Use of Context in Gene Expression and Proteomic Profile Analysis

Limsoon Wong

NUS
National University of Singapore

Preliminaries

- This tutorial assumes you already know a little about what biological networks are. If you don’t, Natasa Przulj’s lecture slides maybe helpful

  http://www.doc.ic.ac.uk/~natasha/341_Lectures_2-3_notes.pdf

- The ppt for this tutorial can be downloaded at

  http://www.comp.nus.edu.sg/~wongls/talks/sstic2013.pdf

- Brief notes for this tutorial can be downloaded at

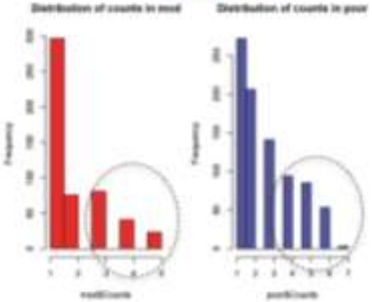
  http://www.comp.nus.edu.sg/~wongls/talks/apbc2012-tutorialnotes.pdf

Outline

Part 1: Delivering reproducible gene expression analysis

- **Some issues in gene expression analysis**
- **Batch effect & normalization**
- **Reproducibility**
  - Law of large numbers
  - Use background info
  - Find more consistent disease subnetworks

Part 2: Delivering more powerful proteomic profile analysis



- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

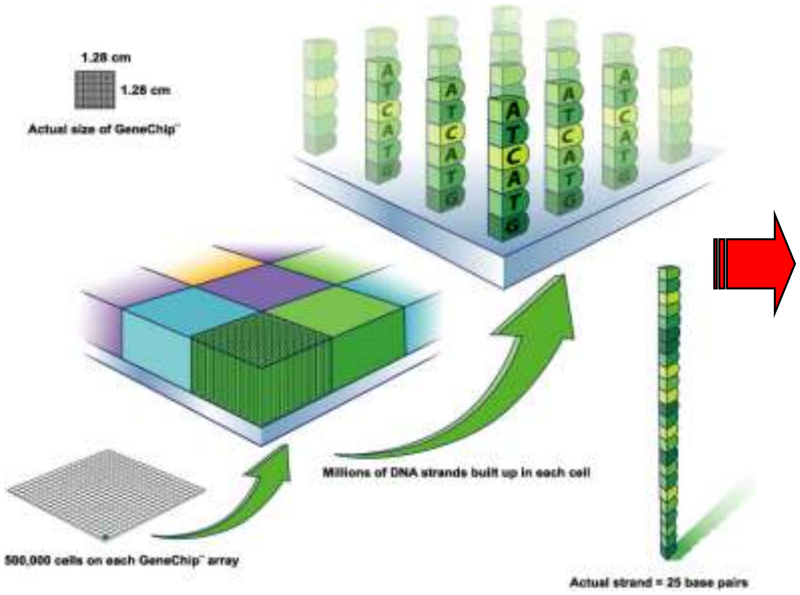
Part 3: How good are available sources of pathway & PPI Network?

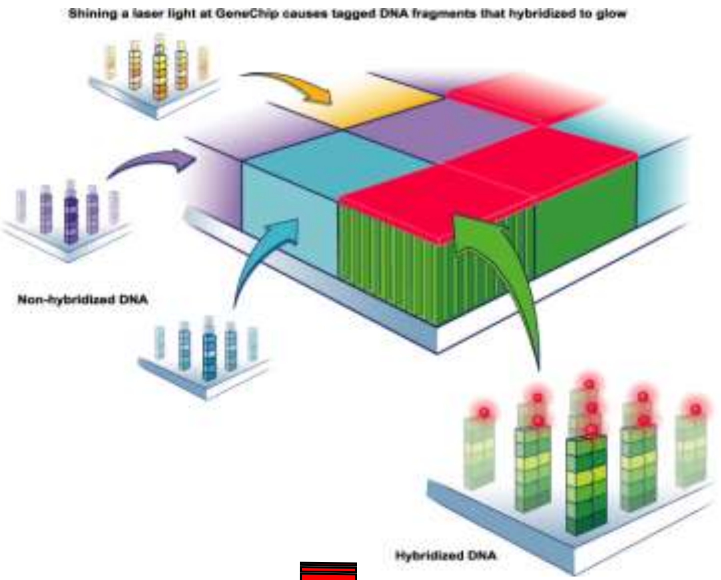
- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**

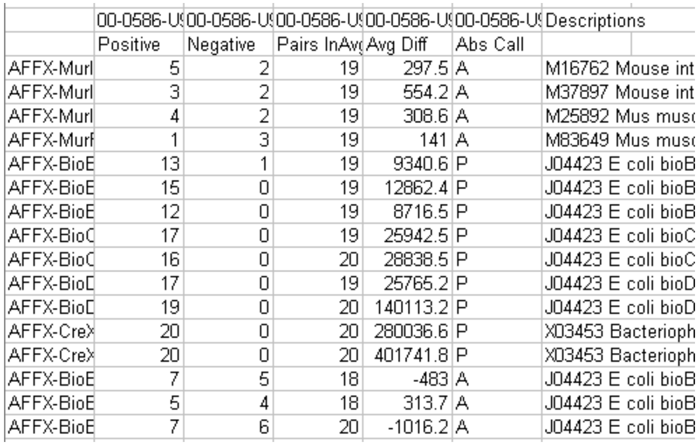
Use of Context in Gene Expression and Proteomic Profile Analysis *Part 1*

Limsoon Wong

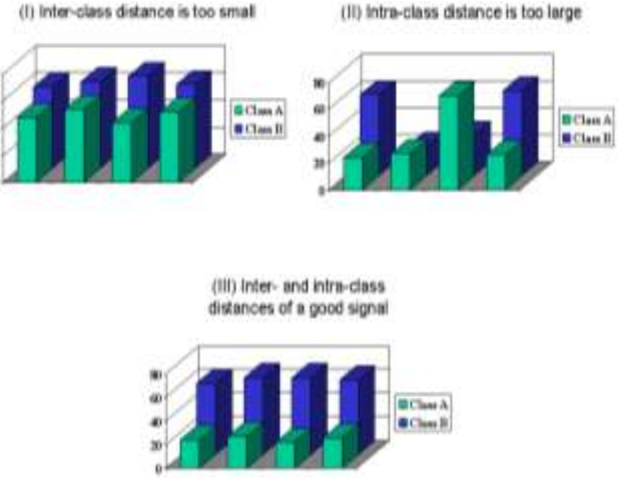
Diagnosis Using Microarray


1.28 cm
1.28 cm
Actual size of GeneChip™
Millions of DNA strands built up in each cell
500,000 cells on each GeneChip™ array
Actual strand = 25 base pairs
A T C A T G


Shining a laser light at GeneChip causes tagged DNA fragments that hybridized to glow
Non-hybridized DNA
Hybridized DNA



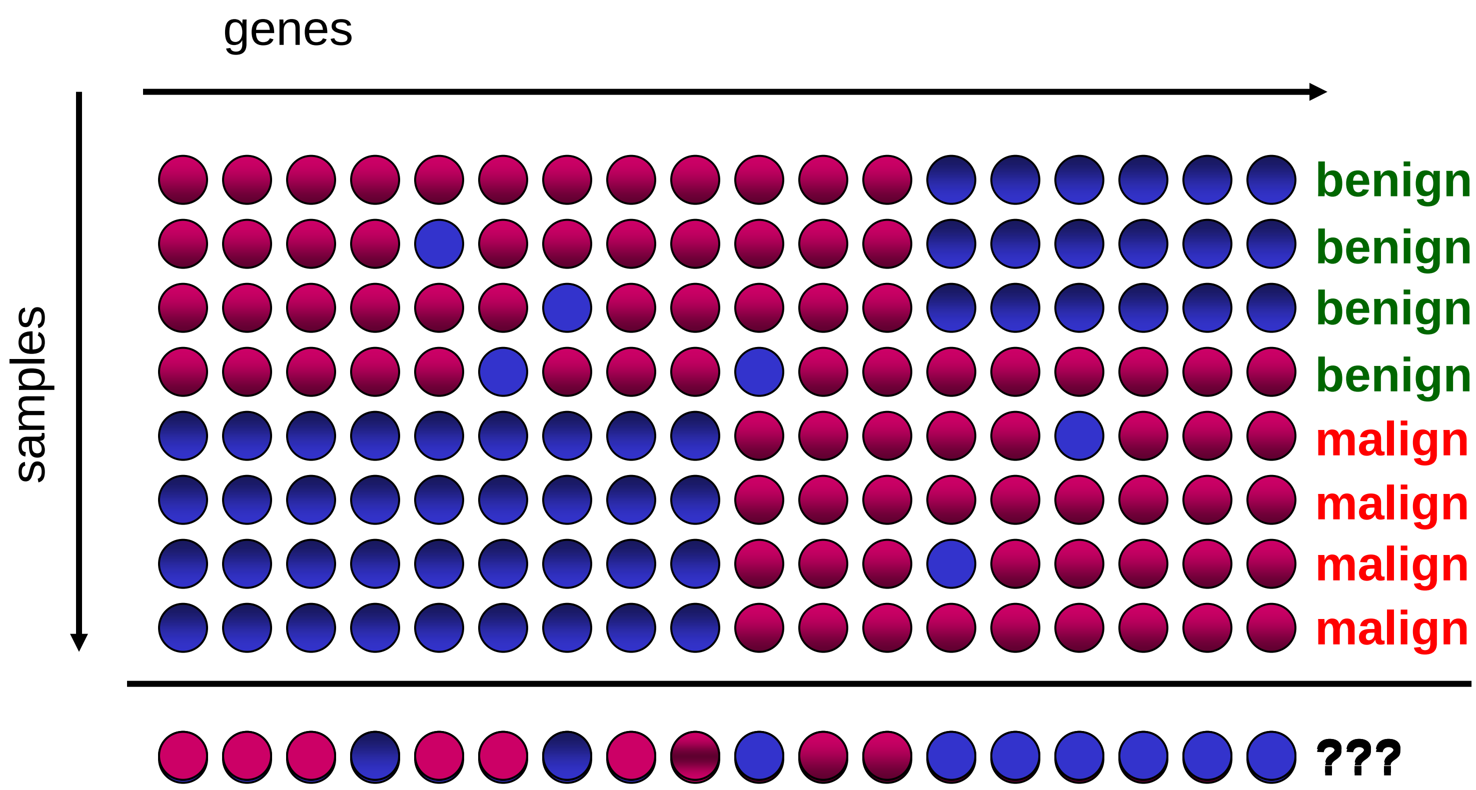
	00-0586-U	00-0586-U	00-0586-U	00-0586-U	00-0586-U	Descriptions
	Positive	Negative	Pairs InAv	Avg Diff	Abs Call	
AFFX-MurI	5	2	19	297.5	A	M16762 Mouse int
AFFX-MurI	3	2	19	554.2	A	M37897 Mouse int
AFFX-MurI	4	2	19	308.6	A	M25892 Mus musc
AFFX-MurF	1	3	19	141	A	M83649 Mus musc
AFFX-BioB	13	1	19	9340.6	P	J04423 E coli bioB
AFFX-BioB	15	0	19	12862.4	P	J04423 E coli bioB
AFFX-BioB	12	0	19	8716.5	P	J04423 E coli bioB
AFFX-BioC	17	0	19	25942.5	P	J04423 E coli bioC
AFFX-BioC	16	0	20	28838.5	P	J04423 E coli bioC
AFFX-BioD	17	0	19	25765.2	P	J04423 E coli bioD
AFFX-BioD	19	0	20	140113.2	P	J04423 E coli bioD
AFFX-CreX	20	0	20	280036.6	P	X03453 Bacterioph
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacterioph
AFFX-BioB	7	5	18	-483	A	J04423 E coli bioB
AFFX-BioB	5	4	18	313.7	A	J04423 E coli bioB
AFFX-BioB	7	6	20	-1016.2	A	J04423 E coli bioB


(I) Inter-class distance is too small

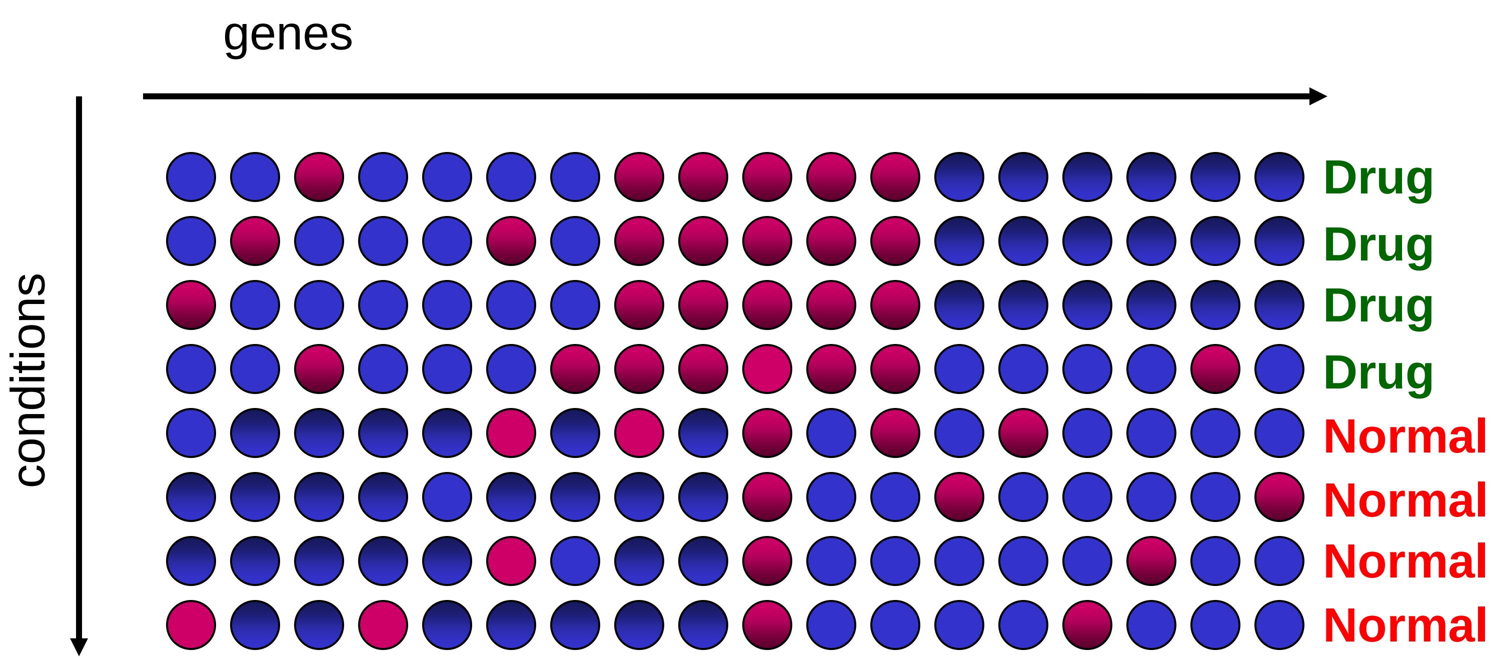
(II) Intra-class distance is too large

(III) Inter- and intra-class distances of a good signal

Application: Disease Subtype Diagnosis



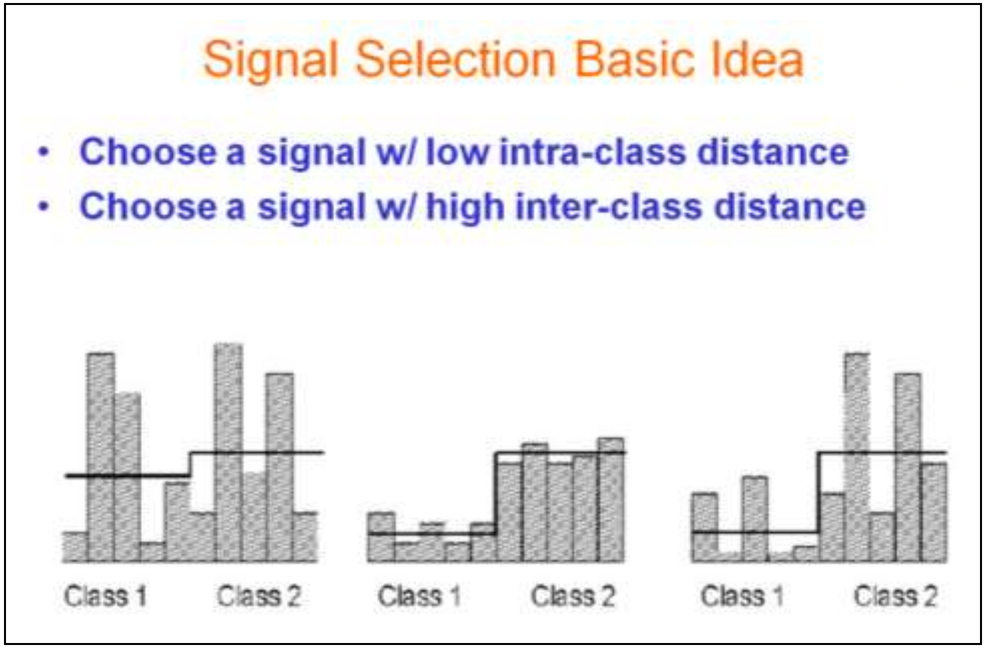
Application: Drug Action Detection



Which group of genes are the drug affecting on?

Typical Analysis Workflow

- **Gene expression data collection**
- **DE gene selection by, e.g., t-statistic**
- **Classifier training based on selected DE genes**
- **Apply the classifier for diagnosis of future cases**



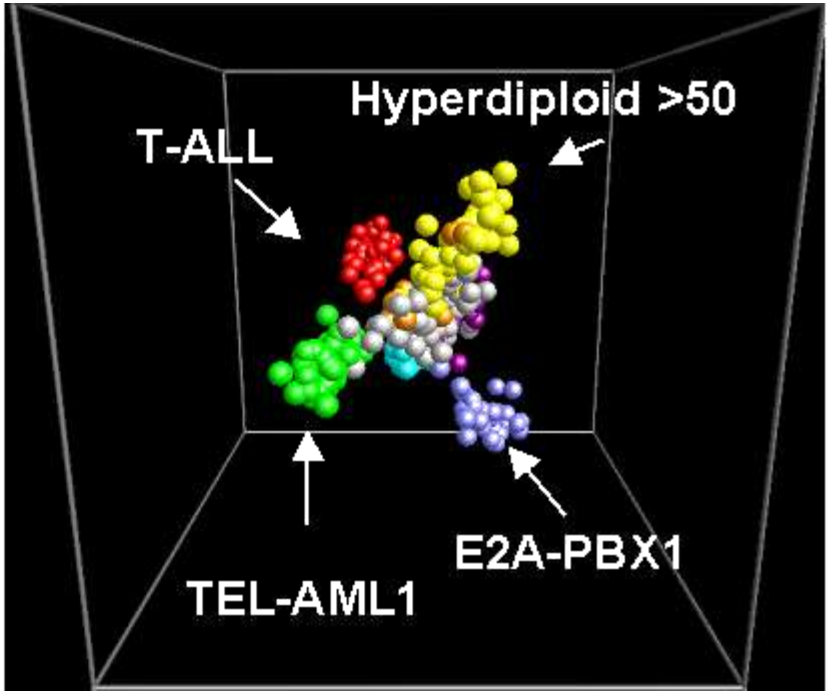
Signal Selection Basic Idea

- Choose a signal w/ low intra-class distance
- Choose a signal w/ high inter-class distance

Image credit: Golub et al., *Science*, 286:531–537, 1999

Terminology: DE gene = differentially expressed gene

PCA Plots



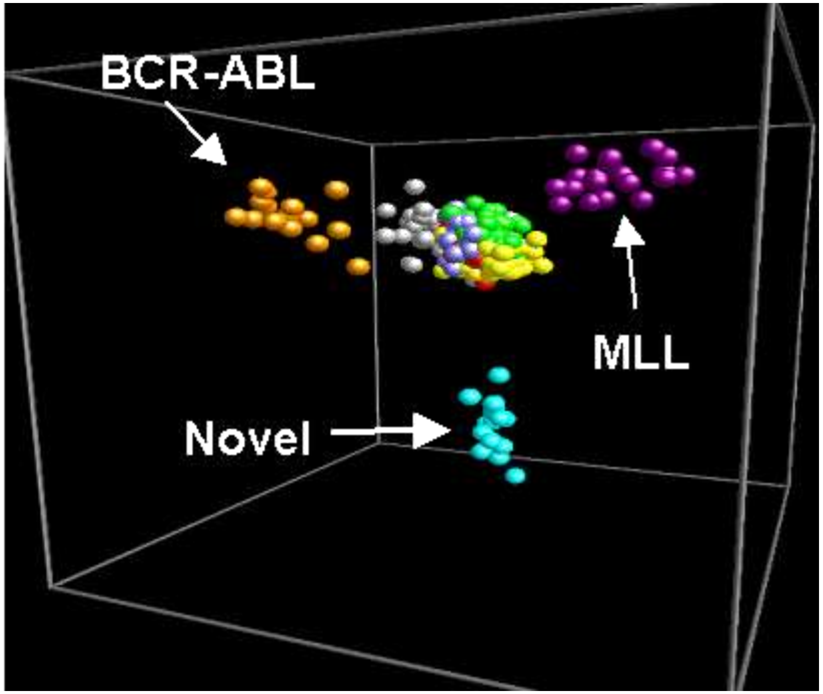


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

Part 1: Delivering reproducible gene expression analysis

- **Some issues in gene expression analysis**
- **Batch effect & normalization**
- **Reproducibility**
  - Law of large numbers
  - Use background info
  - Find more consistent disease subnetworks

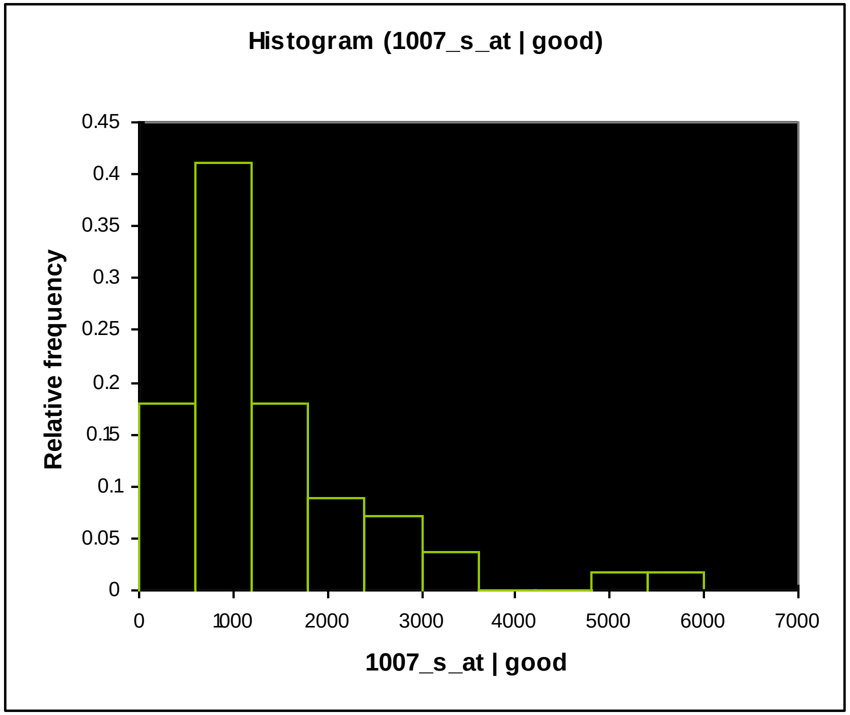
Some Headaches

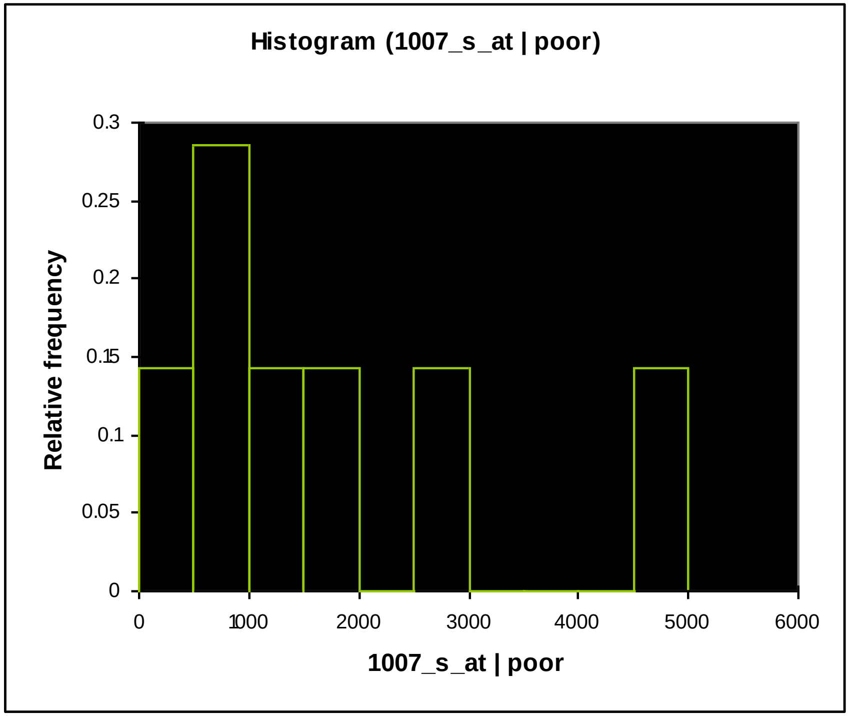
- **Natural fluctuations of gene expression in a person**

