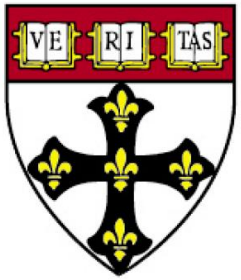
Integrated Exploratory Data Analysis: Biclustering & Meta Gene Set Analysis

Aedín Culhane

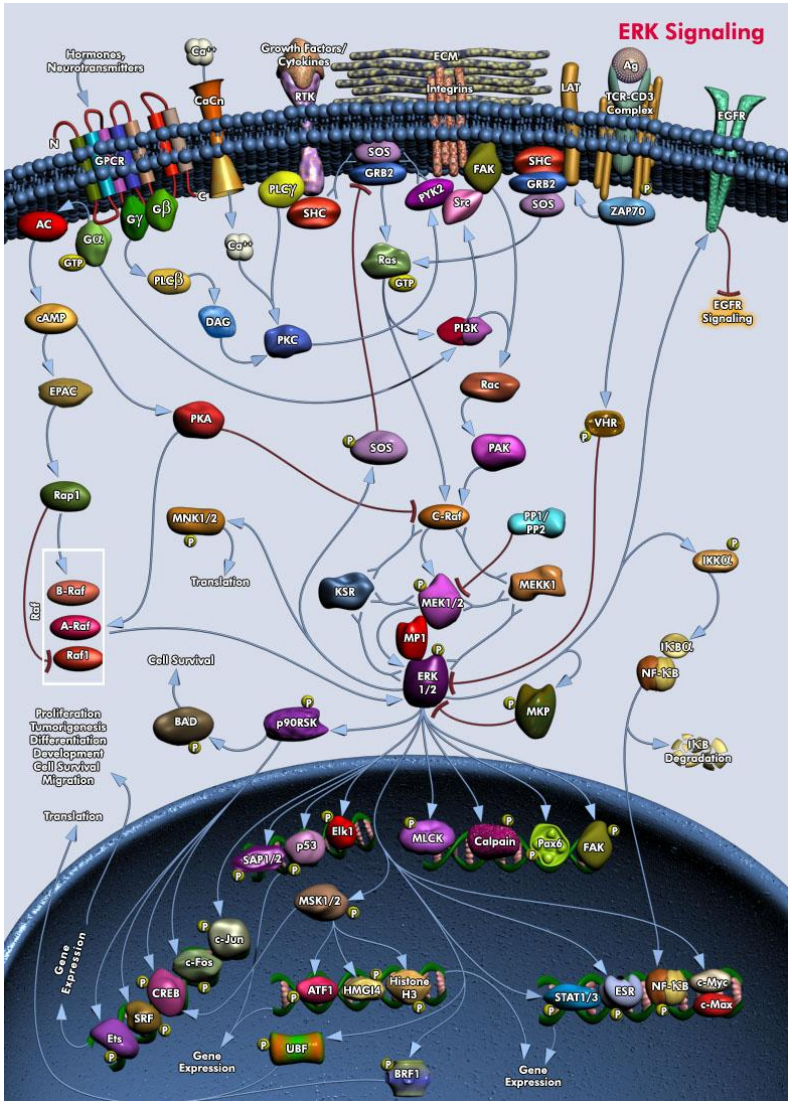
aedin@jimmy.harvard.edu

DANA-FARBER
CANCER INSTITUTE



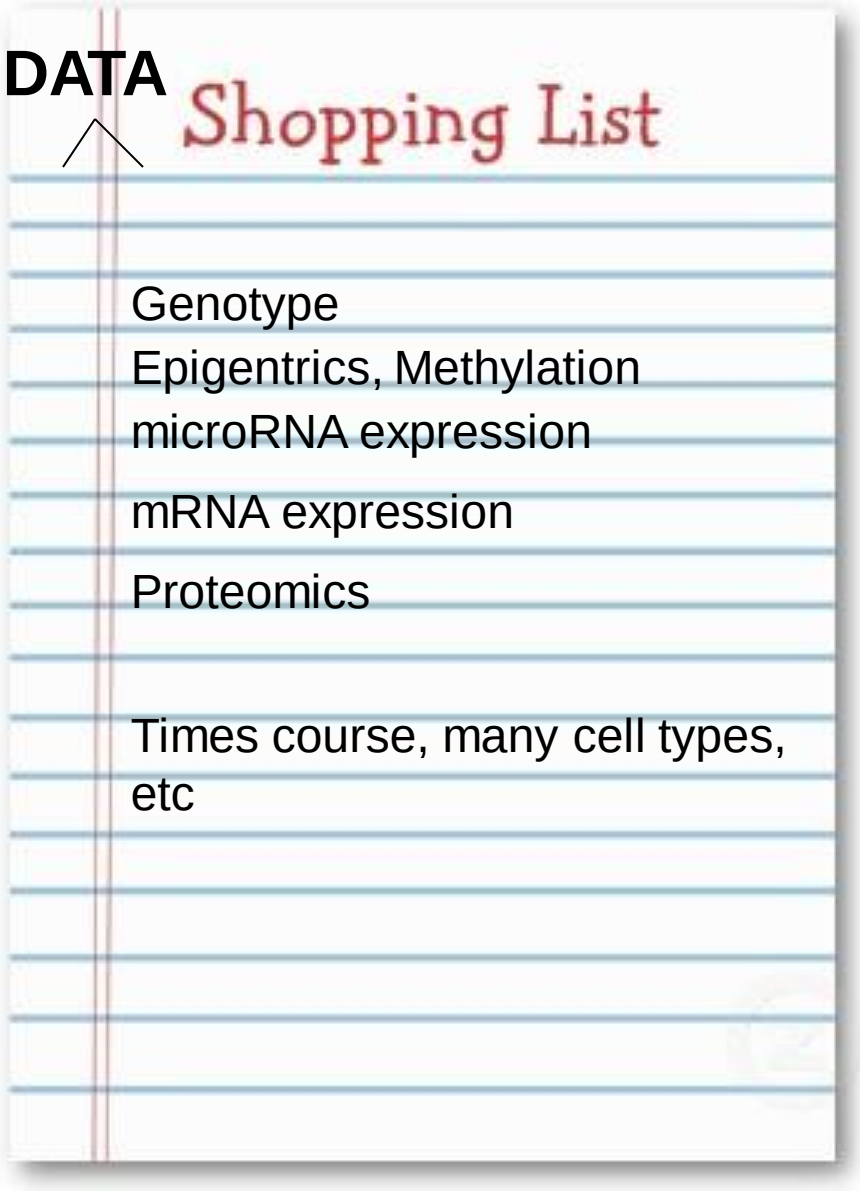
Data Analysis Challenges

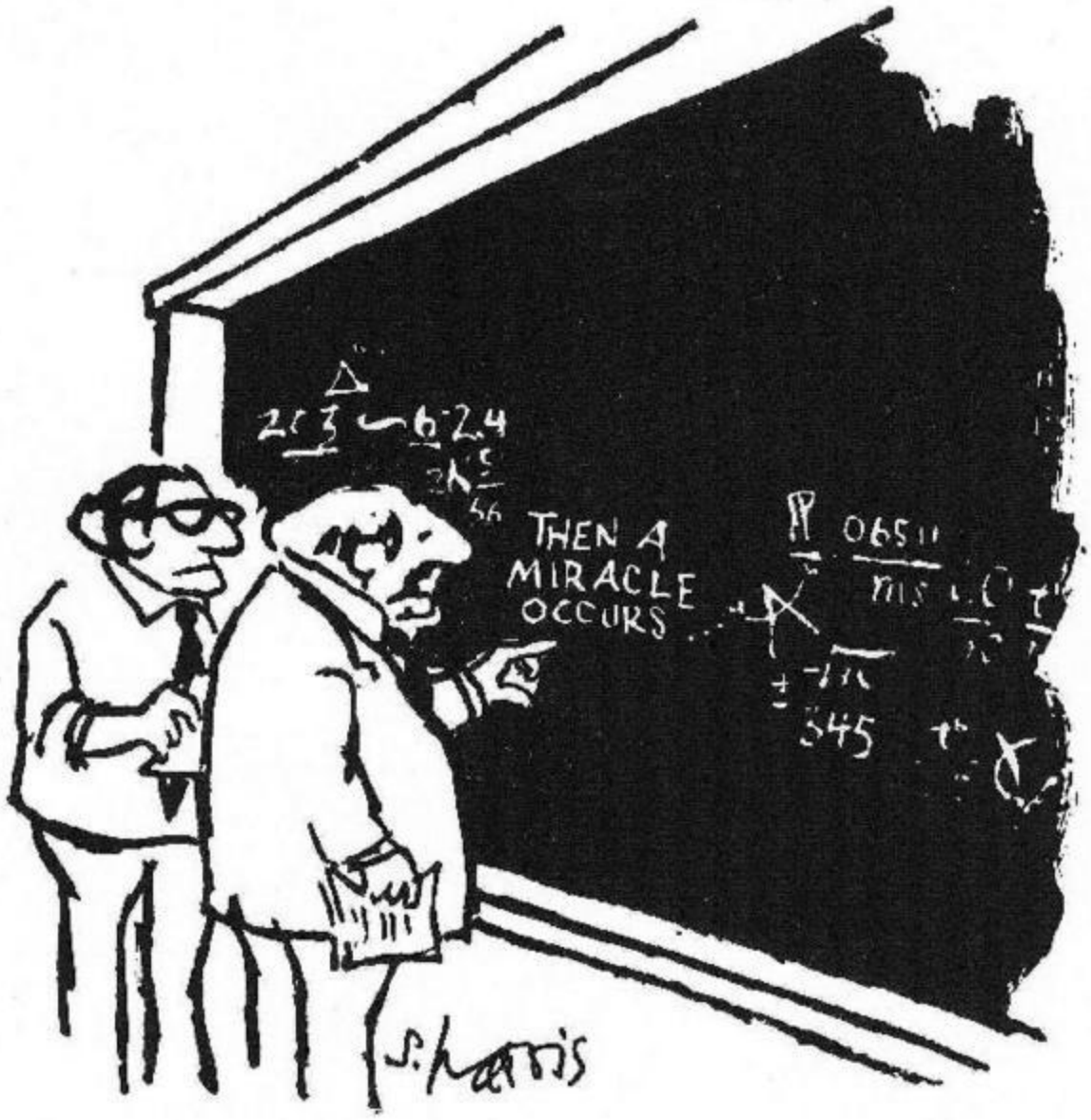
- Thousands of variables, few cases
- Noise, few calibration standards
- Limited knowledge of “correct” model
- Redundancy, crossover, feedback inhibition
- Most experiments capture a one data type (gene, protein, miRNA, etc)



GOAL:

Model of cell





"I think you should be more explicit here in step two."

No simple solution to integrated data analysis

- Combine P-values of individual analyses
  - Does rank complex complex system. Many genes with marginal effect if acting in cohort also significant impact on system
- Analyze a meta-dataset
  - Difficult to match (on genes?), loose data, imputation,
- Use a probablistic, Bayesian framework to direct integration
  - Computational expensive, maybe sensitive to initial seed or active module analysis, solution maybe difficult to interpret if analysis is not focused

.

Multivariate Methods to detecting co-related trends in data

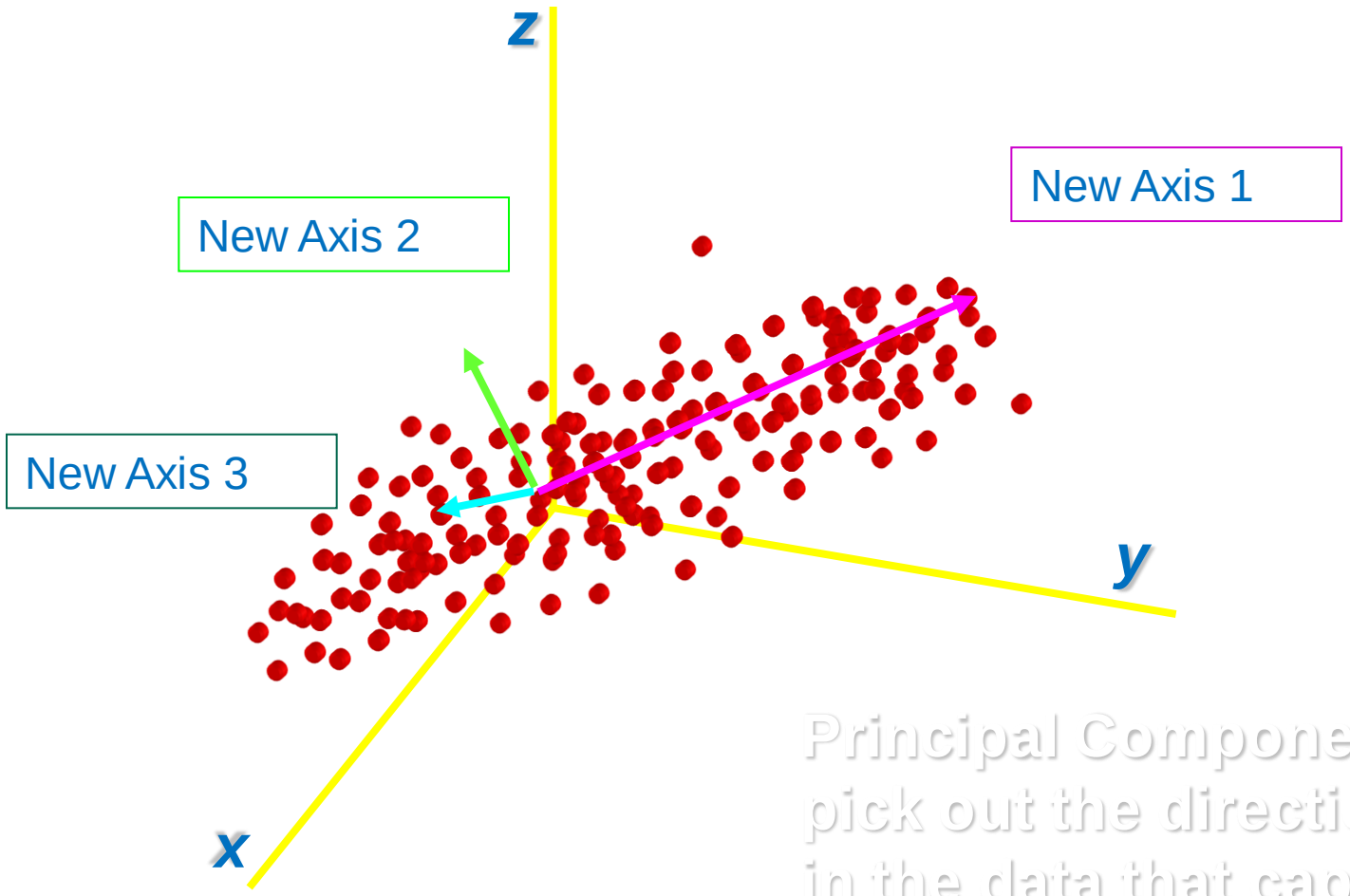
- Canonical correlation analysis
- Partial least squares
- Co-inertia analysis

Coinertia Analysis

- Useful for cross-platform comparison where the same samples have been arrayed.
- Identifies correlated “trends” in data
- Consensus and divergence between gene expression profiles from different DNA microarray platforms are graphically visualised.
- Not dependent on annotation thus can extract important genes even when there are NOT present across all datasets.

