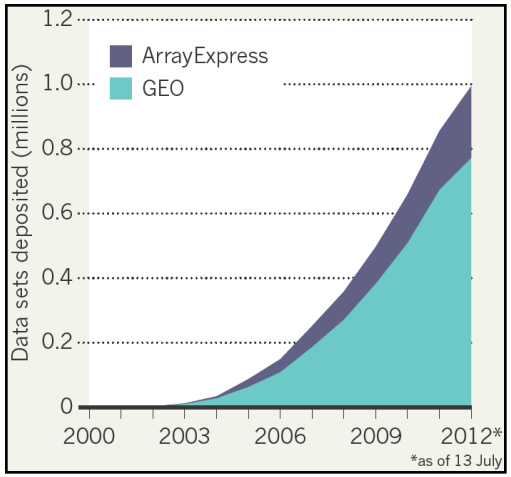
Analysis of microarray data and interpretation of gene signature

Agenda

- How to pre-process and perform statistical analysis of microarray data
- How to interpret gene signature, once it is obtained

Importance of learning about microarray in RNA-Seq era

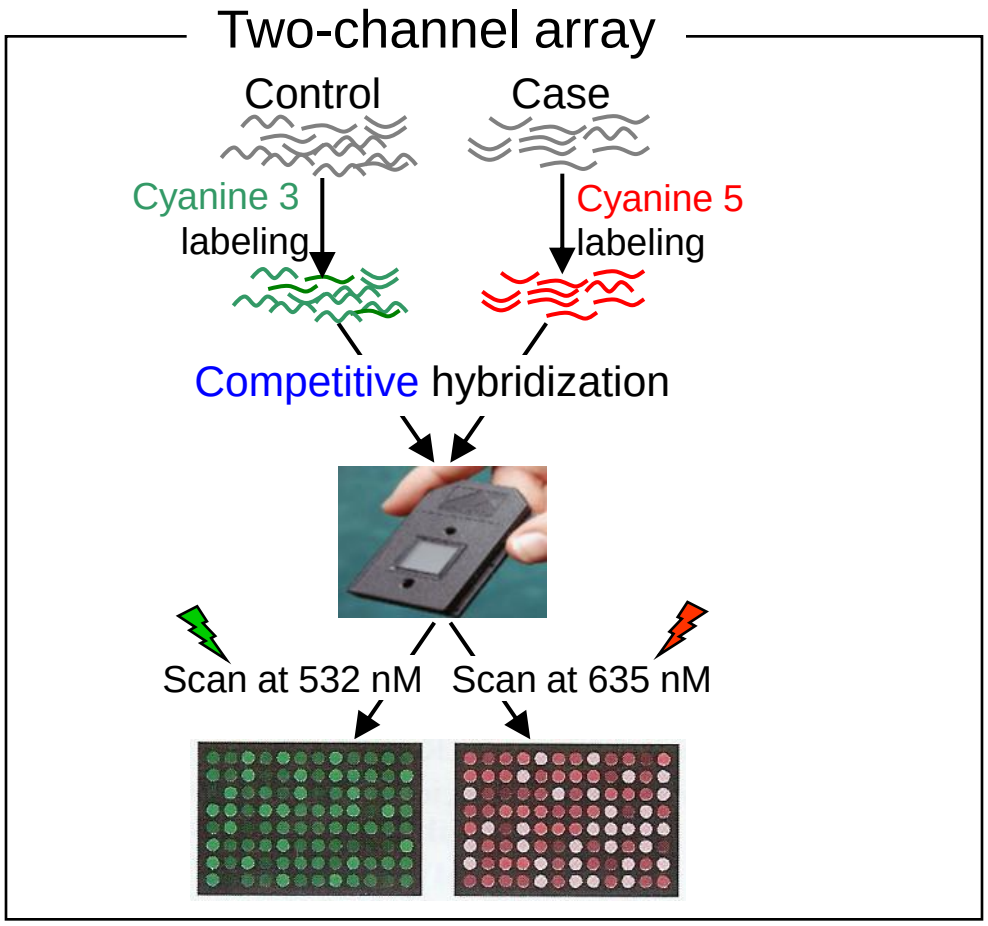
- With respect to dealing with expression values, many aspects of microarray analysis can be applied to RNA-Seq analysis.
- Large volume of publicly available microarray data

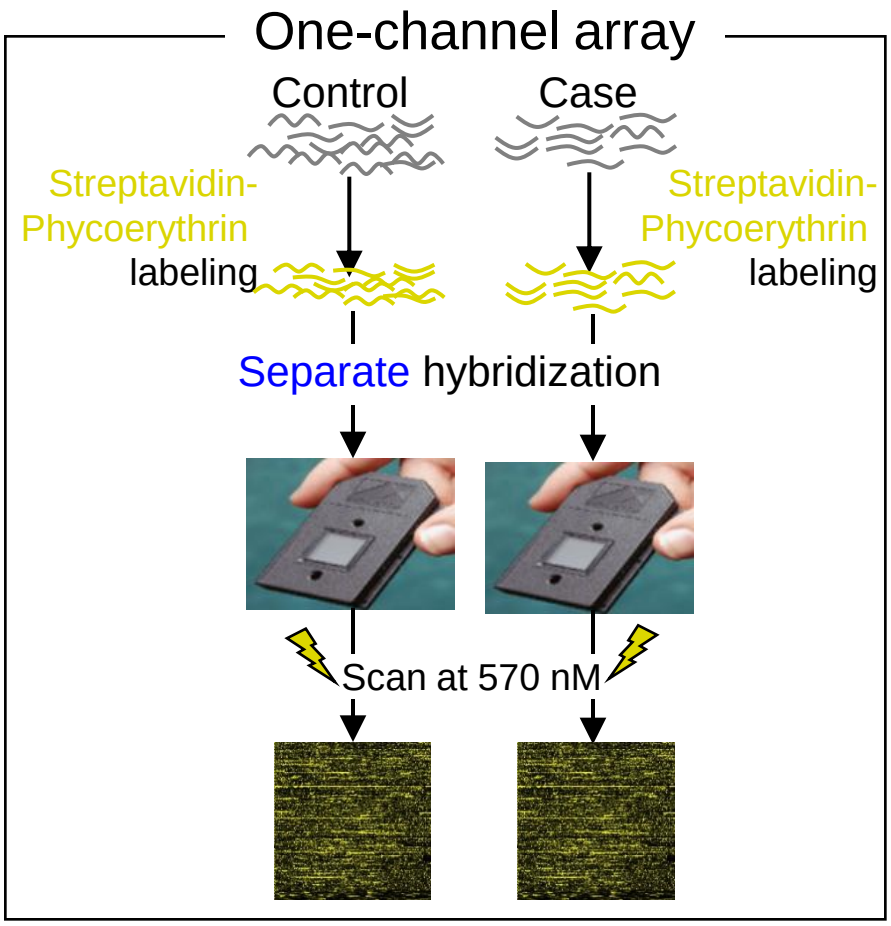


Seungwoo Hwang, swhwang@kribb.re.kr
Korean Bioinformation Center (KOBIC)
Korea Research Institute of Bioscience and Biotechnology (KRIBB)

Two-channel array vs. One-channel array

Channel: The number of fluorescent dyes used in the experiment





cDNA chip: an outdated technology.
Just ignore it.

Affymetrix GeneChip, etc.

Preprocessing of Affymetrix array: Robust Multi-Chip Average (RMA) Method

Scanned chip image

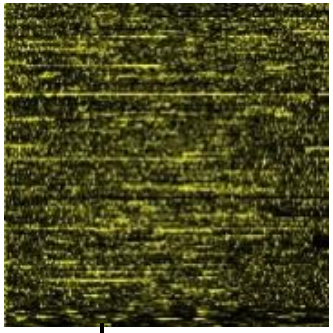


Image processing

- A gene is represented by a ***probe set*** , but a probe set consists of several ***probes***
- Need to pool the intensity values of all the ***probes*** into a single representative value of a ***probe set***

A data calibration process to make the intensity ranges of all the chips in a given dataset comparable

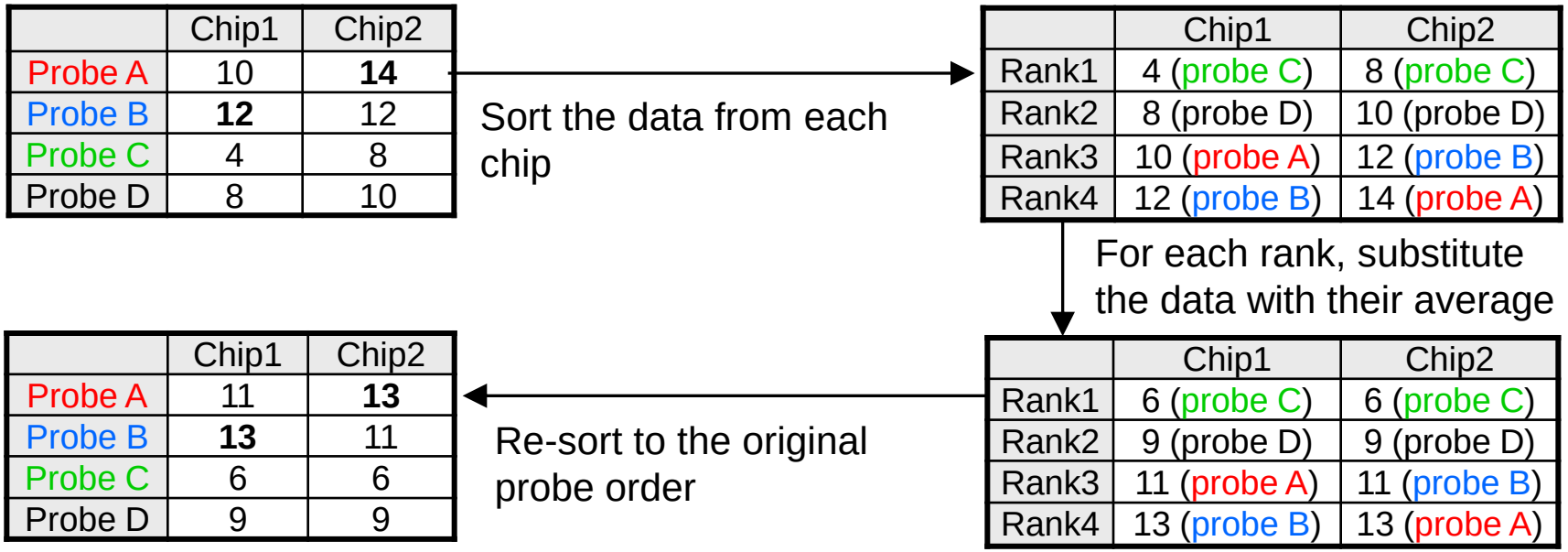
CEL file → Preprocessing = Probe set summarization + Normalization

RMA = Tukey's median polish + Quantile normalization → Gene expression data table

An easy reading: http://www.plexdb.org/modules/documentation/RMAexplained.pdf

Quantile normalization

• What it does: Makes the replicate arrays to have equal distribution.
• Rationale: It can be assumed that **overall distribution** of expression values does not vary much across all samples under most types of studies.
• How it works: Values for each column are ranked, then the average per rank is taken and is reattributed to each column according to the original rank.



Through quantile normalization, all chips in a dataset get the same distribution. That is,
• The highest intensity values in each chip become identical.
• The second highest intensity values in each chip become identical.
• …
• The lowest intensity values in each chip become identical.