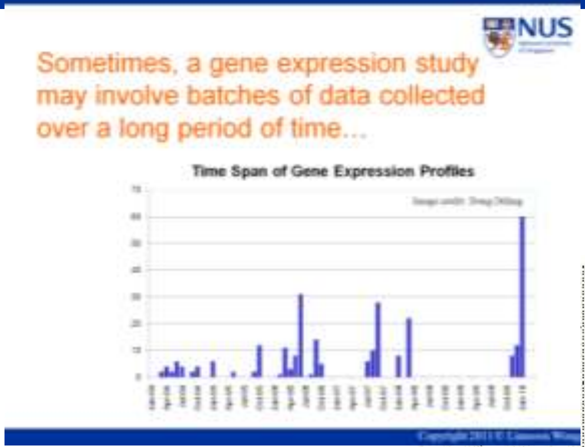
- **Noise in experimental protocols**
  - – Numbers mean diff things in diff batches
  - – Numbers mean diff things in data obtained from diff platforms

⇒ **Selected genes may not be meaningful**
  - – Diff genes get selected in diff expts

Natural Fluctuations



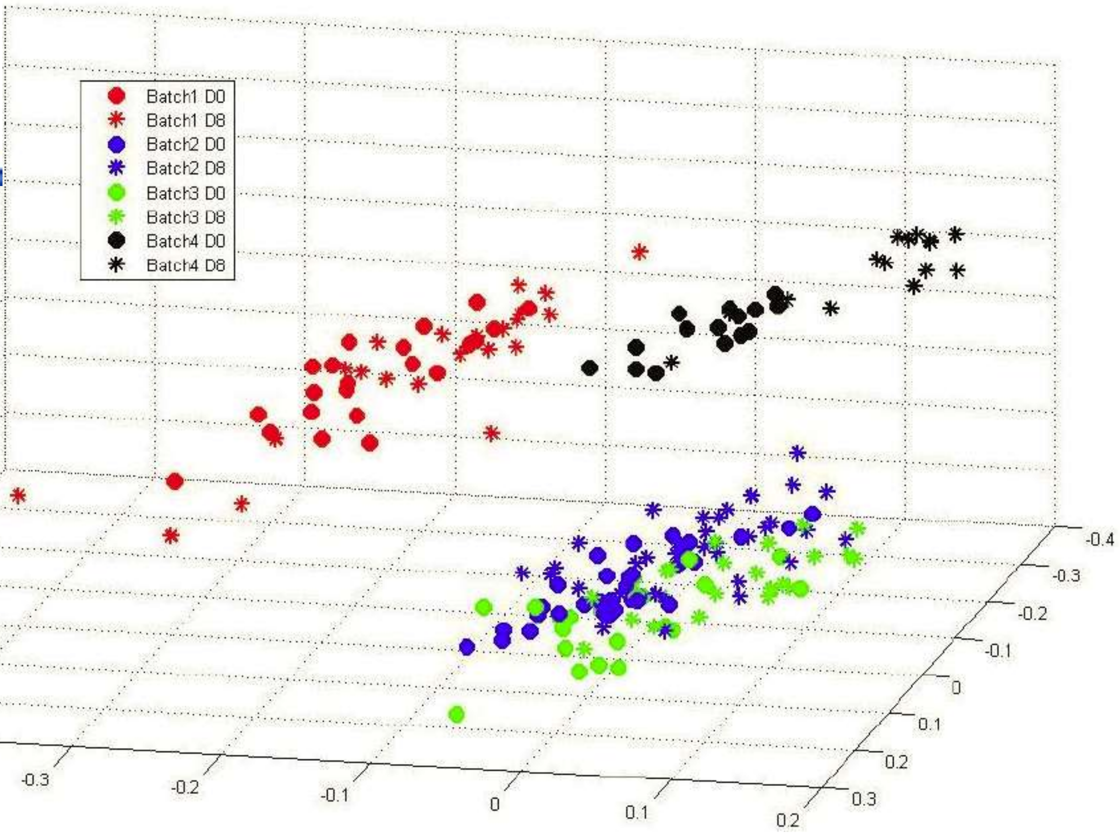




Sometimes, a gene expression study may involve batches of data collected over a long period of time…

Time Span of Gene Expression Profiles

Batch Effects



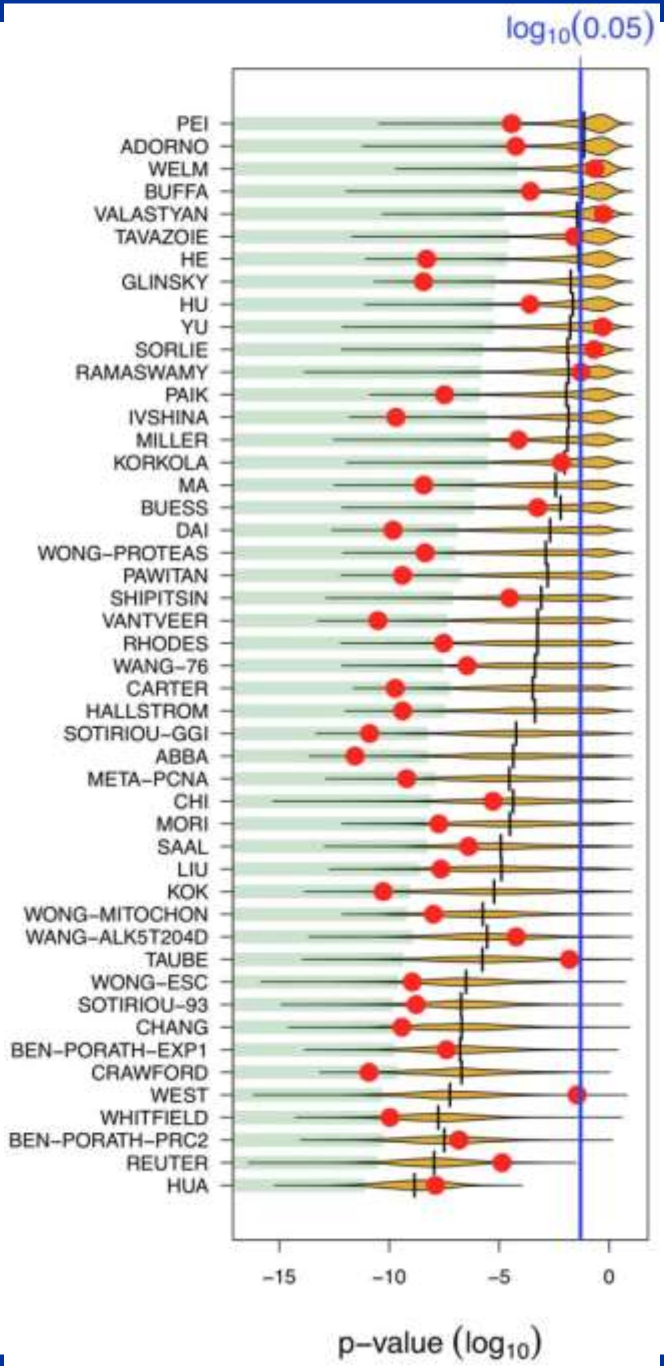
- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

Percentage of Overlapping Genes

- **Low % of overlapping genes from diff expt in general**
  - Prostate cancer
    - **Lapointe et al, 2004**
    - **Singh et al, 2002**
  - Lung cancer
    - **Garber et al, 2001**
    - **Bhattacharjee et al, 2001**
  - DMD
    - **Haslett et al, 2002**
    - **Pescatori et al, 2007**

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

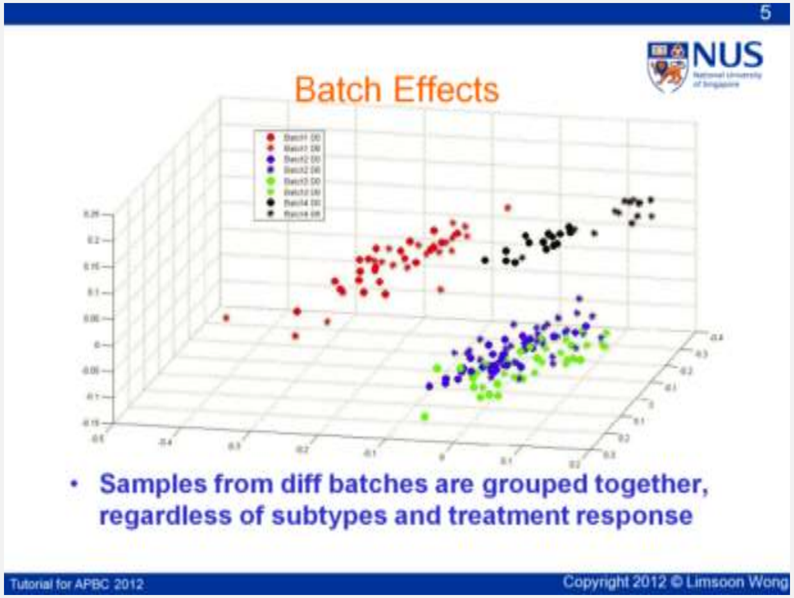
Zhang et al, Bioinformatics, 2009



“Most random gene expression signatures are significantly associated with breast cancer outcome”

Venet et al., *PLoS Comput Biol*, 7(10):e1002240, 2011.

Part 1: Delivering reproducible gene expression analysis



5

Batch Effects

- Samples from diff batches are grouped together, regardless of subtypes and treatment response

Tutorial for APBC 2012

Copyright 2012 © Limsoon Wong

- Some issues in gene expression analysis
- **Batch effect & normalization**
- Reproducibility
  - – Law of large numbers
  - – Use background info
  - – Find more consistent disease subnetworks

Approaches to Normalization

- **Aim of normalization:**

  Reduce variance w/o increasing bias

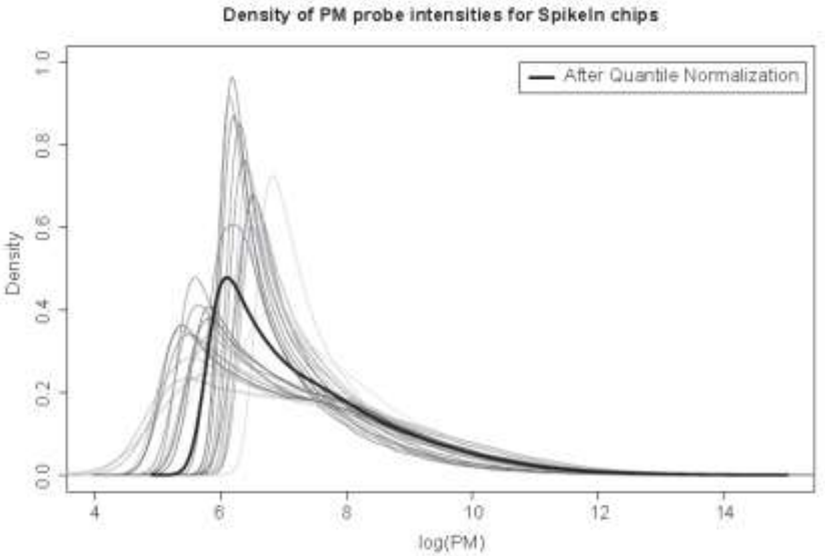
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

- **Transform data so that distribution of probe intensities is same on all arrays**
  - E.g., $(x - \mu) / \sigma$

- **Quantile normalization**

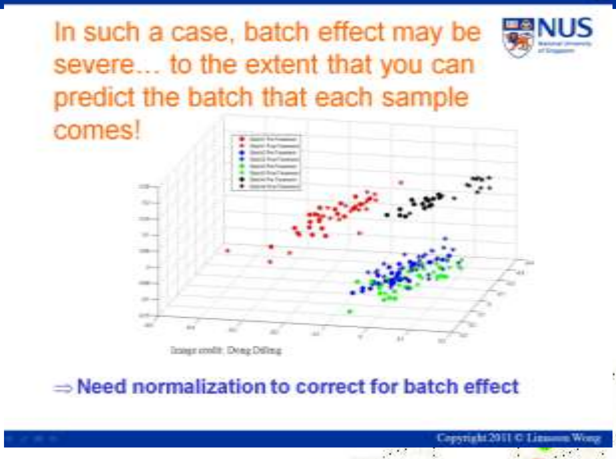
Quantile Normalization

- **Given *n arrays of length p, form X of size p × n where each array is a column***
- **Sort each column of X *to give* X_{sort}**
- **Take means across rows of X_{sort} *and assign this* mean to each elem in the row to get X'_{sort}**
- **Get $X_{normalized}$ *by arranging each column of* X'_{sort} to have same ordering as X**



- Implemented in some microarray s/w, e.g., EXPANDER

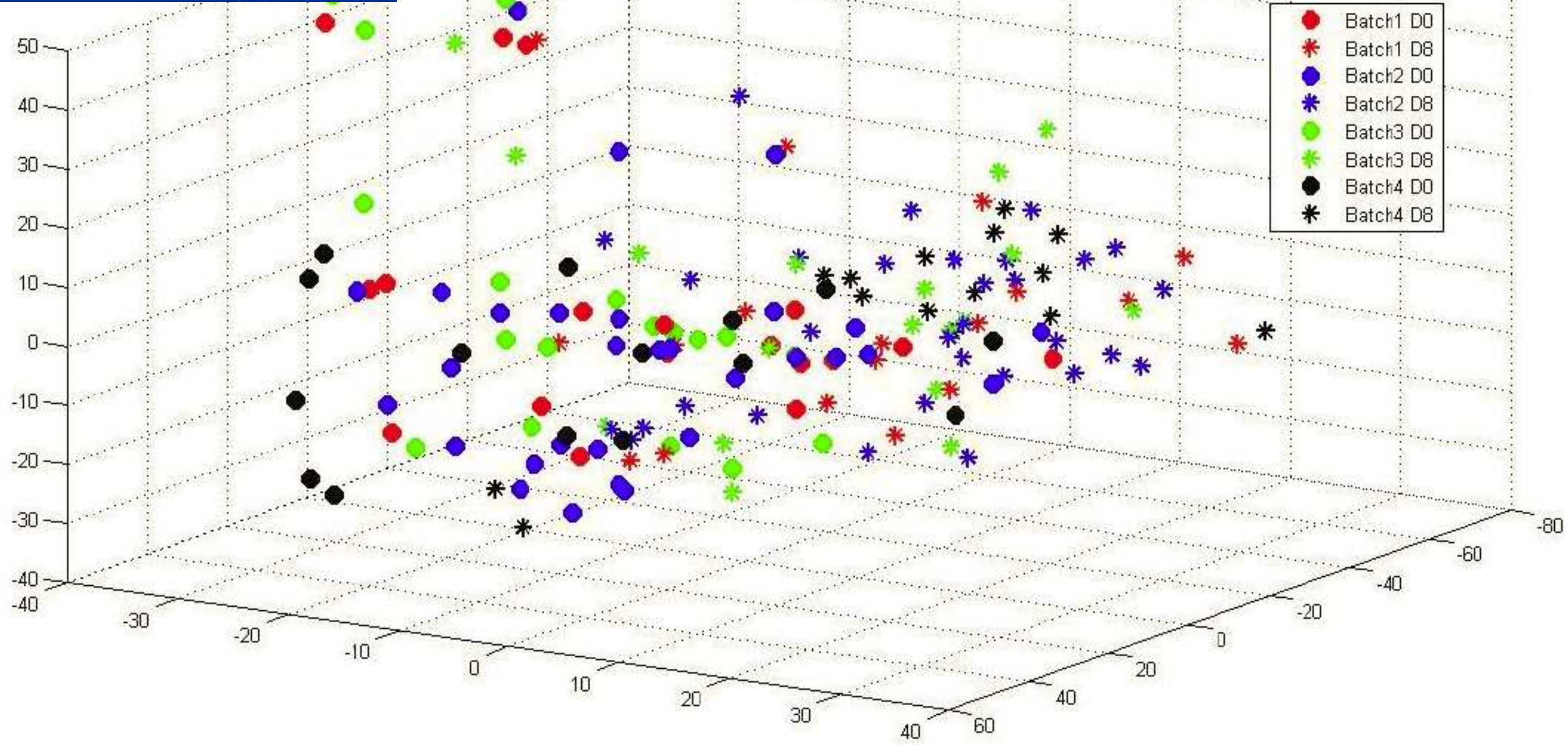
After quantile normalization



In such a case, batch effect may be severe… to the extent that you can predict the batch that each sample comes!

Image credit: Dong Difeng

⇒ Need normalization to correct for batch effect



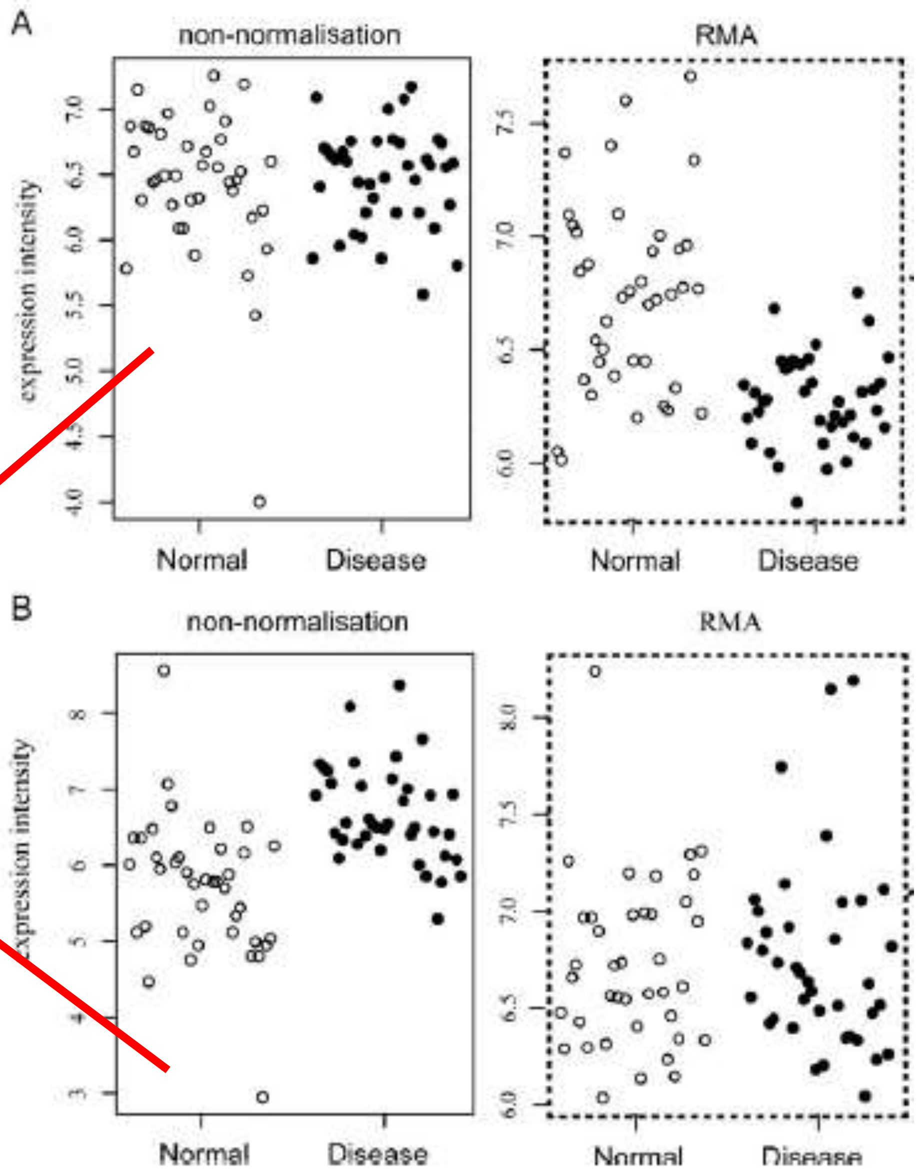
GEP after removing batch effect by quantile normalization

Caution: “Over normalize” signals in cancer samples

A gene normalized by quantile normalization (RMA) was detected as down-regulated DE gene, but the original probe intensities in cancer samples were higher than those in normal samples

A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile nornmalized data

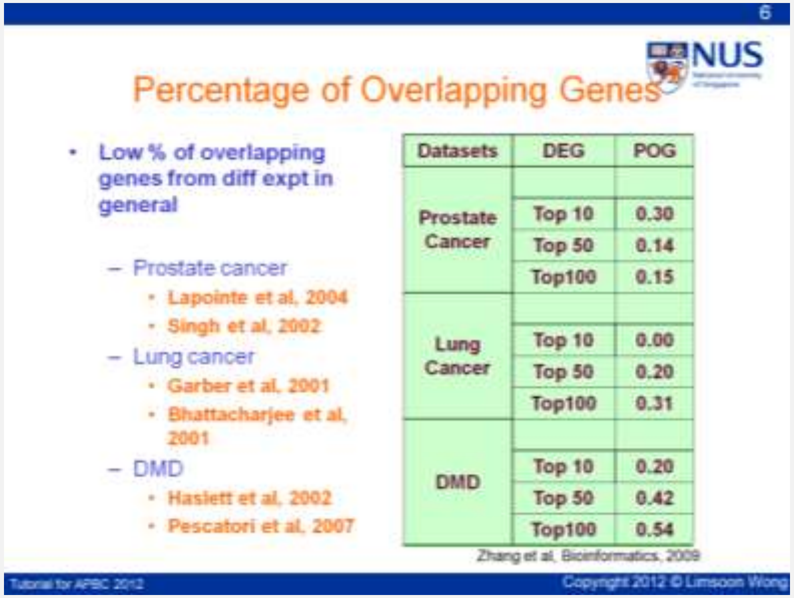
Genes are extensively upregulated in cancers. Normalizing them mislead them to be considered downregulated!