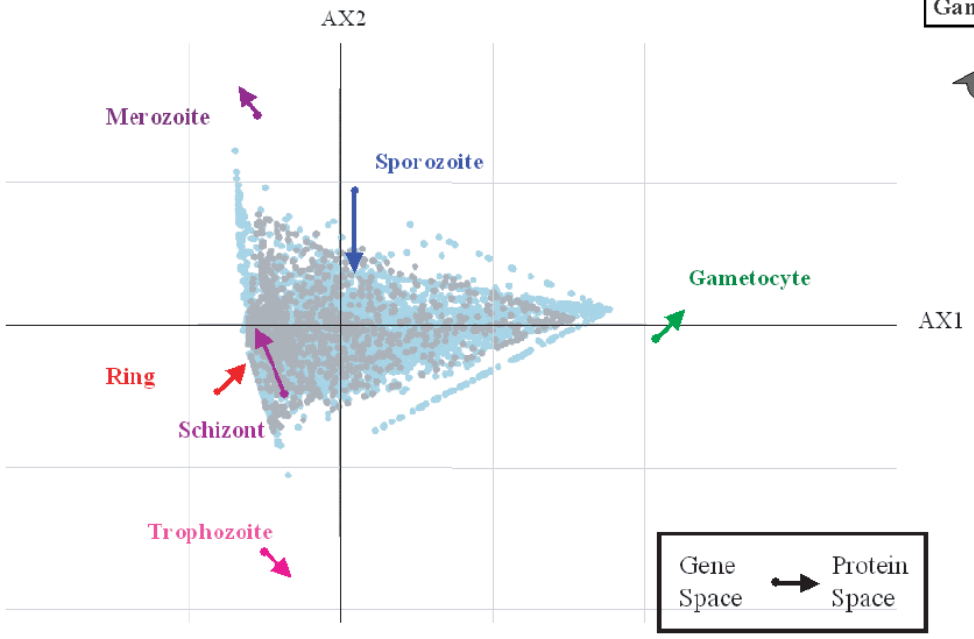
Culhane, A.C., Perriere, G., Higgins D.G., (2003) Cross platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4:59

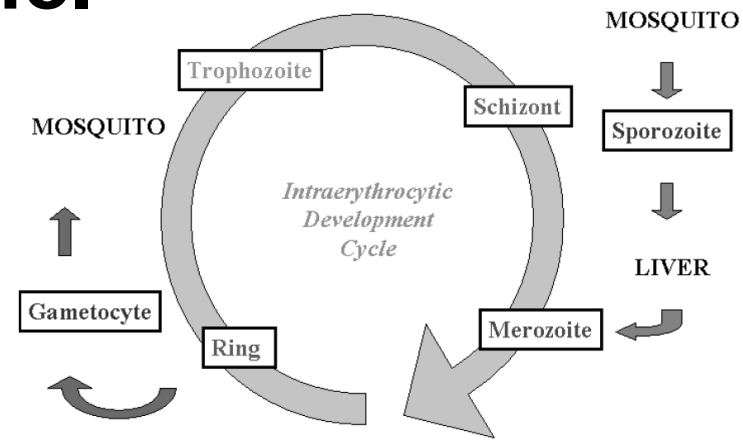
Dimension Reduction (Ordination)



Principal Components pick out the directions in the data that capture the greatest variability

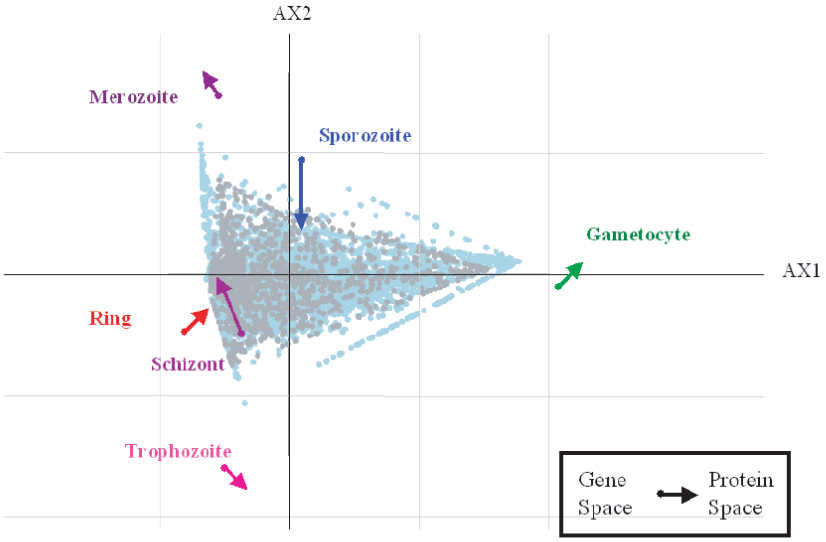
Gene expression and proteomics data from the life cycle of the malarial parasitic.



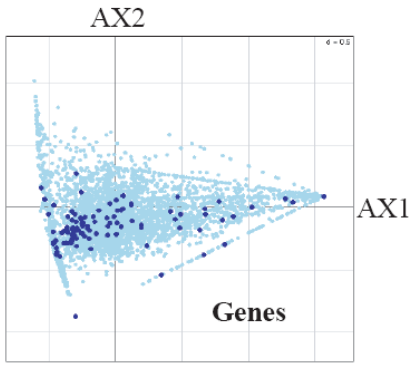


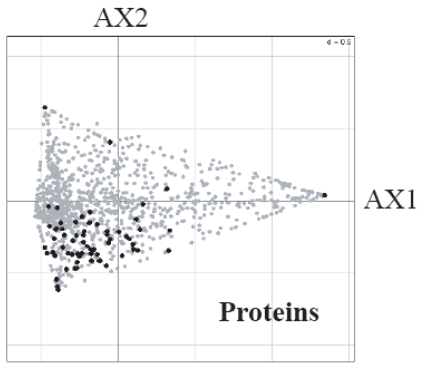
Project GO terms on Genes & Proteins space

Sample with variables (tri-plot)

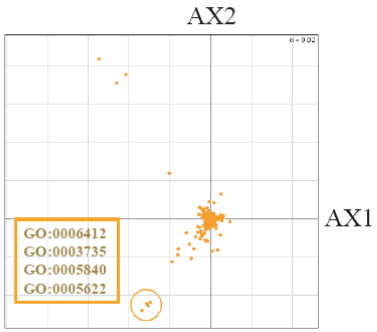


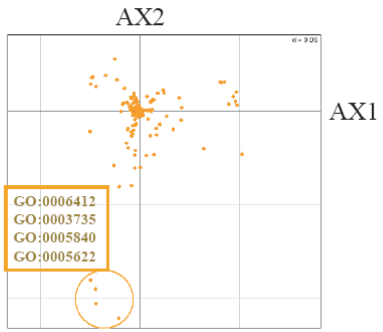
Variables





GO Terms





Axis 1 (horizontal) Accounts for 24.6% variance. Splits sexual & asexual life stages

Axis 2 (vertical) 4.8% variance. Splits invasive stages (Merozoite and Sporozoite stages which invade red blood)

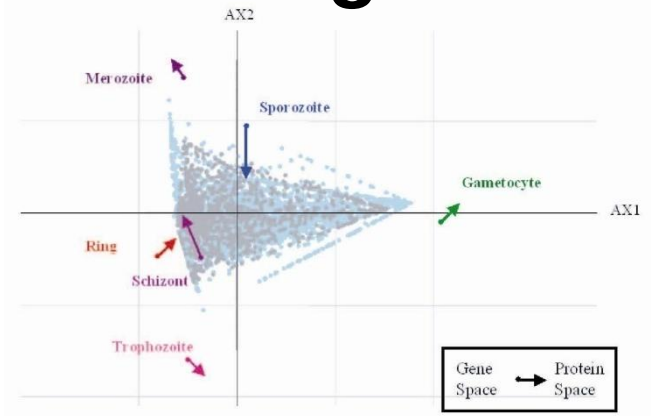
Package: made4

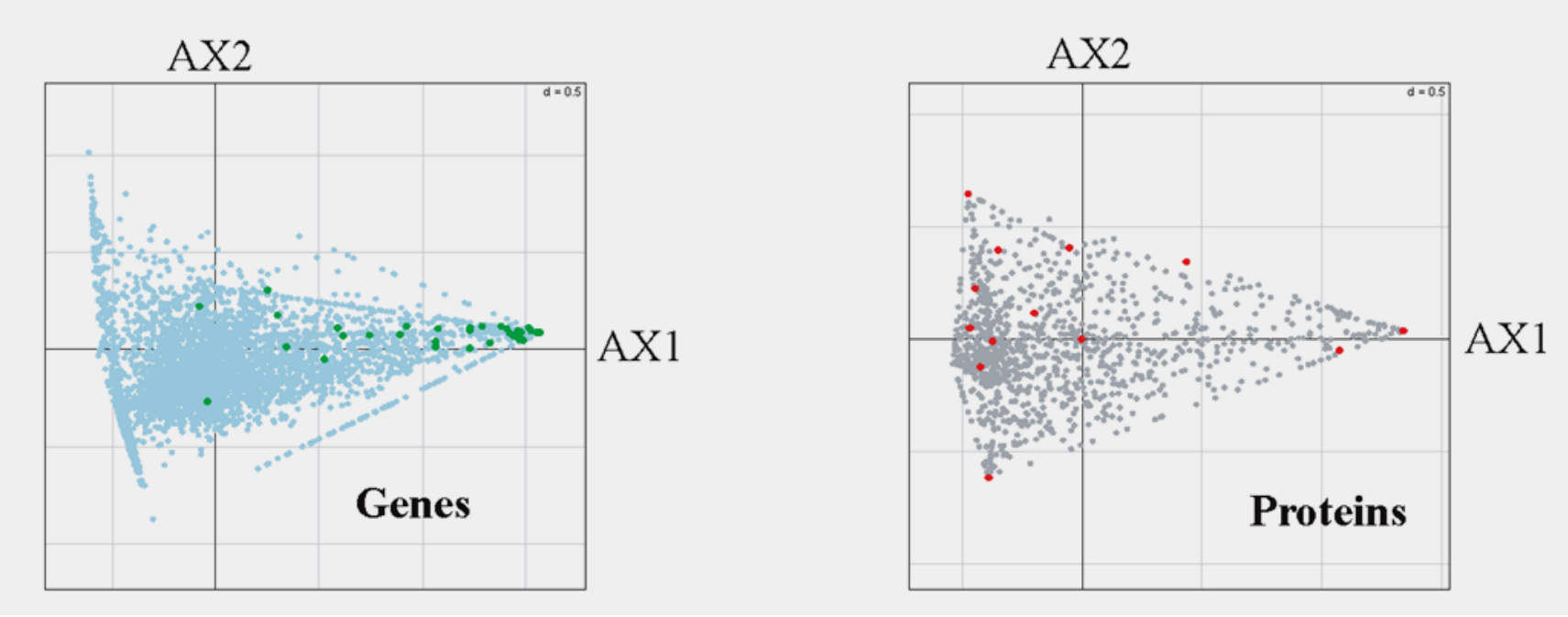
Detecting translationally repressed genes

Known: translationally repressed in female Gametocyte stage of *Plasmodium berghei*. These genes silence in the gametocyte stage but once ingested by mosquito, undergo translation into their respective proteins.

Examined *Plasmodium falciparum* orthologs

CIA: See genes transcriptionally active but their protein product is absent in the gametocyte stage.

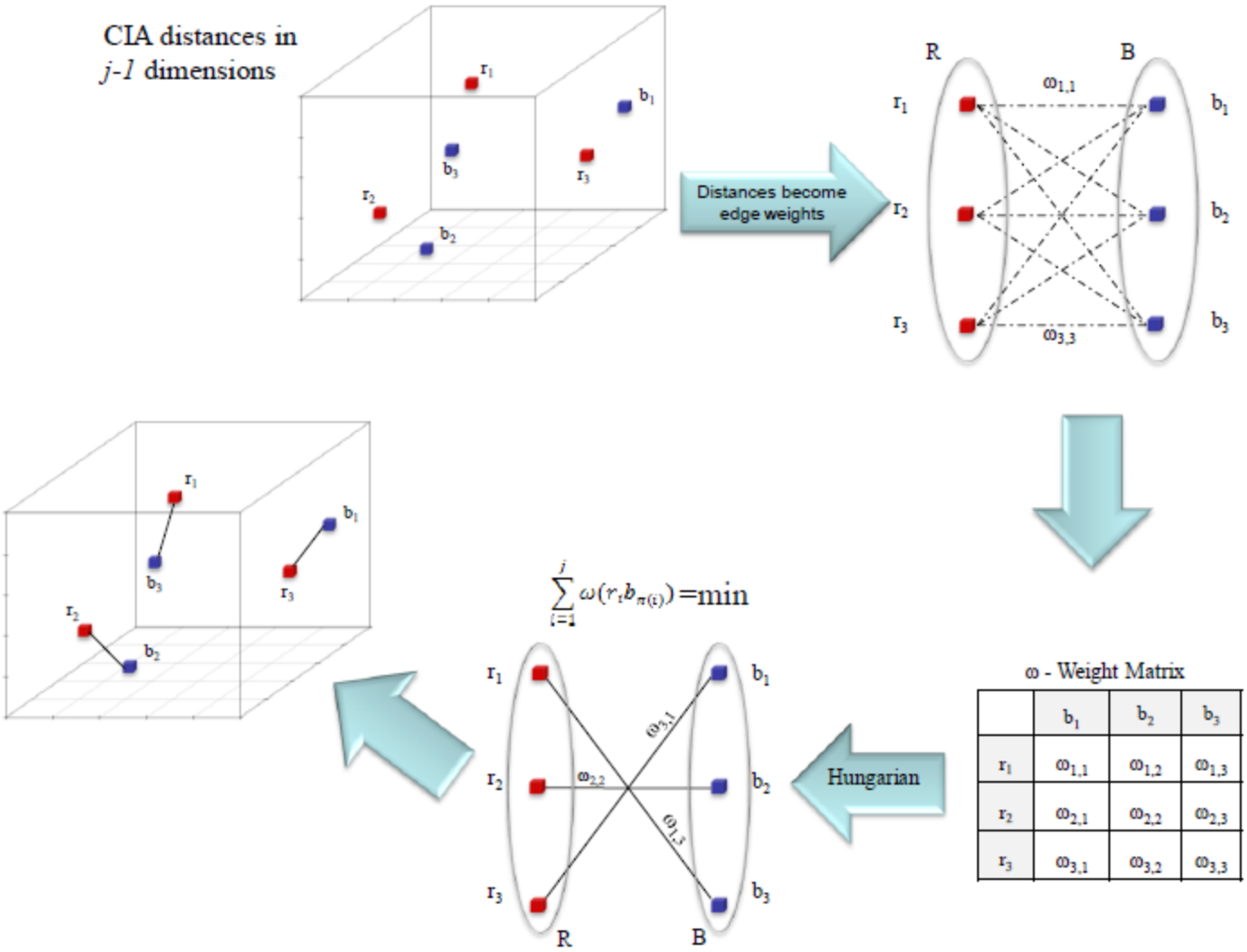




Visualising Genes, Proteins and GO terms

- CIA useful particularly to visualize variant “opposing” trends
- Addition of GO terms may assist when lack protein annotation (MS/MS data)
- Can be extended to supplement any annotation terms.

Fagan A, **Culhane AC**, Higgins DG. (2007) A Multivariate Analysis approach to the Integration of Proteomic and Gene Expression Data. *Proteomics.* 7(13):2162-71.