Normalize all the samples in a dataset together

Control samples | Case samples

	Rep 1	Rep 2	…	Rep 1	Rep 2	…
Probe 1	5	5.2	…	6.1	6.3	…
Probe 2	4.1	3.9	…	4.1	4.0	…
Probe 3	6.5	6.3	…	4.5	4.3	…
…	…	…	…	…	…	…
Probe N	2.2	2.4	…	2.5	2.2	…

All group normalization: Normalize all these samples together?

Or split into two sets, and,

	Rep 1	Rep 2	…
Probe 1	5	5.2	…
Probe 2	4.1	3.9	…
Probe 3	6.5	6.3	…
…	…	…	…
Probe N	2.2	2.4	…

	Rep 1	Rep 2	…
Probe 1	6.1	6.3	…
Probe 2	4.1	4.0	…
Probe 3	4.5	4.3	…
…	…	…	…
Probe N	2.5	2.2	…

Separate group normalization: Normalize the two sets of samples separately?

- **All group normalization:** presumes that the sample groups are similar enough.
- **Separate group normalization:** presumes that the sample groups are very different.
- In most situations (e.g., tumor vs. normal // treated vs. untreated), all group normalization is to be used.
- In some rare situations (e.g., brain vs. liver), do something else.

Gene-level collapsing of duplicate probe sets' intensities

- Some probe sets correspond to identical gene. For example, in Affymetrix U133A, there are ~21,000 probe sets that have Entrez Gene annotation, but they are mapped to only ~13,000 Entrez Gene IDs. Thus ~8,000 probe sets are duplicates.
- Many of the signature interpretation methods require the genes in a signature to be unique.

Probeset-level data table

Probeset ID	Entrez Gene	Expression values			
		Sample1	Sample2	Sample3	....
217757_at	2				
207268_x_at	10152				
209856_x_at	10152				
211793_s_at	10152				
216113_at	10152				
201746_at	7157				
211300_s_at	7157				
....	....				

Collapsing
Condensing
Aggregation

Gene-level data table

Entrez Gene	Expression values			
	Sample1	Sample2	Sample3	....
2				
10152				
7157				
....				

Collapsing methods: Miller (2011) BMC Bioinformatics PMID:21816037

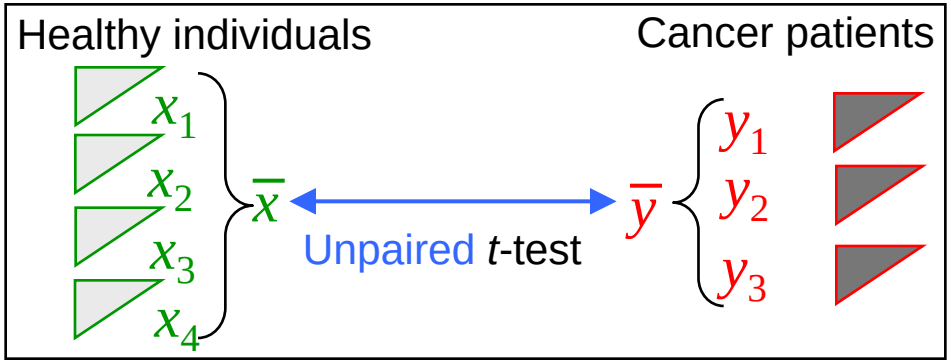
- Simple averaging (most popular): Mean or Median
- Choosing a representative: Probe set with highest overall intensity, or largest variance

Also it is the most important to use the most recent probe set annotation

Unpaired *t*-test

Test whether, on average, gene expression levels are different between control group and case group

x_i: log expression level from *i*th control individual



y_i: log expression level from *i*th case individual

Take the means first and then take their difference

General form: $t = \dfrac{(\bar{x} - \bar{y})}{SE_{\bar{X}-\bar{Y}}}$ where $SE_{\bar{X}-\bar{Y}}$ = standard error of the difference between means

Student's *t*:

(assuming equal variance between X and Y)

$$SE_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{m}}$$ where

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$$ Pooled sample variance

$$df = n + m - 2$$

Welch's *t*:

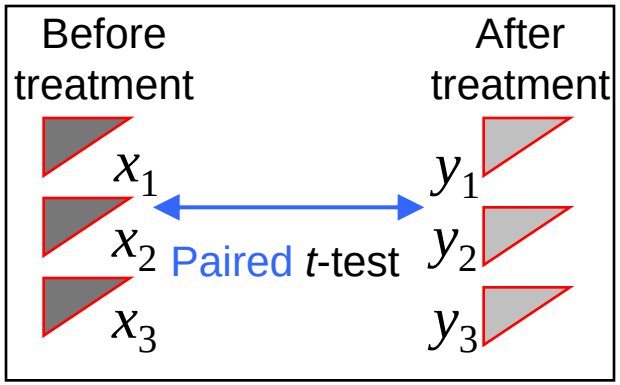
(not assuming equal variance)

$$SE_{\bar{X}-\bar{Y}} = \sqrt{\frac{s_x^2 + s_y^2}{n+m}}$$

$$df = \left(\frac{s_x^2 + s_y^2}{n+m}\right)^2 \Bigg/ \left(\frac{s_x^4}{n^2(n-1)} + \frac{s_y^4}{m^2(m-1)}\right)$$

Paired *t*-test

Test whether gene expression levels are different, for example, before and after treatment for each and every individual.



In paired *t*-test,

(1) Take difference first for each pair: $d_i = x_i - y_i$

(2) Then take the mean of differences: $\bar{d}$

t-statistic is calculated as $t = \dfrac{\bar{d}}{SE_{\bar{D}}}$ where $SE_{\bar{D}} = \dfrac{s_d}{\sqrt{n}}$

df = n-1

The other way around

In unpaired *t*-test,

- you first take means of the two groups ($\bar{x}$ and $\bar{y}$),
- then take the difference between the means ($\bar{x} - \bar{y}$)

Multiple testing correction

1. What we do

- Suppose we have 10K array and obtained P-values through statistical testing

Gene	P-value
gene 1	0.00004
gene 2	0.0001
…	…
gene 2000	0.04
gene 2001	0.06
…	…
gene 10000	0.99

Test result seems to suggest that these 2000 genes are differentially expressed

P-value cutoff = 0.05

- However, we need to use smaller P-value cutoff, or equivalently, adjust the raw P-values by increasing them in order to avoid many false positives

2. Why we do (an analogy)

- Large-scale psychokinesis experiment with 10,000 people from the street

- Task: Flip a coin 10 times. If head turns up more than nine times, you are a psychic. (since Pr(H>=9)=0.0107)
- Result: Astonishingly, 100 people accomplished the task! An earth-shattering discovery?
- Problem statement:
  - Even though none of the 10,000 subjects have any supernatural power, there could be 107 (0.0107x10,000) subjects who accomplish the task, by chance alone.
  - Probability that none of them accomplishes the task is $2x10^{-47}$ $(=(1-0.0107)^{10,000})$
- Solution: Use more stringency (psychic only if head turns up all ten times), or multiple testing correction.

3. How we do

- False-discovery rate (FDR) correction: FDR of 0.05 means 5% of false positives are expected among those identified as positives. Benjamini-Hochberg's procedure is the norm in bioinformatics.
- Family-wise error rate (FWER) correction: FWER of 0.05 means 0.05 genes are expected to be false positive. Bonferroni's procedure is well known but too strict for most bioinformatics tasks.

An apparent paradox of multiple testing correction

Microarray experiment

- Suppose raw P-value of 0.01 was obtained for TP53 gene from microarray.
- The TP53 gene was not called significant since multiple testing corrected P-value was not small enough, even though raw P-value was small.

RT-PCR on a single gene

- Suppose that, upon measuring only TP53 with RT-PCR, raw P-value of 0.01 was obtained.
- This time, TP53 was called significant since raw P-value was small.

Apparent paradox

Why TP53 was penalized only in the large-scale experiment even though they showed the same result in both small- and large-scale experiments?

Large-scale experiment is a screening experiment

- Here the scientist did not have any hypothesis.
- So the scientist should recognize the possibility of the presence of many false findings in his result, which can occur by chance alone, and do the best to prevent these false leads from getting into scientific community, through multiple testing correction.

Small-scale experiment is a confirmatory experiment

- Here the scientist had a clear hypothesis. So he already narrowed down many possibilities before the experiment, and performed the experiment only on that possibility to confirm his hypothesis.
- So the scientist is able to be confident on his finding ("*TP53 is differentially expressed, just as I expected*!")

R for microarray data analysis

EMA - A R package for Easy Microarray data analysis

A combined R package of many individual R packages. Easy to use.

Nicolas Servant[1,2,3*†], Eleonore Gravier[1,2,3,4†], Pierre Gestraud[1,2,3], Cecile Laurent[1,2,3,6,7,8], Caroline Paccard[1,2,3], Anne Biton[1,2,3,5], Isabel Brito[1,2,3], Jonas Mandel[1,2,3], Bernard Asselain[1,2,3], Emmanuel Barillot[1,2,3], Philippe Hupé[1,2,3,5]

Run R code at the command line

Hello world in R

(1) Make a code named **`hello.R`**

emacs: hello.R

```
# hello.R prints a text string on the terminal

args <- commandArgs(trailingOnly=T)
inputString <- args[1]
cat("Hello", inputString, "\n")
```

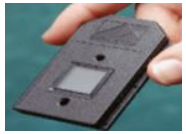
(2) Run the code at UNIX command line as follows

UNIX console, UNIX prompt

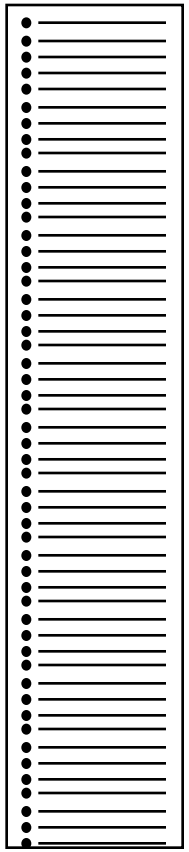
```
UNIX prompt> R --slave < hello.R --args world
Hello world
UNIX prompt>
```

Avoid using R interactive mode. Make a code and run it at UNIX command line.

How to interpret microarray data and gene signature



Statistical test



Gene list (also called gene signature)

What can we do once we obtain a gene list?