Wang et al. *Molecular Biosystems*, 8:818-827, 2012

Part 1: Delivering reproducible gene expression analysis

- Some issues in gene expression analysis
- Batch effect & normalization
- **Reproducibility**
  - Law of large numbers
  - Use background info
  - Find more consistent disease subnetworks



6

Percentage of Overlapping Genes

- Low % of overlapping genes from diff expt in general
  - Prostate cancer
    - Lapointe et al, 2004
    - Singh et al, 2002
  - Lung cancer
    - Garber et al, 2001
    - Bhattacharjee et al, 2001
  - DMD
    - Haslett et al, 2002
    - Pescatori et al, 2007

Datasets	DEG	POG
Prostate Cancer		
	Top 10	0.30
	Top 50	0.14
	Top100	0.15
Lung Cancer		
	Top 10	0.00
	Top 50	0.20
	Top100	0.31
DMD		
	Top 10	0.20
	Top 50	0.42
	Top100	0.54

Zhang et al, Bioinformatics, 2009

Tutorial for APBC 2012

Copyright 2012 © Limsoon Wong

Law of Large Numbers

- Suppose you are in a room with 365 other people
- Q: What is prob that a specific person in the room has the same birthday as you?
- A: 1/365 = 0.3%
- Q: What is prob that there is a person in the room having same birthday as you?
- A: $1 – (364/365)^{365} =$ 63%
- Q: What is prob that there are two persons in the room having same birthday?
- A: 100%

Individual Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples

- **How many genes on a microarray are expected to perfectly correlate to these samples?**

- **Prob(a gene is correlated) = $1/2^6$**
- **# of genes on array = 100,000**

⇒ E(# of correlated genes) = 1,562

⇒ Many false positives

- **These cannot be eliminated based on pure statistics!**

Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

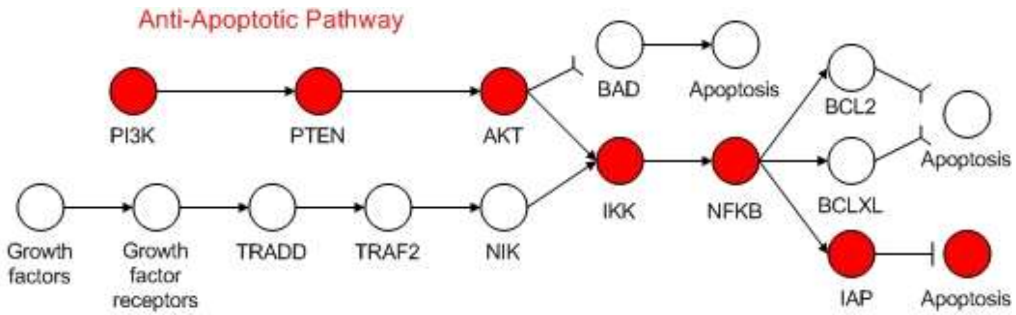
- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# of groups = $^{100000}C_5$**

⇒ **E(# of groups of genes correlated) = $^{100000}C_5$ * $(1/2^6)^5$ = $2.6*10^{12}$**

⇒ **Even more false positives?**

- **Perhaps no need to consider every group**

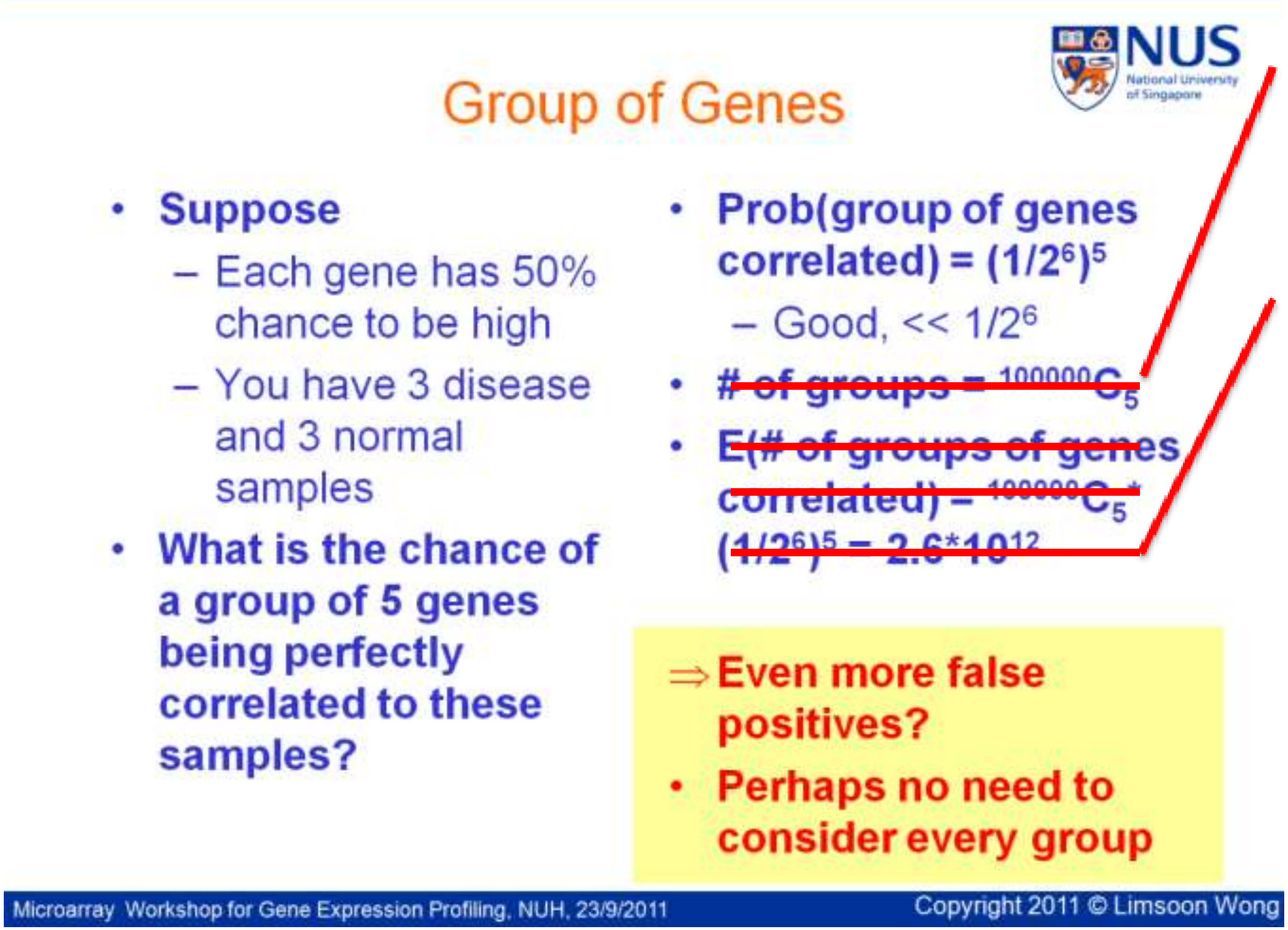
Regulatory Circuits – The Context



- **Each disease phenotype has some underlying cause**
- **There is some unifying biological theme for genes that are truly associated with a disease subtype**

- **Uncertainty in selected genes can be reduced by considering biological processes of the genes**
- **The unifying biological theme is basis for inferring the underlying cause of disease subtype**

Taming false positives by considering pathways instead of all possible groups



Group of Genes

- **Suppose**
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- **What is the chance of a group of 5 genes being perfectly correlated to these samples?**

- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- ~~**# of groups = ${}^{100000}C_5$**~~
- ~~**E(# of groups of genes correlated) = ${}^{100000}C_5$ * $(1/2^6)^5$ = $2.6*10^{12}$**~~

⇒ **Even more false positives?**

- **Perhaps no need to consider every group**

Microarray Workshop for Gene Expression Profiling, NUH, 23/9/2011

Copyright 2011 © Limsoon Wong

# of pathways = 1000

E(# of pathways correlated) = 1000 * $(1/2^6)^5$ = $9.3*10^{-7}$

Towards More Meaningful Genes

- **ORA**
  - Khatri et al
  - *Genomics*, 2002

Overlap Analysis

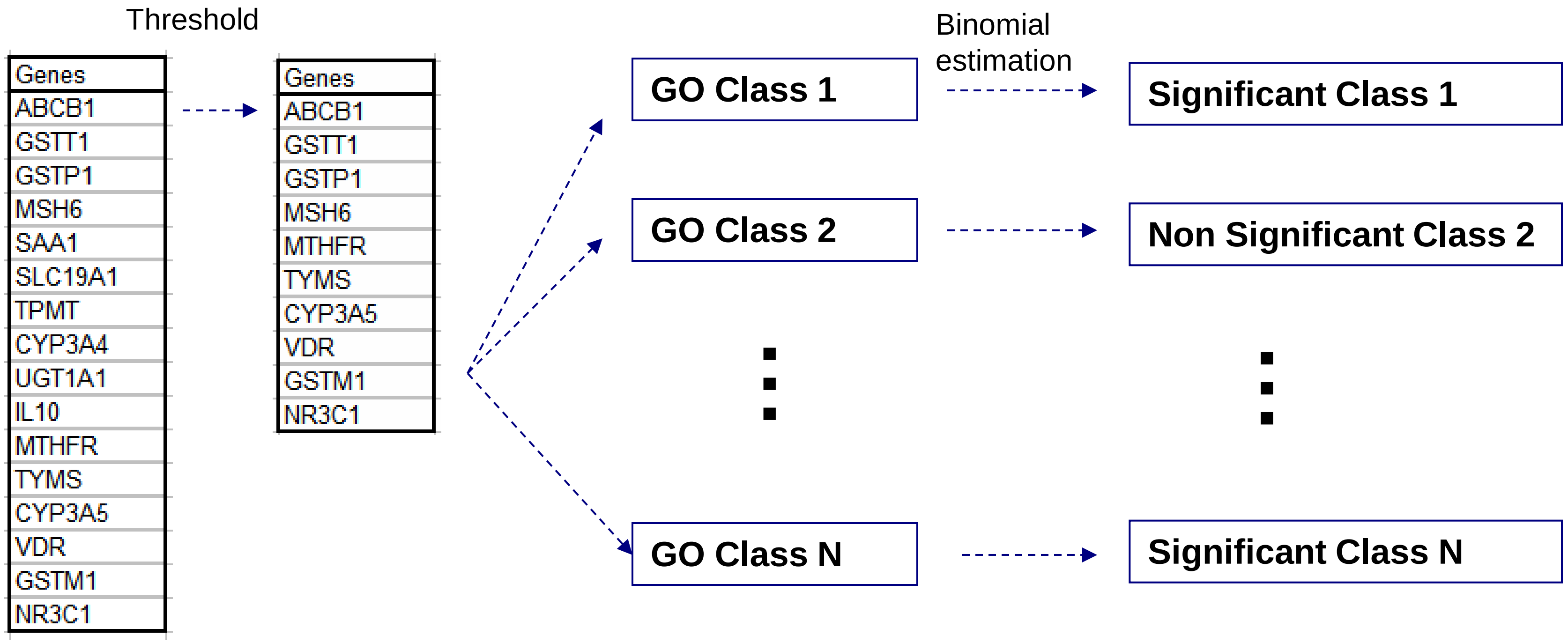
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005

Direct-Group Analysis

- **SNet**
  - Soh et al
  - *BMC Genomics*, 2011

Network-Based Analysis

Overlap Analysis: ORA



Threshold

Genes
ABCB1
GSTT1
GSTP1
MSH6
SAA1
SLC19A1
TPMT
CYP3A4
UGT1A1
IL10
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

Genes
ABCB1
GSTT1
GSTP1
MSH6
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

Binomial estimation

GO Class 1 → Significant Class 1

GO Class 2 → Non Significant Class 2

⋮

GO Class N → Significant Class N

S Draghici et al. “Global functional profiling of gene expression”. *Genomics*, 81(2):98-104, 2003.

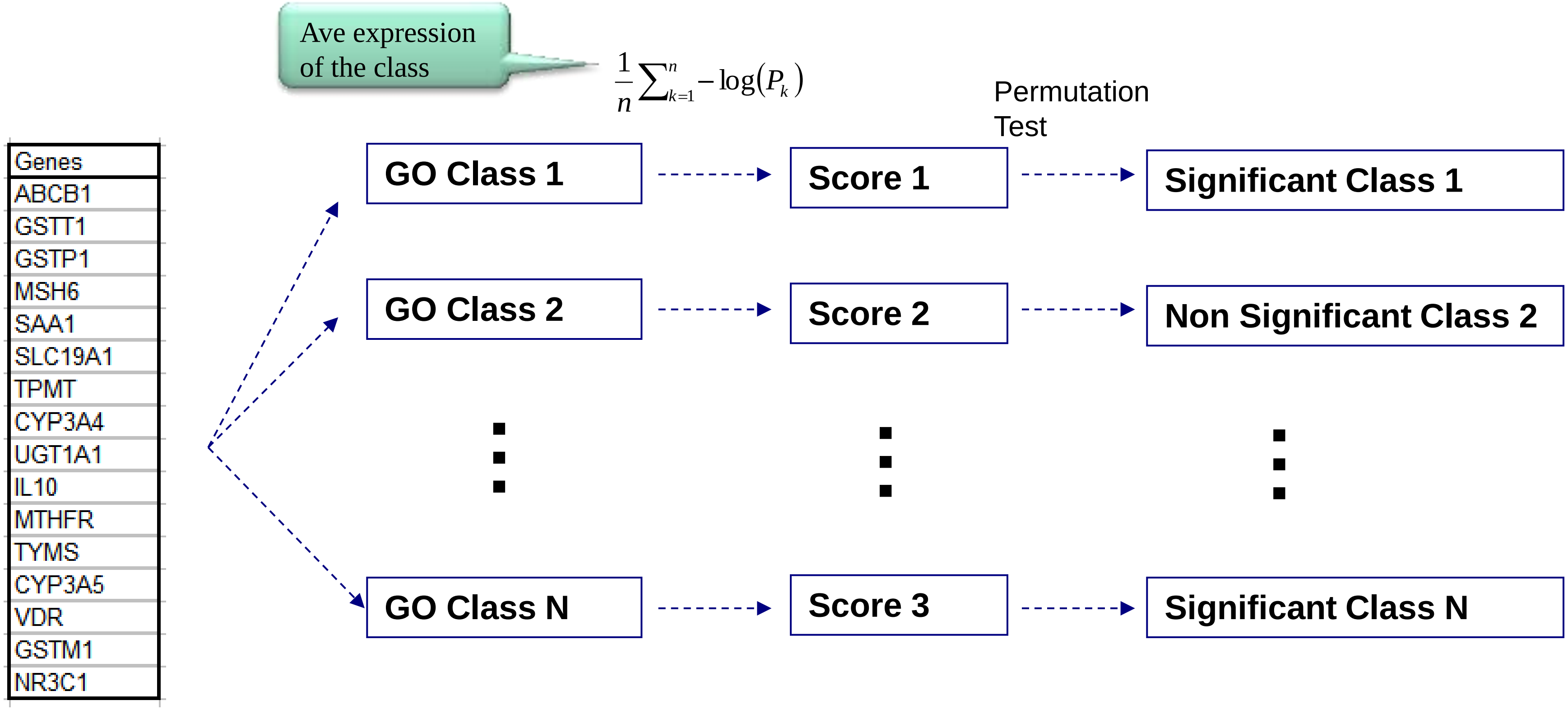
A problem w/ ORA

- **It is essentially testing whether A ∩ B is significant, where**
  - A = the set of differentially expressed genes
  - B = the set of gene in a specified pathway

- **The set of differentially expressed genes is defined by an arbitrary threshold on, e.g., fold change, t-statistic, …**

- **If you change that threshold, you can change A drastically. This has big impact on A ∩ B**

Direct-Group Analysis: FCS



Ave expression of the class

$$\frac{1}{n}\sum_{k=1}^{n} -\log(P_k)$$

Permutation Test

Genes
ABCB1
GSTT1
GSTP1
MSH6
SAA1
SLC19A1
TPMT
CYP3A4
UGT1A1
IL10
MTHFR
TYMS
CYP3A5
VDR
GSTM1
NR3C1

GO Class 1 → Score 1 → Significant Class 1

GO Class 2 → Score 2 → Non Significant Class 2

⋮

GO Class N → Score 3 → Significant Class N

P Pavlidis et al. “Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex”. *Neurochem Res*., 29(6):1213-1222, 2004.

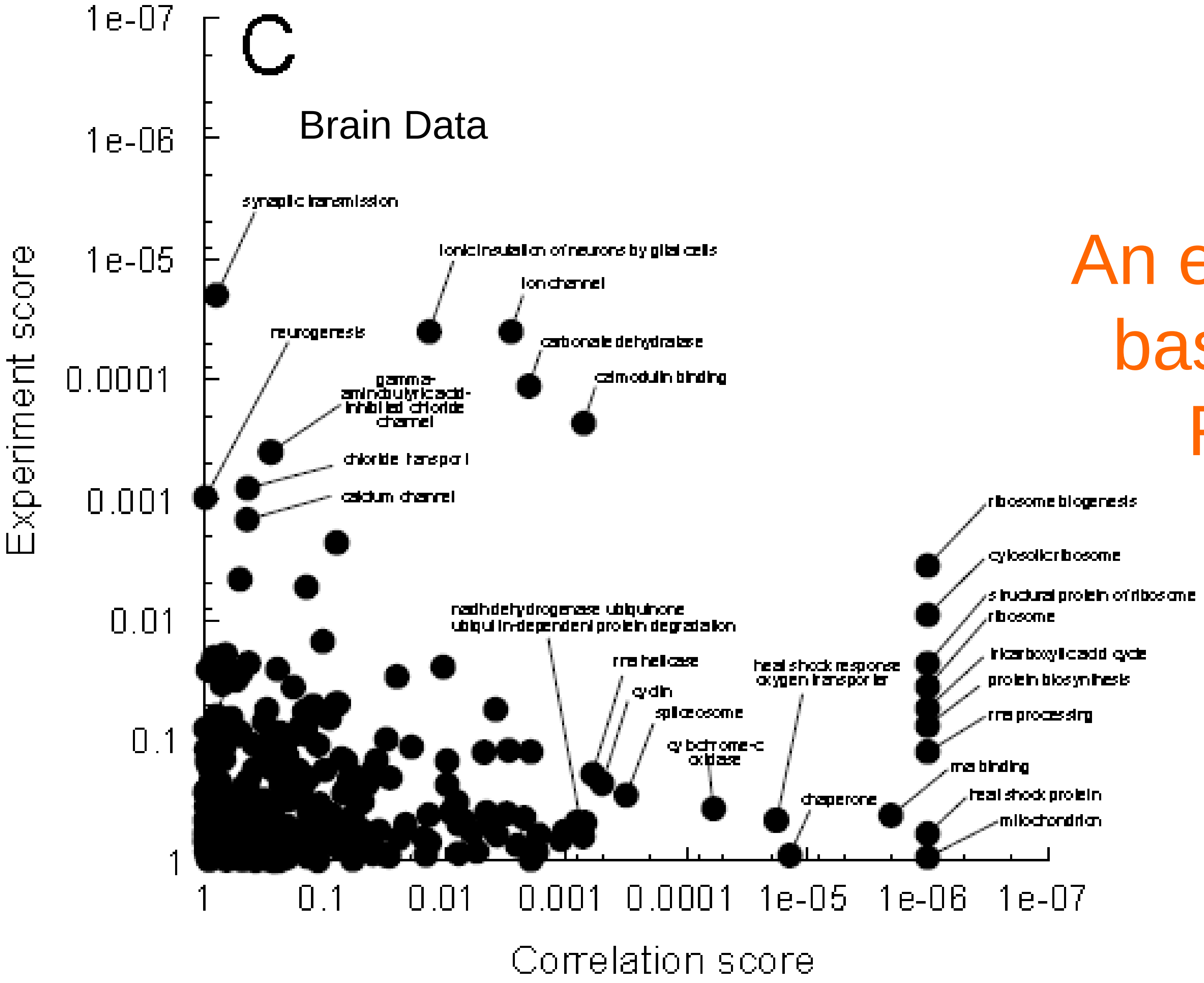
FCS: Key variations

- **"Correlation score"**
  - Score of a class C = average pair-wise correlation of genes in the class C

- **"Experimental score"**
  - Score of a class C = average of log-transformed p-values of genes in the class C

- **Null distribution to estimate the p-value of the scores above is by repeated sampling of random sets of genes of the same size as C**

Pavlidis et al., PSB 2002

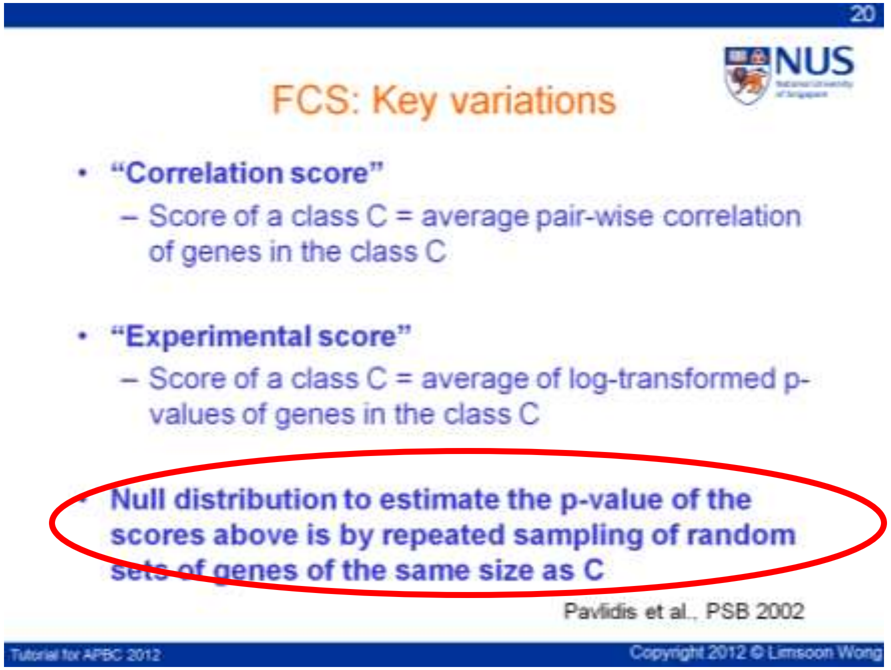


An example based on FCS

Pavlidis et al., PSB 2002

Goeman & Buhlmann. “Analyzing gene expression data in terms of gene sets: Methodological issues”. *Bioinformatics*, 23(8):980-987, 2007

A problem w/ FCS as proposed by Pavlidis et al in PSB 2002



20

FCS: Key variations

- **“Correlation score”**
  - Score of a class C = average pair-wise correlation of genes in the class C
- **“Experimental score”**
  - Score of a class C = average of log-transformed p-values of genes in the class C
- **Null distribution to estimate the p-value of the scores above is by repeated sampling of random sets of genes of the same size as C**

Pavlidis et al., PSB 2002

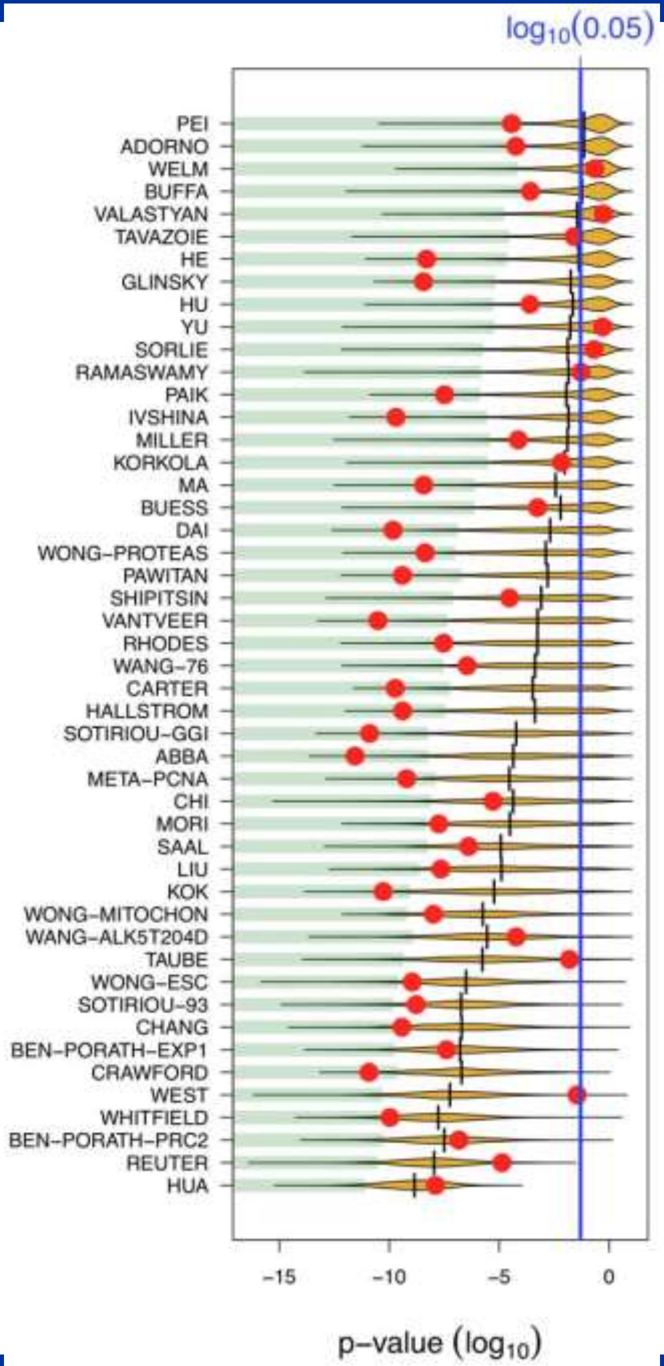
Tutorial for APBC 2012

Copyright 2012 © Limsoon Wong

- **Its null hypothesis:**
  - “genes in C are independently expressed & not diff from other genes
- **But …**
  - Genes in a pathway are not independent

  ⇒ Becomes over sensitive

- **Solution: generate null distribution by randomizing patient class labels**



FCS: Why do we estimate p-value using a null distribution based on repeated sampling of randomized gene sets / patient sets?