Cross-species common regulatory network inference without requirement for prior gene affiliation Gholami & *Fellenberg, Bioinformatics , (2010) 26:8 ,1082–1090*

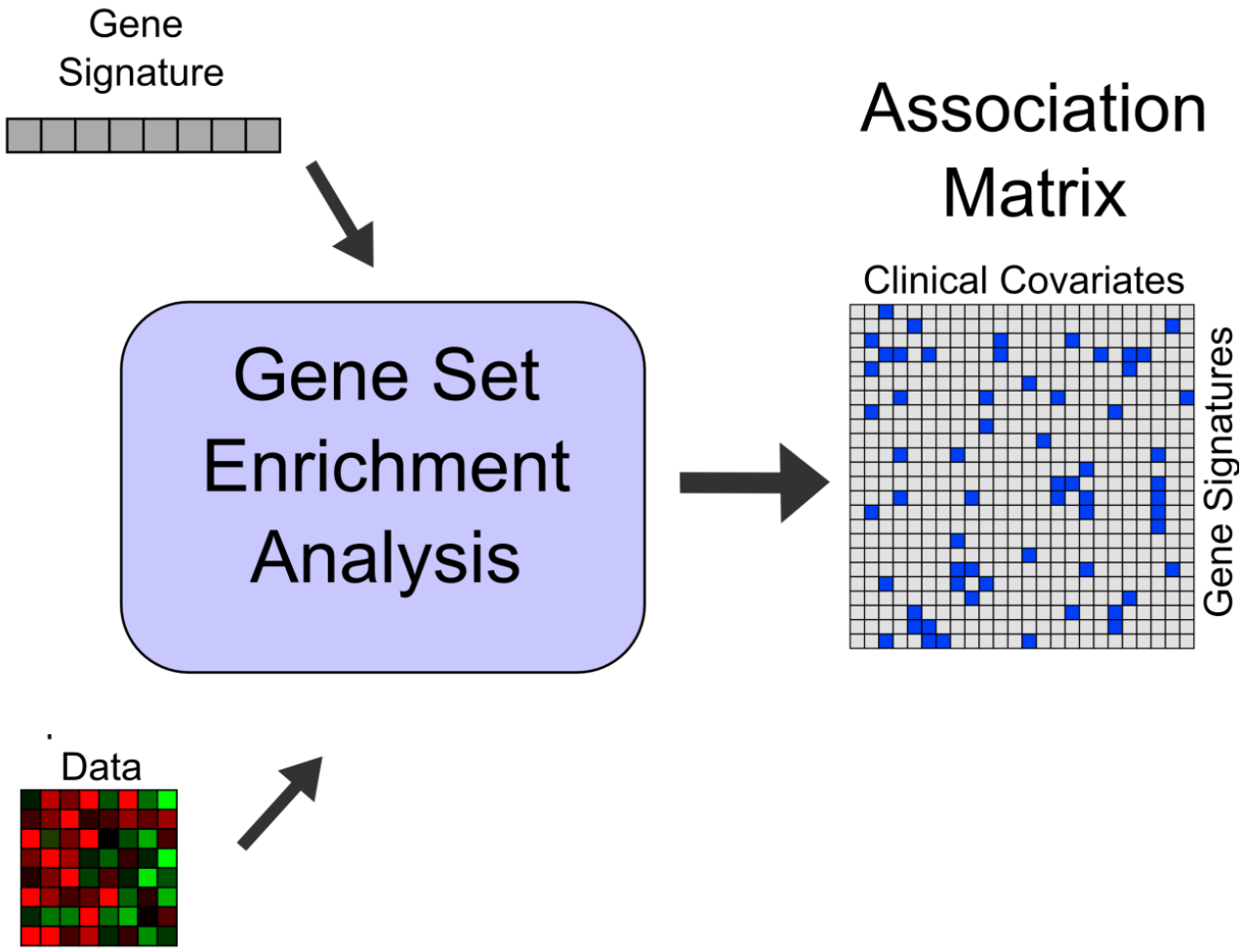
Exercise 1

Collectively Analysis of Genes

- Phenotypic characteristics or clinical diseases can only rarely be defined by one single gene

- Most diseases, are complex and involve multiple genes

Analysis Pipeline



GSA packages in Bioconductor

- GESABase,
- GOseq
- Category, GOstats and topGO
- GSEAlm
- Limma
  - *mean-rank gene-set enrichment* Michaud et al (2008) wilcox. .
  - Rotation- Roast, Romer (ROtation testing using MEan Ranks). Majewski et al (2010). Tests if up, down or both, estimates p-values by simulation
- GlobalTest, GlobalAncova

Per sample GSA

- Simple analysis
  - Order rank list.
  - t-test of genes in geneset to all others
- GSVA
- Outside Bioconductor
  - GiTools “Sample Level Enrichment Analysis”

GeneSets: GSEABase

- MSigDB
- GO
- KEGG
- Reactome
- GeneSigDB

There are different kinds of gene sets

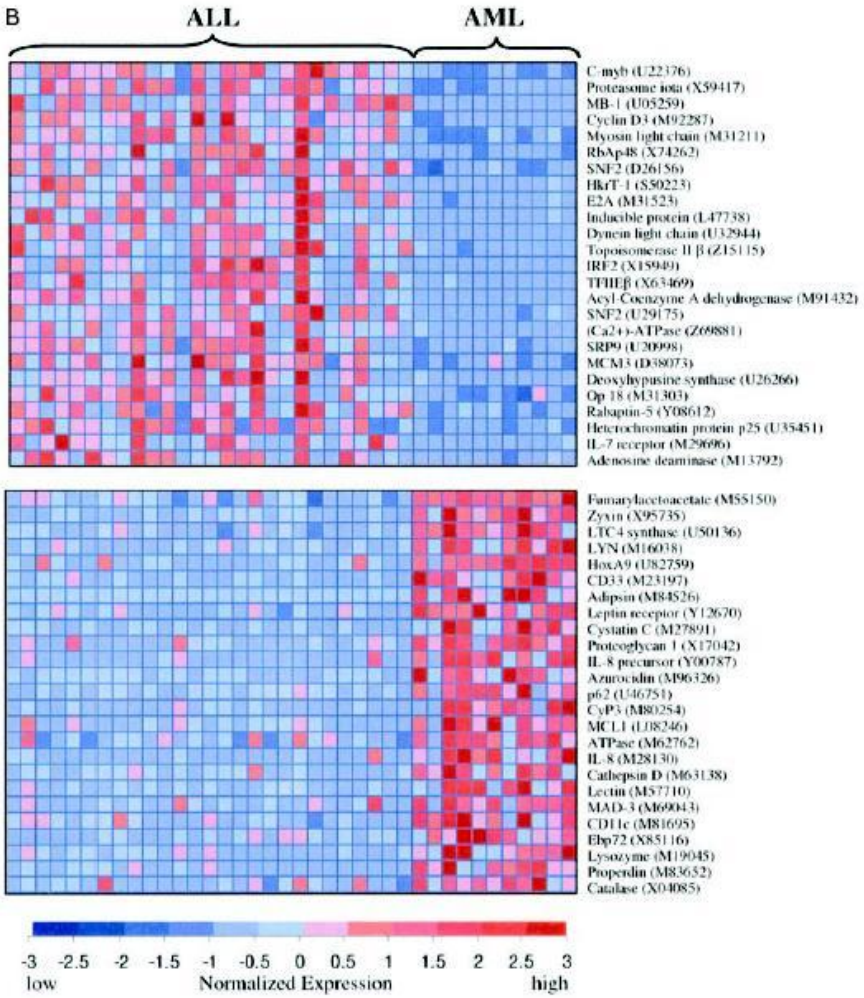
- Knowledge-driven gene sets

  require expert knowledge to construct gene sets. These are usually specific to domains of interest.

- Data-driven gene sets

  usually use high-throughput experiments in order to derive and identify sets of related genes.

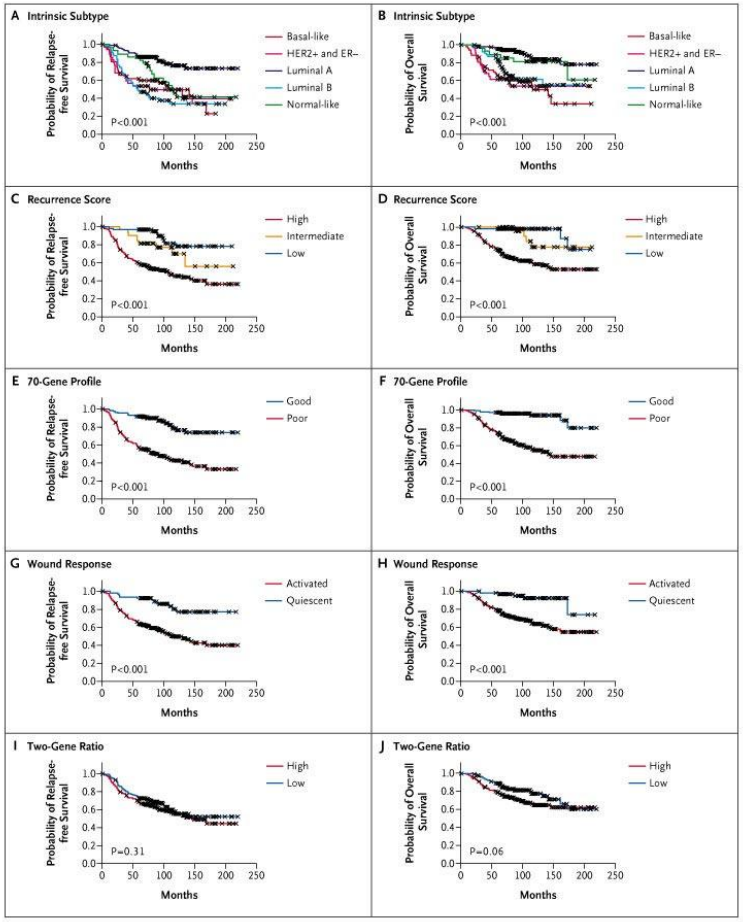
Gene Expression Signatures of cancer



Golub et al., Science 286:531-537. (1999).

Importance of Gene Signatures

- FDA approved
  - Mammaprint 70 gene signature
- Commercially available
  - Mammaprint, Oncotype DX, 76 gene veridex
- Widely used in analysis
  - Re-analyzed
  - Compared
  - GSEA
- No Standards/Public Resource



Fan et al., N Engl J Med. 2006 Aug 10;355(6):560-9.

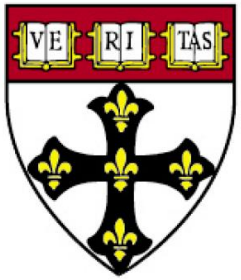
GeneSigDB

Curated Gene Signatures

- >3,500 manually curated gene signatures
- mRNA, miRNA in mouse, rat, human
- Free, to download and use

http://www.GeneSigDB.org

DANA-FARBER
CANCER INSTITUTE



GeneSigDB

Curated Gene Signatures

Publication Search

Search the full text of articles to retrieve a list of publications and the gene signatures they describe. Enter one or more search terms, such as author name, article title, journal name, or keywords.

Search Publications (e.g.: basal breast cancer)

OR

Gene Search

Search gene annotations to retrieve genes listed in GeneSigDB gene signatures.

Search Genes (e.g.: BRCA*, BRCA1)

The **Gene Sig**nature **D**ata**B**ase is a searchable database of fully traceable, standardized, annotated gene signatures which have been manually curated from publications that are indexed in PubMed. Enter a search term above to get started.

News

September, 2011: GeneSigDB Data and Website Update

We continue to expand. So far we have read and processed almost 3,000 publications to extract 3,515 genes signatures from 1,604 publications. See GeneSigDB Release 4 release notes

We have a new tag cloud Browse feature to enable easy browsing of GeneSigDB.

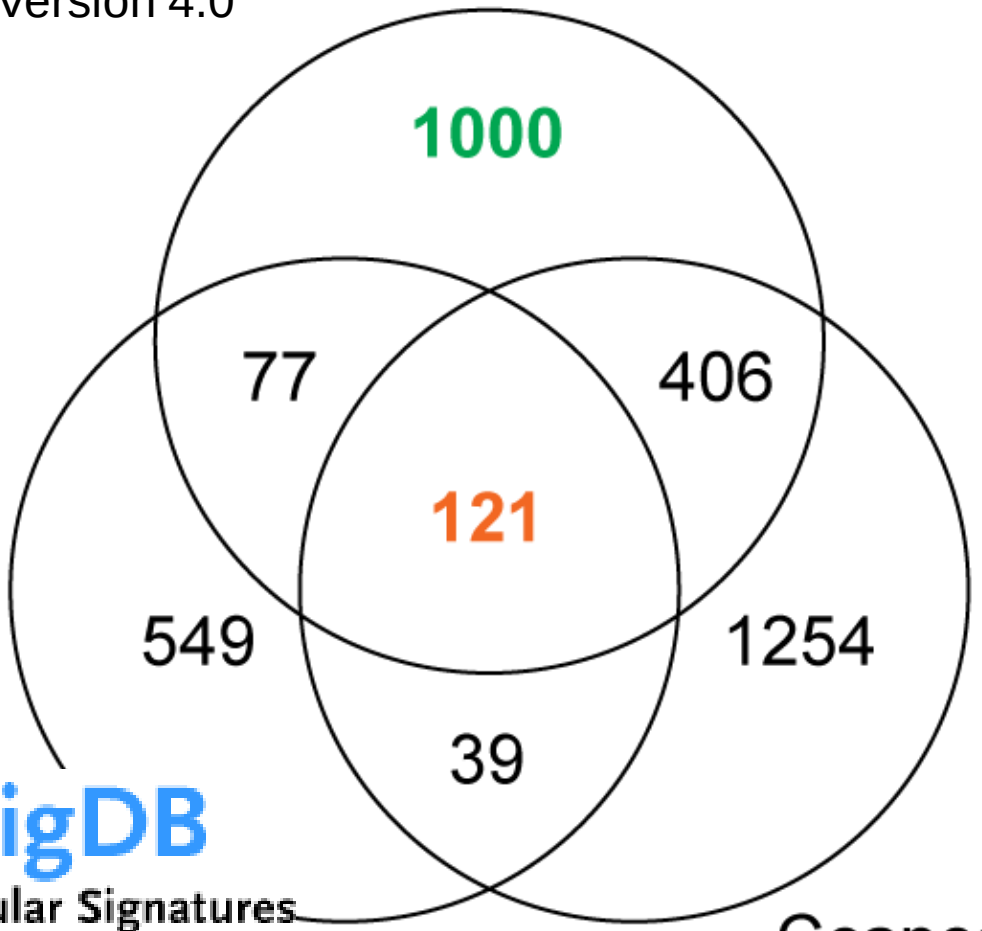
GeneSigDB Data Release 4

Gene Signatures: 3515
Published Articles: 1604
Genes (Human): 20,523
Tissues and Diseases: More than 50
Species: 3