- ❑ Try to publish a paper with a huge gene list table?
- ❑ Try to learn about all genes on the list by reading thousands of papers?
- ❑ The paradox: The more facts we learn, the less we understand the process we study [Lazebnik (2002) Cancer Cell; PMID:12242150]
- ❑ An indication of information overload → poses an interpretation challenge

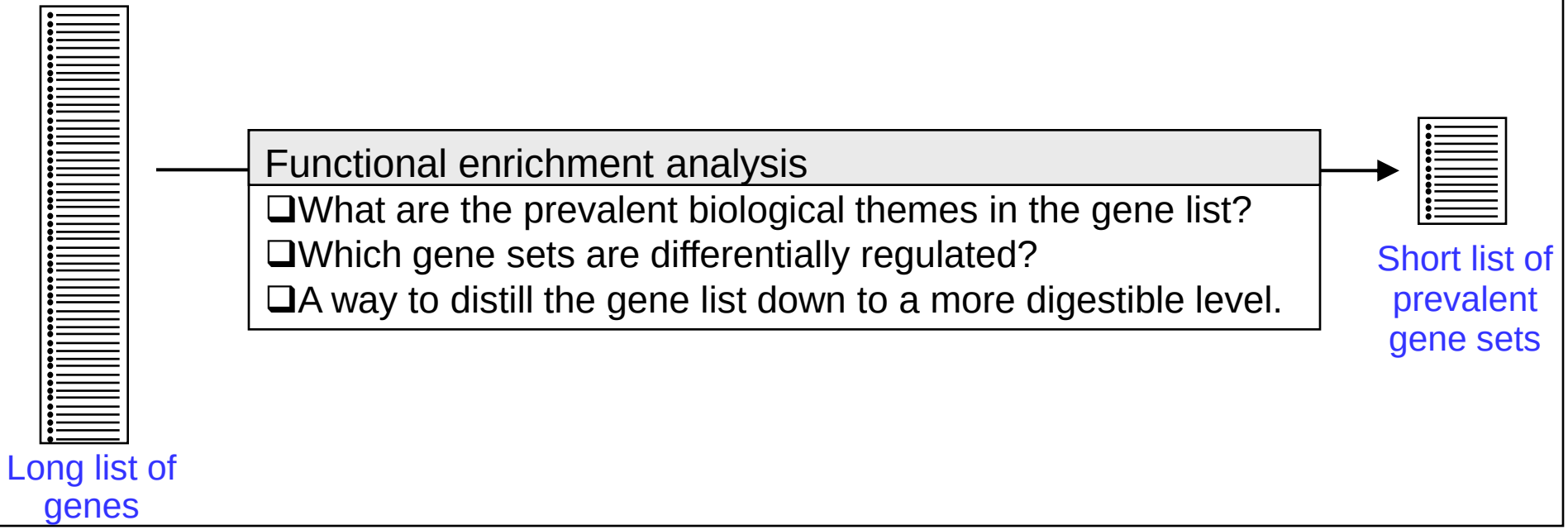
Approaches for interpretation of microarray data and gene list

- ❑ Functional enrichment analysis (GO analysis)
- ❑ Removing redundancy from GO analysis result
- ❑ Identifying differentially expressed PPI subnetworks
- ❑ Signature comparison using signature databases
- ❑ Comparison between two signatures
- ❑ Deriving a consensus signature

- Most of them are not tied to a particular technology
- Can be applied to gene signatures from not only microarray, but also RNA-Seq, and other high-throughput technologies

Functional enrichment analysis

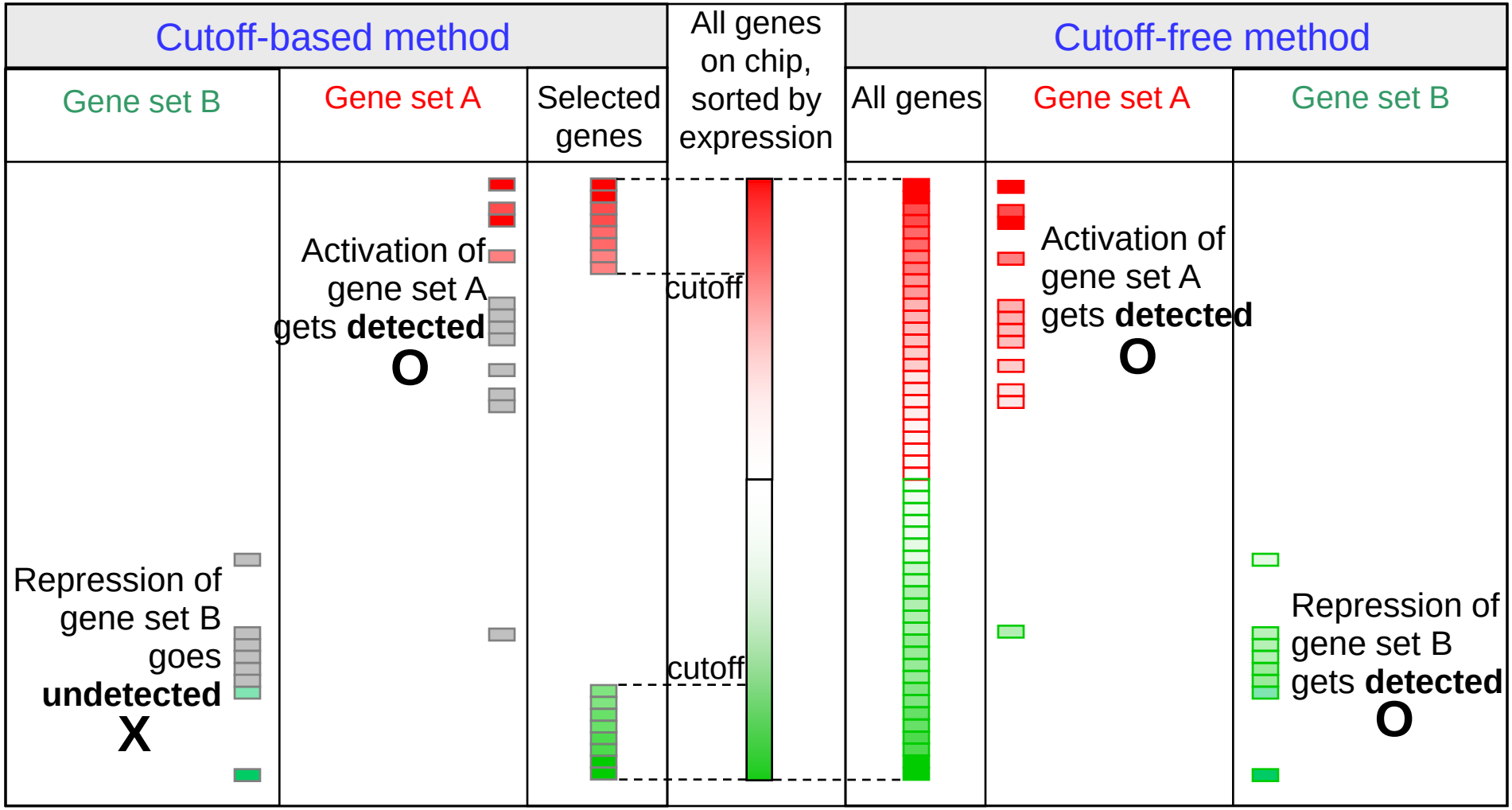
- Also called gene set analysis and GO analysis
- It is a way to see the forest (gene sets—GO terms, KEGG pathways) through the trees (individual genes)



- Two types of functional enrichment analysis (and other interpretation approaches as well)

	Cutoff-based methods	Cutoff-free method
Input	Unordered list of selected genes above a selection cutoff (e.g., p-value<0.05, fold change>2-fold)	Ordered list of all genes, along with their statistics (e.g., p-value, fold change)
Advantage	Simpler and intuitive	Can detect subtle signals

Cutoff-based vs. Cutoff-free functional enrichment analysis



Softwares

- ❑ Cutoff-based enrichment analysis: DAVID [Huang (2008) Nat Protoc; PMID:19131956]
- ❑ Cutoff-free enrichment analysis: GSEA [Subramanian (2007) Bioinformatics; PMID:17644558]
  GeneTrail [Backes (2007) Nucleic Acid Res; PMID:17526521]

Cutoff-based enrichment analysis (Fisher's exact test)

Enrichment of TCA cycle genes in the selected DEGs

Array of 20,000 genes

30 TCA cycle genes

19,970 other genes

Genome-wide, only **0.15%** are TCA cycle genes

Selection of 200 DEGs

5 TCA cycle genes

195 other genes

In the DEGs, **1.67%** are TCA cycle genes

Summarize as contingency table

	DEG	Non-DEG	TOTAL
TCA	5	25	30
Other	195	19,775	19,970
TOTAL	200	19,800	20,000

Calculate the probability of the observed event by hypergeometric distribution

$$\Pr(X=5)=\frac{\binom{30}{5}\binom{19970}{195}}{\binom{20000}{200}}$$

R> dhyper(5, 30, 19970, 200)
→ 1.06 E-06

Calculate *p*-value by enumerating more extreme events

Observed event

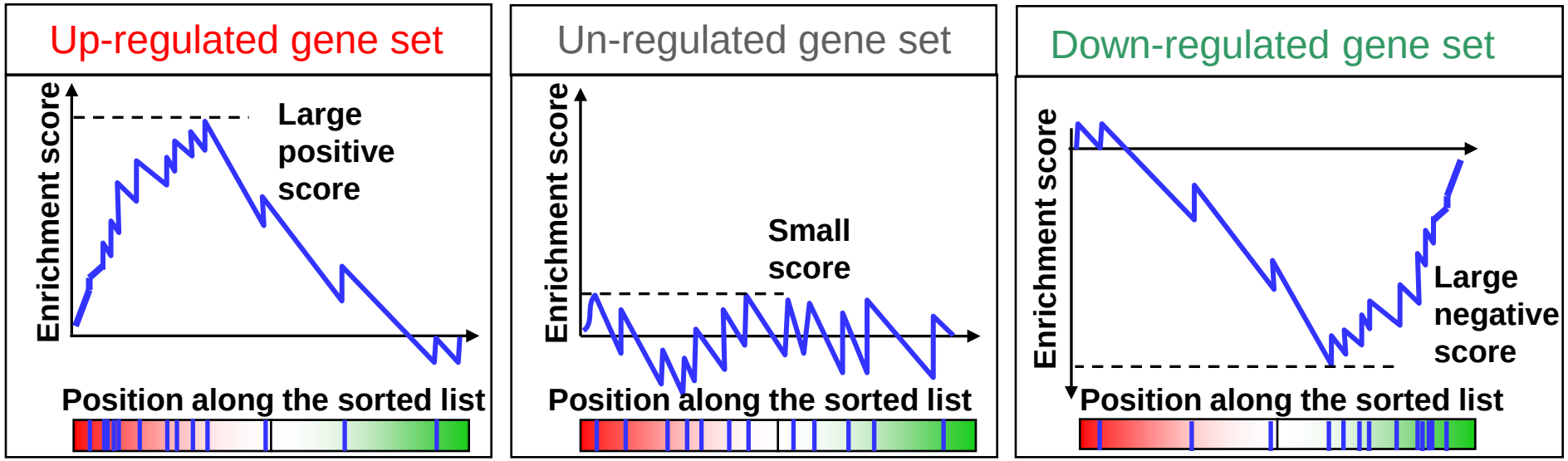
More extreme events

$$p\text{-value}=\Pr(X=5)+\Pr(X=6)+\cdots+\Pr(X=30)$$

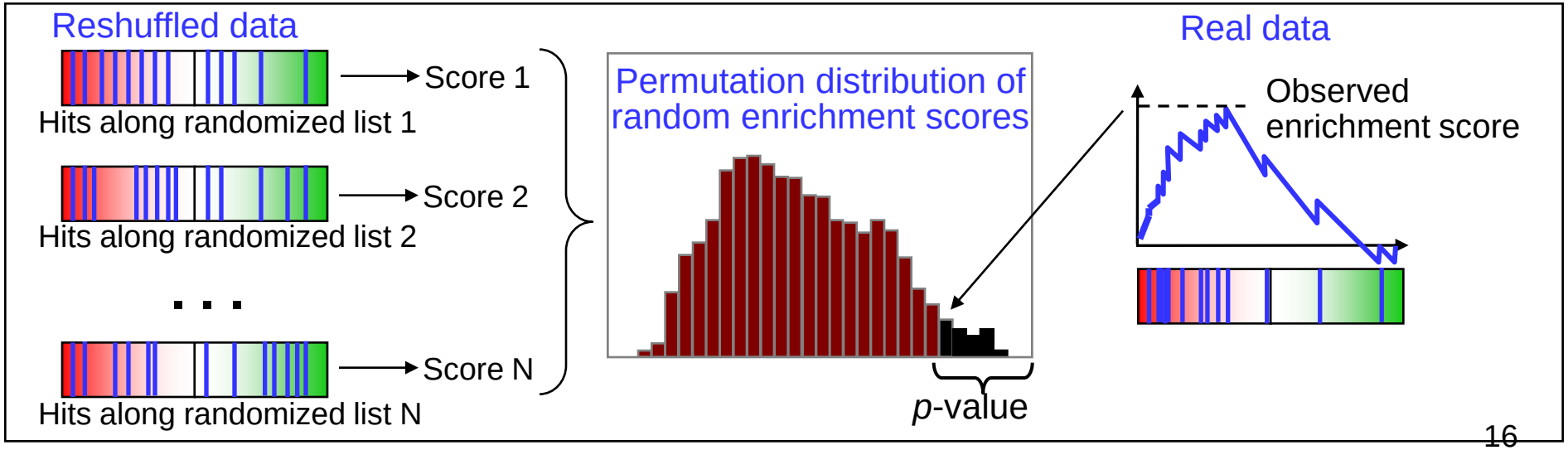
R> sum(dhyper(5:30, 30, 19970, 200))
→ 1.11 E-05