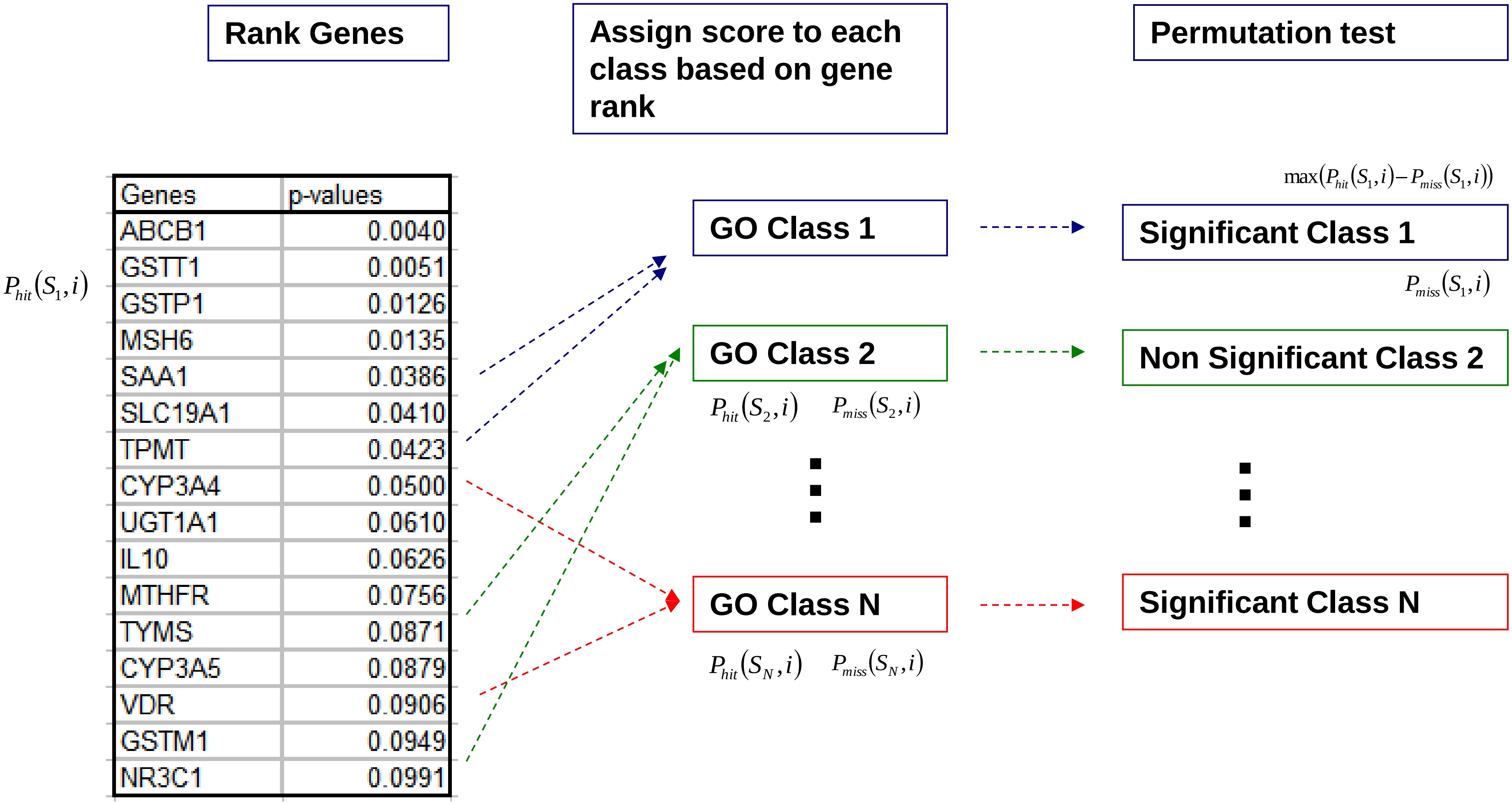
Venet et al. "Most random gene expression signatures are significantly associated with breast cancer outcome". *PLoS Computational Biology*, 7(10):e1002240, 2011.

An expt by a student on the nominal and empirical p-values for t-test

- **“I performed permutation test on the DMD dataset and obtained a null distribution. Then I computed two p-values (nominal and empirical) and took the genes at 5% threshold.**

- **Out of 8,867 genes, 2,091 were significant under nominal and 482 were significant under empirical. The significant genes had 0.13 overlap between two methods (309 intersect and 2265 union).”**

Direct-Group Analysis: GSEA



Rank Genes

Assign score to each class based on gene rank

Permutation test

Genes	p-values
ABCB1	0.0040
GSTT1	0.0051
GSTP1	0.0126
MSH6	0.0135
SAA1	0.0386
SLC19A1	0.0410
TPMT	0.0423
CYP3A4	0.0500
UGT1A1	0.0610
IL10	0.0626
MTHFR	0.0756
TYMS	0.0871
CYP3A5	0.0879
VDR	0.0906
GSTM1	0.0949
NR3C1	0.0991

$P_{hit}(S_1,i)$

GO Class 1

GO Class 2

$P_{hit}(S_2,i)$ $P_{miss}(S_2,i)$

⋮

GO Class N

$P_{hit}(S_N,i)$ $P_{miss}(S_N,i)$

$\max(P_{hit}(S_1,i)-P_{miss}(S_1,i))$

Significant Class 1

$P_{miss}(S_1,i)$

Non Significant Class 2

⋮

Significant Class N

A Subramanian et al. “Gene set enrichment analysis: A knowledge-based approach for interpreting genome wide expression profiles”. *PNAS*, 102(43):15545-15550, 2005

GSEA: Key Points

- **"Enrichment score"**
  - The degree that the genes in gene set C are enriched in the extremes of ranked list of all genes
  - Measured by Komogorov-Smirnov statistic

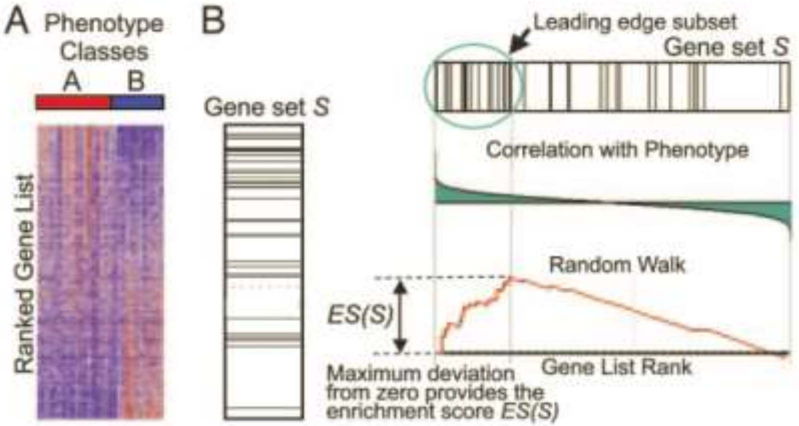


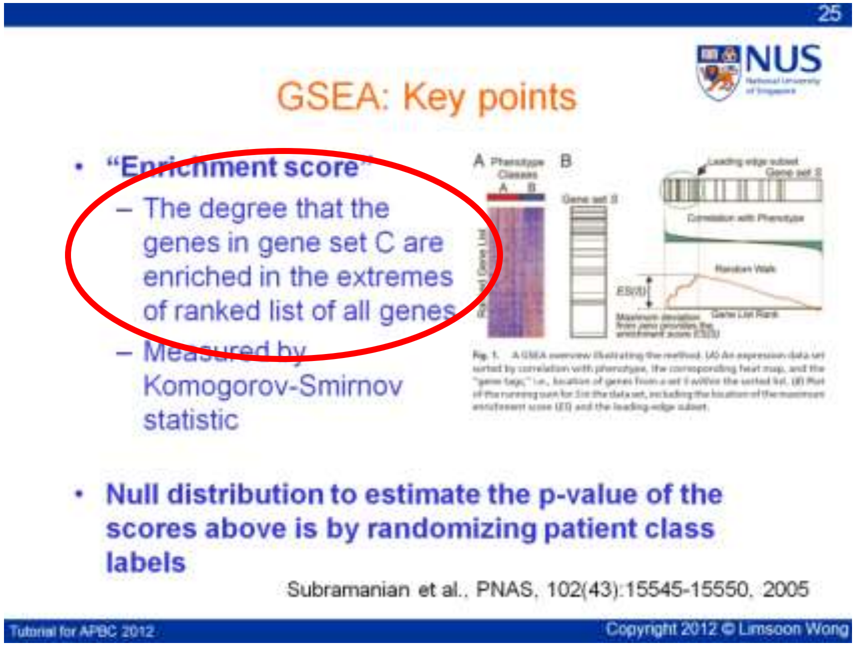
Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location of the maximum enrichment score (*ES*) and the leading-edge subset.

Subramanian et al., *PNAS*, 102(43):15545-15550, 2005

- **Null distribution to estimate the p-value of the scores above is by randomizing patient class labels**

Wong. "Using Biological Networks in Protein Function Prediction and Gene Expression Analysis". *Internet Mathematics*, 7(4):274--298, 2011.

A problem w/ GSEA



GSEA: Key points

- "Enrichment score"
  - The degree that the genes in gene set C are enriched in the extremes of ranked list of all genes
  - Measured by Komogorov-Smirnov statistic

Fig. 1. A GSEA overview illustrating the method. (A) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set S within the sorted list. (B) Plot of the running sum for S in the data set, including the location of the maximum enrichment score (ES) and the leading-edge subset.

- Null distribution to estimate the p-value of the scores above is by randomizing patient class labels

Subramanian et al., PNAS, 102(43):15545-15550, 2005

- **Its enrichment score considers all genes in C**
- **But …**
  - Not all branches of a large pathway have to "go wrong"

  ⇒ Cannot detect if only a small part of a pathway malfunctions
- **Solution: Break pathways into subnetworks**

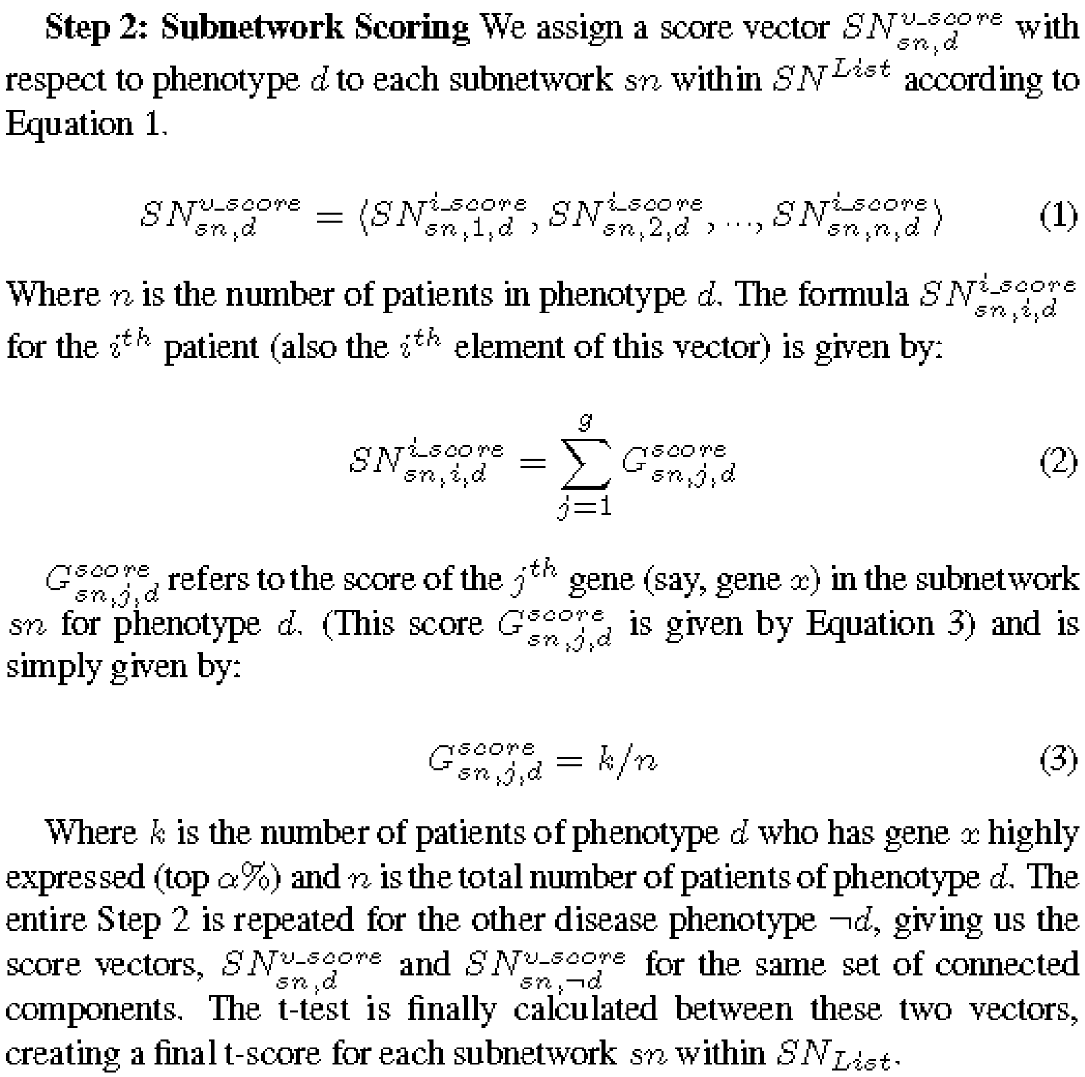
Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets". *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

Network-Based Analysis: SNet

- **Group samples into type D and ¬D**
- **Extract & score subnetworks for type D**
  - Get list of genes highly expressed in most D samples
    - **These genes need not be differentially expressed!**
  - Put these genes into pathways
  - Locate connected components (ie., candidate subnetworks) from these pathway graphs
  - Score subnetworks on D samples and on ¬D samples
- **For each subnetwork, compute t-statistic on the two sets of scores**
- **Determine significant subnetworks by permutations**

Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets".
***BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.**

SNet: Score Subnetworks



Step 2: Subnetwork Scoring We assign a score vector $SN_{sn,d}^{v\_score}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

$$SN_{sn,d}^{v\_score} = \langle SN_{sn,1,d}^{i\_score}, SN_{sn,2,d}^{i\_score}, \ldots, SN_{sn,n,d}^{i\_score} \rangle \qquad (1)$$

Where n is the number of patients in phenotype d. The formula $SN_{sn,i,d}^{i\_score}$ for the i^{th} patient (also the i^{th} element of this vector) is given by:

$$SN_{sn,i,d}^{i\_score} = \sum_{j=1}^{g} G_{sn,j,d}^{score} \qquad (2)$$

$G_{sn,j,d}^{score}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G_{sn,j,d}^{score}$ is given by Equation 3) and is simply given by:

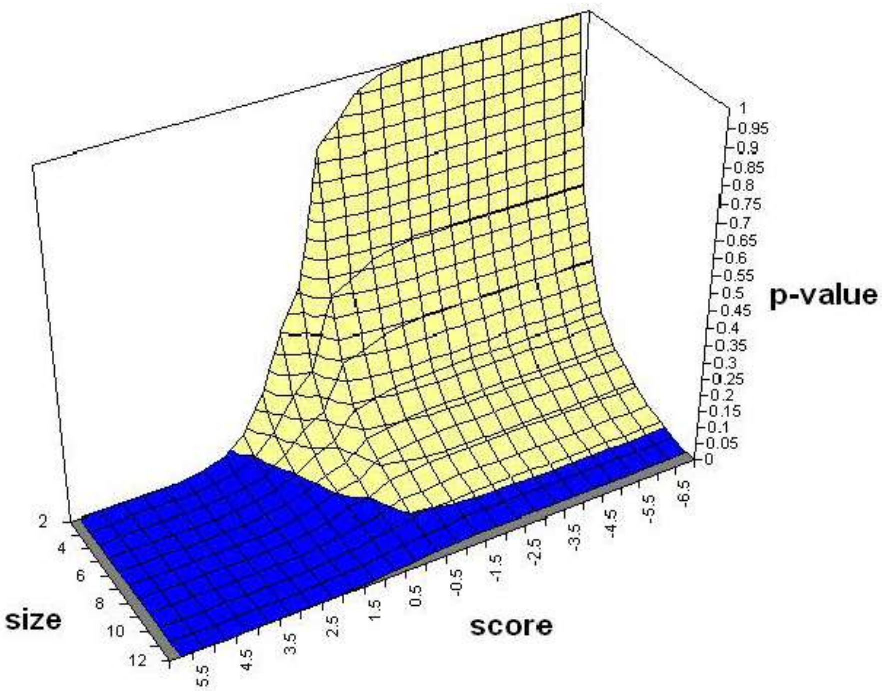
$$G_{sn,j,d}^{score} = k/n \qquad (3)$$

Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN_{sn,d}^{v\_score}$ and $SN_{sn,\neg d}^{v\_score}$ for the same set of connected components. The t-test is finally calculated between these two vectors, creating a final t-score for each subnetwork sn within SN_{List}.

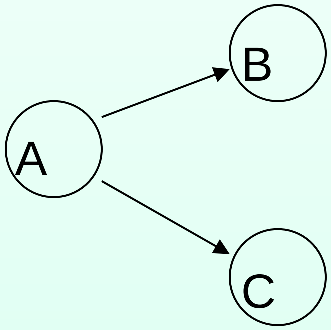
Soh et al. “Finding Consistent Disease Subnetworks Across Microarray Datasets”. *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

SNet: Significant Subnetworks

- **Randomize patient samples many times**
- **Get t-score for subnetworks from the randomizations**
- **Use these t-scores to establish null distribution**
- **Filter for significant subnetworks from real samples**



Key Insight # 1



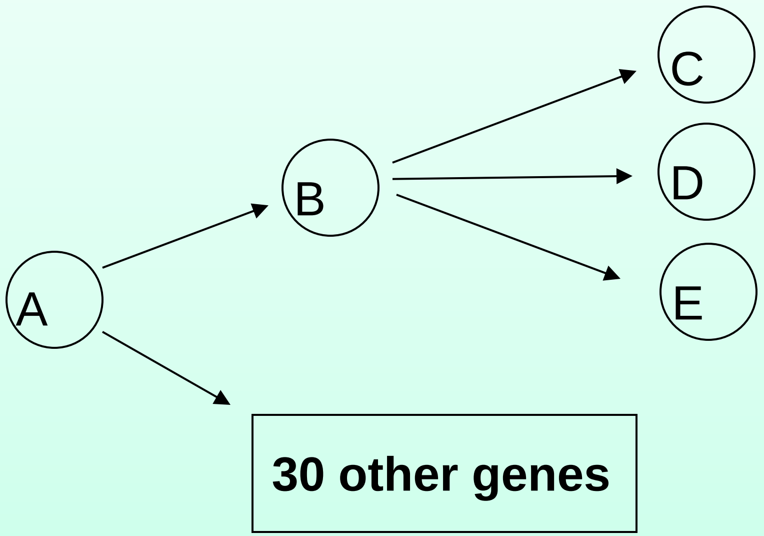
Genes A, B, C are high in phenotype *D*

A is high in phenotype ~*D* but B and C are not

Conventional techniques: Gene B and Gene C are selected. Possible incorrect postulation of mutations in gene B and C

- SNet does not require all the genes in subnet to be diff expressed
- It only requires the subnet as a whole to be diff expressed
- Able to capture entire relationship, postulating a mutation in gene A

Key Insight # 2



A

B

C

D

E

30 other genes

A branch within pathway consisting of genes A, B, C, D and E are high in phenotype *D*

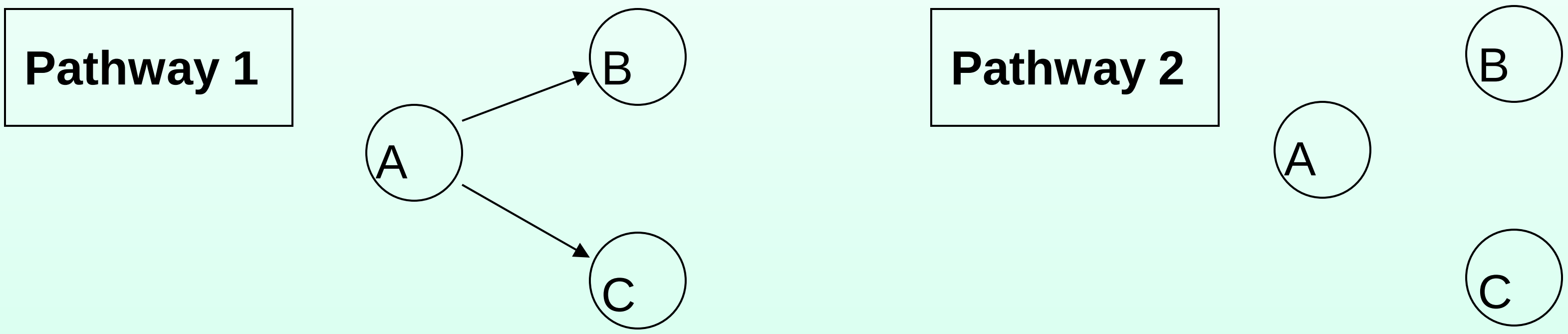
Genes C, D and E not high in phenotype ~*D*

30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

- **SNet: Able to capture the subnetwork branch within the pathway**

Key Insight # 3


Pathway 1

Pathway 2

Genes A, B and C are present in two separate pathways

A, B and C are high in phenotype *D*, but not high in phenotype *~D*

Conventional techniques:

Both pathways are scored equally. So both got selected, resulting in pathway 2 being a false positive

- **SNet: Able to select only pathway 1, which has the relevant relationship**

Let’s see whether SNet gives us subnetworks that are

(i) more consistent between datasets of the same types of disease samples

(ii) larger and more meaningful

Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets". ***BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.**

Better Subnetwork Overlap

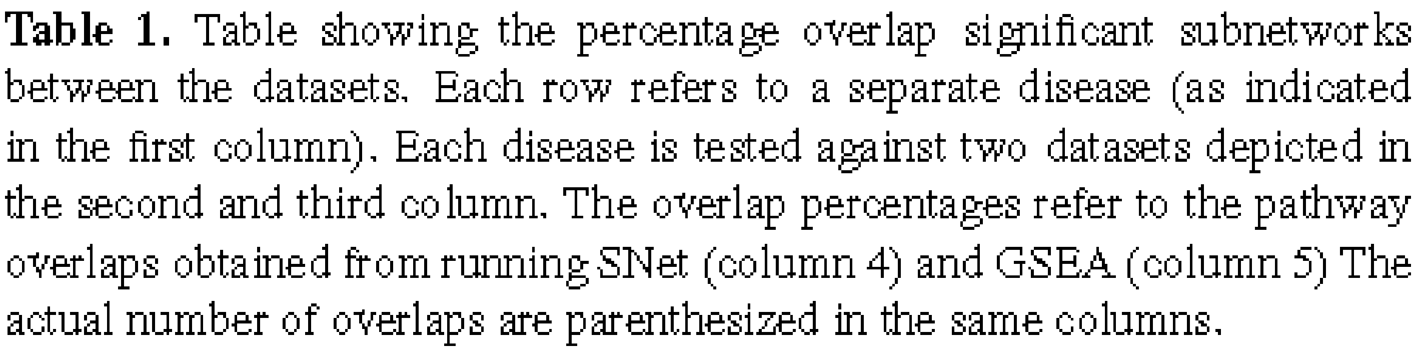
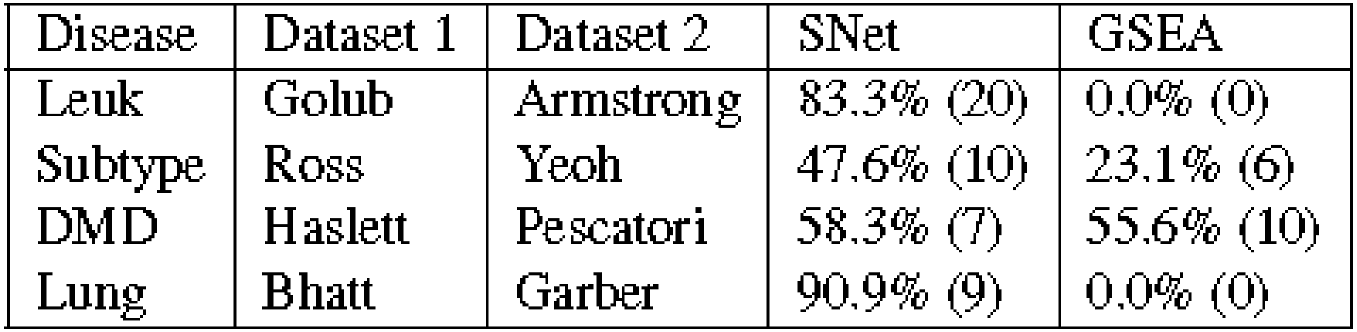


Table 1. Table showing the percentage overlap significant subnetworks between the datasets. Each row refers to a separate disease (as indicated in the first column). Each disease is tested against two datasets depicted in the second and third column. The overlap percentages refer to the pathway overlaps obtained from running SNet (column 4) and GSEA (column 5) The actual number of overlaps are parenthesized in the same columns.