Comparison of Gene Set Resources

GeneSigDB

Version 4.0



MSigDB

Molecular Signatures Database

Version 3.0

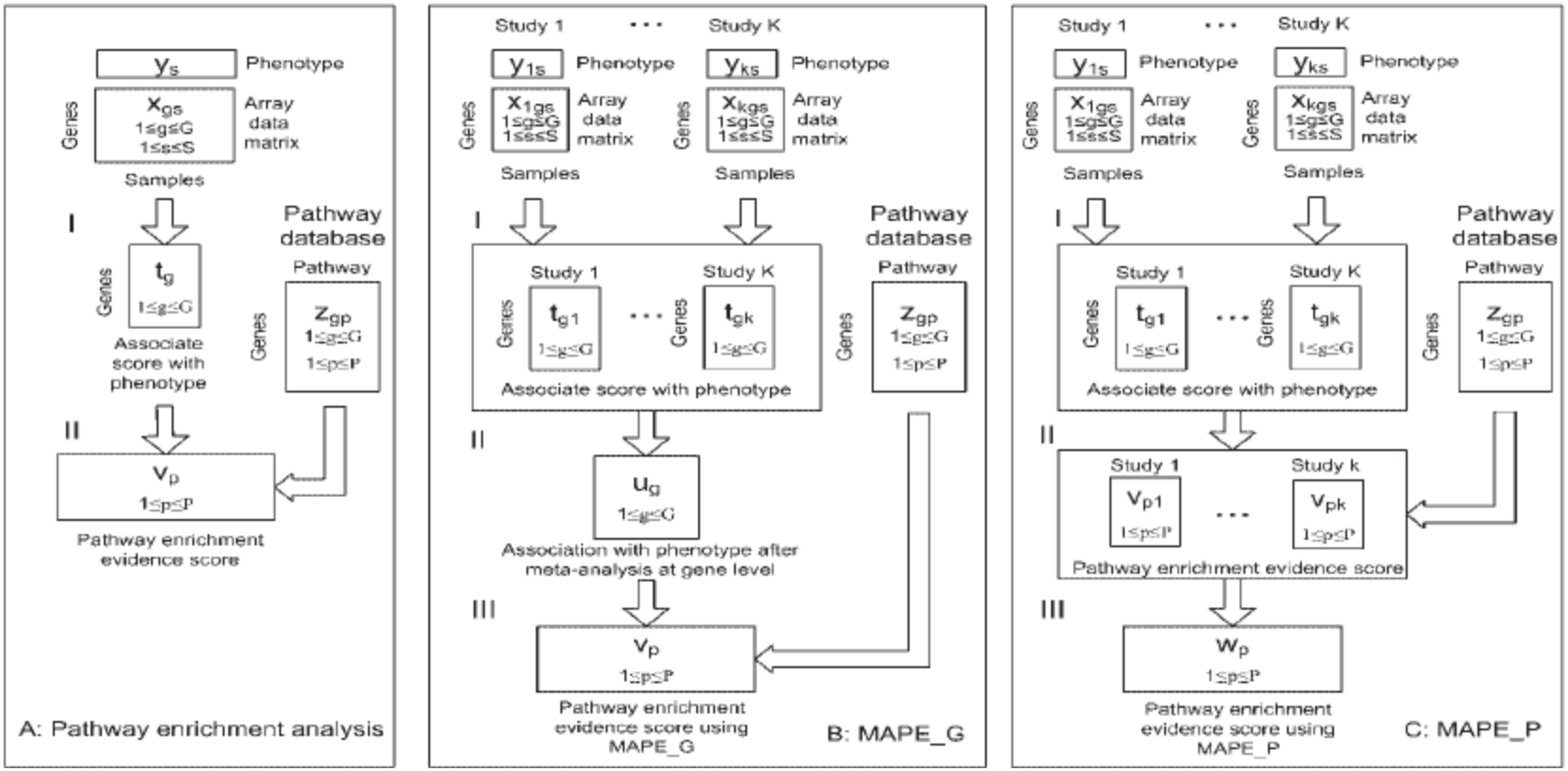
Exercise 2

Collective Analysis

- Test if a collection (of genes) are more highly ranked eg Gene Set Enrichment Analysis
- Individual genes measurements “merged” ignoring missing data
- When applied to >1 datasets, provides means to merge data without need to match individual genes

GeneSet 1	GeneSet 4	GeneSet 8	GeneSet 1
GeneSet 2	GeneSet 3	GeneSet 2	GeneSet 4
GeneSet 3	GeneSet 2	GeneSet 1	GeneSet 3
GeneSet 4	GeneSet 1	GeneSet 3	GeneSet 6
GeneSet …	GeneSet …	GeneSet …	GeneSet …
GeneSet n	GeneSet n	GeneSet n	GeneSet n

Common Meta-GSA Approaches



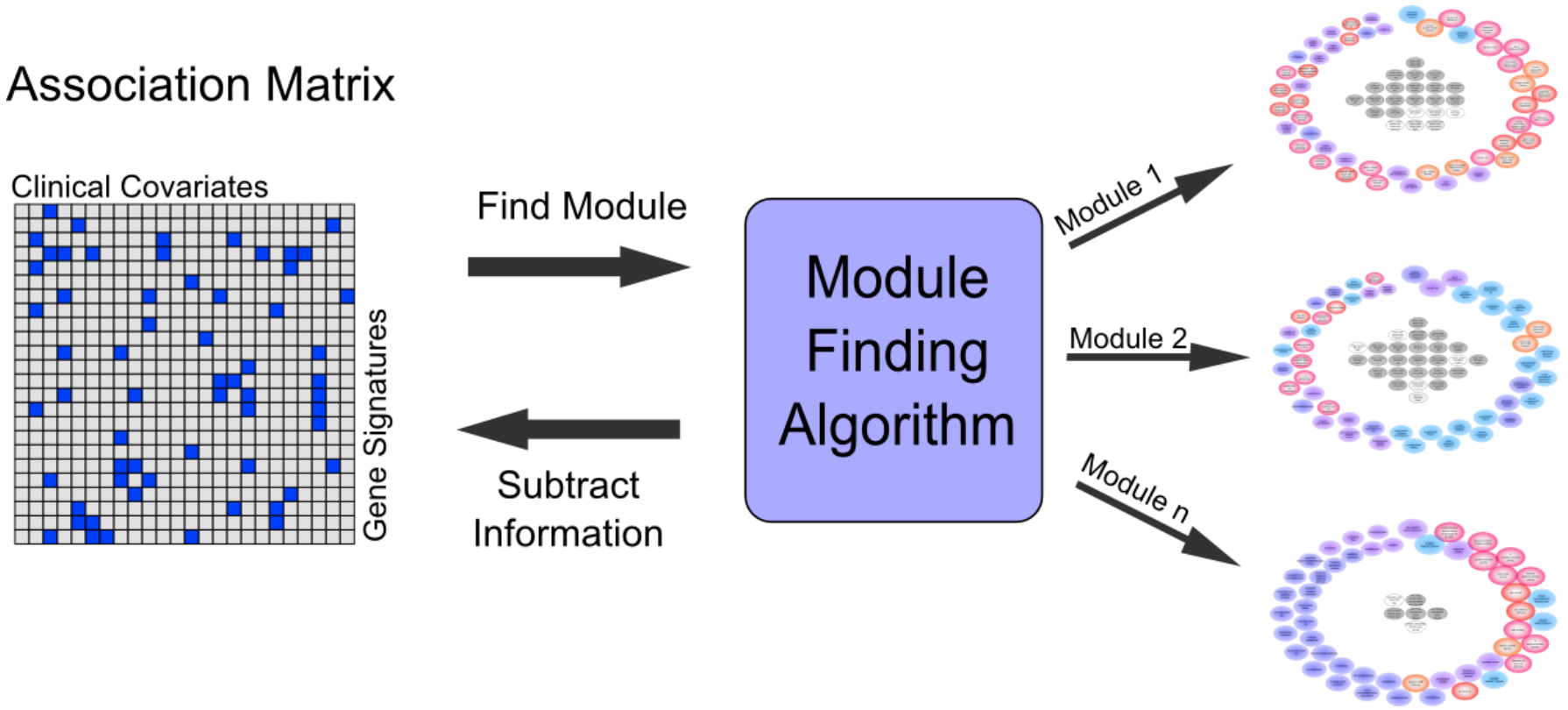
MAPE_G, MAPE_P Shen & Tseng 2010

RTopper

- Similar approach
  - Integrate individual genes -> GSA
  - GSA -> meta score (logistic regression)
- Example
  - TCGA data, 2x gene expression, 2x CNV
  - Limited to case, where all datasets have same genes and patients. Genes measured only in a subset of platforms are filtered

- Svitlana Tyekucheva, Luigi Marchionni, Rachel Karchin, and Giovanni Parmigiani. "Integrating diverse genomic data using gene sets. Genome Biology 2011, **12**:R105

Module Extracting



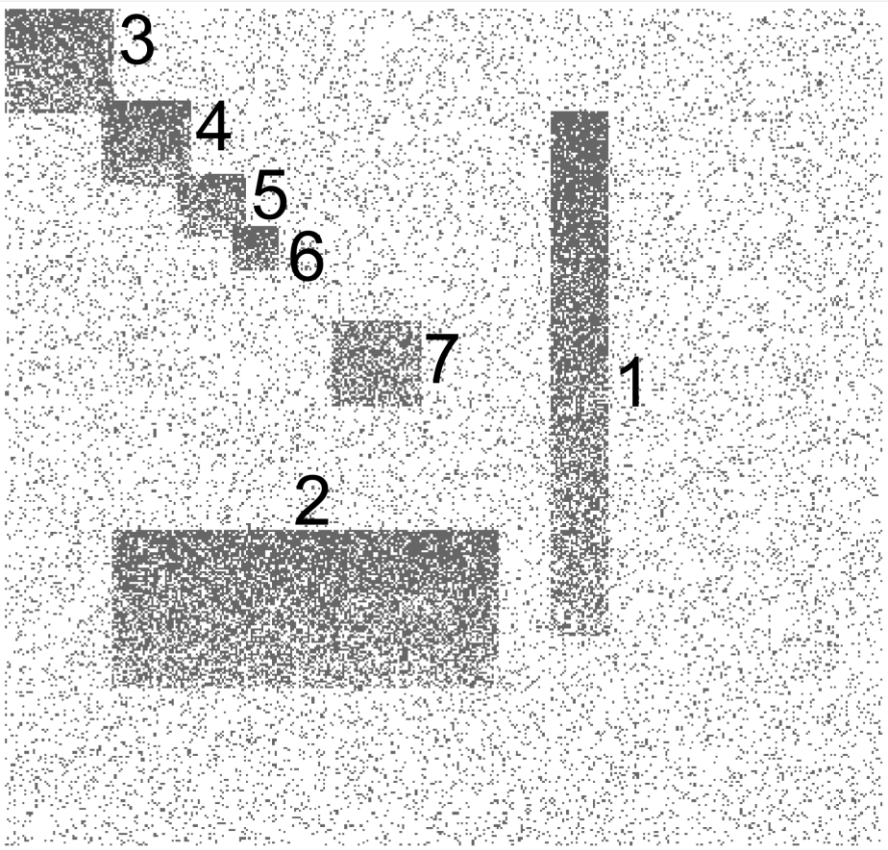
<u>A module is</u> *a group of phenotypes that are described by a ranked list of gene sets*

Simulated gene set modules

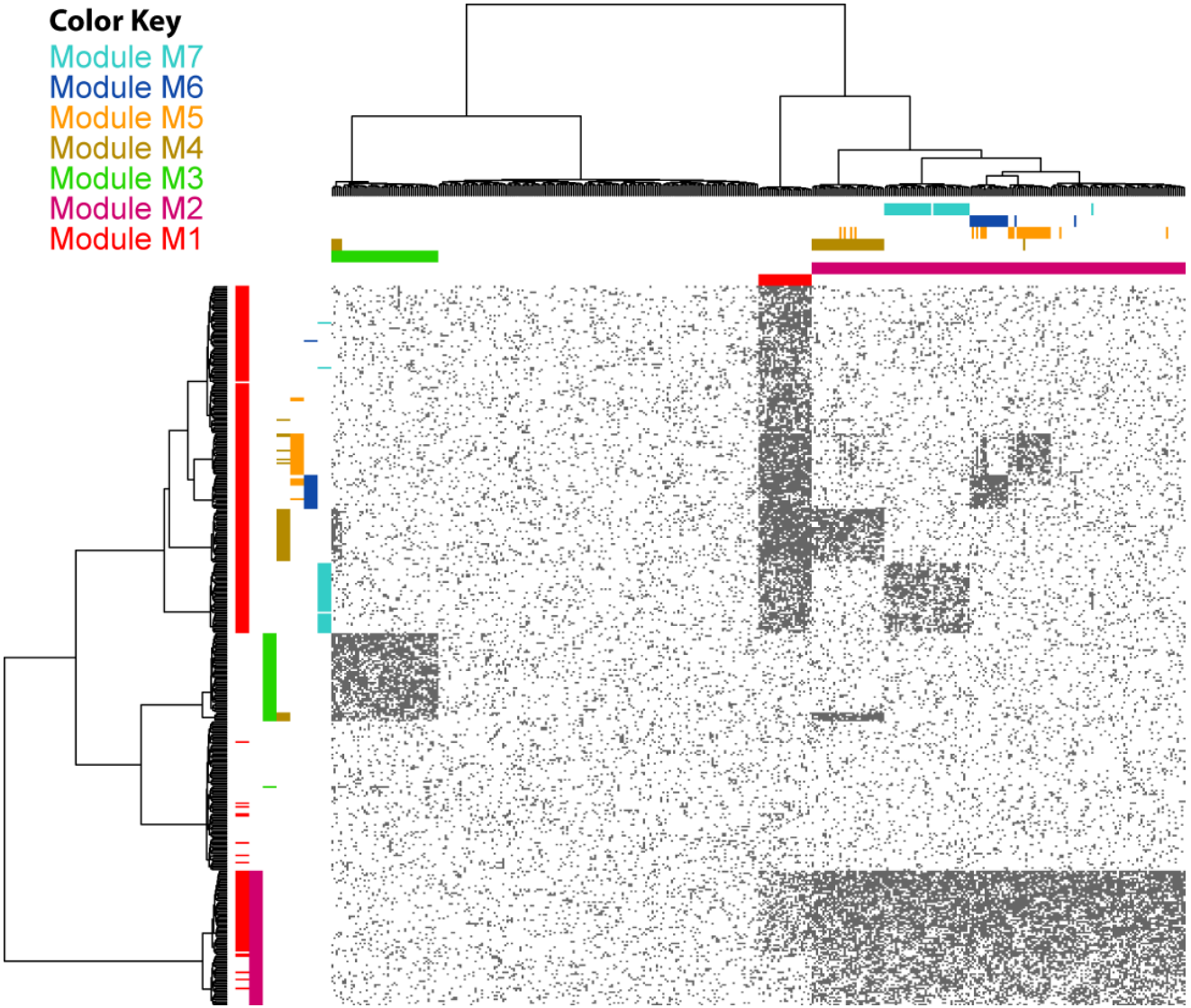
- 7 overlapping signals of various sizes
- 10% background noise