Cutoff-free functional enrichment analysis (GSEA)

Step 1: Calculation of enrichment score of a gene set



Step 2: Calculation of *p*-value by permutation test



GSEA can detect subtle but coordinate expression changes

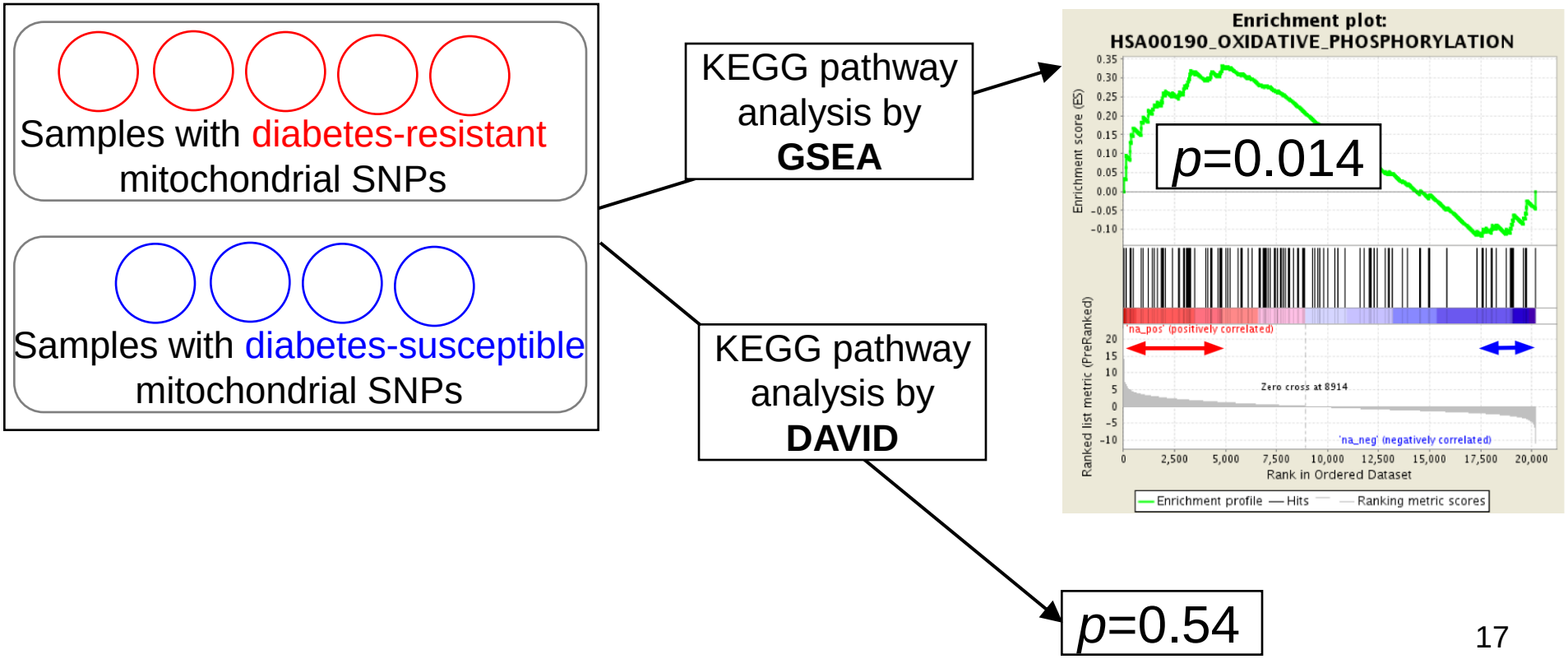
OPEN ACCESS Freely available online

PLoS one

Gene Expression Pattern in Transmitochondrial Cytoplasmic Hybrid Cells Harboring Type 2 Diabetes-Associated Mitochondrial DNA Haplogroups

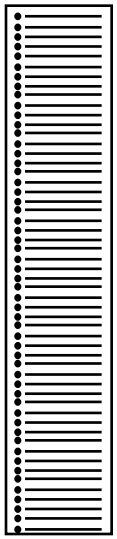
Seungwoo Hwang[1☯], Soo Heon Kwak[2☯], Jong Bhak[3], Hae Sun Kang[2], You Ri Lee[2], Bo Kyung Koo[2], Kyong Soo Park[2], Hong Kyu Lee[4], Young Min Cho[2*]

- Other than the mitochondrial SNPs, the two groups were identical
- Therefore, expression profile differences between the two groups were very subtle



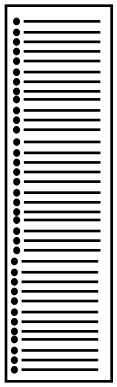
Removing redundancy from GO analysis result (GO-Module)

Long list of genes



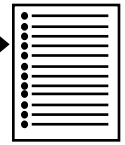
Functional enrichment analysis (DAVID, GSEA, etc)

Long list of significant GO terms (due to hierarchical nature of GO)



Remove redundant GO terms (GO-Module)

Compact list of significant GO terms



BIOINFORMATICS APPLICATIONS NOTE

Vol. 27 no. 10 2011, pages 1444–1446
doi:10.1093/bioinformatics/btr142

Databases and ontologies

Advance Access publication March 17, 2011

GO-Module: functional synthesis and improved interpretation of Gene Ontology patterns

Xinan Yang[1], Jianrong Li[1], Younghee Lee[1] and Yves A. Lussier[1,2,*]

Removing redundancy from GO analysis result (GO-Module)

1. Find Key Terms

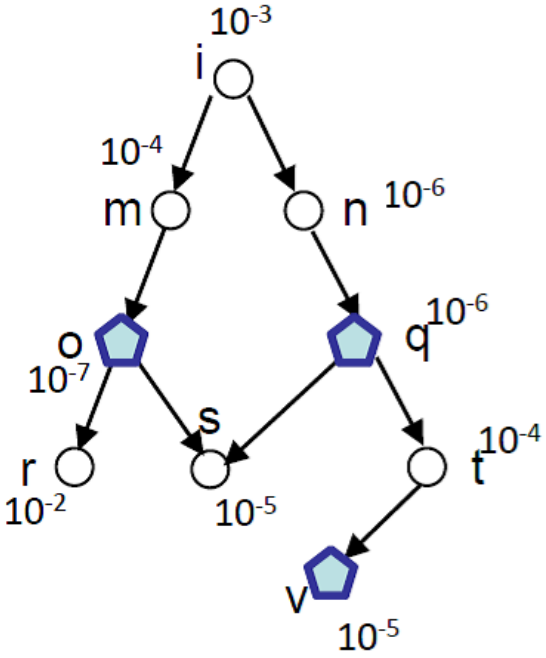
More significant than all other neighboring terms in the GO tree

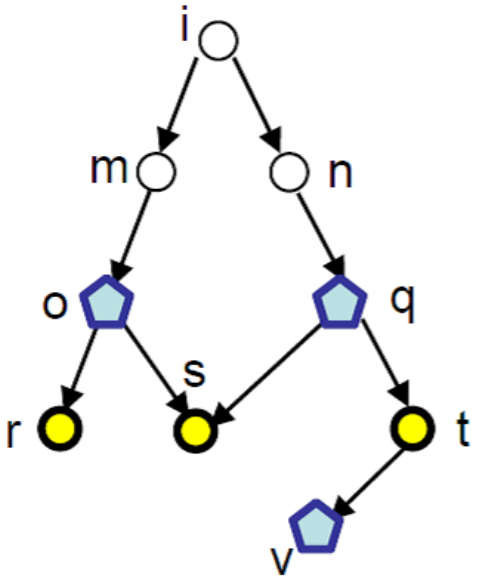
2. Find True Positive Terms

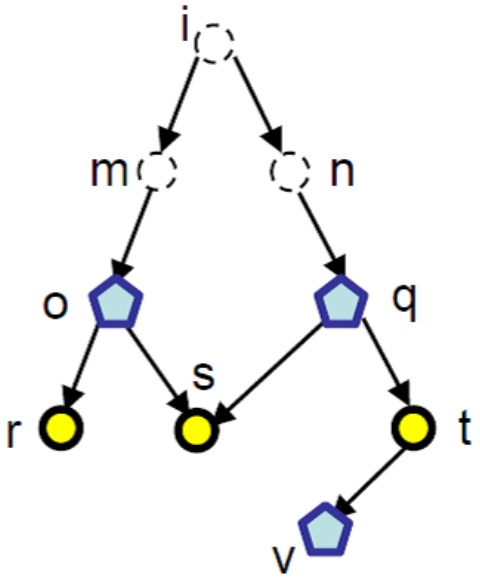
All child terms of the key terms

3. Find False Positive Terms

All the rest







4. Report only key terms or both key terms and true positive terms +

Removing redundancy from GO analysis result (GO-Module)

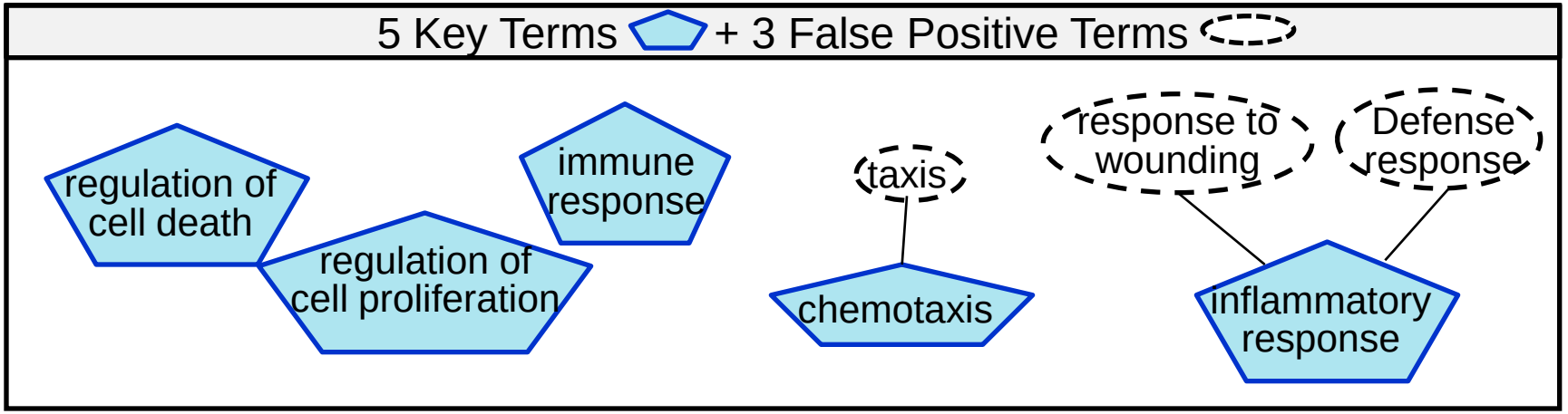
An example run

27 genes that are differentially expressed under free fatty acid treatment

DAVID

8 significant GO terms	
immune response	4.38E-06
inflammatory response	3.51E-04
defense response	5.00E-04
chemotaxis	6.01E-04
taxis	6.01E-04
regulation of cell proliferation	0.00115
regulation of cell death	0.00140
response to wounding	0.00177