Overlap = $|A \cap B| / \min(|A|,|B|)$



Disease	Dataset 1	Dataset 2	SNet	GSEA
Leuk	Golub	Armstrong	83.3% (20)	0.0% (0)
Subtype	Ross	Yeoh	47.6% (10)	23.1% (6)
DMD	Haslett	Pescatori	58.3% (7)	55.6% (10)
Lung	Bhatt	Garber	90.9% (9)	0.0% (0)

- **For each disease, take significant subnetworks from one dataset and see if it is also significant in the other dataset**

Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets".
***BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.**

Better Gene Overlaps

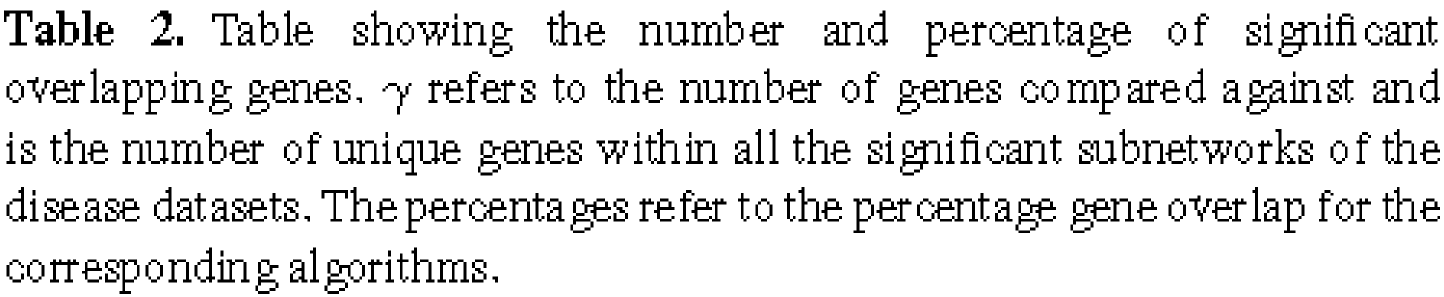
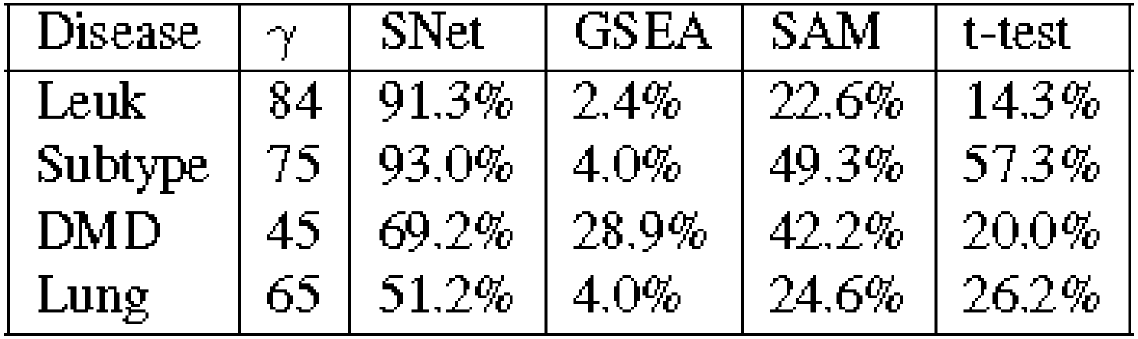


Table 2. Table showing the number and percentage of significant overlapping genes. γ refers to the number of genes compared against and is the number of unique genes within all the significant subnetworks of the disease datasets. The percentages refer to the percentage gene overlap for the corresponding algorithms.

$\text{Overlap} = |A \cap B| / \min(|A|,|B|)$



Disease	γ	SNet	GSEA	SAM	t-test
Leuk	84	91.3%	2.4%	22.6%	14.3%
Subtype	75	93.0%	4.0%	49.3%	57.3%
DMD	45	69.2%	28.9%	42.2%	20.0%
Lung	65	51.2%	4.0%	24.6%	26.2%

- **For each disease, take significant subnetworks extracted independently from both datasets and see how much their genes overlap**

Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets".
***BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.**

Larger Subnetworks

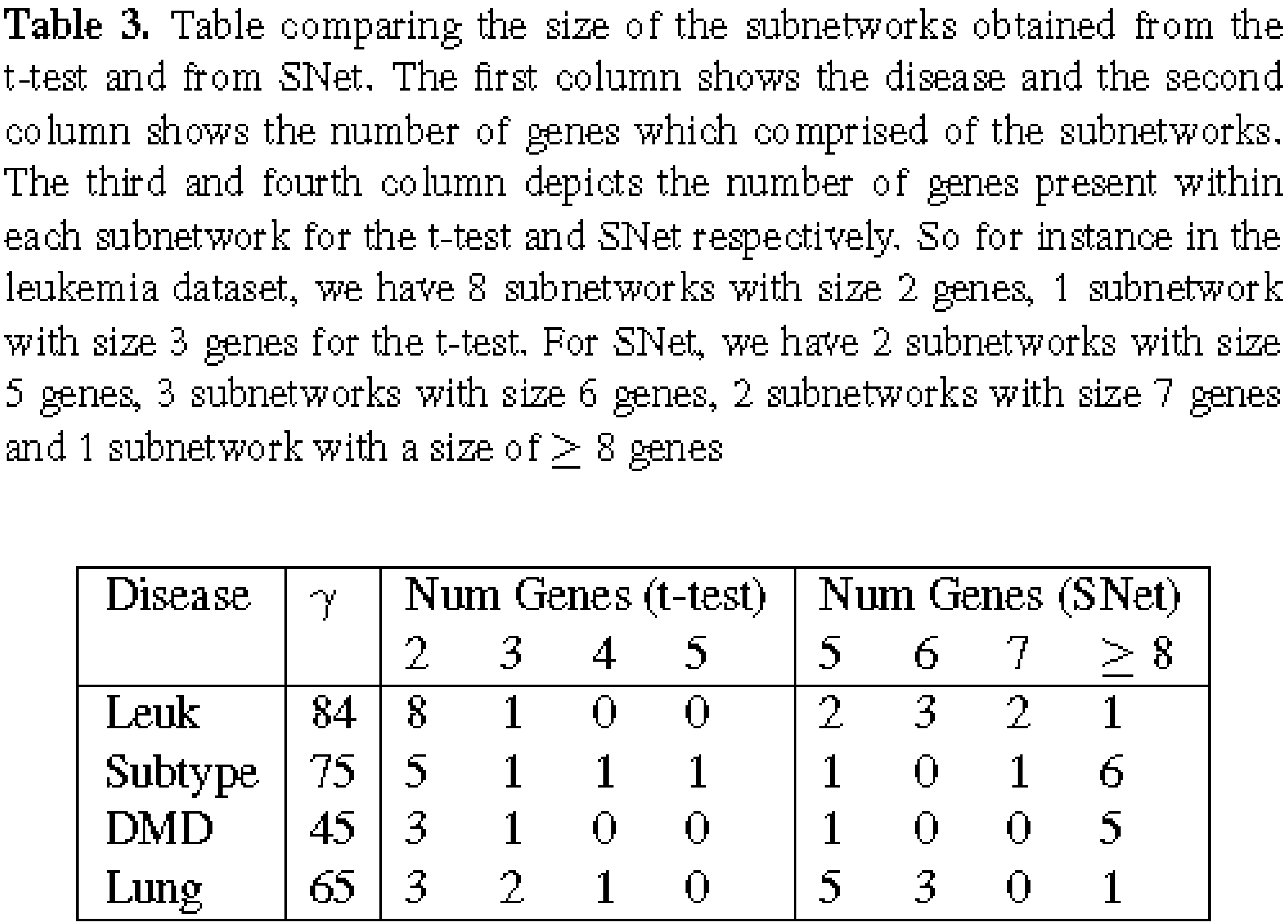


Table 3. Table comparing the size of the subnetworks obtained from the t-test and from SNet. The first column shows the disease and the second column shows the number of genes which comprised of the subnetworks. The third and fourth column depicts the number of genes present within each subnetwork for the t-test and SNet respectively. So for instance in the leukemia dataset, we have 8 subnetworks with size 2 genes, 1 subnetwork with size 3 genes for the t-test. For SNet, we have 2 subnetworks with size 5 genes, 3 subnetworks with size 6 genes, 2 subnetworks with size 7 genes and 1 subnetwork with a size of $\geq$ 8 genes

Disease	γ	Num Genes (t-test)				Num Genes (SNet)			
		2	3	4	5	5	6	7	$\geq$ 8
Leuk	84	8	1	0	0	2	3	2	1
Subtype	75	5	1	1	1	1	0	1	6
DMD	45	3	1	0	0	1	0	0	5
Lung	65	3	2	1	0	5	3	0	1

Issue #1 with SNet

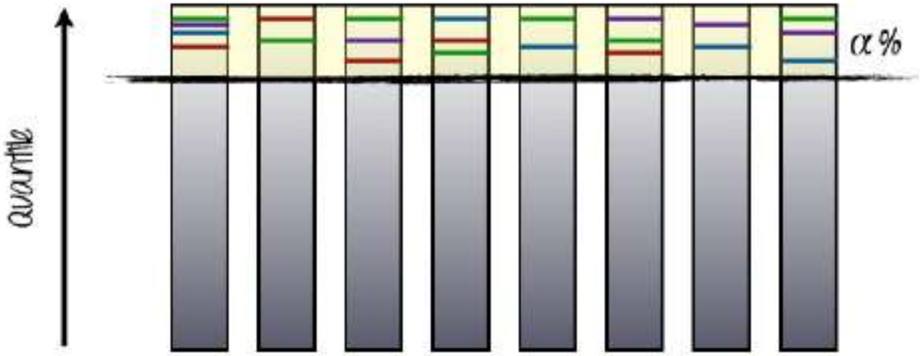
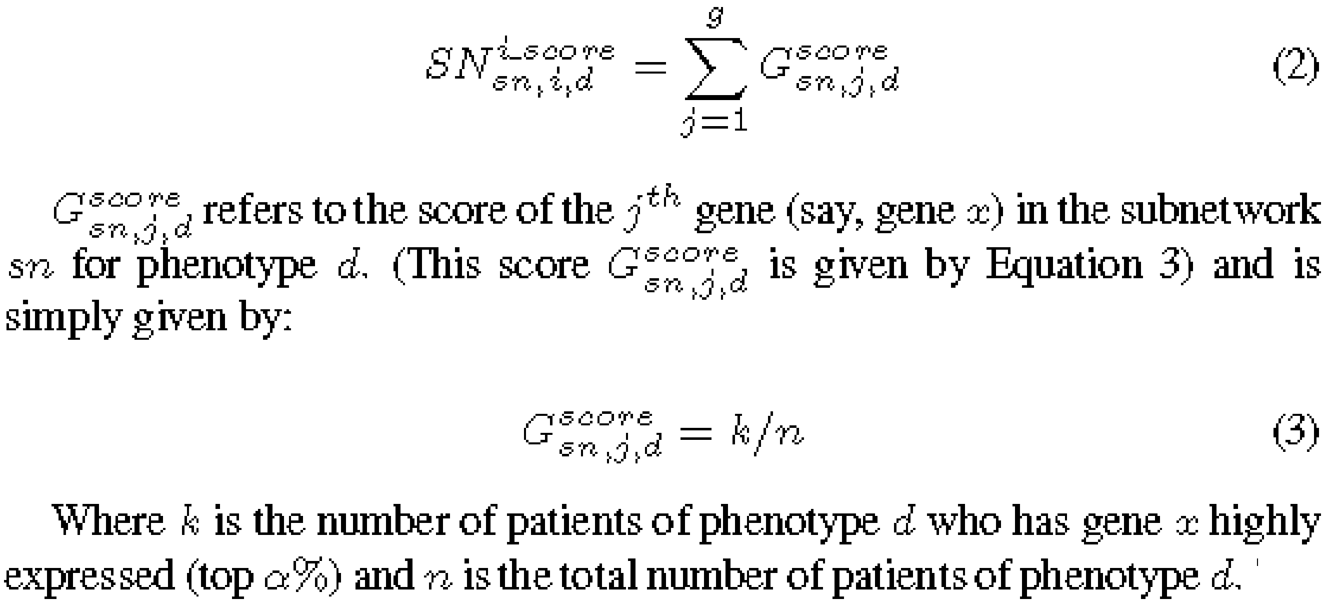


Fig. 2. In SNet, the top $\alpha\%$ of genes of each sample in phenotype D is highlighted in yellow. A subset of these genes that are thus highlighted in at least 50% of the samples are then taken to induce subnetworks.

- What if the real important genes are close to, but not in, the top α% most highly expressed genes?

- Blindly increasing α does not help, as this will bring in lots of false-positive genes

Issue #2 with SNet



$$SN^{i\_score}_{sn,i,d} = \sum_{j=1}^{g} G^{score}_{sn,j,d} \qquad (2)$$

$G^{score}_{sn,j,d}$ refers to the score of the j^{th} gene (say, gene x) in the subnetwork sn for phenotype d. (This score $G^{score}_{sn,j,d}$ is given by Equation 3) and is simply given by:

$$G^{score}_{sn,j,d} = k/n \qquad (3)$$

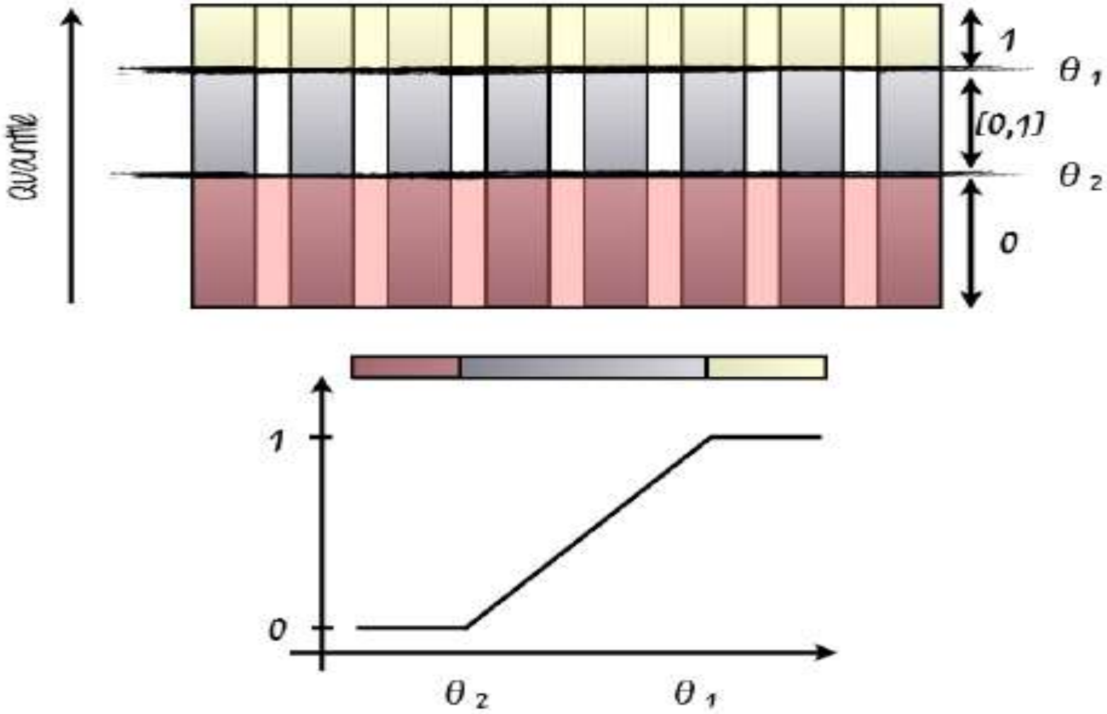
Where k is the number of patients of phenotype d who has gene x highly expressed (top α%) and n is the total number of patients of phenotype d. '

- **SNet weighs genes & scores subnetworks only on the basis of phenotype D**
- **Why not consider phenotype ~D as well?**

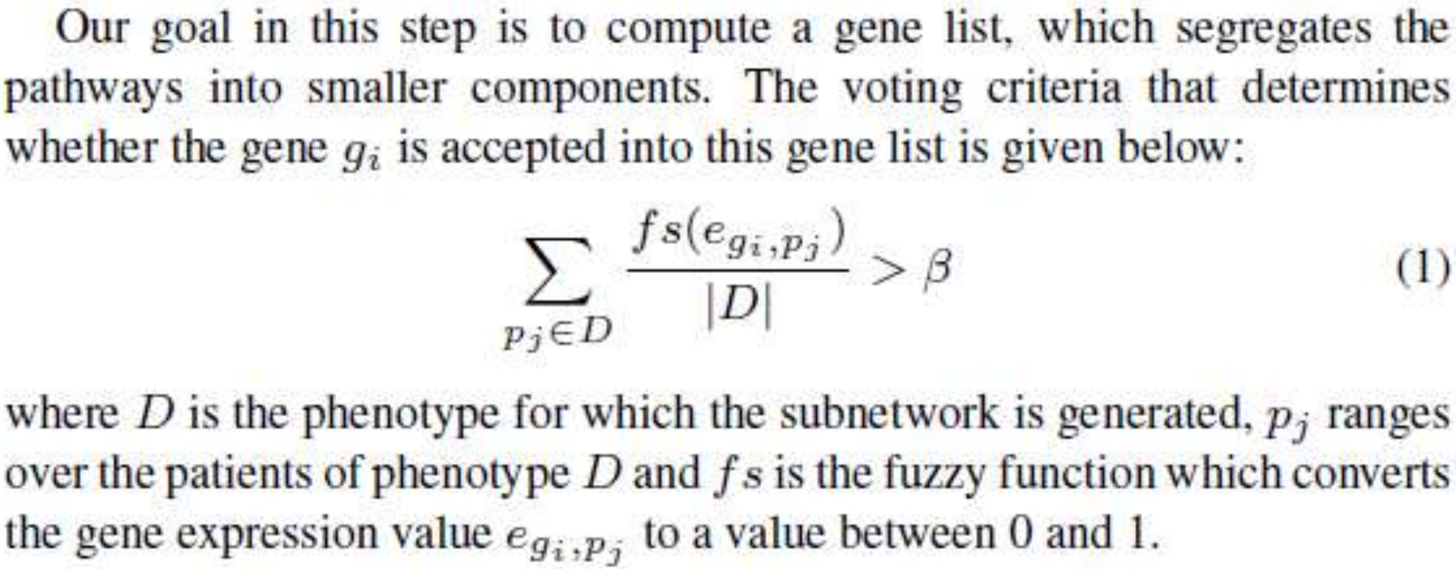
PFSNet

- **Deal with issue #1 of SNet using “fuzzification”**
- **Deal with issue #2 of SNet using paired t-test**

⇒ **PFSNet – Paired Fuzzy SNet**



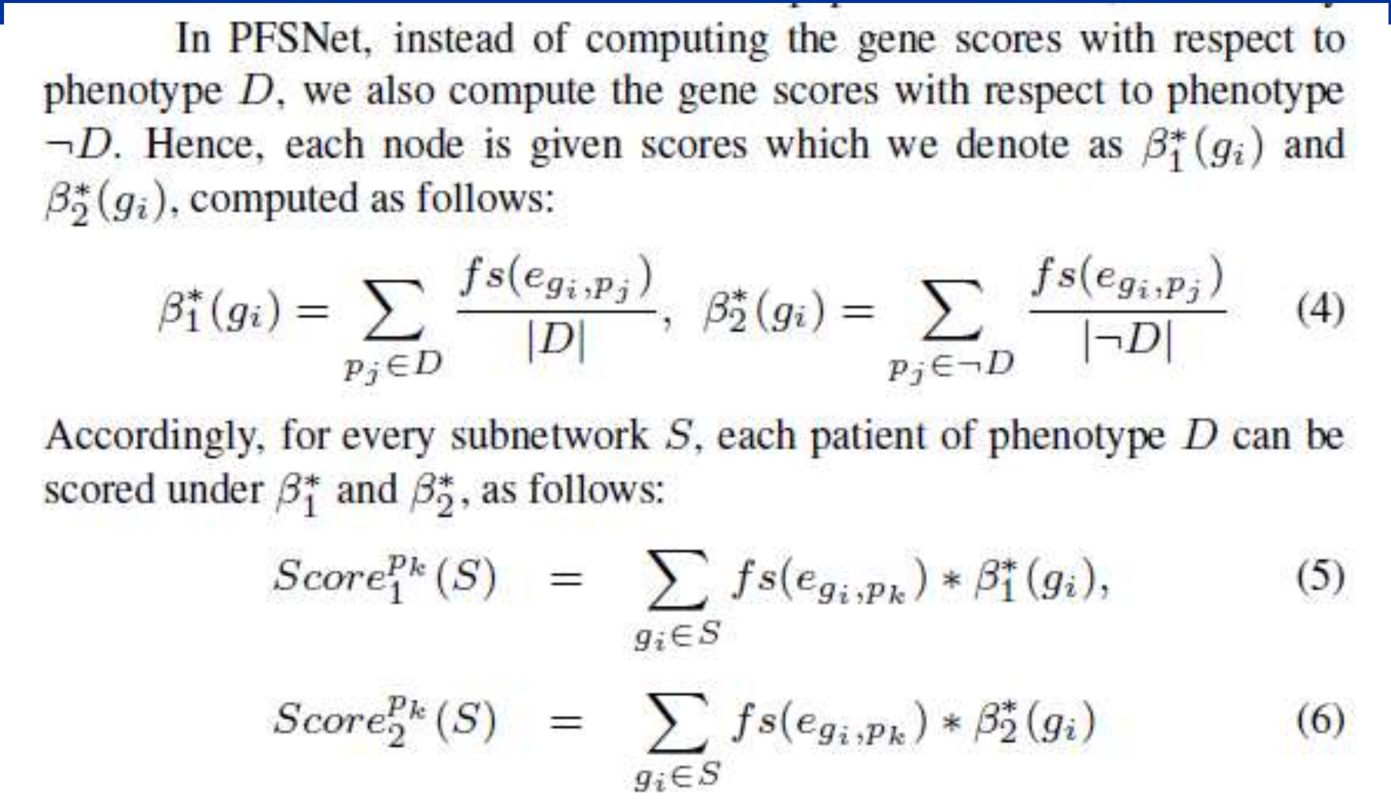
Fuzzification



Our goal in this step is to compute a gene list, which segregates the pathways into smaller components. The voting criteria that determines whether the gene g_i is accepted into this gene list is given below:

$$\sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|} > \beta \qquad (1)$$

where D is the phenotype for which the subnetwork is generated, p_j ranges over the patients of phenotype D and fs is the fuzzy function which converts the gene expression value e_{g_i,p_j} to a value between 0 and 1.