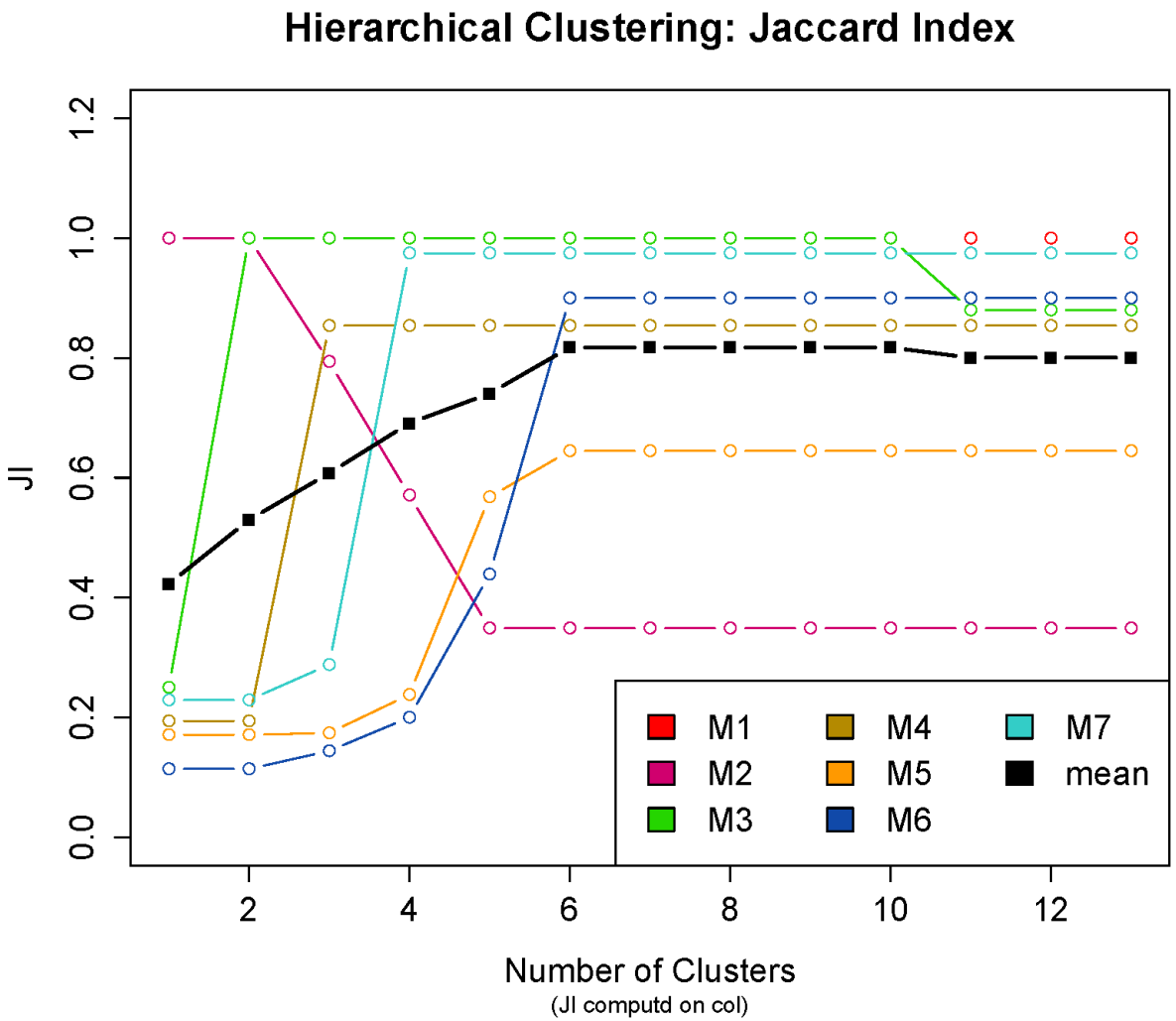
- include overlaps of clinical covariates (columns)
- overlaps of gene sets (rows)
- and overlaps in both dimensions.

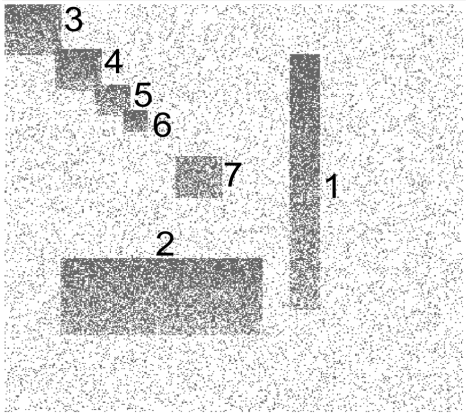
- Signals contains artificially induced noise that varies from 10% up to 60%.



Results from Hierarchical







Problem: Does not allow for overlapping membership

Objective

- Identify the pathways or molecular states that characterize cancer across different tissues

Daniel Gusenleitner

Results of BiMax, Fabia

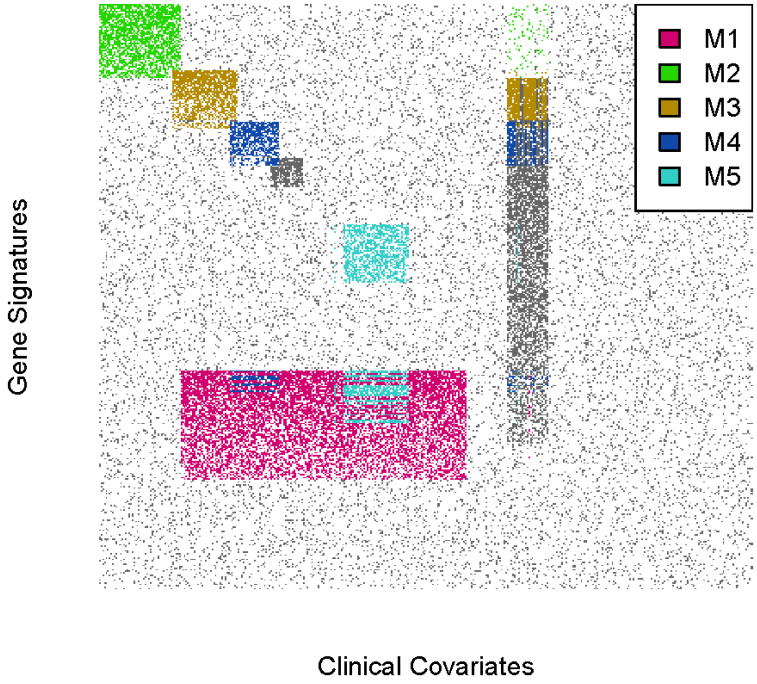
Table 13. Results of Bimax with parameters R22 and C4, which had highest JI sum over the 7 modules. JI between columns and rows are provided. Accur. Sens amd Spec are accuracy, sensitivity and specificity respectively. 200 clusters were identified in each run and the cluster which had the highest JI to each module is listed.

Module	Best	JI		Cluster Size		Covariate (Col)					Gene Set (Row)				
	Cluster	(col)	(row)	nRow	nCol	Accur	Sens	Spec	PPV	NPV	Accur	Sens	Spec	PPV	NPV
M1	168	0.24	0.30	41	6	0.95	0.24	1.00	1.00	0.95	0.48	0.16	1.00	1.00	0.42
M2	101	0.02	0.30	22	4	0.57	0.02	1.00	1.00	0.57	0.86	0.28	1.00	0.95	0.86
M3	1	0.08	0.50	24	4	0.88	0.08	1.00	1.00	0.88	0.94	0.48	1.00	1.00	0.93
M4	33	0.10	0.37	26	4	0.91	0.10	1.00	1.00	0.91	0.88	0.25	0.96	0.38	0.92
M5	110	0.10	0.40	23	4	0.93	0.10	1.00	0.75	0.93	0.92	0.37	0.97	0.48	0.95
M6	109	0.04	0.14	22	4	0.94	0.05	0.99	0.25	0.95	0.90	0.00	0.94	0.00	0.95
M7	101	0.02	0.17	22	4	0.90	0.02	0.99	0.25	0.90	0.84	0.00	0.94	0.00	0.89

Table 9. Results of FABIA with parameters p=8, alpha=0.2, cyc=1000. Alpha is the spareness loading, p is the number of clusters and cyc is the number of cycles. Accur, Sens, Spec, PPV and NPV are accuracy, sensitivity, specificity positive predictive value (precision) and negative predictive value respectively FABIA identified large clusters with many false positives.

Module	JI	Cluster Size		Covariate (Col)					Gene Set (row)				
		nRow	nCol	Accur	Sens	Spec	PPV	NPV	Accur	Sens	Spec	PPV	NPV
M1	0.15	83	352	0.18	1.00	0.13	0.07	1.00	0.39	0.18	0.74	0.53	0.35
M2	0.09	10	182	0.98	1.00	0.97	0.96	1.00	0.82	0.09	0.99	0.70	0.83
M3	0.90	45	97	0.88	1.00	0.87	0.52	1.00	0.99	0.90	1.00	1.00	0.99
M4	0.80	32	105	0.84	1.00	0.82	0.38	1.00	0.98	0.80	1.00	1.00	0.98
M5	0.77	23	127	0.76	1.00	0.74	0.24	1.00	0.98	0.77	1.00	1.00	0.98
M6	1.00	20	145	0.69	1.00	0.67	0.14	1.00	1.00	1.00	1.00	1.00	1.00
M7	0.82	33	111	0.82	1.00	0.80	0.36	1.00	0.98	0.82	1.00	1.00	0.98

Biclustering - COALESCE



Iterative Binary Bi-clustering of Gene Set Analyses: iBBiG

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

Bioinformatics. In Review

iBBiG Algorithm

I.) Run genetic algorithm

a) Initialize population with random covariate groupings (modules)

b) Calculate fitness score for every individual based on the module size and the entropy of the associations for every single signature

c) Select parents for the next generation

d) Create children using recombination and mutation

e) Repeat b-d) until the population converges

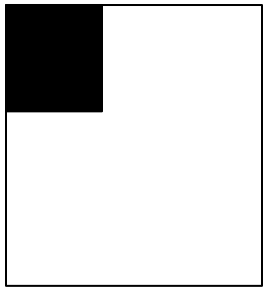
f) The individual with the highest fitness score describes the strongest module: a list of covariates that belong together described by a ranked list of signatures

II.) Extract information used in the strongest individual from the association matrix.

III.) Rerun Algorithm

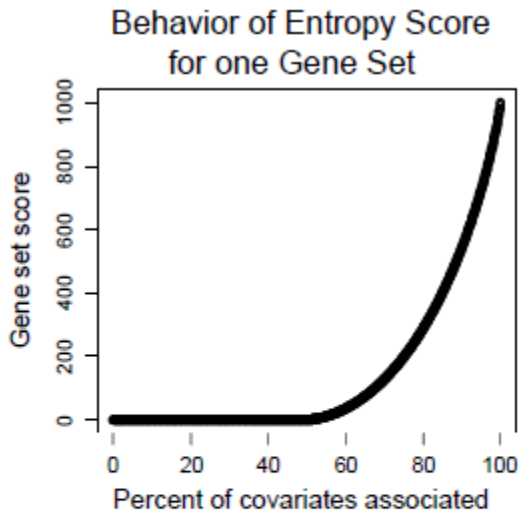
iBBiG: Score (GA)

$$p_i = \frac{\sum_{j=1}^{k} M_{ij}}{k}$$



$$H_i = -p_i \log_2 p_i - (1 - p_i) \log_2 (1 - p_i)$$

$$S_i = \begin{cases} \sum_{j=1}^{k} W_{ij}(1 - H_i)^{\alpha} & \text{if } p_i > 0.5 \\ 0 & \text{if } p_i \leq 0.5 \end{cases}$$



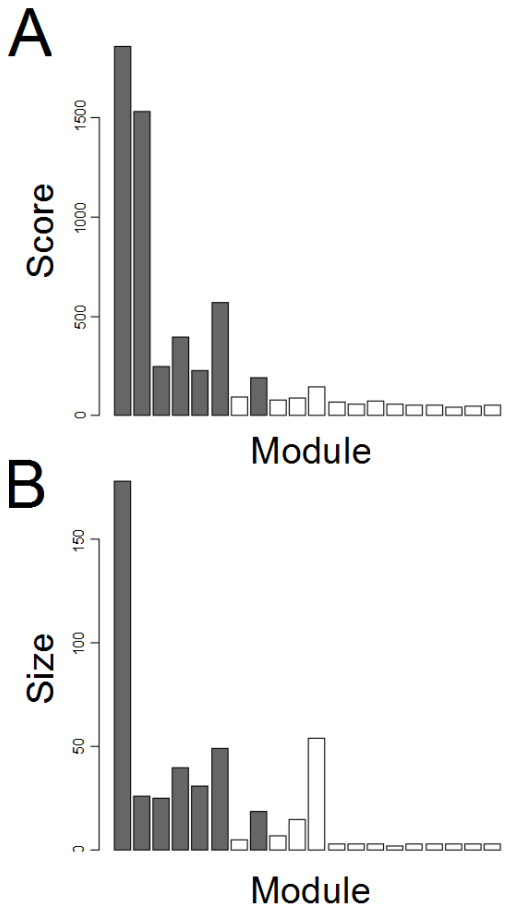
Supplementary Fig. 1. Behavior of the entropy based score for one single gene set and 1000 clinical covariates. The left side represents a situation in which the gene set is not associated with any of the clinical covariates within the chosen grouping, whereas the right side indicates a strong association with all clinical covariates.

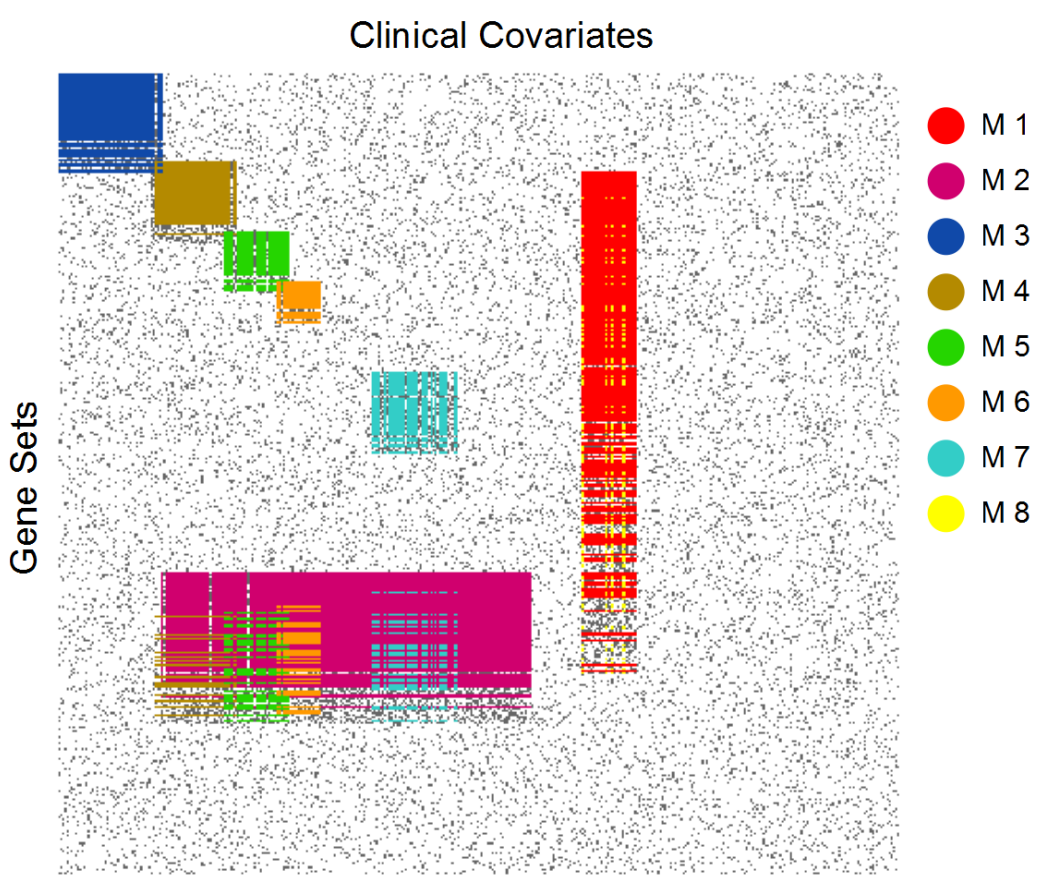
iBBiG

Table 14. Average results of iBBiG analyses on artificial dataset. Default parameters; alpha of 0.3, a selection pressure of 1.2, a population size of 100, a mutation rate of 0.8 and a success ratio of 0.6 was used for the GA. 100 runs of iBBiG were performed to test the robustness of iBBiG. Results of the best run are given in Table 15. Accur, Sens and Spec are accuracy, sensitivity and specificity respectively.