GO-Module

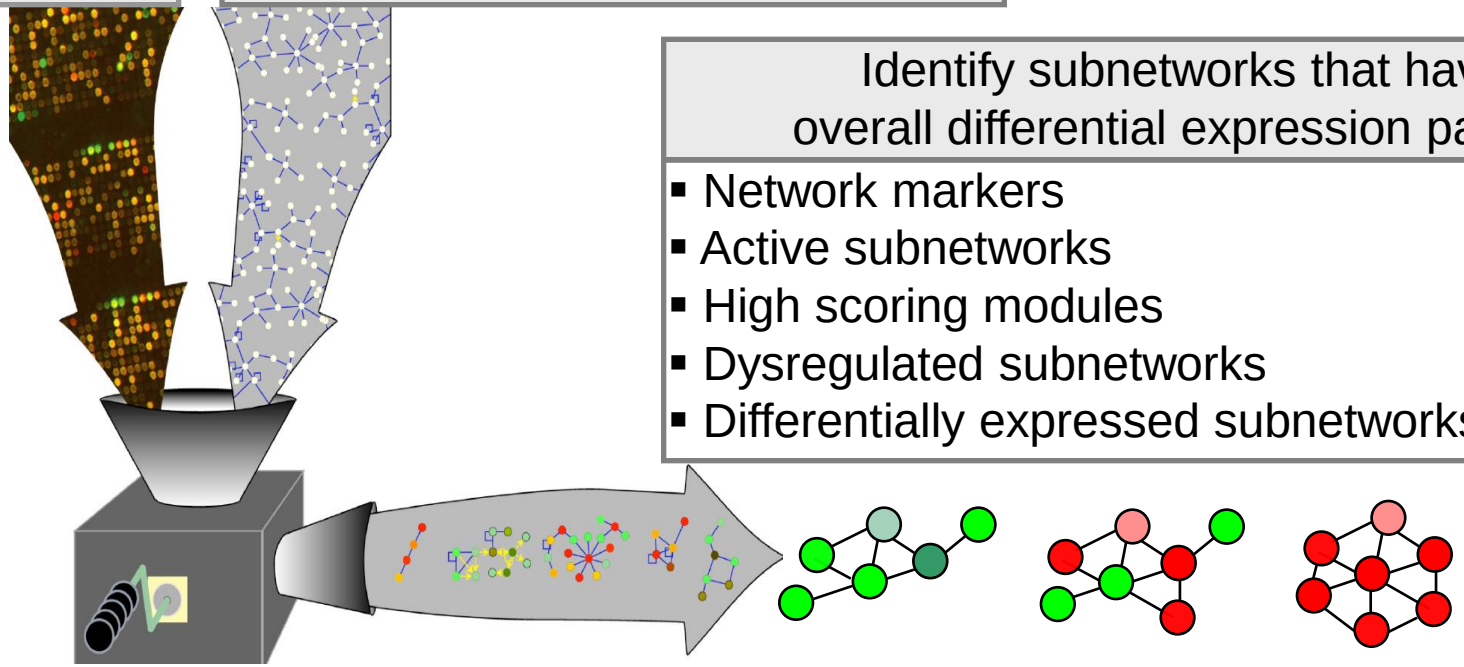


Identifying differentially expressed PPI subnetworks

Overall scheme

Gene signature

Protein-protein interactome (PPI)



Identify subnetworks that have overall differential expression pattern

- Network markers
- Active subnetworks
- High scoring modules
- Dysregulated subnetworks
- Differentially expressed subnetworks

Why identify differentially expressed subnetworks?

❑ Current set of pre-defined pathways (e.g., GO, KEGG) is based on current biological knowledge, which is far from complete.

❑ Only part of the pathway is usually altered during biological processes.

} Functional enrichment analysis on pre-defined pathways is not sufficient

❑ Solution: To identify differentially expressed subnetworks de novo.

Interactome mapping of DEGs versus DE subnetworks

Interactome mapping of DEGs (cutoff-based)

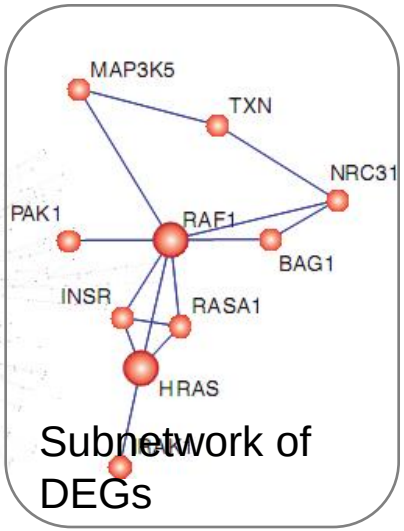
Expression data

Selection of DEGs

List of DEGs

Overlay DEGs onto global PPI

Identify connected subnetworks



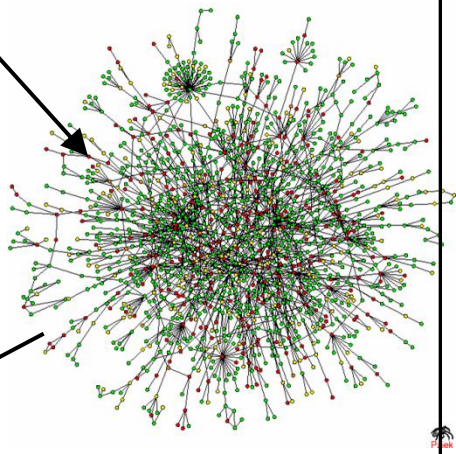
Subnetwork of DEGs

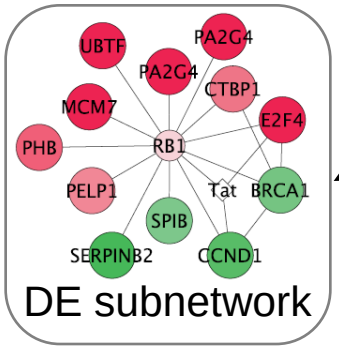
- Usually called list-to-network method
- Many web server softwares exist

Differentially expressed subnetworks (cutoff-free)

Expression data

Full list of all genes + statistics





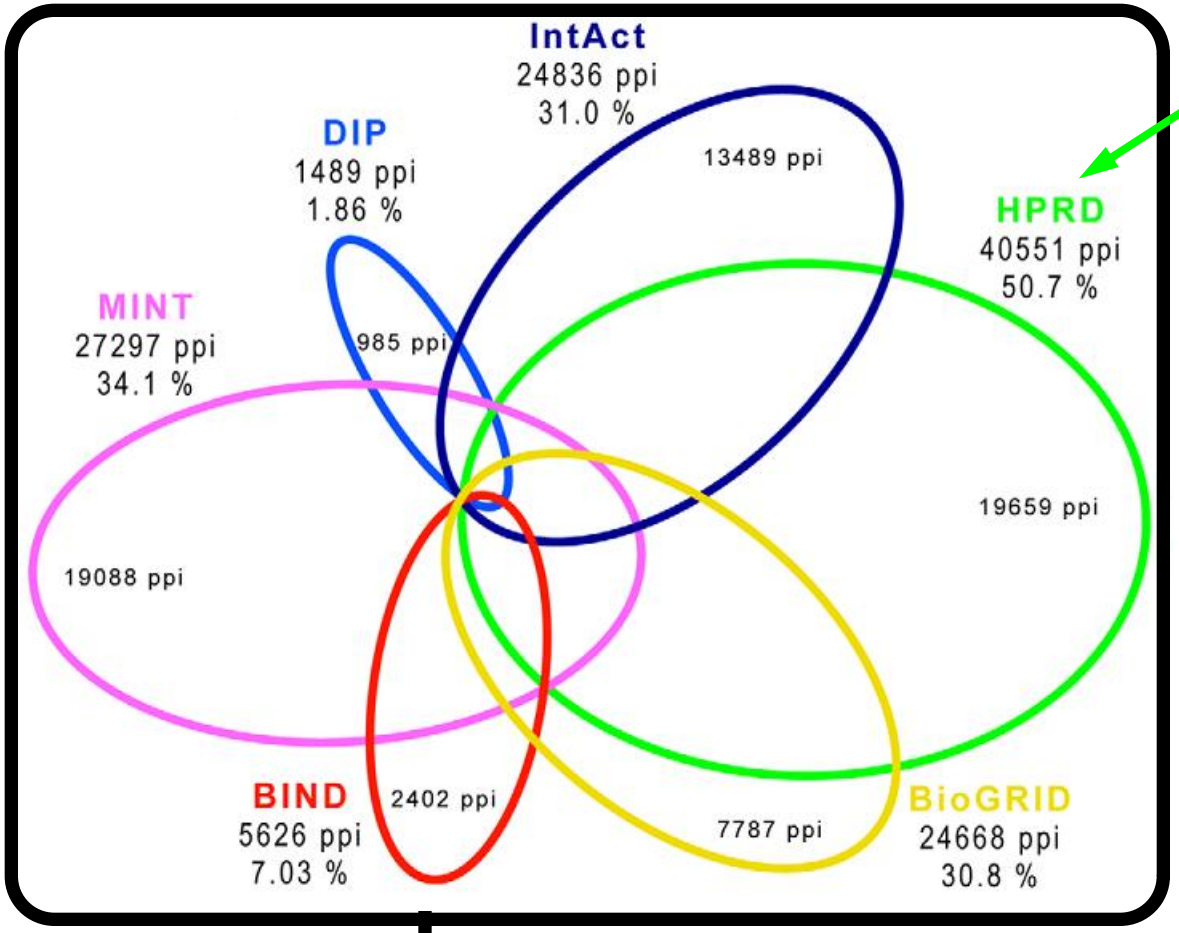
DE subnetwork

Identify subnetworks with overall significant differential expression

- Covered in this lecture

Primary databases of PPI

Coverage of human PPIs on major primary databases
[De Las Rivas (2010) PLoS Comput Biol; PMID:20589078]



- HPRD shows the largest coverage.
- Nevertheless, there are many PPIs that are not covered by HPRD.
- In addition, overlap between databases is low.
- The low overlap is intentional because these DBs form a consortium (IMEx; International Molecular Interaction Exchange).
- The IMEx consortium coordinates their curation efforts to avoid unnecessary redundancy.

