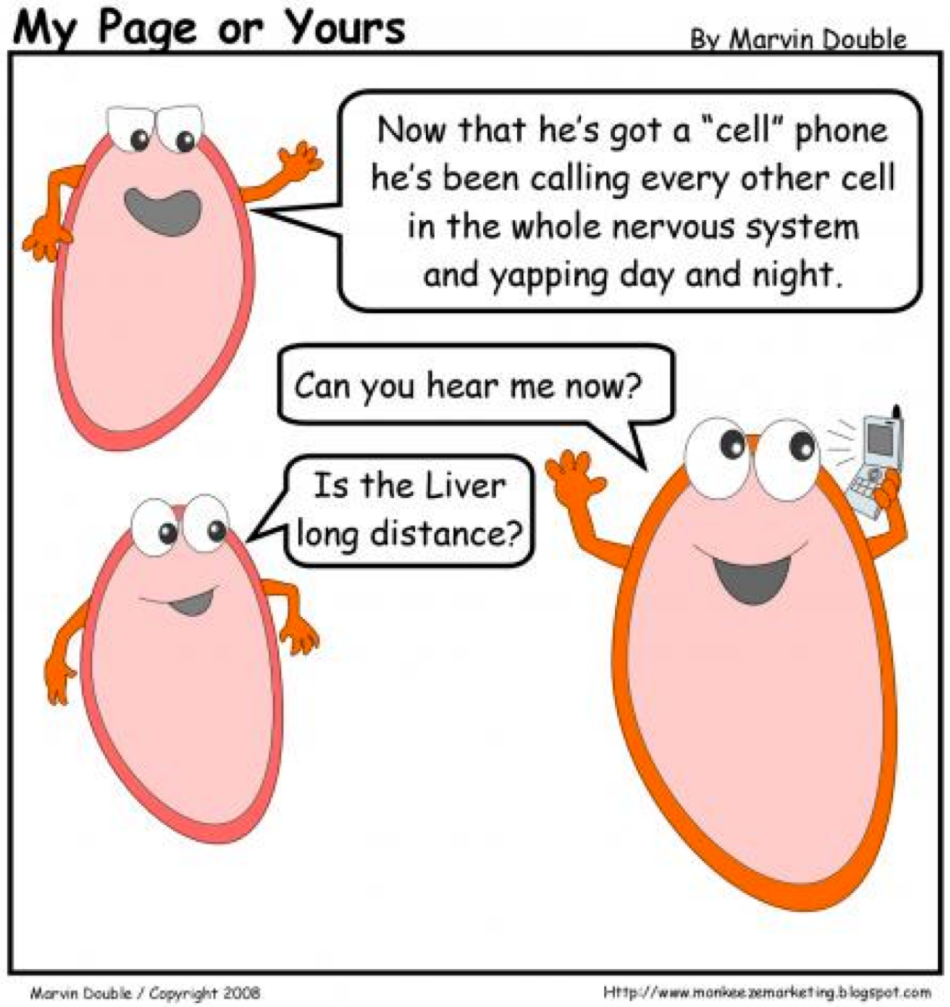
Matthew Schwede
Azfar Basunia
Chen Meng * (with Amin, Bernard)

Eugen Dhimolea, Constantine Mitsiades (DFCI)

Technische Universitaet Muenchen, Germany
Amin Moghaddas Gholami, Bernard Kuster

PanCanAtlas Immune Response Working Group
Vésteinn Thorsson
Ilya Shmulevich
Benjamin Vincent

Levi Waldron (CUNY)
Vince Carey (Channing)



@AedinCulhane

CHAN ZUCKERBERG INITIATIVE

NIH NATIONAL CANCER INSTITUTE

Congressionally Directed Medical Research Programs CDMRP Department of Defense