Module	JI	Cluster Size		Covariate (Col)					Gene Set (Row)				
		nRow	nCol	Accur	Sens	Spec	PPV	NPV	Accur	Sens	Spec	PPV	NPV
M1	0.99	114.88	24.73	1.00	0.99	1.00	1.00	1.00	0.66	0.46	1.00	1.00	0.53
M2	0.98	39.04	173.79	0.99	0.99	1.00	1.00	0.99	0.91	0.52	1.00	1.00	0.90
M3	0.99	33.40	49.34	1.00	0.99	1.00	1.00	1.00	0.96	0.67	1.00	1.00	0.95
M4	0.97	22.57	39.41	1.00	0.97	1.00	0.99	1.00	0.95	0.56	1.00	0.99	0.95
M5	0.82	19.80	34.96	0.97	0.87	0.98	0.95	0.99	0.96	0.54	0.99	0.87	0.96
M6	0.70	19.42	36.96	0.94	0.82	0.95	0.76	0.99	0.96	0.57	0.98	0.70	0.98
M7	0.74	27.19	29.82	0.97	0.74	1.00	1.00	0.97	0.95	0.58	0.99	0.86	0.96

Clustering with iBBiG





Gusenleitner et al., Bioinformatics, in review

Summary iBBiG

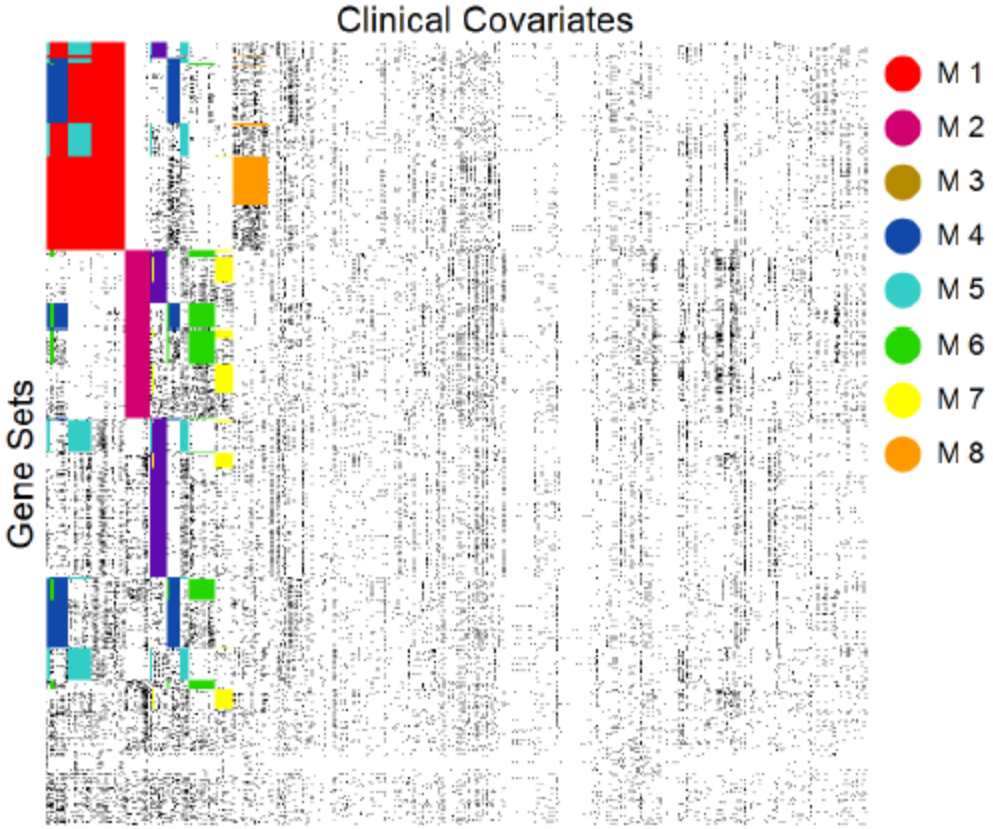
- Biclustering algorithm optimised for sparse binary data
- No requirement to pre-specifiy number or size of clusters
- Finds overlapping clusters
- When applied to our simulated data, outperforms FABIA, bimax

Packages: iBBiG (in prepartion)

Exercise 3

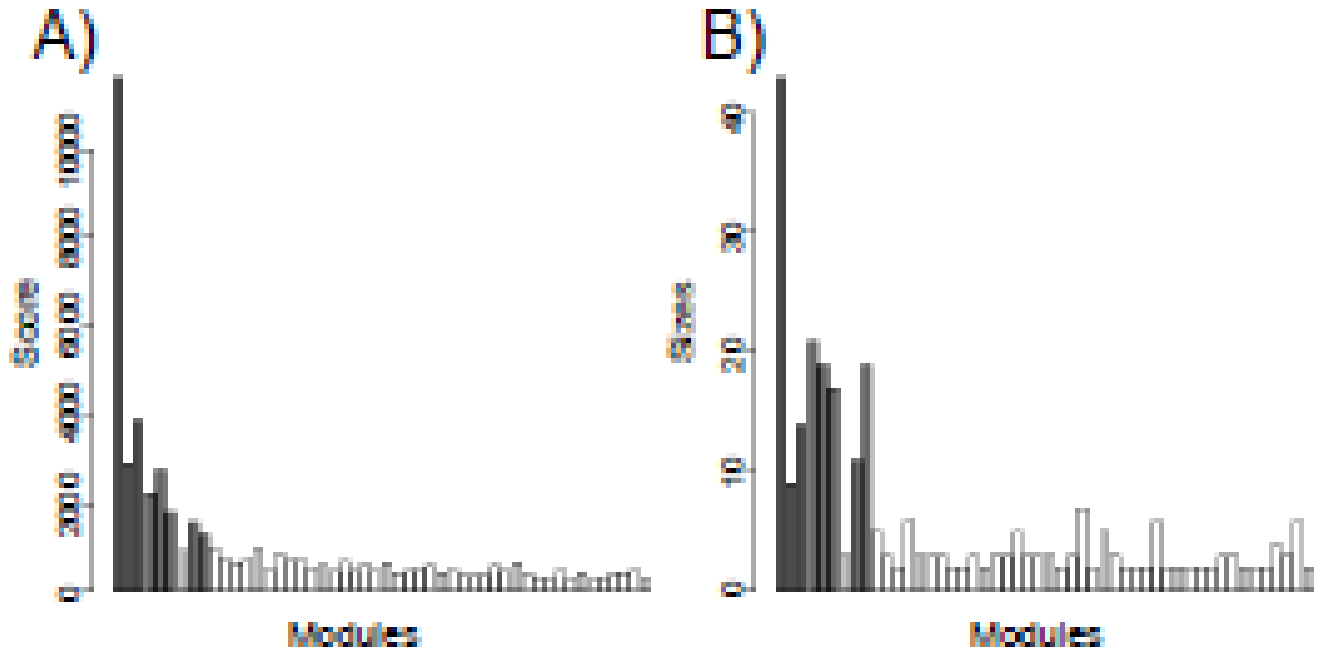
Application to 21 Breast Cancer Datasets

(3875 profiles, 446 covariates, 2,853 gene sets)



Application to 21 Breast Cancer Datasets

(3875 profiles, 446 covariates, 2,853 gene sets)

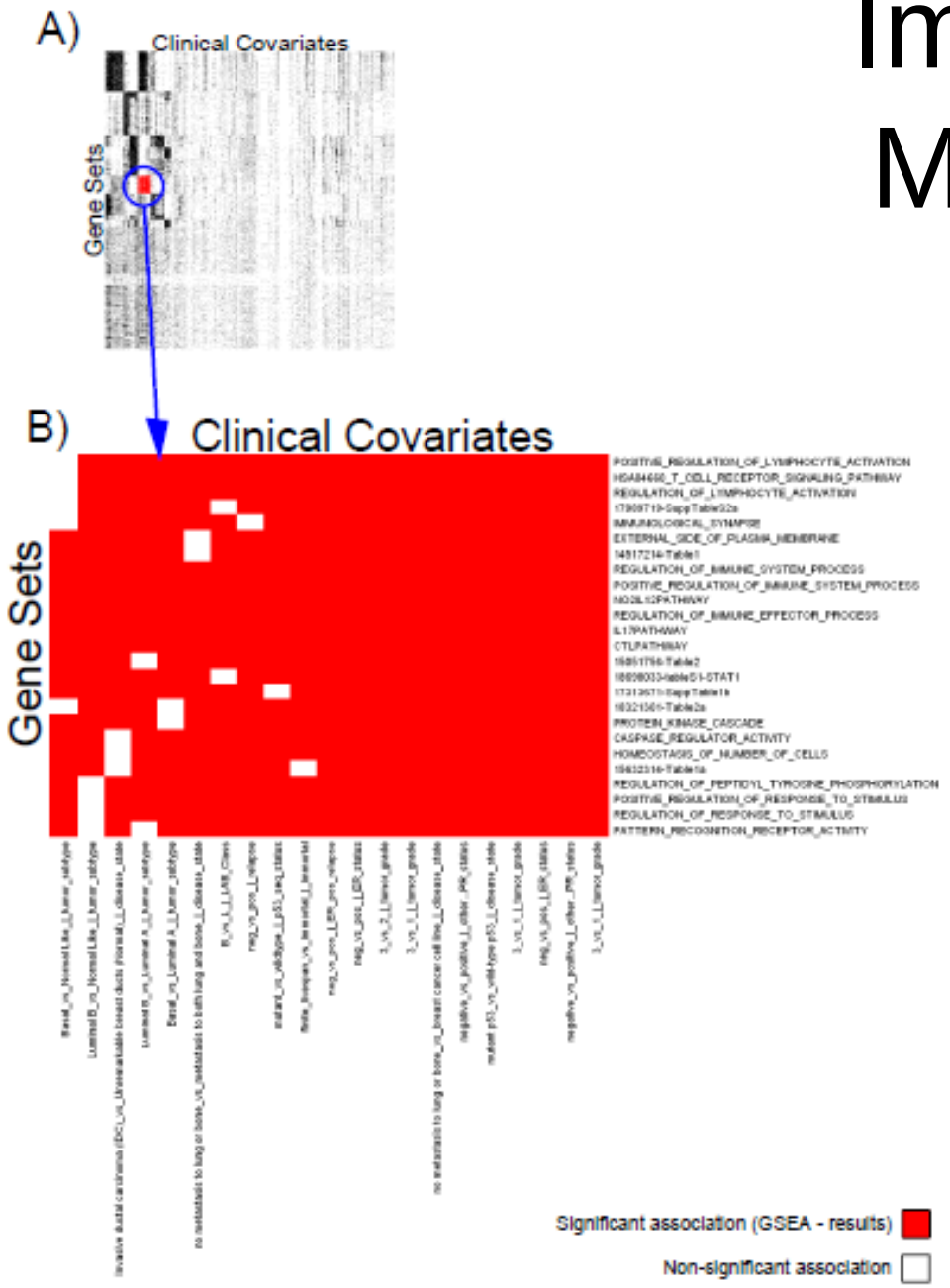


Breast Cancer Modules

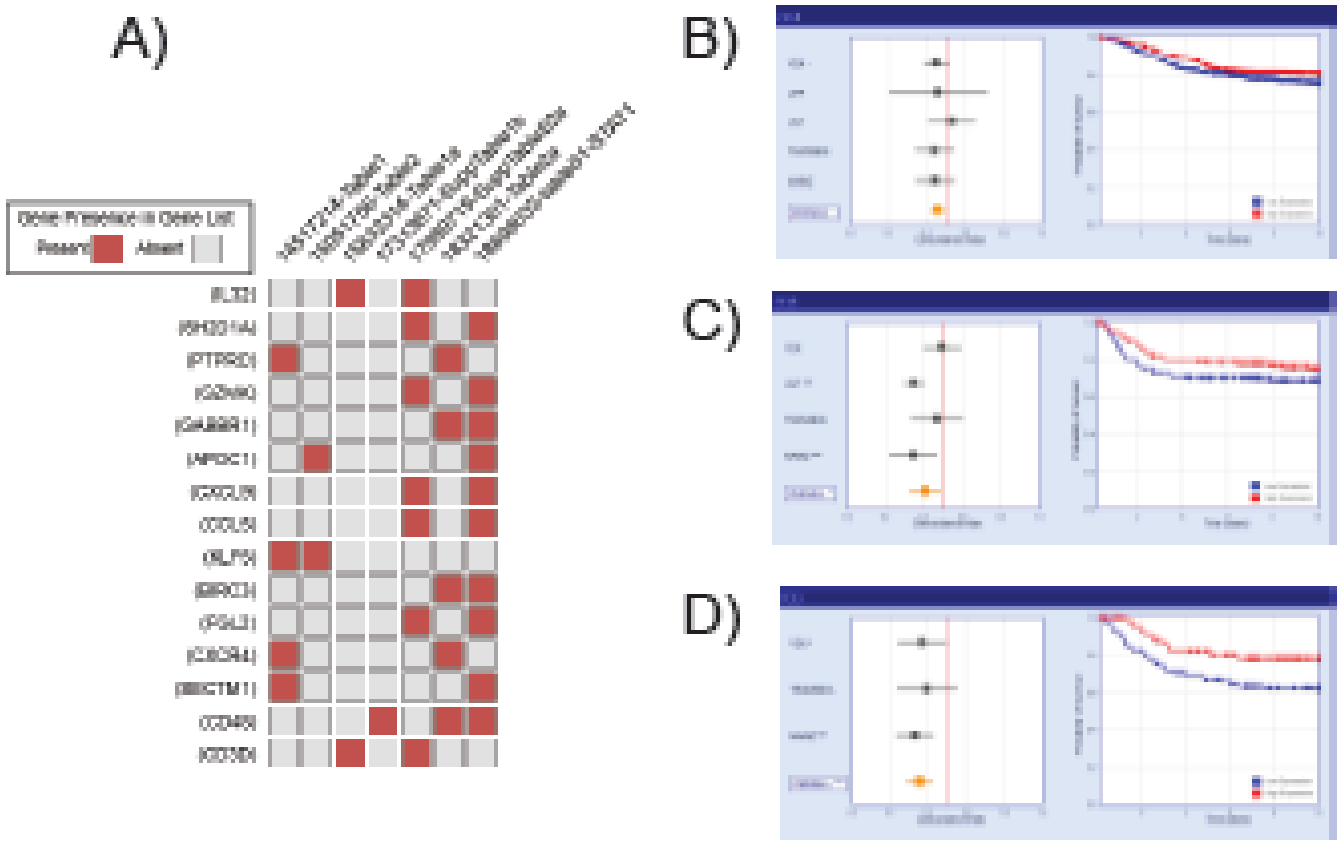
Table 2. Summary of the eight resulting breast cancer modules B1-B8.