Need a collective database of all these primary databases of PPI
→ Called a meta-database (or consolidated database)

Meta-databases of PPI

- MiMI (Michigan Molecular Interactions)

Michigan molecular interactions r2: from interacting proteins to pathways

V. Glenn Tarcea, Terry Weymouth, Alex Ade, Aaron Bookvich, Jing Gao, Vasudeva Mahavisno, Zach Wright, Adriane Chapman, Magesh Jayapandian, Arzucan Özgür, Yuanyuan Tian, Jim Cavalcoli, Barbara Mirel, Jignesh Patel, Dragomir Radev, Brian Athey, David States and H. V. Jagadish*

- I2D (Interologous Interaction Database)

Unequal evolutionary conservation of human protein interactions in interologous networks

Kevin R Brown*† and Igor Jurisica*†‡

- iRefWeb (Interaction Reference Web)

iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence

Brian Turner[1], Sabry Razick[2,3], Andrei L. Turinsky[1], James Vlasblom[4], Edgard K. Crowdy[5], Emerson Cho[1], Kyle Morrison[1], Ian M. Donaldson[2,6] and Shoshana J. Wodak[1,4,7,*]

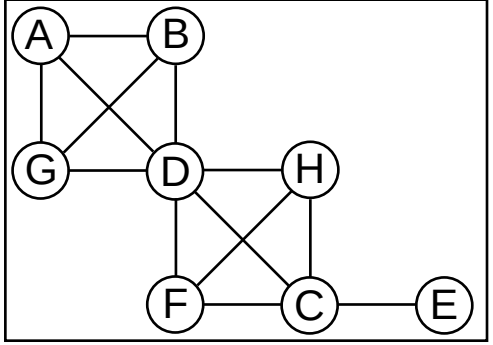
Identifying differentially expressed subnetworks with GiGA

Graph-based iterative Group Analysis enhances microarray interpretation

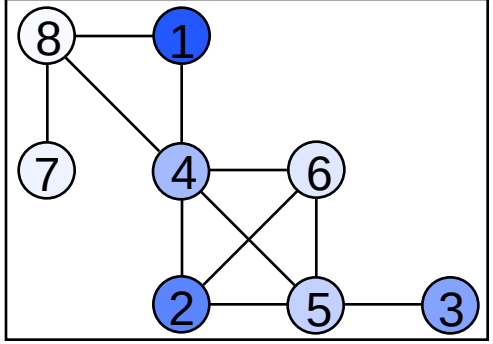
Rainer Breitling*[1,2], Anna Amtmann[1] and Pawel Herzyk[2,3]

Step 1: Assign the ranks to all the nodes in PPI

PPI



Assign the ranks to all the nodes in PPI

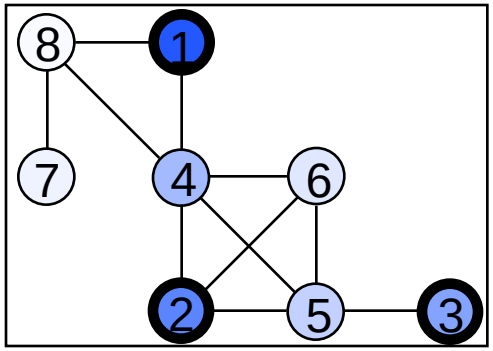


Gene signature

Sorted list of all genes

Rank	Gene	*p*-value
1	B	0.000001
2	F	0.00003
3	E	0.0004
4	D	0.002
5	C	0.06
6	H	0.1
7	G	0.5
8	A	0.8

Step 2: Find local minimum nodes



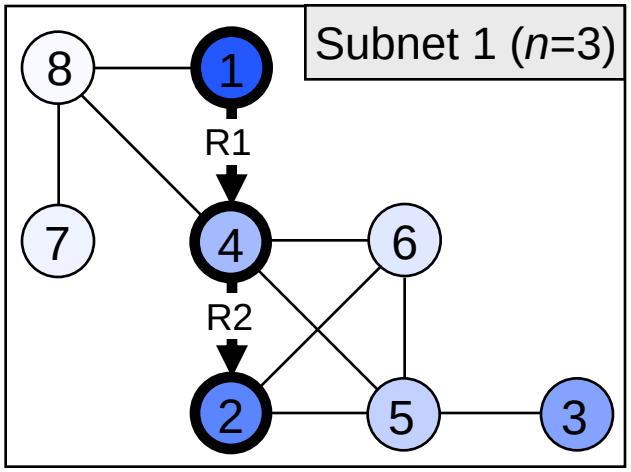
Local minima:

- ❑ nodes that have a higher rank than their direct neighbors
- ❑ serve as seed nodes for subnetwork extension

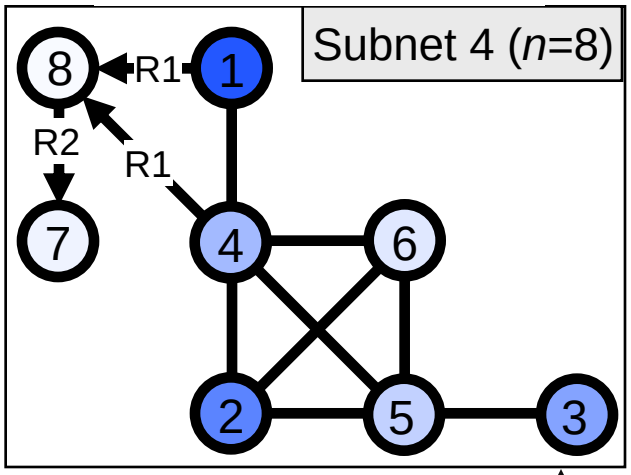
Identifying differentially expressed subnetworks with GiGA

Step 3: Subnetwork extension from rank 1 seed node

Extension round 1



Extension round 4

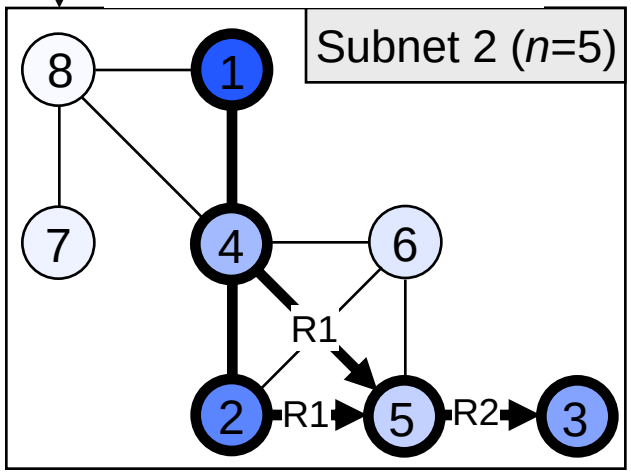


Extension rules

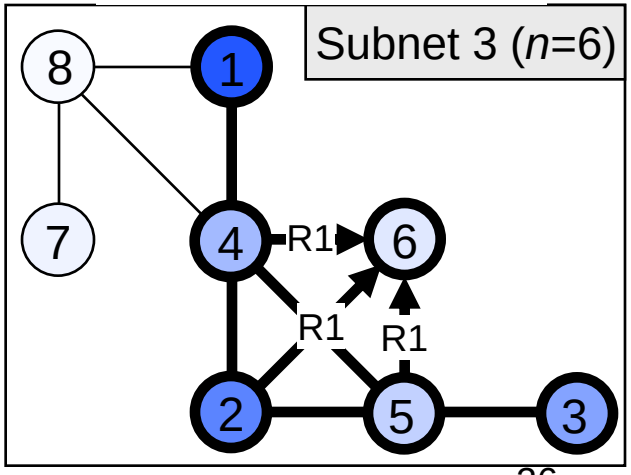
Rule 1: Extend to the direct neighbor with the highest rank

Rule 2: If the extended node has a neighbor with even higher rank, extend further to it

Extension round 2

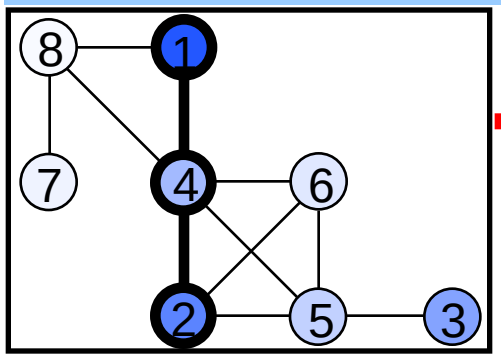


Extension round 3

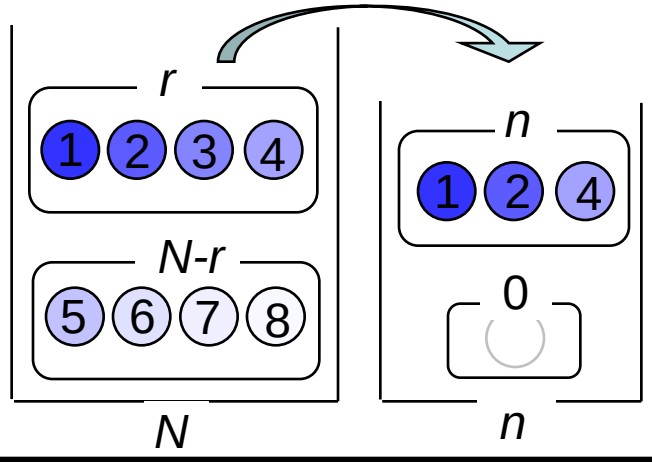


Identifying differentially expressed subnetworks with GiGA

Step 4: Calculation of *p*-value of subnetwork by hypergeometric distribution



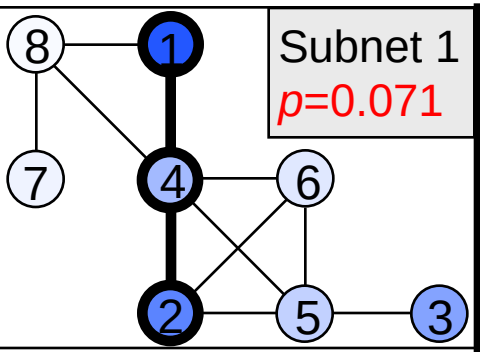
N (=8) # nodes in global net
n (=3) # nodes in subnet
r (=4) Largest node rank in subnet



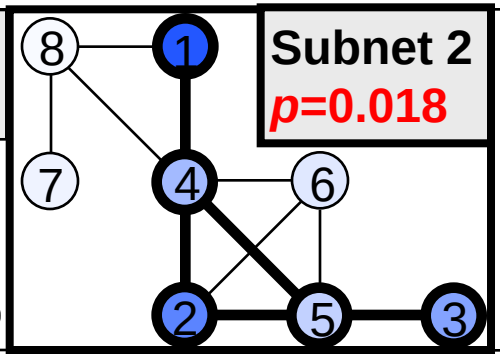
$$P(X=n)=\frac{\binom{r}{n}\binom{N-r}{0}}{\binom{N}{n}}$$

R> N=8; n=3; r=4
R> dhyper(n, r, N-r, n)
→ 0.071

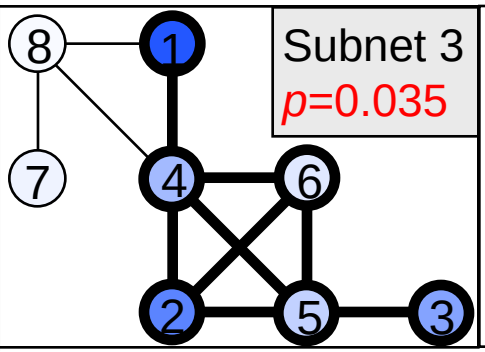
Step 5: Choose the most significant subnetwork