In PFSNet, instead of computing the gene scores with respect to phenotype D, we also compute the gene scores with respect to phenotype $\neg D$. Hence, each node is given scores which we denote as $\beta_1^*(g_i)$ and $\beta_2^*(g_i)$, computed as follows:

$$\beta_1^*(g_i) = \sum_{p_j \in D} \frac{fs(e_{g_i,p_j})}{|D|}, \quad \beta_2^*(g_i) = \sum_{p_j \in \neg D} \frac{fs(e_{g_i,p_j})}{|\neg D|} \qquad (4)$$

Accordingly, for every subnetwork S, each patient of phenotype D can be scored under β_1^* and β_2^*, as follows:

$$Score_1^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_1^*(g_i), \qquad (5)$$

$$Score_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i) \qquad (6)$$

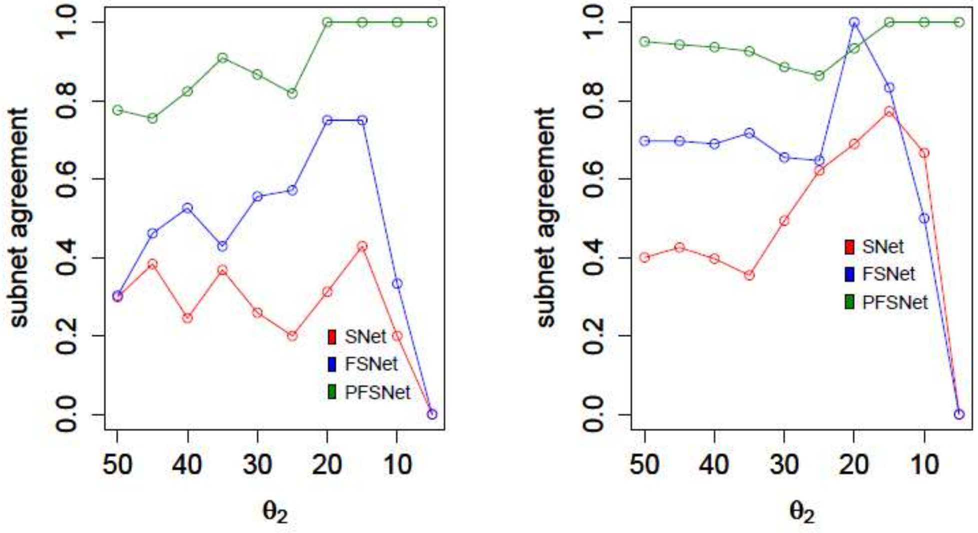
Paired T-Test

- **Score^{Pk}_1(S) and Score^{Pk}_2(S) are computed for the same sample Pk and subnetwork S**

⇒ **Can do paired t-test**

  – Null hypothesis: If S is irrelevant to D vs ~D, we expect Score^{Pk}_1(S) – Score^{Pk}_2(S) to be around 0

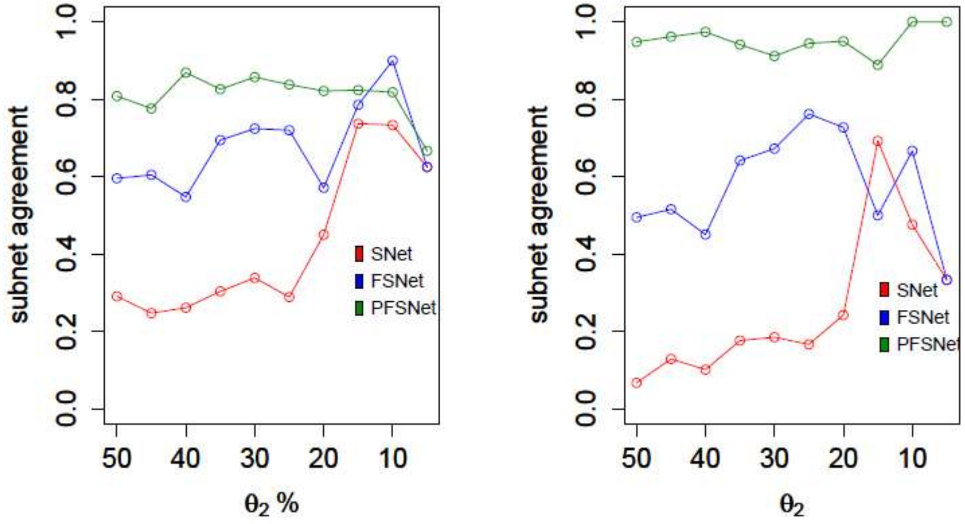
PSFNet vs SNet: Subnet Agreement



upregulated in ALL

upregulated in AML

Fig. 4: Consistency of subnetworks in Leukemia dataset



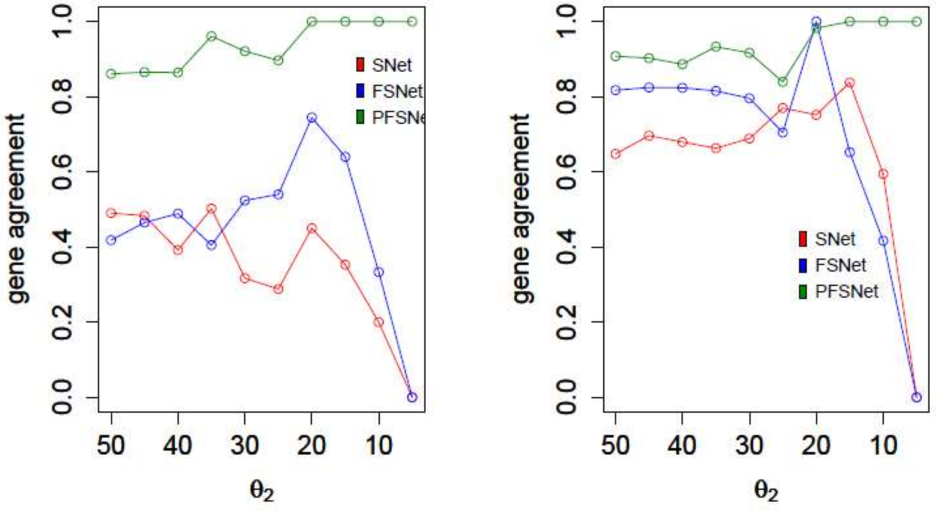
upregulated in DMD

upregulated in NORM

Fig. 6: Consistency of subnetworks in DMD dataset

Overlap = $|A \cap B| / |A \cup B|$

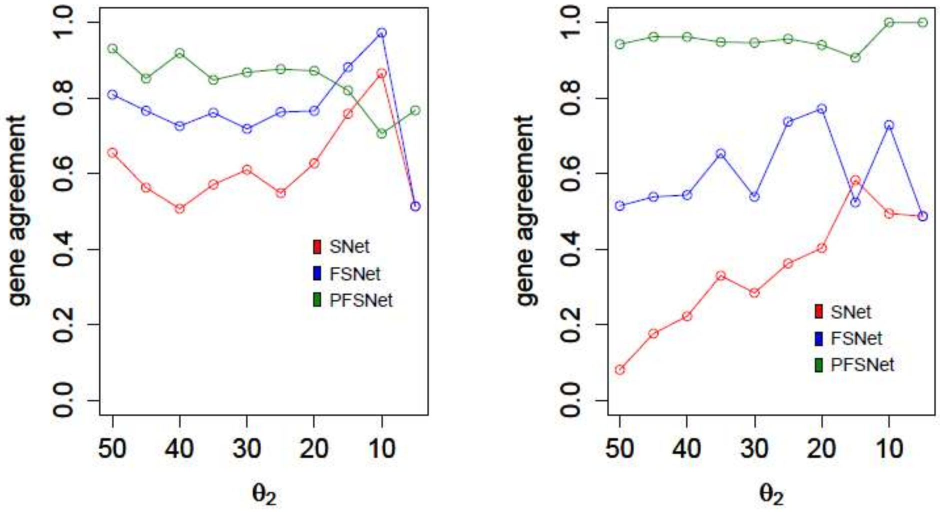
PSFNet vs SNet: Gene Agreement



upregulated in ALL

upregulated in AML

Fig. 7: Consistency of genes in Leukemia dataset



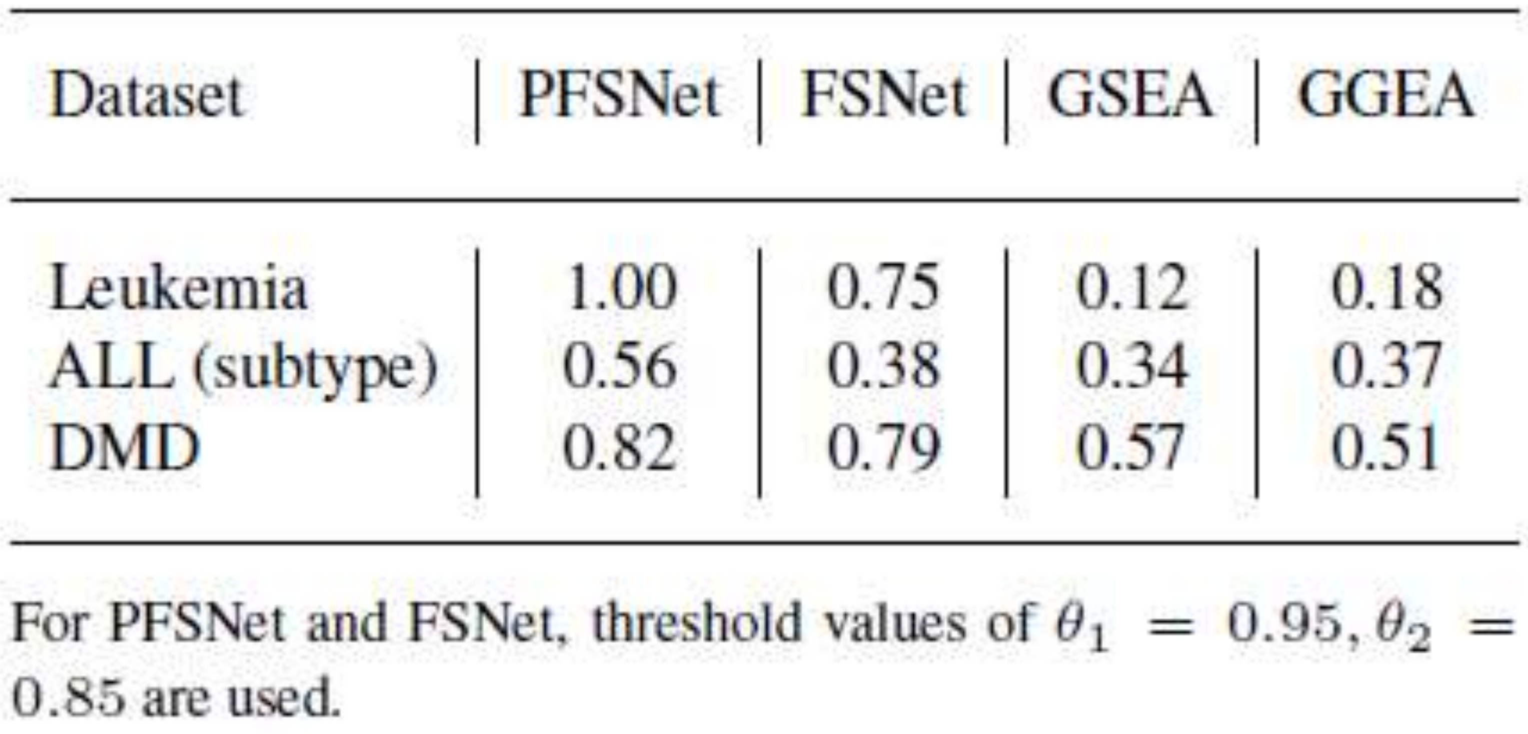
upregulated in DMD

upregulated in NORM

Fig. 9: Consistency of genes in DMD dataset

Overlap = |A∩B| / |A∪B|

PFSNet vs GSEA & GGEA: Pathway Agreement

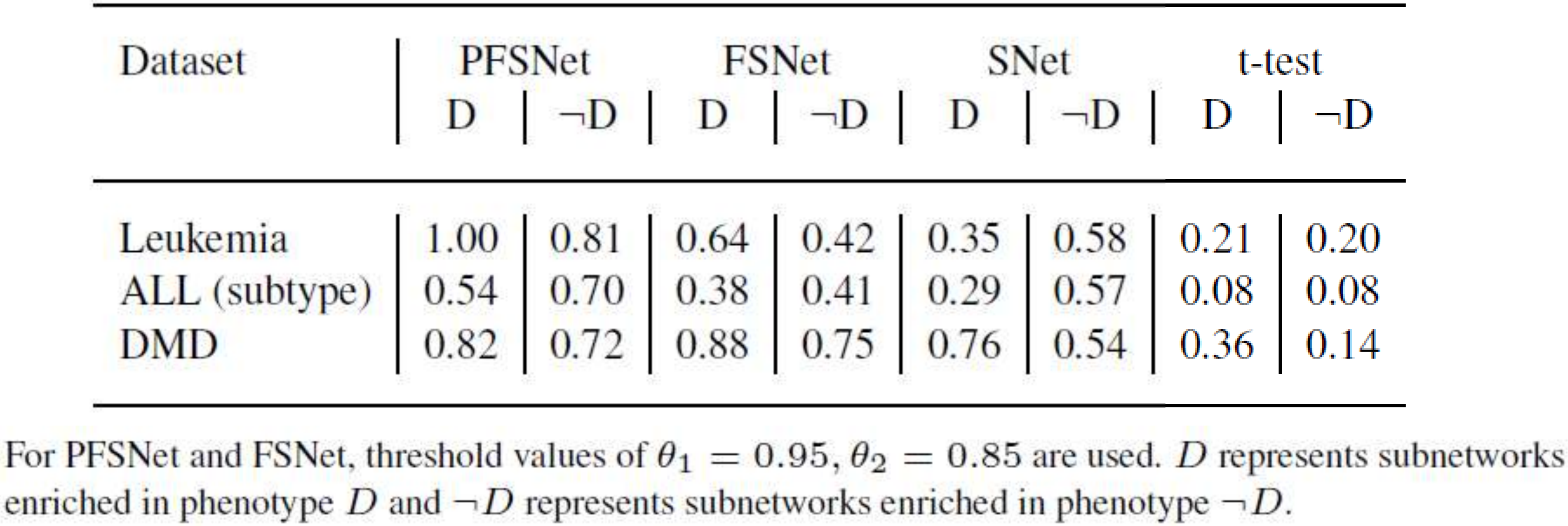


Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

For PFSNet and FSNet, threshold values of $\theta_1 = 0.95$, $\theta_2 = 0.85$ are used.

Overlap = |A∩B| / |A∪B|

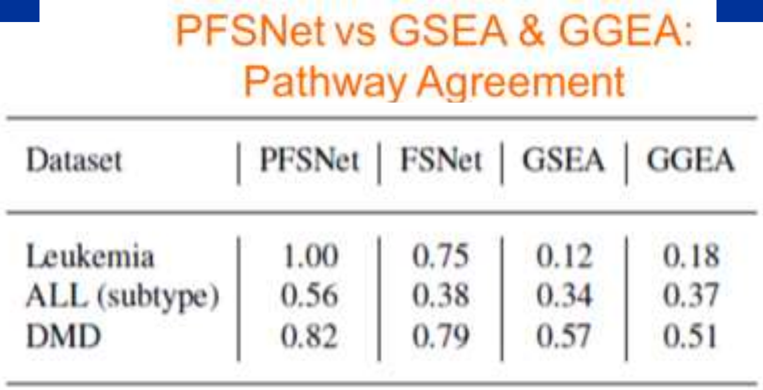
PFSNet vs T-Test: Gene Agreement



Dataset	PFSNet		FSNet		SNet		t-test	
	D	¬D	D	¬D	D	¬D	D	¬D
Leukemia	1.00	0.81	0.64	0.42	0.35	0.58	0.21	0.20
ALL (subtype)	0.54	0.70	0.38	0.41	0.29	0.57	0.08	0.08
DMD	0.82	0.72	0.88	0.75	0.76	0.54	0.36	0.14

For PFSNet and FSNet, threshold values of $\theta_1 = 0.95$, $\theta_2 = 0.85$ are used. D represents subnetworks enriched in phenotype D and $\neg D$ represents subnetworks enriched in phenotype $\neg D$.

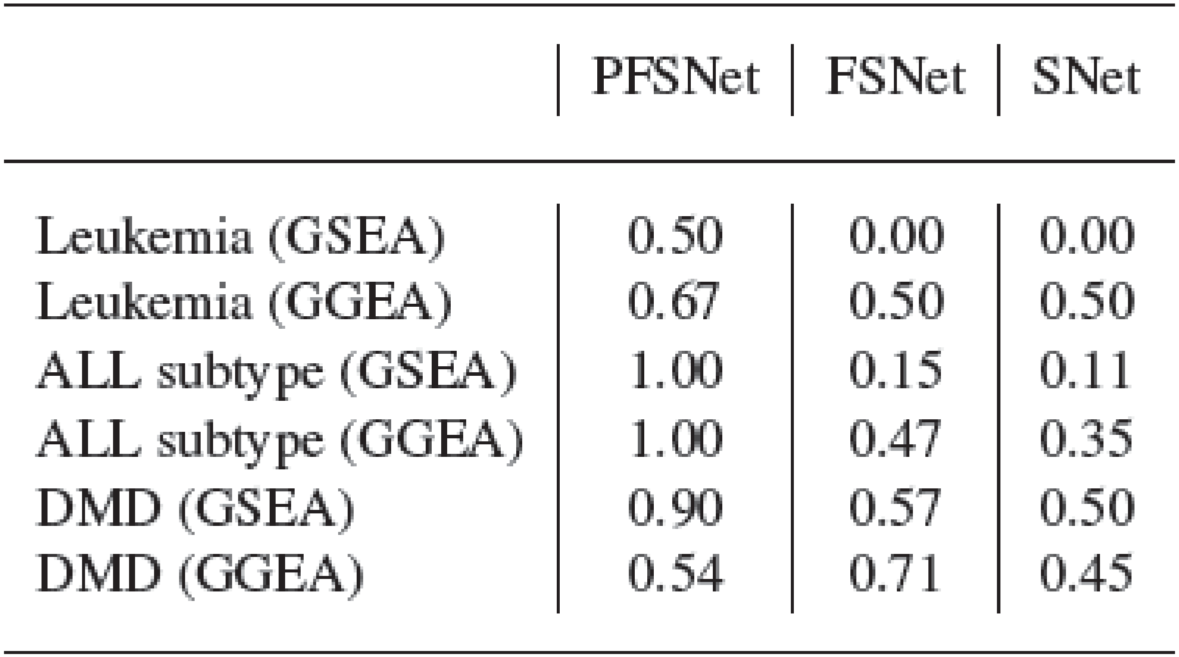
Overlap = |A∩B| / |A∪B|



PFSNet vs GSEA & GGEA: Pathway Agreement

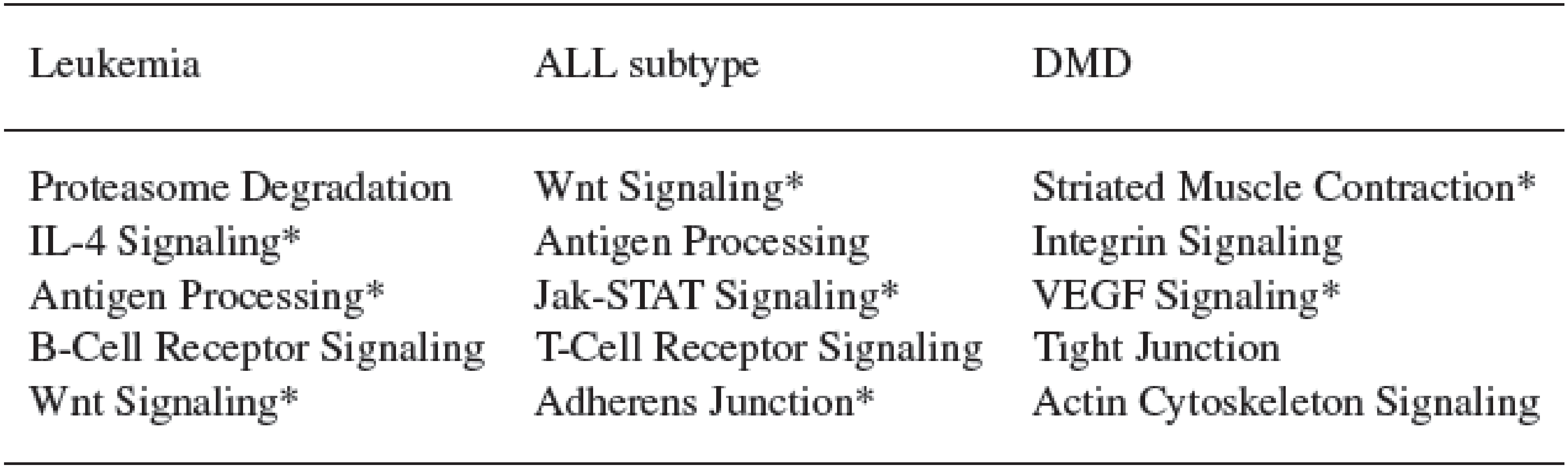
Dataset	PFSNet	FSNet	GSEA	GGEA
Leukemia	1.00	0.75	0.12	0.18
ALL (subtype)	0.56	0.38	0.34	0.37
DMD	0.82	0.79	0.57	0.51

Testing subnets from PFSNet using GSEA & GGEA



	PFSNet	FSNet	SNet
Leukemia (GSEA)	0.50	0.00	0.00
Leukemia (GGEA)	0.67	0.50	0.50
ALL subtype (GSEA)	1.00	0.15	0.11
ALL subtype (GGEA)	1.00	0.47	0.35
DMD (GSEA)	0.90	0.57	0.50
DMD (GGEA)	0.54	0.71	0.45

Top 5 Subnets



Leukemia	ALL subtype	DMD
Proteasome Degradation	Wnt Signaling*	Striated Muscle Contraction*
IL-4 Signaling*	Antigen Processing	Integrin Signaling
Antigen Processing*	Jak-STAT Signaling*	VEGF Signaling*
B-Cell Receptor Signaling	T-Cell Receptor Signaling	Tight Junction
Wnt Signaling*	Adherens Junction*	Actin Cytoskeleton Signaling

The asterisk indicates subnetworks that were not found in SNet

DMD: Striated Muscle Contraction

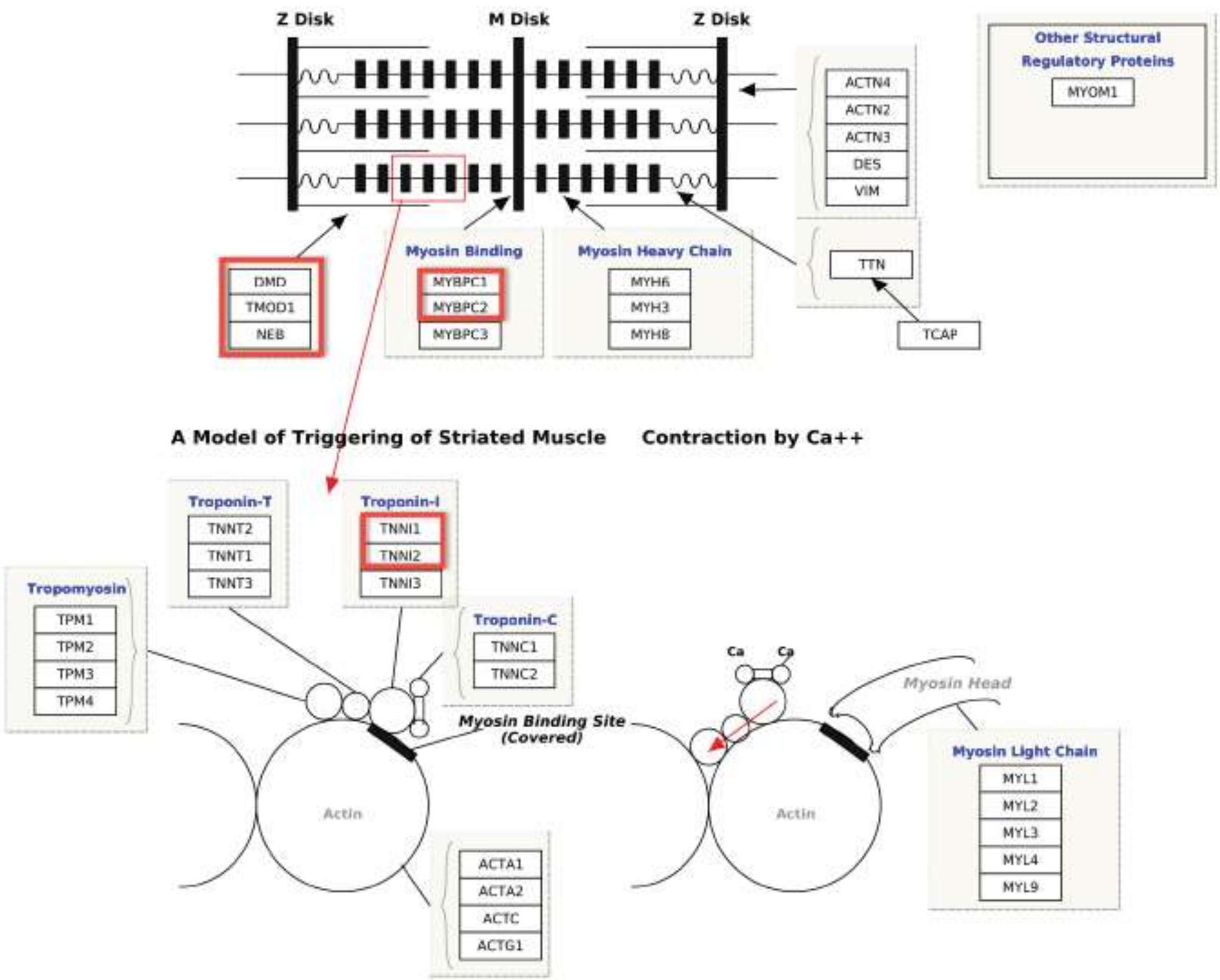
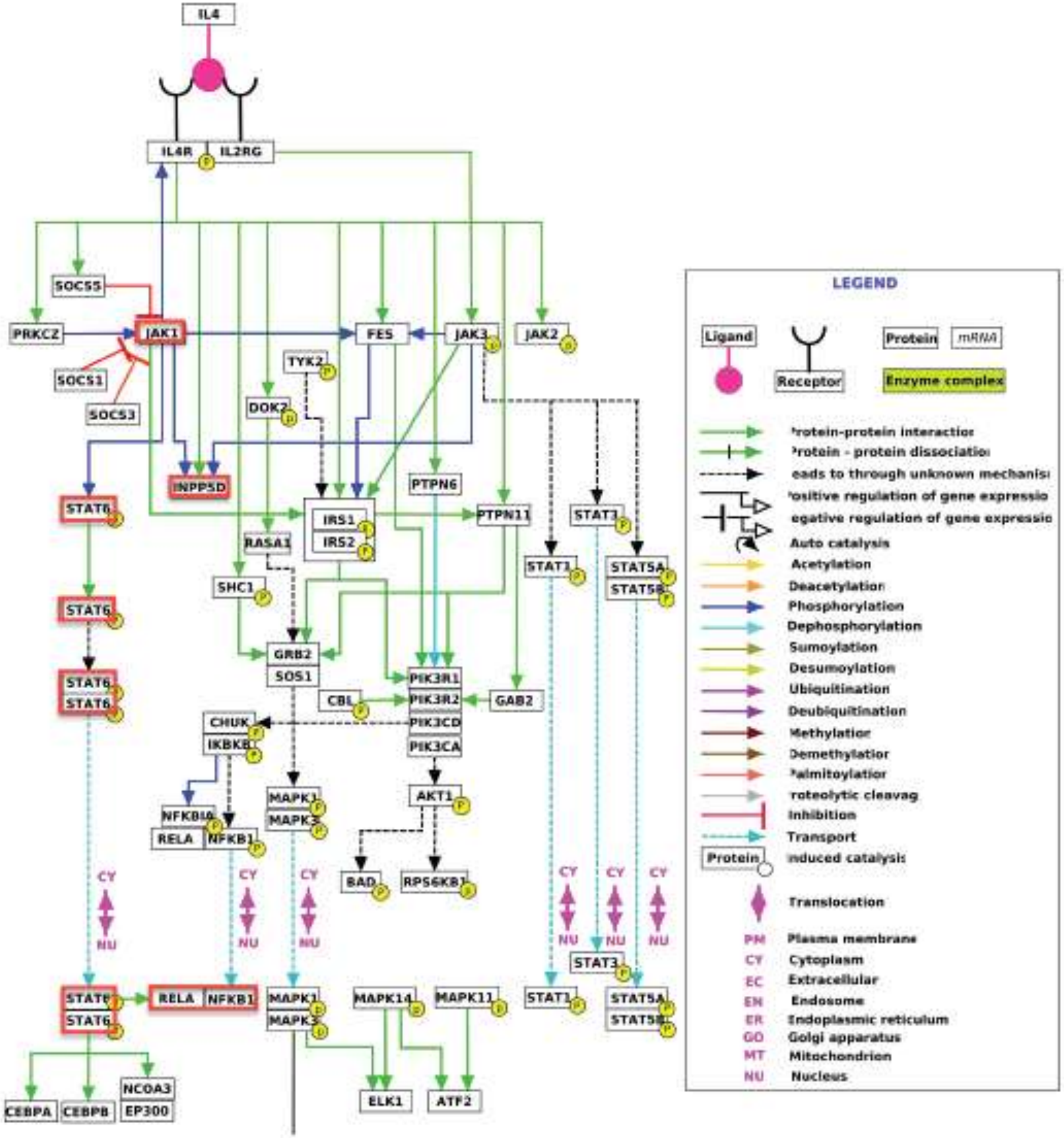


Fig. 5. An example of a biologically relevant pathway for DMD. The nodes from the induced subnetwork identified by PFSNet is highlighted with red boxes.