	Module Size			
	nCov	nGeneSets	Clinical Covariates	Gene Sets
B1	43	247	High grade, basal / luminal B, mutant p53, ER-, PR-, immortal, relapse, cell line	DNA replication, cell cycle, mitosis, M-phase, spindle, DNA metabolic process and apoptotic mitochondrial changes
B2	9	262	High grade, untreated, resistant, immortal	Wound healing, coagulation, response to light stimulus, cell-cell signaling, excretion, ion channel activity, transmembrane receptor activity and plasma membrane
B3	14	270	Low grade, wild-type p53, normal like breast tissue	Developmental maturation, enzyme linked receptor protein signaling, cell maturation, basolateral plasma membrane, basal lamina, negative regulation of cell differentiation and extracellular matrix
B4	21	75	High grade, basal / luminal B, mutant p53, ER-, PR-, no metastasis	Regulation if immune system process, IL17 pathway, protein kinase cascade, T-cell receptor signaling, lymphocyte activation and chemokine activation
B5	19	89	Cell line, luminal, low stage, metastatic	DNA directed RNA polymerase, endoplasmic reticulum, protein catabolic process, RNA splicing, ubiquitin protein ligase activity, cellular protein catabolic process, secondary metabolic process and citrate cycle
B6	17	74	Tamoxifen treated, luminal, ER+, PR+	Synaptic vesicle, secretion by cell, vesicle mediated transport, intrinsic to Golgi membrane, sphingoid metabolic process, Golgi vesicle transport and Golgi stack
B7	11	115	No relapse, no subtype, high grade	Insulin like growth factor receptor binding, extracellular matrix, myoblast differentiation, actin binding, muscle cell differentiation, focal adhesion, muscle development and cell matrix junction
B8	19	62	High stage, ER-, metastatic, basal, ductal	Cell cycle process, chromosome segregation, mitosis, cell cycle checkpoint, interphase and, condensed chromosome

Immune Module



Genes in Immune Module



Summary: iBBiG of Breast Cancer

- Discovered 8 modules in analysis of gene expression profiles of 21 datasets
  - (446 covariates, 2,853 genesets)
- Strongest signal – proliferation
- Others: Immune, Extracellular matrix
- Each associated with different covariates
- Discovered immune module associated with better outcome in high grade breast cancer

Summary: meta-GSA and iBBiG

- This pilot study shows meta-GSA can uncover common themes or cellular processes across large number of diseases, studies and platforms
- Data integration by building modules of phenotypes which share common features
- Process can easily be apply to other types of data (NGS, proteomics, miRNA etc.)

Acknowledgements

Daniel Gusenleitner

John Quackenbush

Survcomp

- Benjamin Hains-Kaibe,
- Markus Schroder,

GXA meta-GSA

- Misha Kapushesky (EBI)
- Alvis Brazma (EBI)

GeneSigDB

- Shaita Piccard, Kerm Piccard, Enzo Martinelli, Benjamin Hains-Kaibe
- Razvan Sultana, Thomas Schwarzl
- Jerry Papenhausen, Niall O'Connor, Mick Correll

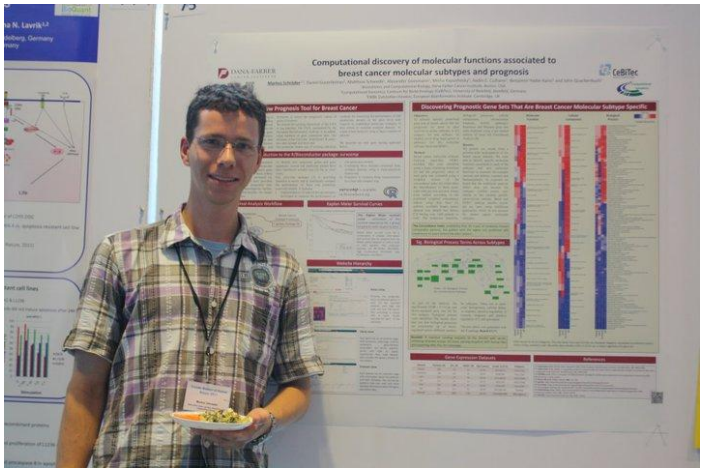
Ovarian Cancer

- Matthew Schwede
- David Harrington, Dimitrios Spentzos, Win Hide, Oliver Hoffman,
- Ronny Drapkin, Hui-Ying Piao

DANA-FARBER CANCER INSTITUTE | WOMEN'S CANCERS PROGRAM

Survival Analysis and Breast Cancer

- Benjamin Hains-Kaibe
- Markus Schroder



Packages: survcomp, genefu, RamiGO and breast cancer datasets

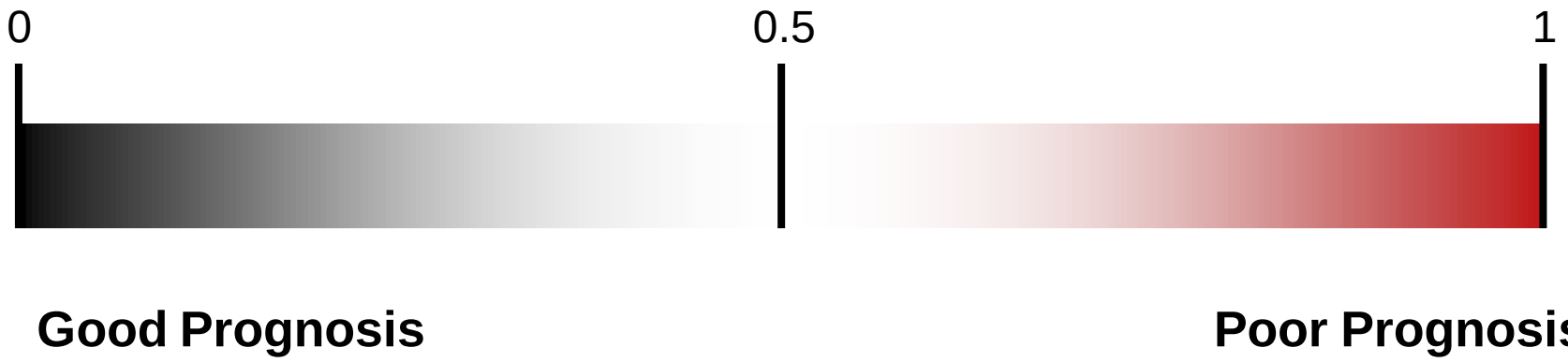
Breast Cancer Data Sets

Dataset	Patients [#]	ER+ [#]	HER2+ [#]	Age [years]	Grade [1/2/3]	Platform
MAINZ	200	155	23	25-90	29/136/35	HGU133A
TRANSBIG	198	123	35	24-60	30/83/83	HGU133A
UPP	251	175	46	28-93	67/128/54	HGU133AB
UNT	137	94	21	24-73	32/51/29	HGU133AB
VDX	344	186	57	26-83	7/42/148	HGU133A
NKI	337	212	53	26-62	79/109/149	Rosetta
Overall	**1467**	**945**	**235**	**24-93**	**244/549/498**	**Affy/Agilent**

Available on Bioconductor.org as experimental data packages: "**breastCancer***"

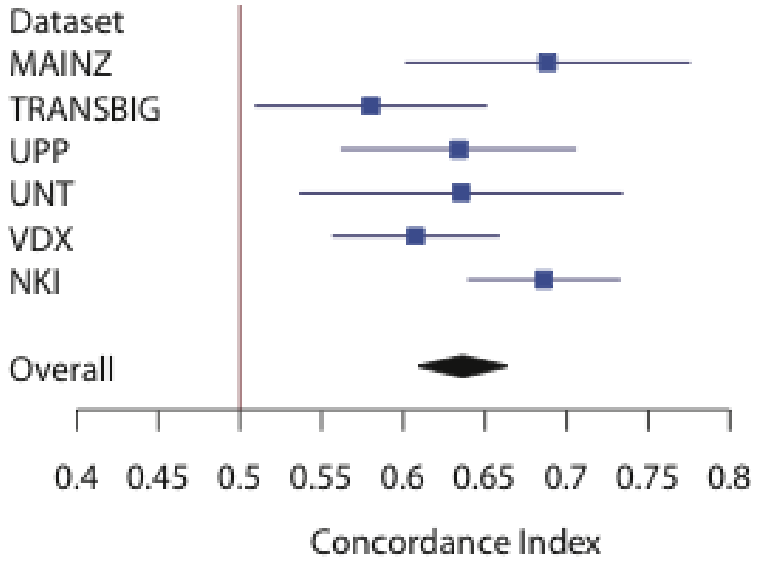
Concordance Index

- A generalization of the AUC to survival data
- Probability that, for a pair of randomly chosen compairable patients, the patient with the higher risk prediction will experience an event before the other patient

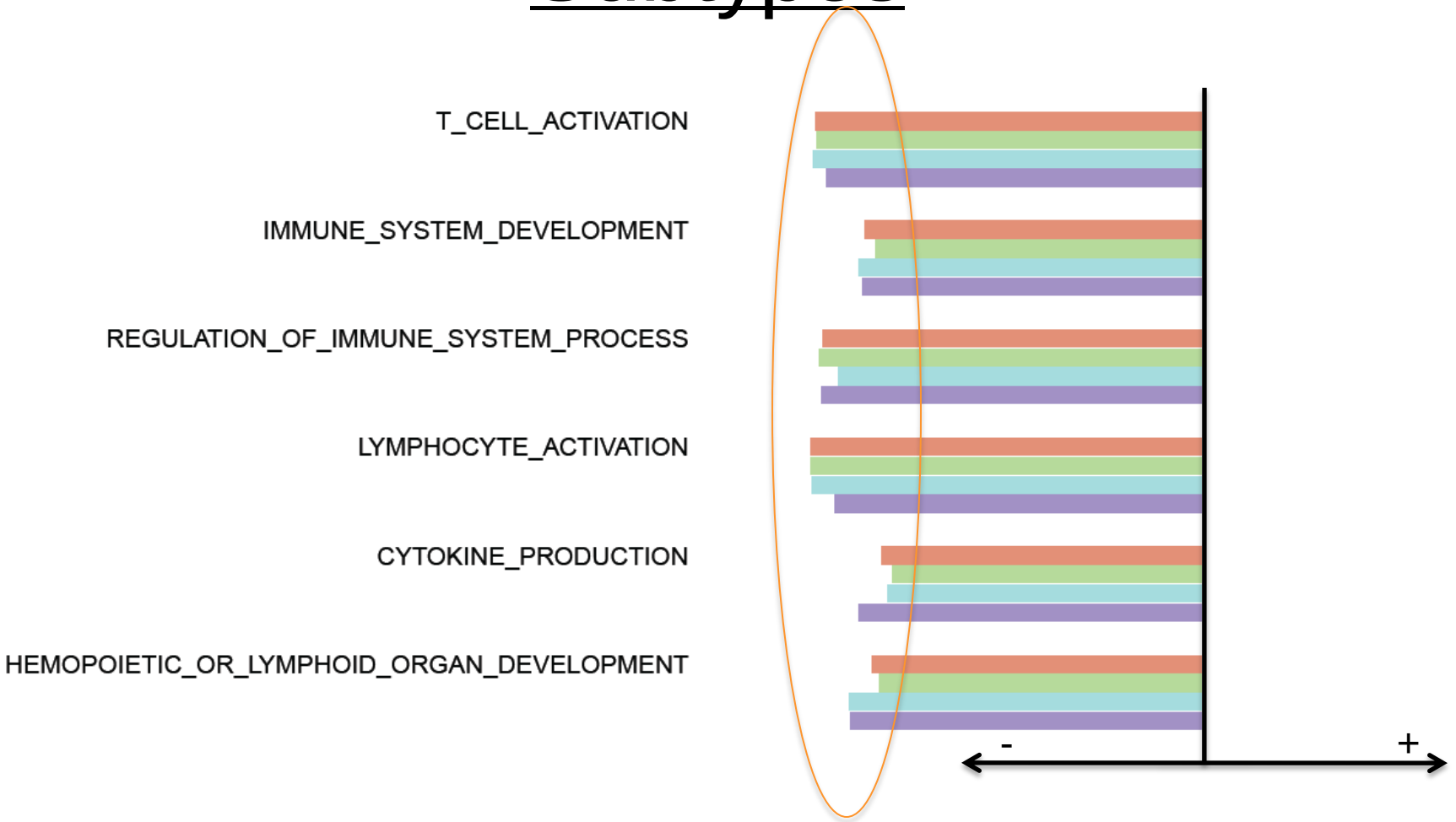


Combine Concordance Indices

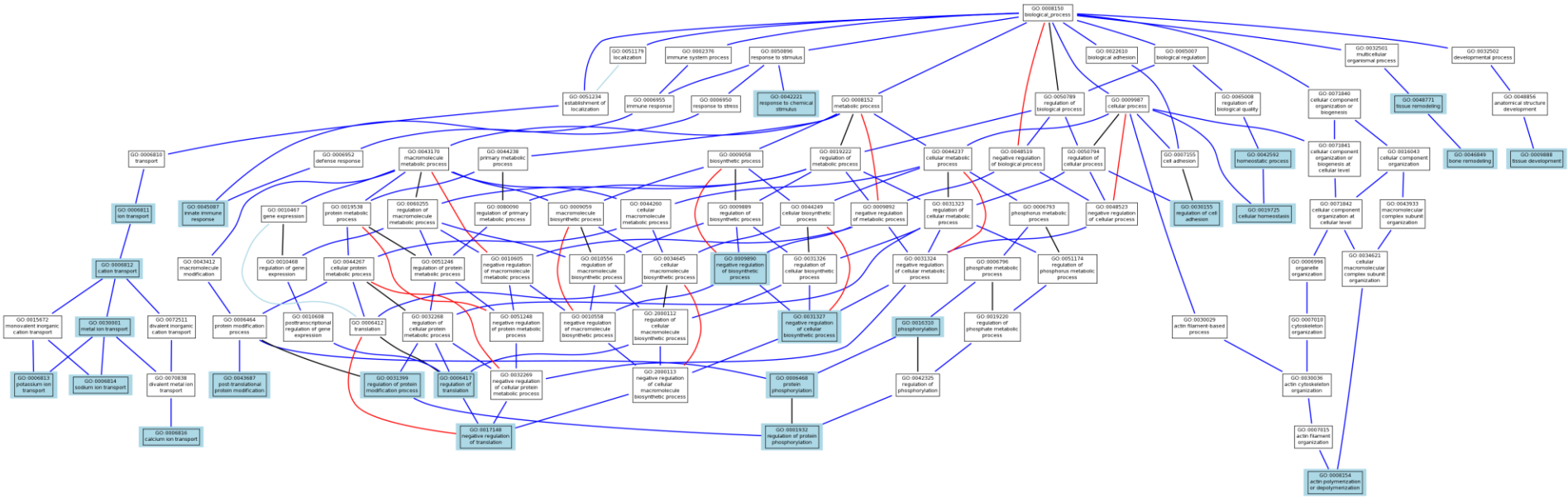
- Combine several estimators using meta-analytical formula to compute a meta-estimate
- Fixed or random effect model
- Use Case:
  - One gene and six datasets