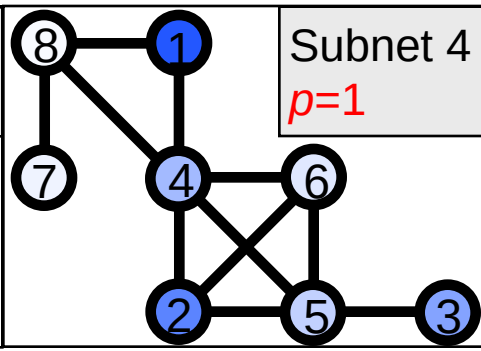
Subnet 1
p=0.071



Subnet 2
***p*=0.018**



Subnet 3
p=0.035



Subnet 4
p=1

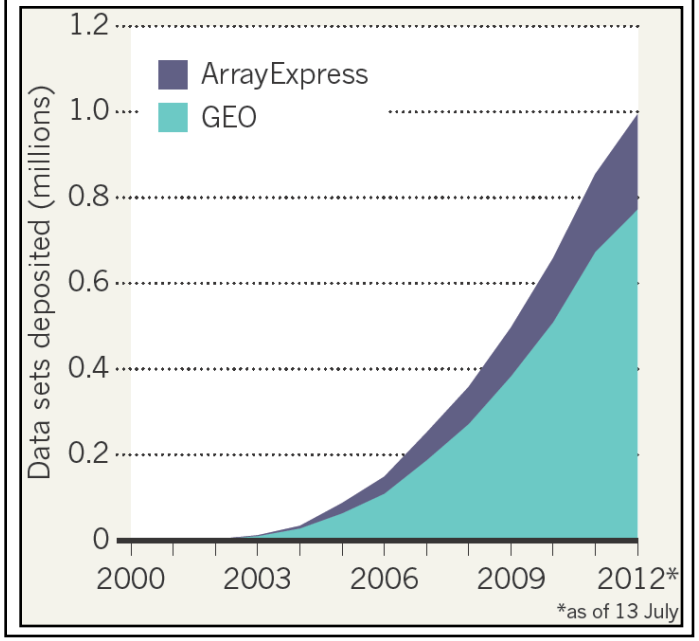
Step 6: Repeat the procedure from next seed node

Signature comparison using signature databases

DB of gene expression data

EMBL-EBI ArrayExpress

GEO Gene Expression Omnibus



DB of signatures

Differential expression analysis pipeline → Gene signature database

Identify related signatures from DB

Input your signature

- ❑ Confirm the findings from your data
- ❑ Identify common DEGs
- ❑ Generate new hypothesis

MARQ: an online tool to mine GEO for experiments with similar or opposite gene expression signatures

Miguel Vazquez[1], Ruben Nogales-Cadenas[2], Javier Arroyo[3], Pedro Botías[4], Raul García[3], Jose M. Carazo[5], Francisco Tirado[2], Alberto Pascual-Montano[5] and Pedro Carmona-Saez[2,*]

Signature comparison using signature databases

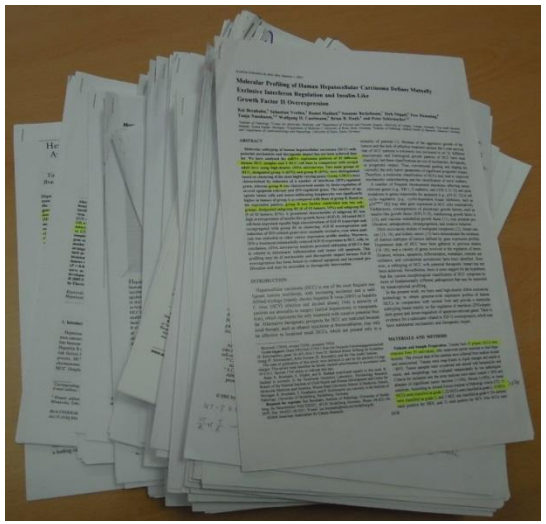
Gene signature tables from papers

Table II. Top 30 Up-regulated Genes Distinguishing MD from WD

Predictor genes	*P* value
proteasome 26S subunit, ATPase, 5	2.774
cytochrome *c* oxidase subunit VIa polypeptide 1	1.992
chaperonin containing TCP1, subunit 3	**1.983**
prohibitin	1.803
human D9 splice variant B mRNA	**1.753**
proteasome subunit, β, type 4	1.733
hydroxyacyl-coenzyme A dehydrogenase, type II	1.697
peptidylprolyl isomerase A	1.662
adenosine deaminase, RNA-specific	**1.654**
GCN5-like 1	1.591
mitochondrial ribosomal protein L12	1.493

2.40E-05	100	1.425	0.974	163235

CGI-69 −1.5 1.5



DB of signatures

Text extraction and curation → Gene signature database

Advantages of publication-derived signature DBs

- ❑ Directly import the end results from expert original analysis of individual studies
- ❑ Can always obtain signatures from papers
- ❑ Can obtain signatures from complex experimental design beyond simple two-class comparison (e.g., survival analysis, tissue specificity analysis)

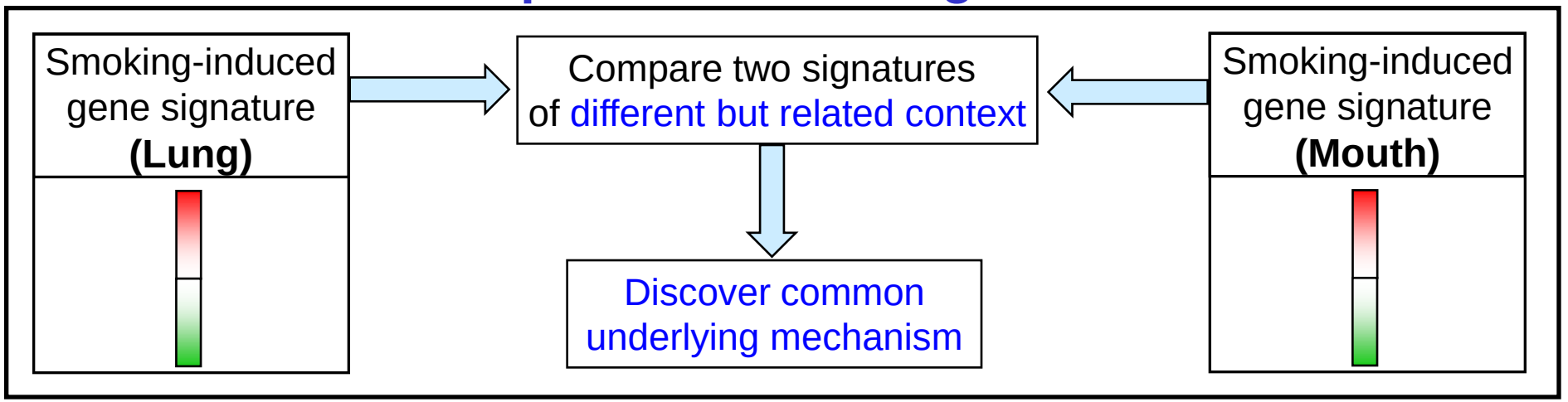
GeneSigDB: a manually curated database and resource for analysis of gene expression signatures

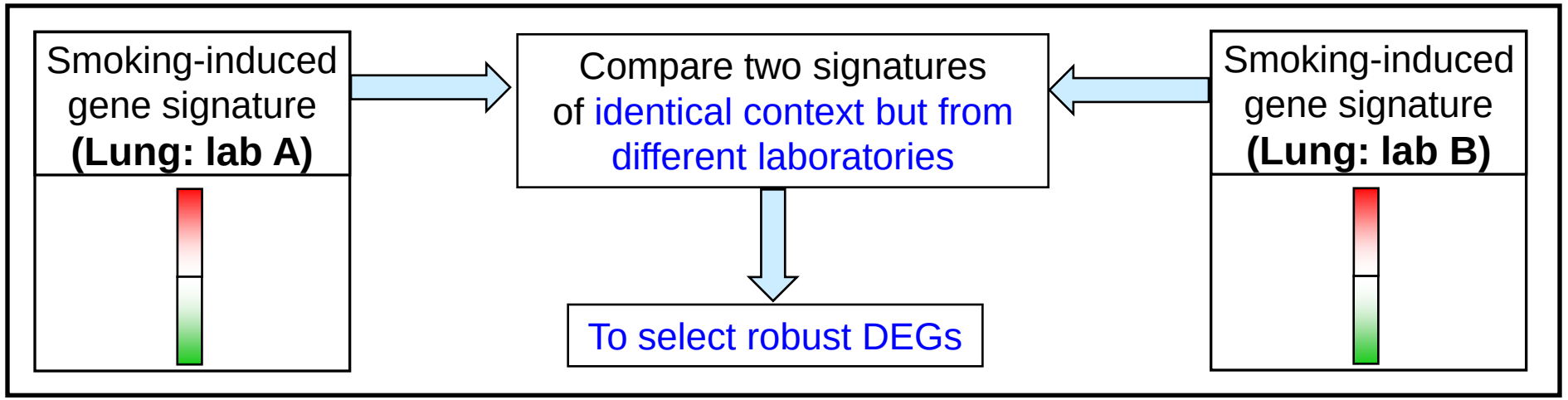
Aedín C. Culhane[1,2,*], Markus S. Schröder[1], Razvan Sultana[1], Shaita C. Picard[1], Enzo N. Martinelli[1], Caroline Kelly[1], Benjamin Haibe-Kains[1,2], Misha Kapushesky[3], Anne-Alyssa St Pierre[1], William Flahive[1], Kermshlise C. Picard[1], Daniel Gusenleitner[1], Gerald Papenhausen[1], Niall O'Connor[1], Mick Correll[1] and John Quackenbush[1,3,4,*]

Liverome: a curated database of liver cancer-related gene signatures with self-contained context information

Langho Lee[1], Kai Wang[2], Gang Li[2], Zhi Xie[2], Yuli Wang[2], Jiangchun Xu[2], Shaoxian Sun[2], David Pocalyko[2], Jong Bhak[3], Chulhong Kim[3], Kee-Ho Lee[4], Ye Jin Jang[5], Young Il Yeom[5], Hyang-Sook Yoo[6*], Seungwoo Hwang[1*]

Comparison of two signatures





***OrderedList*—a bioconductor package for detecting similarity in ordered gene lists**

Claudio Lottaz[1,*], Xinan Yang[1,2], Stefanie Scheid[1] and Rainer Spang[1]