Leukemias: IL-4 Signaling in ALL



What have we learned?

- **Common headaches in gene expression analysis**
  - Natural fluctuation, protocol noise, batch effect

- **Use of biological background info to tame false positives**

- **Overlap analysis → direct-group analysis → network-based analysis**

- **Subnetwork-based methods yield more consistent and larger disease subnetworks**

From pathways to models, From static to dynamic:

A couple of very recent papers that are worth your leisure reading…

- Geistlinger et al. **From sets to graphs: Towards a realistic enrichment analysis of transcriptomic systems**. *Bioinformatics*, 27(13):i366—i373, 2011
- Zampieri et al. **A system-level approach for deciphering the transcriptional response to prion infection**. *Bioinformatics*, 27(24): 3407--3414, 2011

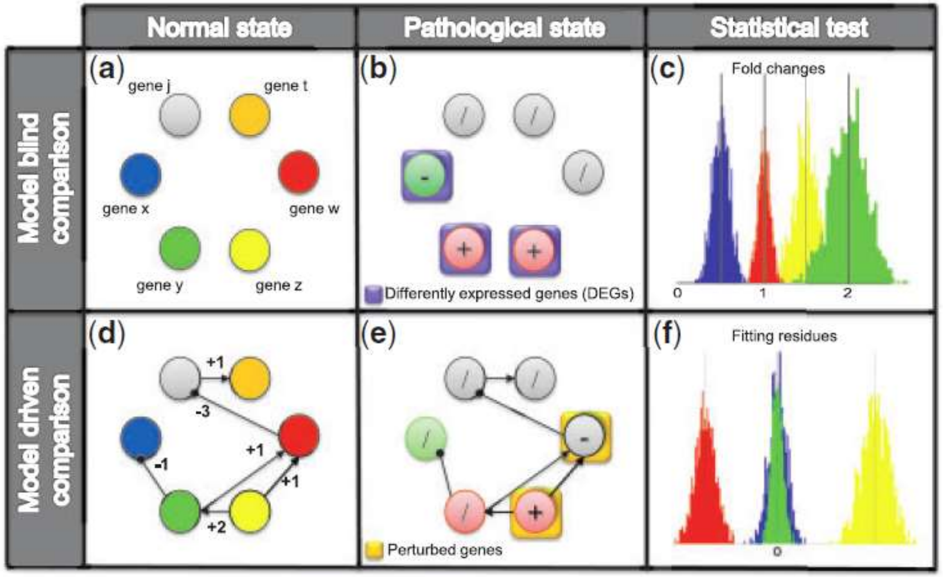


Fig. 1. System response inference: a toy genetic network consisting of six genes exemplifies the advantages of using a system-level data comparison (**a**). Standard statistical tests (i.e. *t*-test) unveil significant fold change in gene expression variations for each transcript individually (**b**), neglecting the underlying regulatory network. Such statistical test can identify whether the expression level of a transcript is significantly changed with respect to a reference. Putative gene expression changes are reported in panel (**c**). In this specific example, two genes are identified to be overexpressed [red/+ nodes] and one downregulated (green/- node), while the remaining three do not show any changes (grey nodes). By knowing the corresponding genetic regulatory network (**d**), we can discriminate the coherent variations from the unexpected ones. As shown in the example, two of the genes that showed a significant expression variations are consistent with model predictions i.e. the expression changes of genes *x* and *y* can be explained by the variation of gene *z*. This is reflected by a skew distribution of discrepancies (i.e. residues), between model predictions and observed data, centered around 0 (**f**). At the same time, one transcript, *w*, is not responding coherently to the initial model. The fact that its expression is unchanged, when it should have been increased, might relate to an anomalous direct effect of the pathology, preventing a synergistic response between all the genes in the system. Hence, the list of 'perturbed genes' can be sensibly different from the standard DEGs identified from individual fold change analysis (**b**/**e**).

Still a major challenge

- **Suppose there are very few samples, so few that you cannot estimate the p-value by permuting class labels**

- **What do you do?**

References

- Zhang et al. **Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes**. *Bioinformatics*, 25(13):1662-1668, 2009
- [ORA] Khatri & Draghici**. Ontological analysis of gene expression data: Current tools, limitations, and open problems**. *Bioinformatics*, 21(18):3587-3595, 2005
- [FCS] Goeman et al. **A global test for groups of genes: Testing association with a clinical outcome**. *Bioinformatics*, 20(1):93-99, 2004
- [GSEA] Subramanian et al. **Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles**. *PNAS*, 102(43):15545-15550, 2005
- [NEA] Sivachenko et al. **Molecular networks in microarray analysis**. *JBCB*, 5(2b):429-546, 2007
- [SNet] Soh et al. **Finding consistent disease subnetworks across microarray datasets**. *BMC Genomics*, **12(Suppl. 13):S15, 2011**

Use of Context in Gene Expression and Proteomic Profile Analysis
Part 2

Limsoon Wong

National University of Singapore

Typical Proteomic MS Experiment

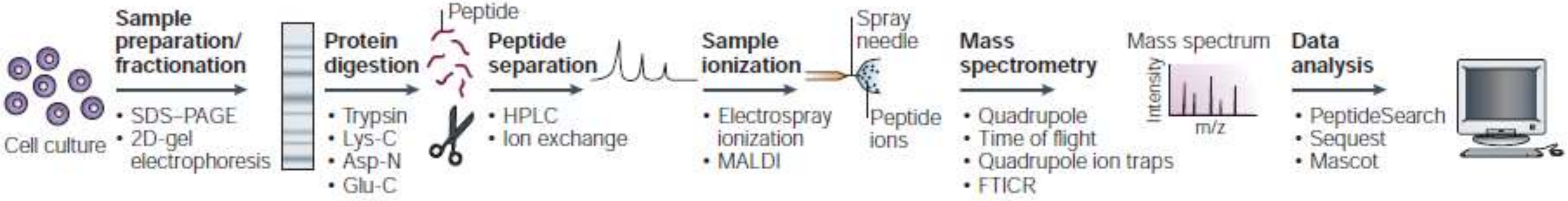
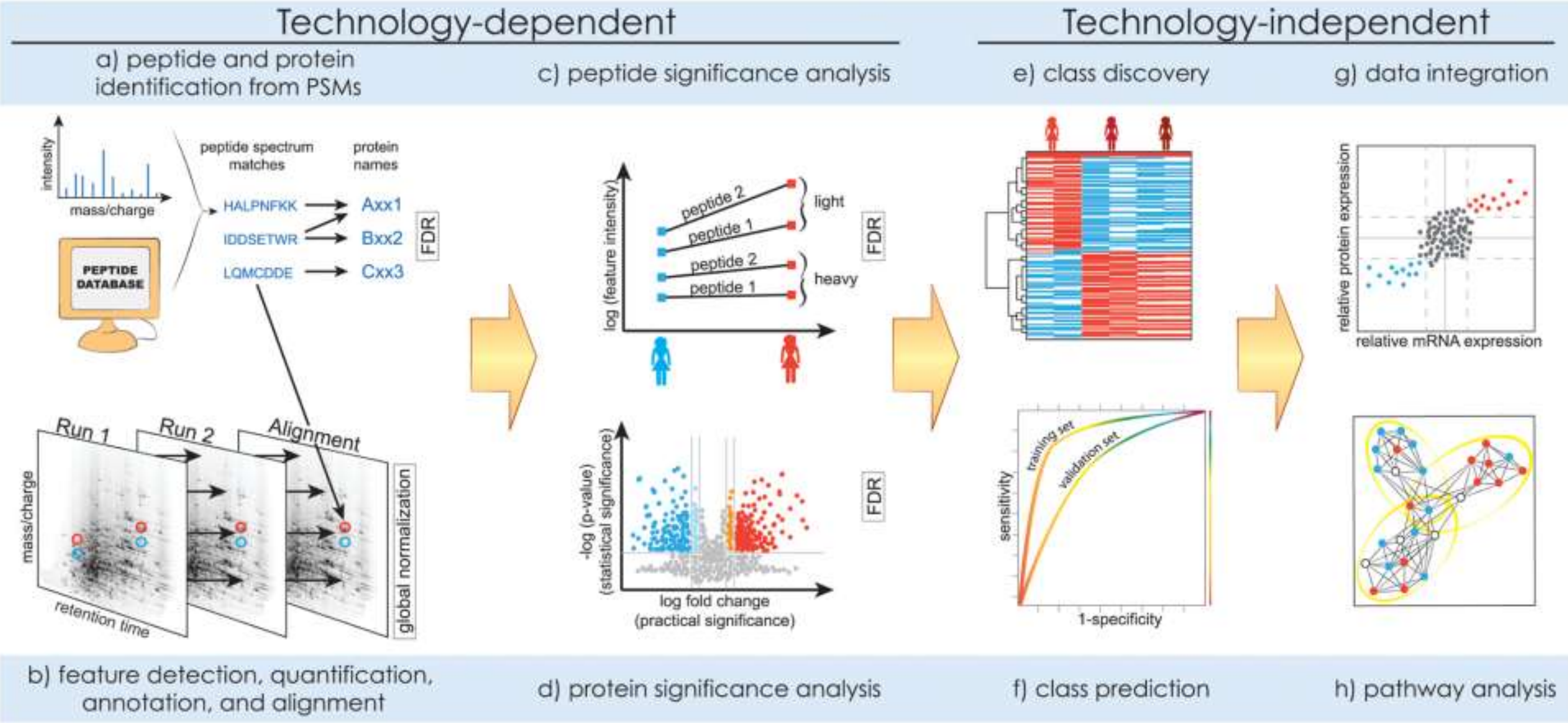


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous different enzymes and/or chemicals are available for this step. The generated peptide mixture is separated on- or off-line using single or multiple dimensions of peptide separation. Peptides are then ionized by electrospray ionization (depicted) or matrix-assisted laser desorption/ionization (MALDI) and can be analysed by various different mass spectrometers. Finally, the peptide-sequencing data that are obtained from the mass spectra are searched against protein databases using one of a number of database-searching programmes. Examples of the reagents or techniques that can be used at each step of this type of experiment are shown beneath each arrow. 2D, two-dimensional; FTICR, Fourier-transform ion cyclotron resonance; HPLC, high-performance liquid chromatography.

Source: Steen & Mann. The ABC’s and XYZ’s of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

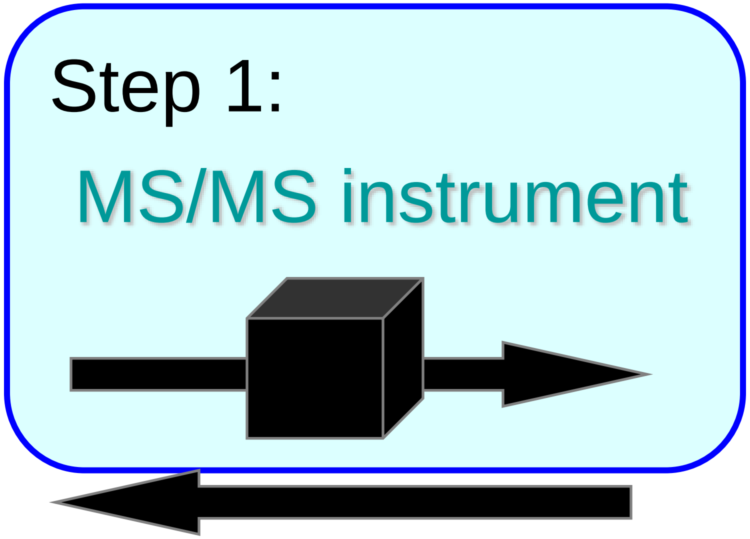
Diagnosis Using Proteomics



Kall and Vitek, 2011

Protein Identification by Mass Spec

Sequence



Step 1:

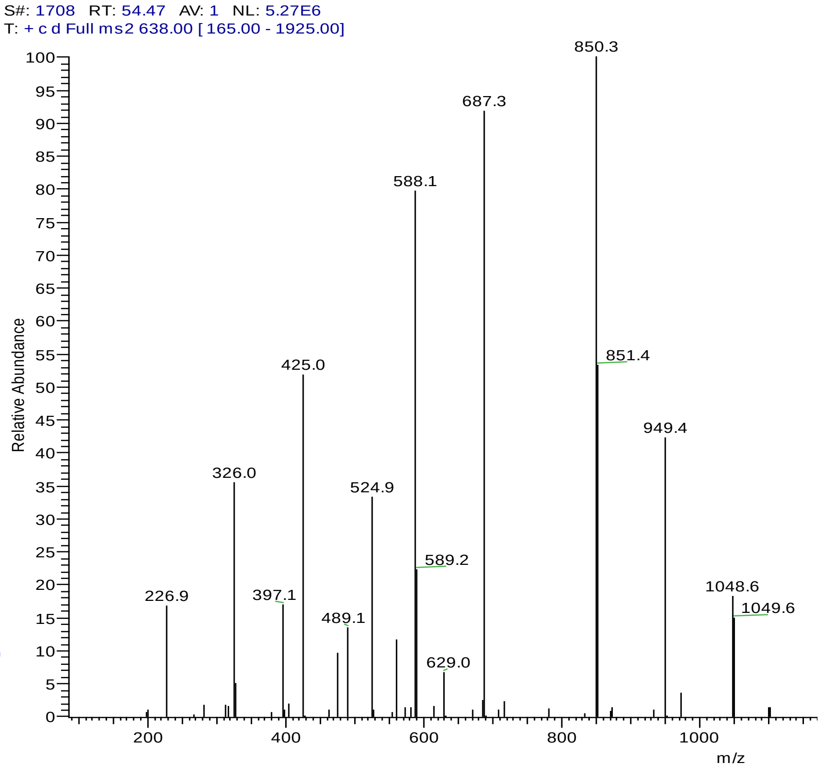
MS/MS instrument

Database search

- Sequest, Mascot, InSpect

de Novo interpretation

- Lutefisk, Peaks, PepNovo



Source: Leong Hon Wai

Breaking Protein into Peptides, and Peptides into Fragment Ions

- **Proteases, e.g. trypsin, break protein into peptides**
- **A Tandem Mass Spectrometer further breaks the peptides down into fragment ions and measures the mass of each piece**
- **Mass Spectrometer accelerates the fragmented ions; heavier ions accelerate slower than lighter ones**
- **Mass Spectrometer measures mass/charge ratio of an ion**

Source: Leong Hon Wai

A rather nice set of proteomic profiles of leukemia patients

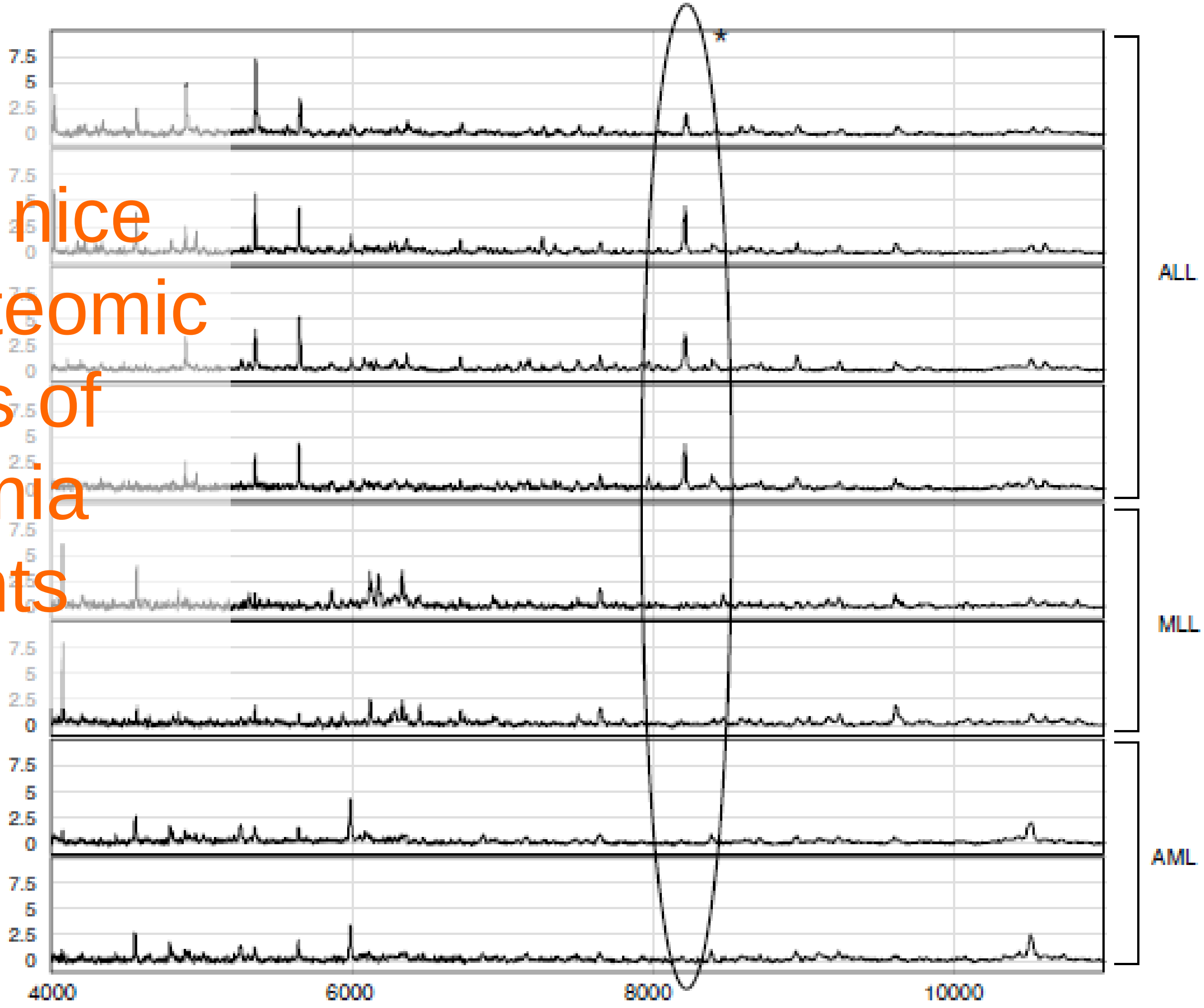


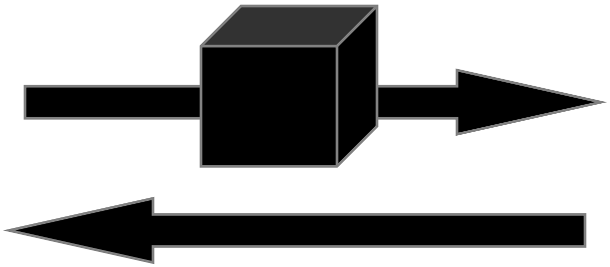
Figure 1 Spectra from SELDI-TOF MS analysis of REH, 697, MV4;11, and Kasumi cell lines. Protein (4 μg) from each cell type was analyzed on SAX2 ProteinChip® Arrays. ALL cell lines shown are REH and 697, the MLL cell line is MV4;11, and the AML cell line is Kasumi. The asterisk indicates the differentially expressed protein at 8.3 kDa.