Gene Sets Prognostic Across Subtypes



HER2+ specific BP GO Terms



Blue Nodes: negative NES (Normalized Enrichtment Score)
Red Nodes: positive NES

Package: RamiGO

Subtype Specific Gene Sets

Basal

RESPONSE_TO_LIGHT_STIMULUS

SULFUR_METABOLIC_PROCESS

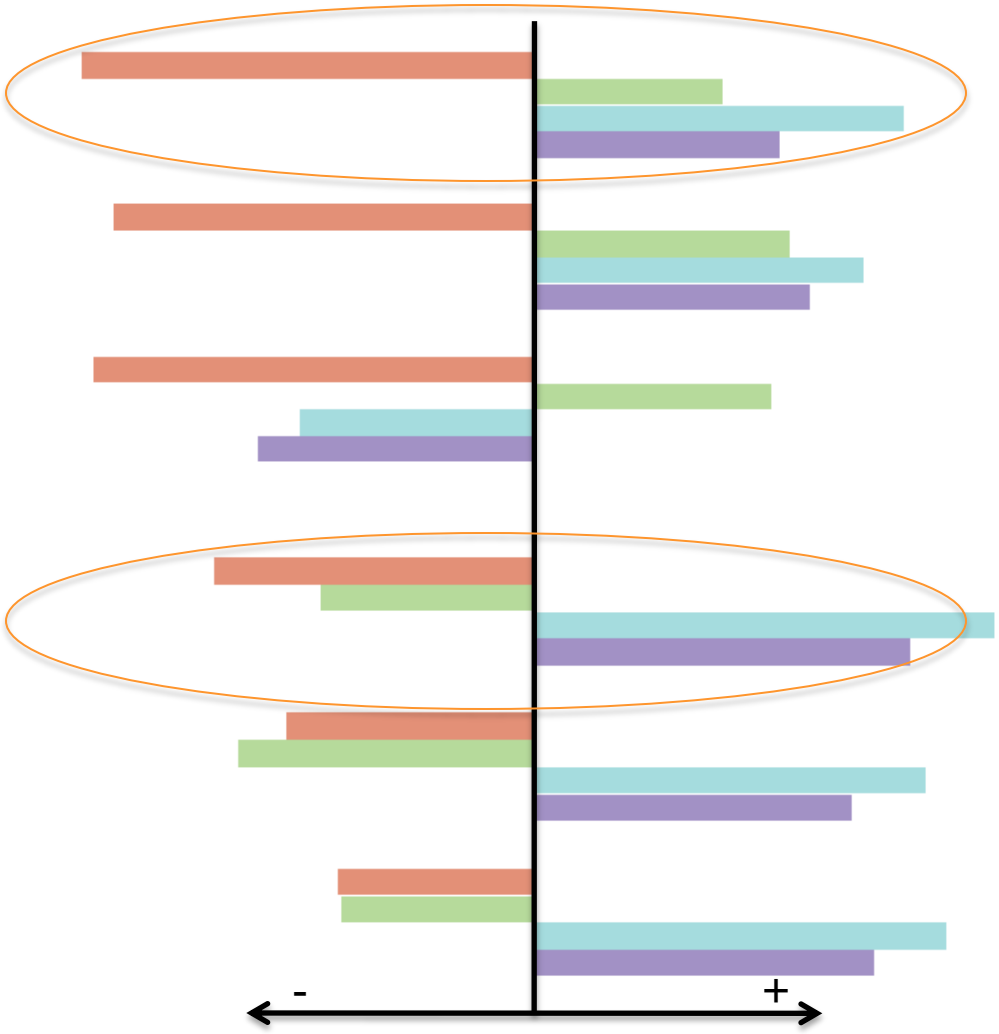
REGULATION_OF_ANGIOGENESIS

Luminal

MEIOTIC_RECOMBINATION

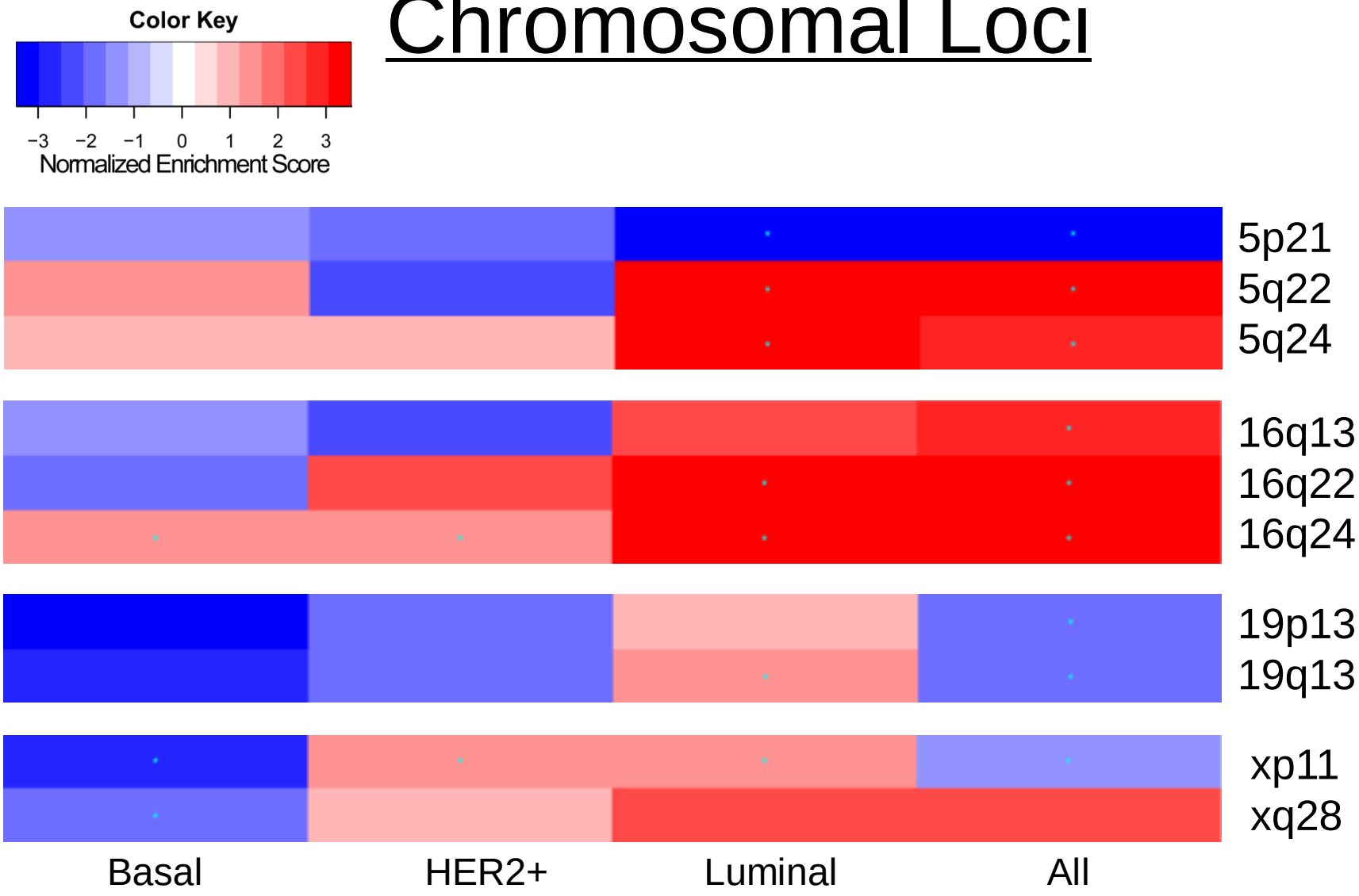
CHROMATIN_MODIFICATION

ORGANELLE_LOCALIZATION

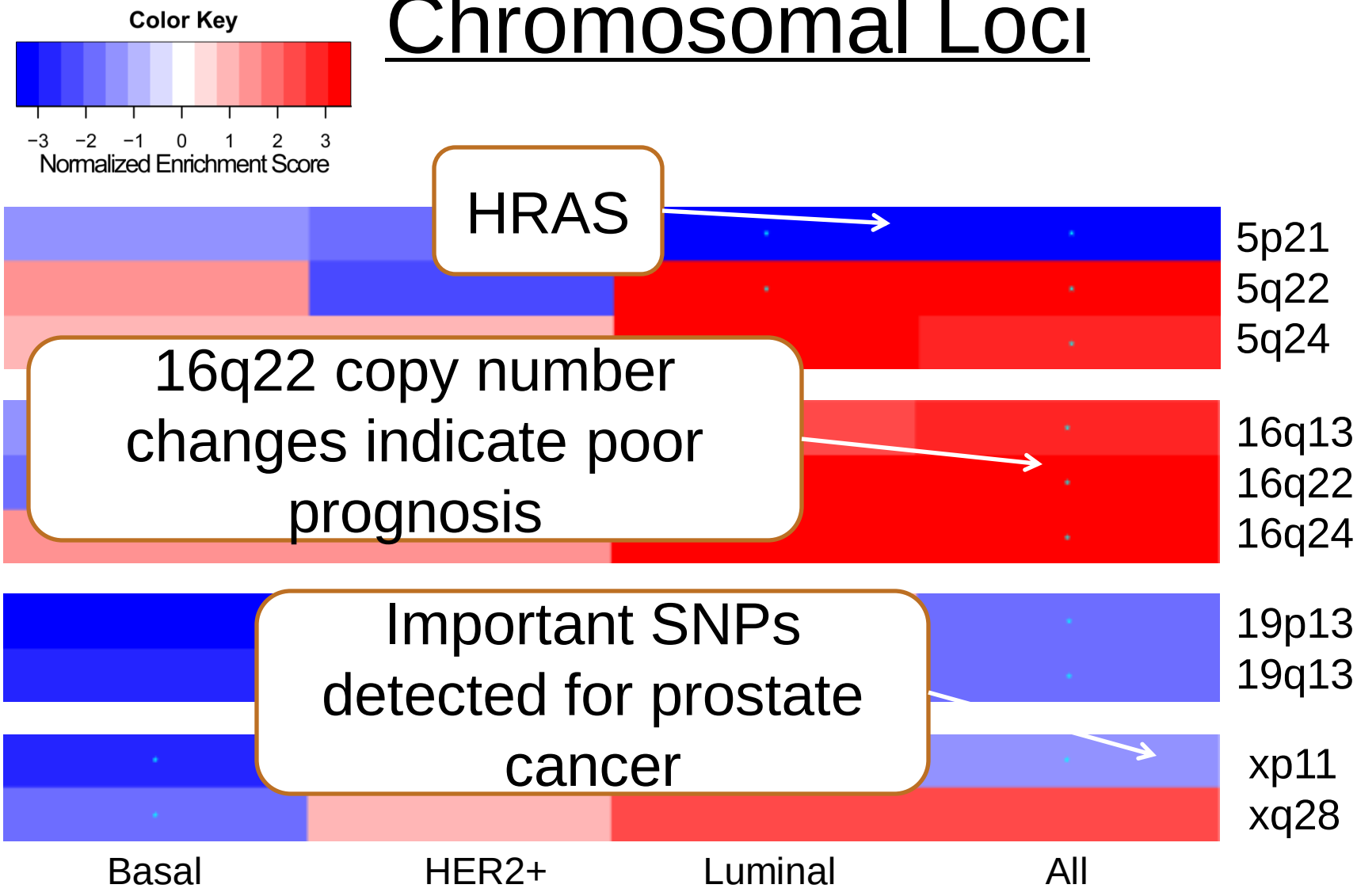


Basal HER2+ Luminal All

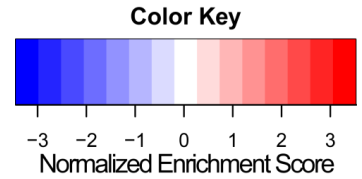
Subtype Specific Prognostic Chromosomal Loci

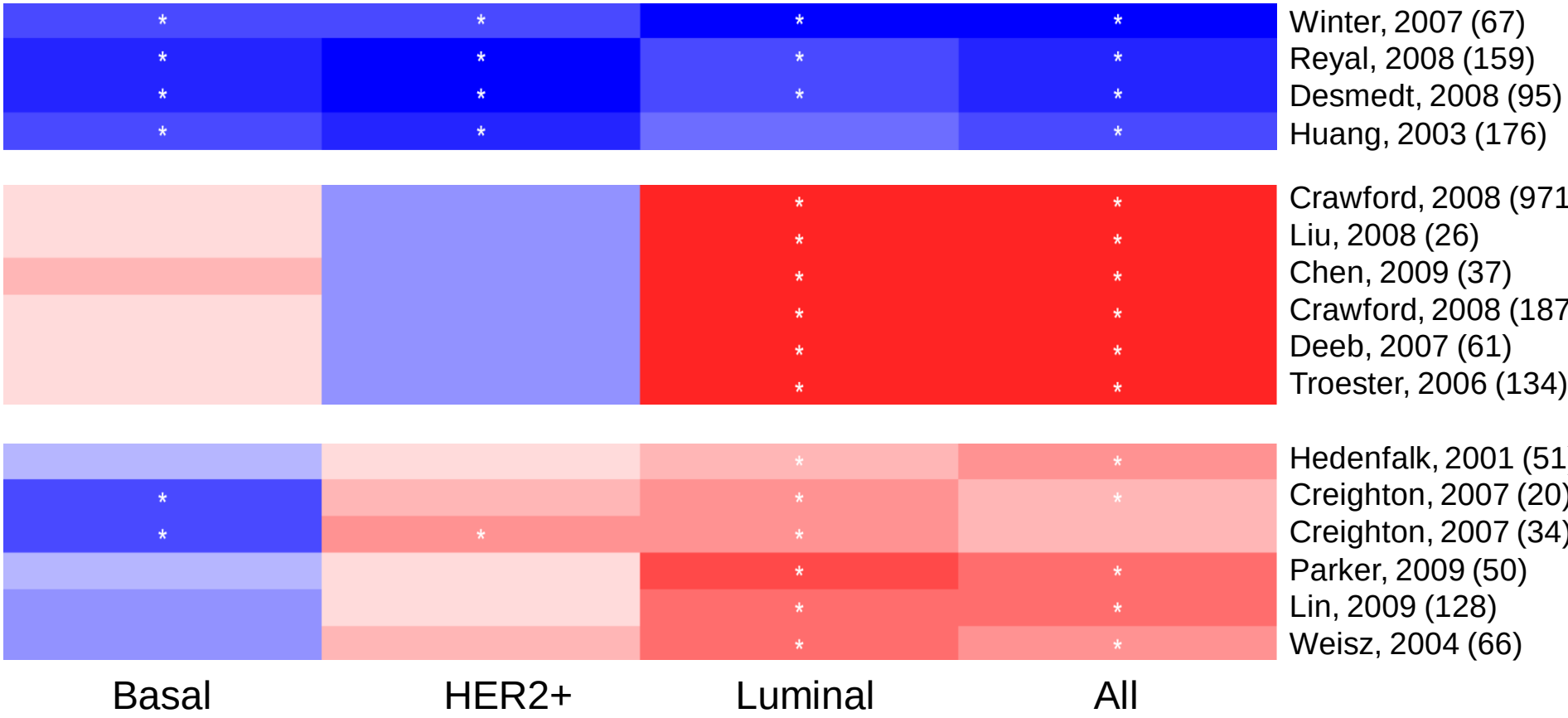


Subtype Specific Prognostic Chromosomal Loci



GeneSigDB Breast Cancer Signatures





Lesson #6: P-values are complicated

Slide From Mathew Schwede