Calculating similarity score between signatures and *p*-value

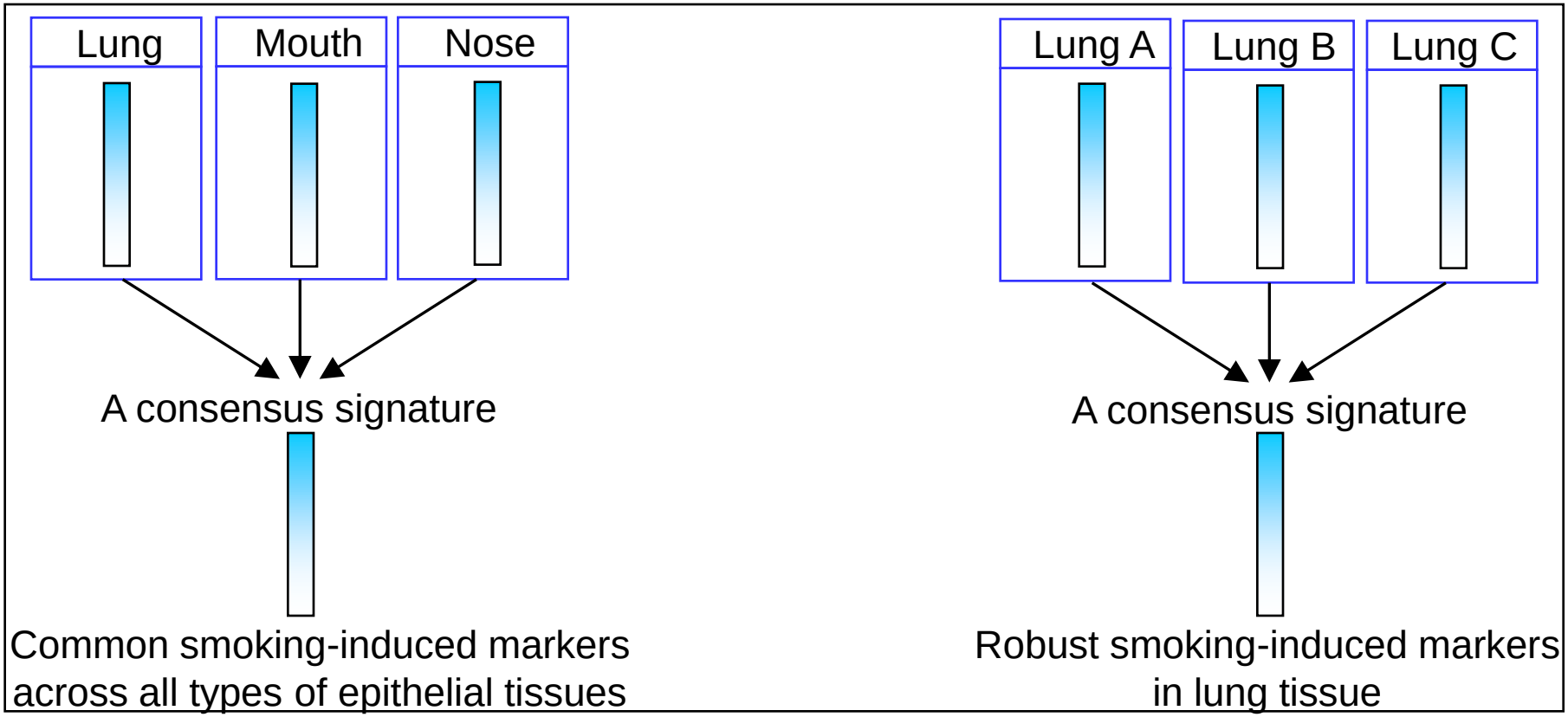
Sorted list 1	Sorted list 2	Rank *i*	Cumulative no. of matches at rank *i*	Weight at rank *i* $w_i = 1/e^{\alpha \cdot i}$	Cumulative weighted no. of matches at rank *i*
COL1A2	MMP14	1	0	1/exp(*α*)	0 x 1/exp(*α*)
MMP14	A2M	2	1	1/exp(2*α*)	1 x 1/exp(2*α*)
SPARC	SPARC	3	2	1/exp(3*α*)	2 x 1/exp(3*α*)
VIM	MST1	4	2	1/exp(4*α*)	2 x 1/exp(4*α*)
A2M	COL1A2	5	4	1/exp(5*α*)	4 x 1/exp(5*α*)
…	…	…	…	…	… …
MST1	VIM	*N*	Up-match sum	1/exp(*Nα*)	Sum x 1/exp(*Nα*)
MT3	SOD1	*N*	Down-match sum	1/exp(*Nα*)	Sum x 1/exp(*Nα*)
…	…	…	…	…	… …
RHOB	RHOB	5	2	1/exp(5*α*)	2 x 1/exp(5*α*)
RND3	SERPINE1	4	1	1/exp(4*α*)	1 x 1/exp(4*α*)
SERPINE1	GAPDH	3	0	1/exp(3*α*)	0 x 1/exp(3*α*)
SOD1	MT3	2	0	1/exp(2*α*)	0 x 1/exp(2*α*)
VTN	PLG	1	0	1/exp(*α*)	0 x 1/exp(*α*)

Obtain overall similarity score between signatures 1 and 2

Calculate *p*-value of the similarity score by permutation

List of major common genes that contribute to the score the most

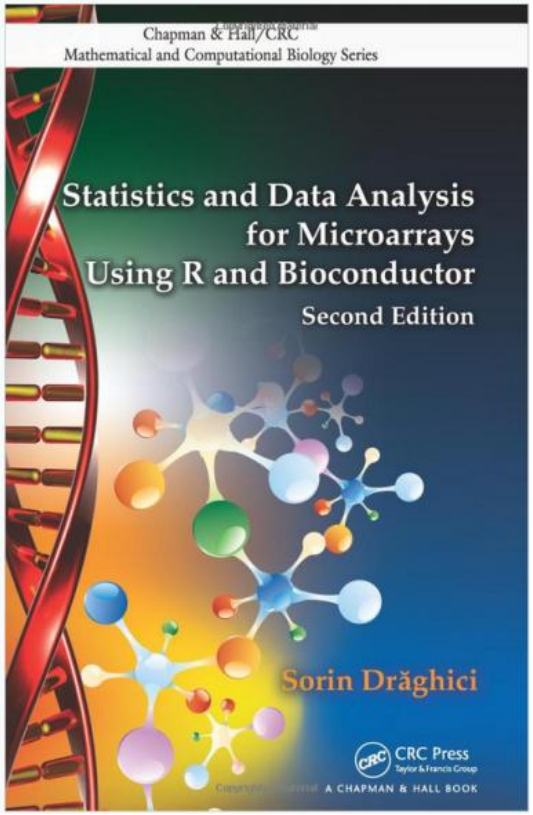
Deriving a consensus signature from several signatures

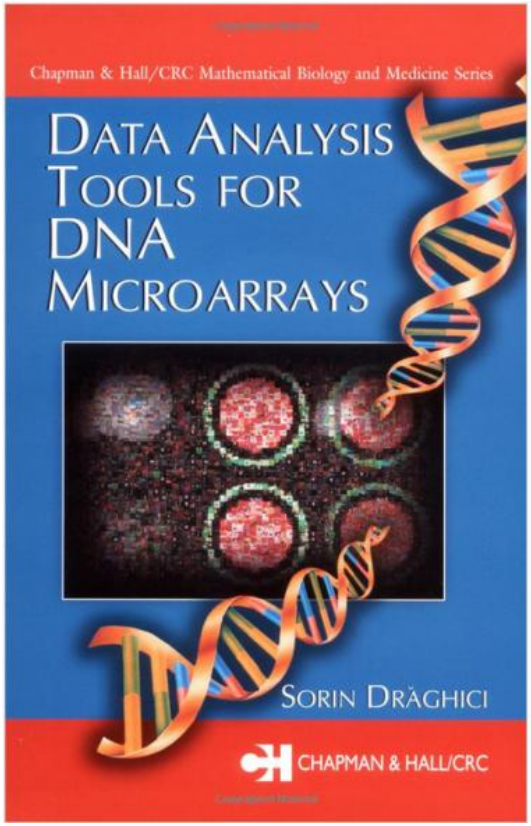


RankAggreg, an R package for weighted rank aggregation
Vasyl Pihur, Susmita Datta and Somnath Datta*

Rank aggregation methods
Shili Lin*

Book recommendations: Microarray





O

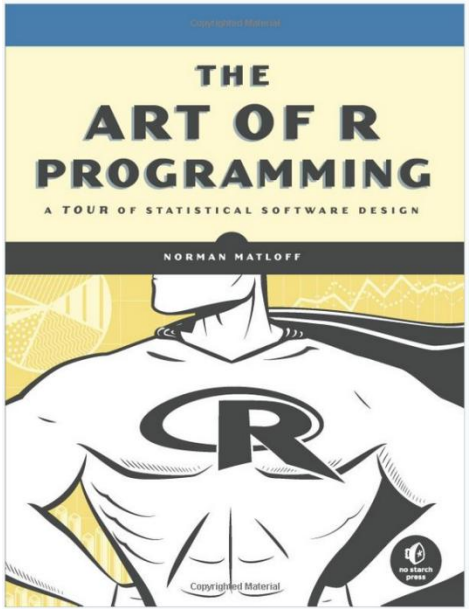
Statistics and Data Analysis for Microarrays using R and Bioconductor (**2011, 2nd Edition**)

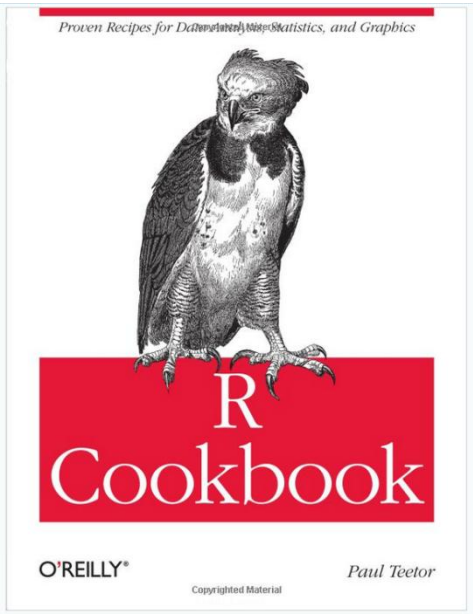
X

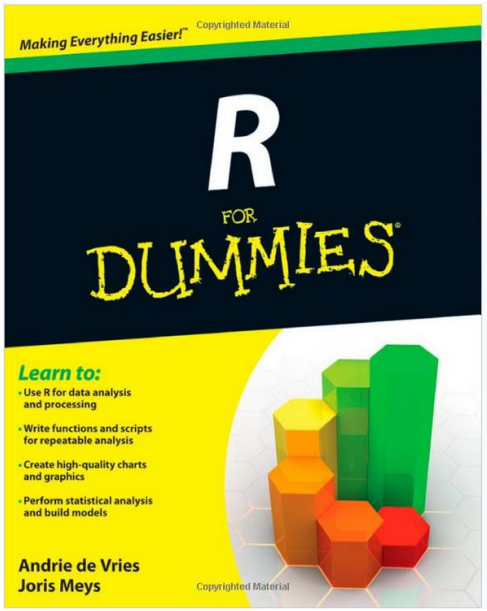
Data Analysis Tools for DNA Microarrays (**2003, 1st Edition**)

Book recommendations: R

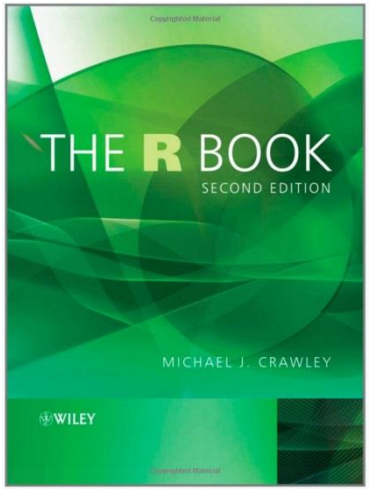
Teach how to program in R (Useful for bioinformaticians)

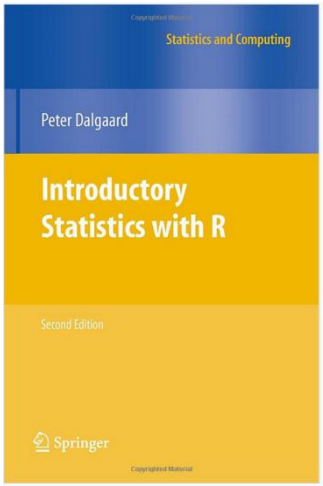






Teach how to do statistics in R (Not appropriate for bioinformaticians)





. . . And may others