Source: Hegedus et al. Proteomic analysis of childhood leukemia. Leukemia, 19:1713-1718, 2005

Peptide Identification by Mass Spec

S
e
q
u
e
n
c
e

MS/MS instrument



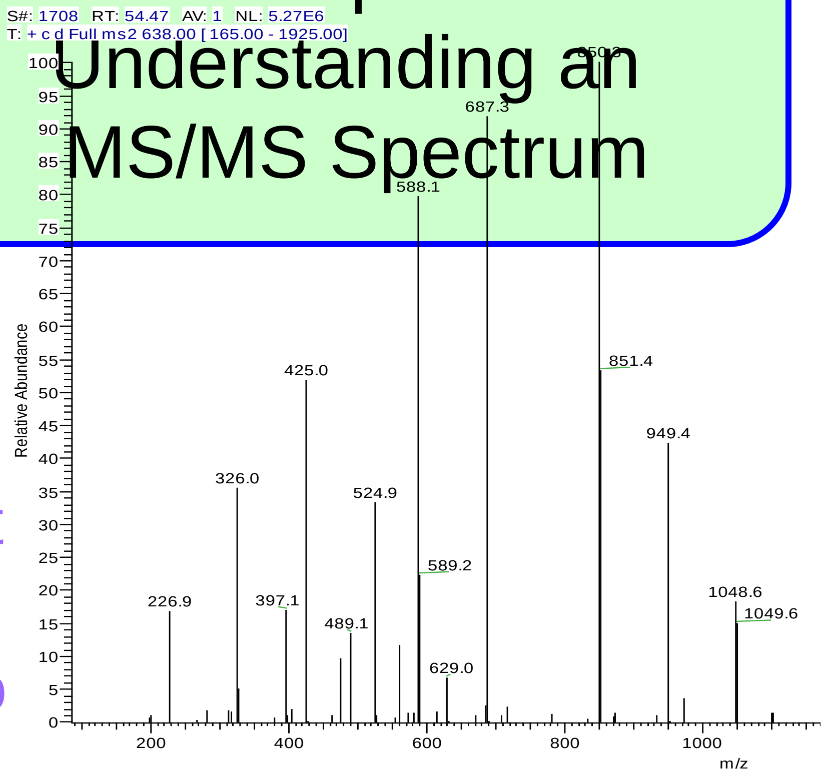
Database search

- Sequest, Mascot, InSpect

de Novo interpretation

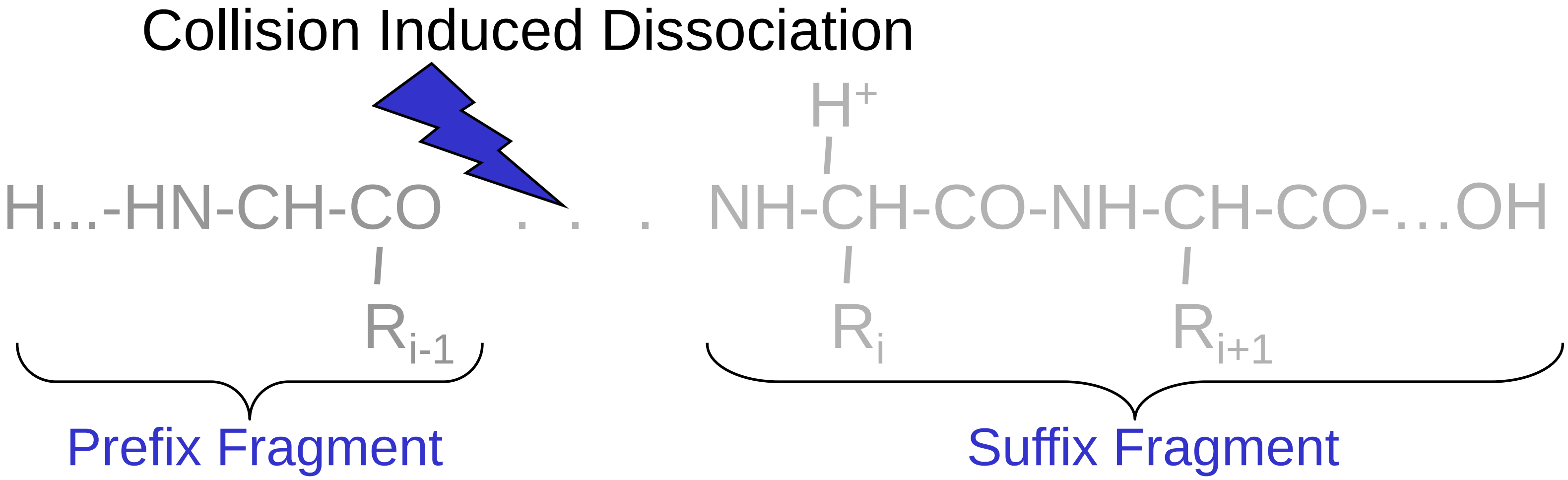
- Lutefisk, Peaks, PepNovo

Step 2:
Understanding an MS/MS Spectrum



Source: Leong Hon Wai

Peptide Fragmentation



Collision Induced Dissociation

H^+

H...-HN-CH-CO . . . NH-CH-CO-NH-CH-CO-…OH

R_{i-1} R_i R_{i+1}

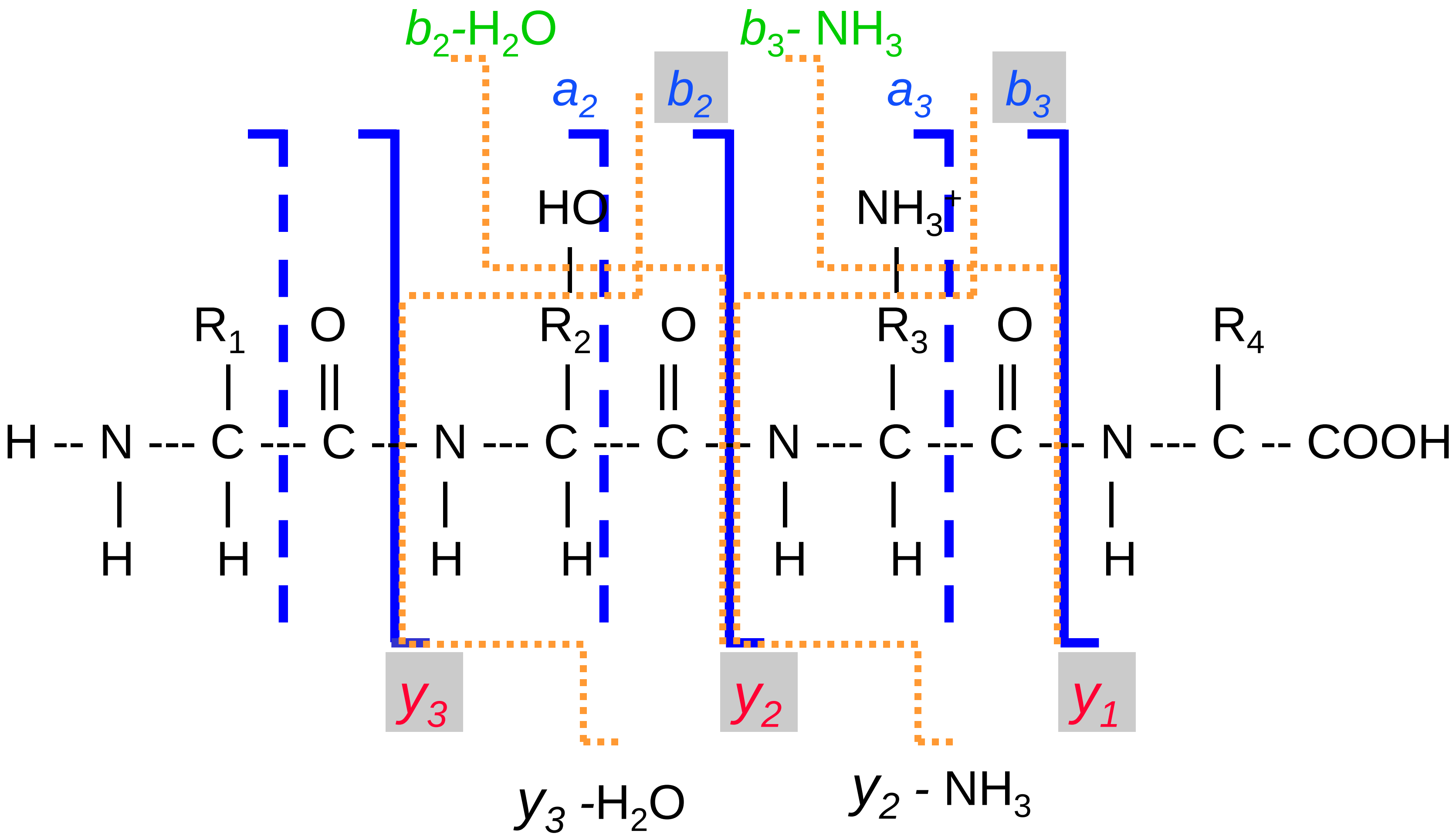
Prefix Fragment

Suffix Fragment

- **Peptides tend to fragment along the backbone**
- **Fragments can also loose neutral chemical groups like NH_3 and H_2O**

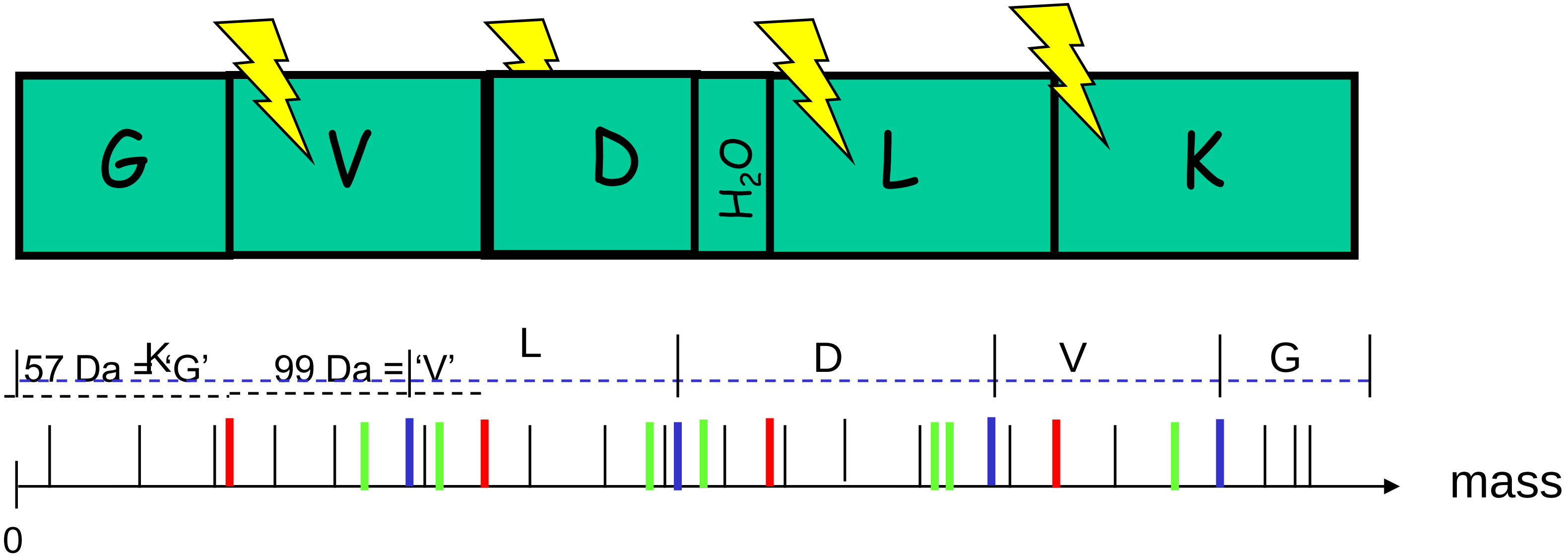
Source: Leong Hon Wai

… and fragments due to neutral losses



Source: Leong Hon Wai

Mass Spectra



- **The peaks in the mass spectrum:**
  - Prefix and Suffix Fragments
  - Fragments with neutral losses (-H_2O, -NH_3)
  - Noise and missing peaks

Source: Leong Hon Wai

Bafna & Edwards. “On de novo interpretation of tandem mass spectra for peptide identification”. RECOMB 2003, pp. 9-18

Example MS/MS Spectrum

88	145	292	405	534	663	778	924	b-ions
S	G	F	L	E	E	D	K	
924	837	780	633	520	391	262	141	y-ions

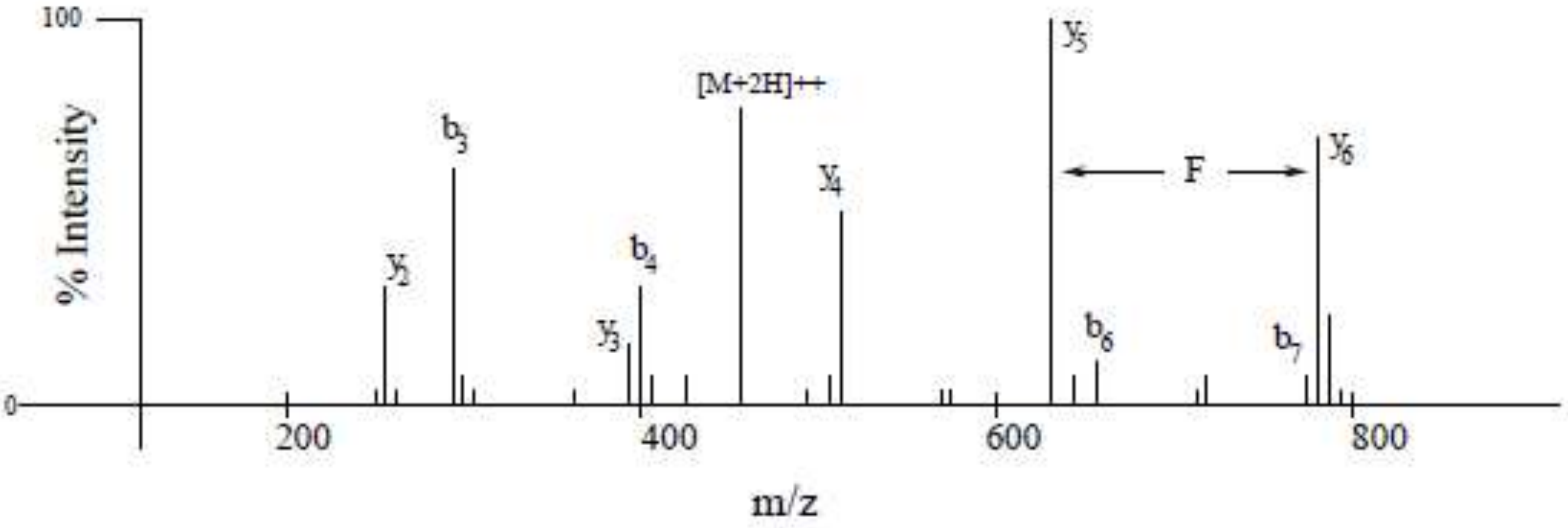
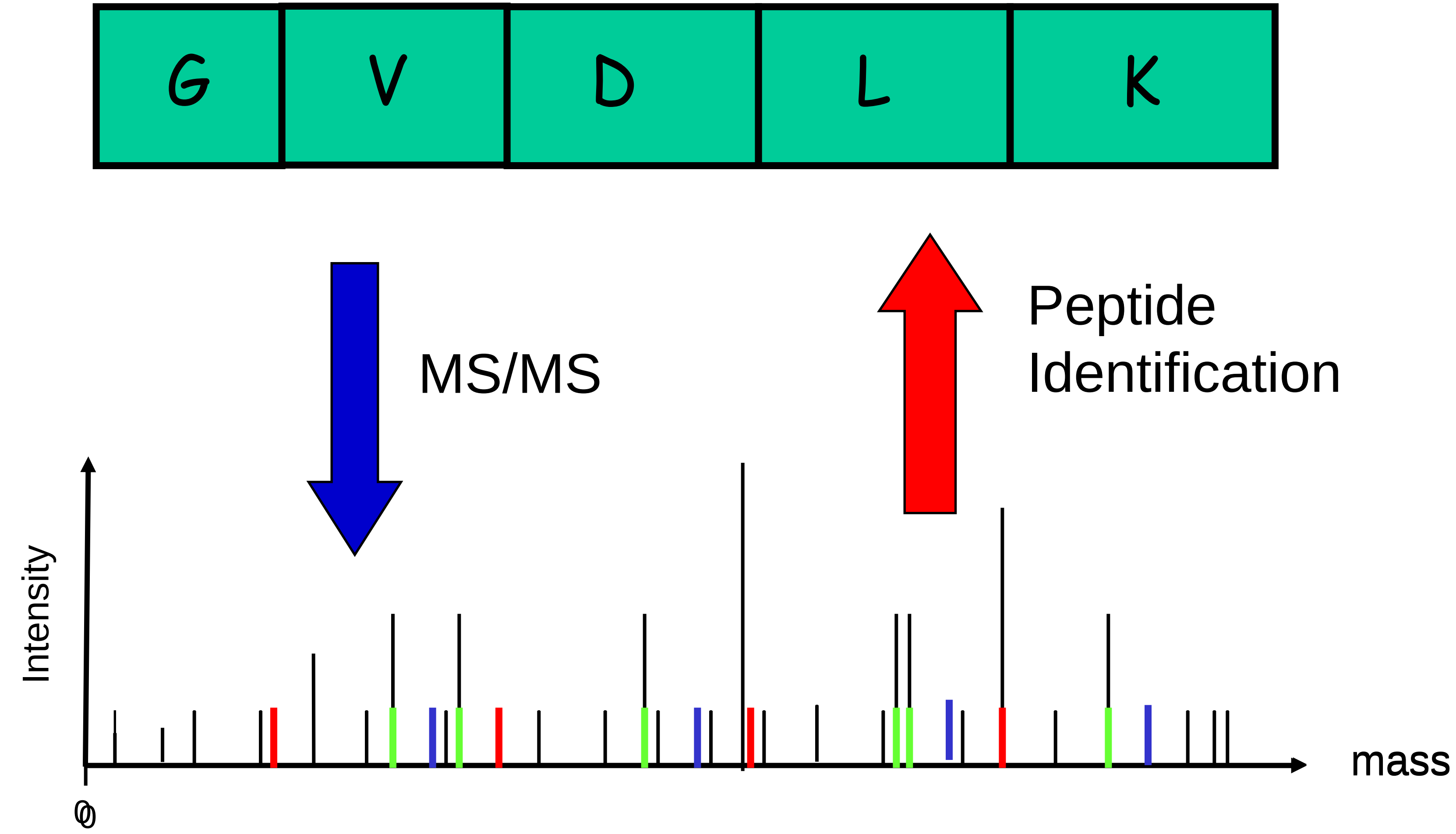


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

Protein Identification with MS/MS

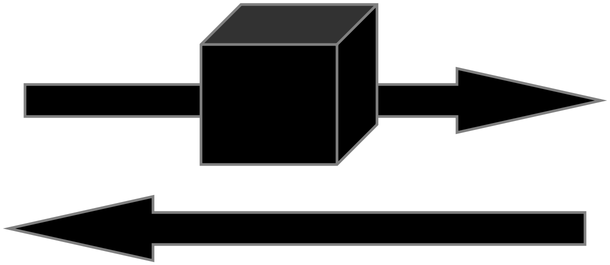


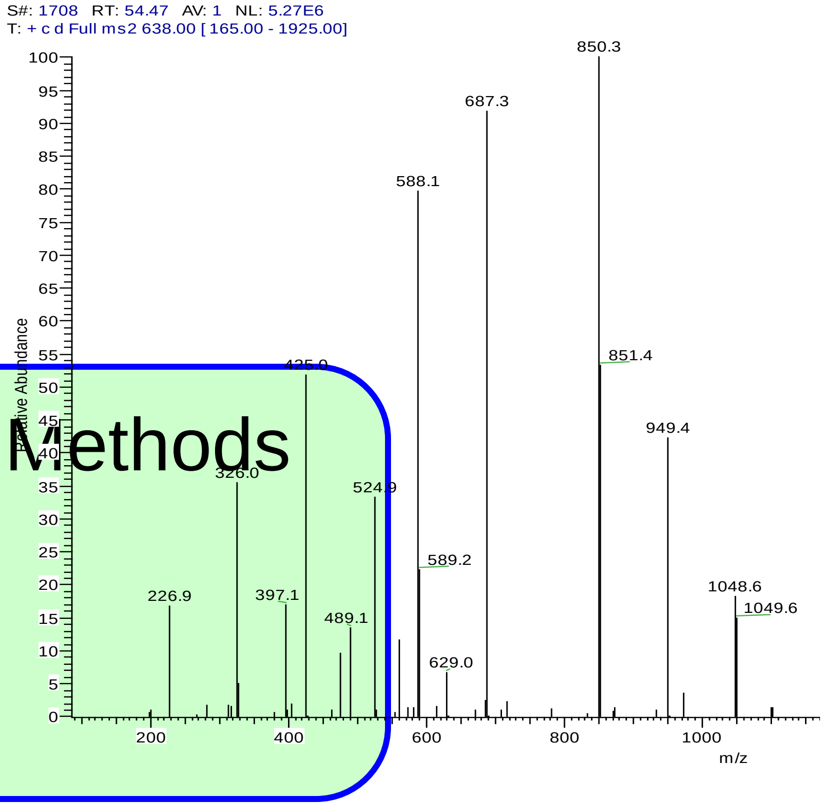
Source: Leong Hon Wai

Peptide Identification by Mass

Sequence

MS/MS instrument





Step 3: Computational Methods

Database search

Sequest, Mascot

de Novo interpretation

Lutefisk, Peaks, PepNovo

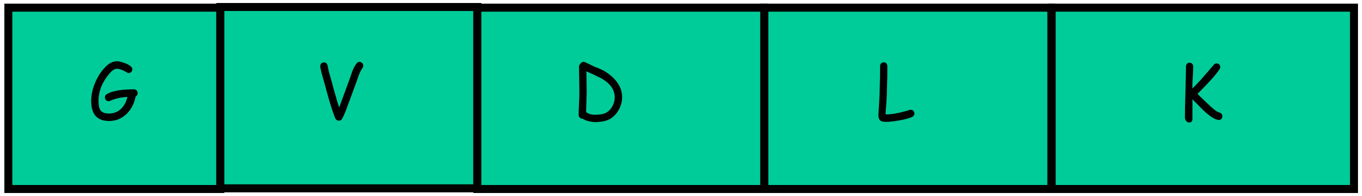
Source: Leong Hon Wai

Database Search Algorithms

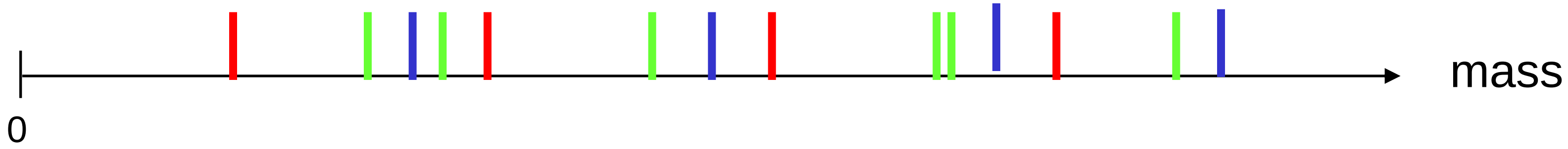
- **Database search**
  - Used for spectrum from known peptides
  - Rely on completeness of database
- **General Approach**
  - Match given spectrum with known peptide
  - Enhanced with advanced statistical analysis and complex scoring functions
- **Methods**
  - SEQUEST, MASCOT, InsPecT, Paragon

Theoretical Spectrum for a Peptide

- **Given this peptide**


G V D L K

- **Its theoretical spectrum is**

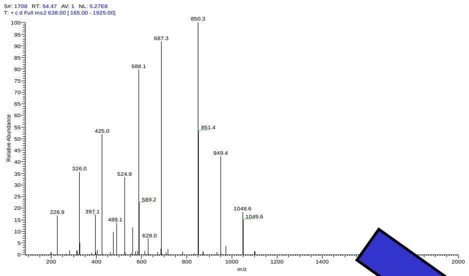


- **Theoretical spectrum is dependent on**
  – Set of ion-types considered
  – Larger if multi-charge ions are considered

Source: Leong Hon Wai

Database Search Algorithm

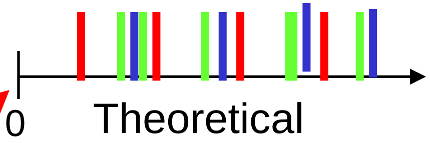
Database Search



Match

Database of known peptides

MDERHILNM, KLQWVCSDL, PTYWASDL, ENQIKRSACVM, TLACHGGEM, NGALPQWRT, HLLERTKMNVV, GGPASSDA, GGLITGMQSD, MQPLMNWE, ALKIIMNVRT, AVGELTK, HEWAILF, GHNLWAMNAC, GVFGSVLRA, EKLNKAATYIN..



Theoretical spectrum

Matching Score for this peptide

Repeat for all the peptides in the Database

Source: Leong Hon Wai

- There are also approaches for de novo peptide identification. ..
- But I will omit these here

Protein Identification

- **After all the peptides have been identified, they are grouped into protein identifications**
- **Peptide scores are added up to yield protein scores**
- **Confidence of a particular peptide identification increases if other peptides identify the same protein and decreases if no other peptides do so**
- **Protein identifications based on single peptides should only be allowed in exceptional cases**

Source: Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

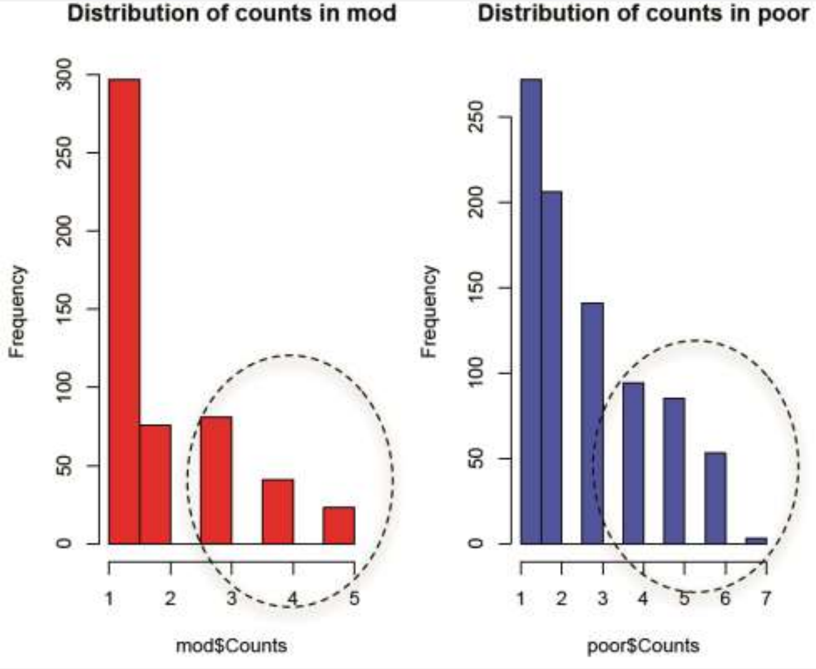
Cf. Gene Expression Profile Analysis

- **Once the proteins are identified, the proteomic profile of a sample can be constructed**
  - I.e., which protein is found in the sample and how abundant it is

- **Similar to gene expression profile. So gene expression profile analysis techs can be applied**

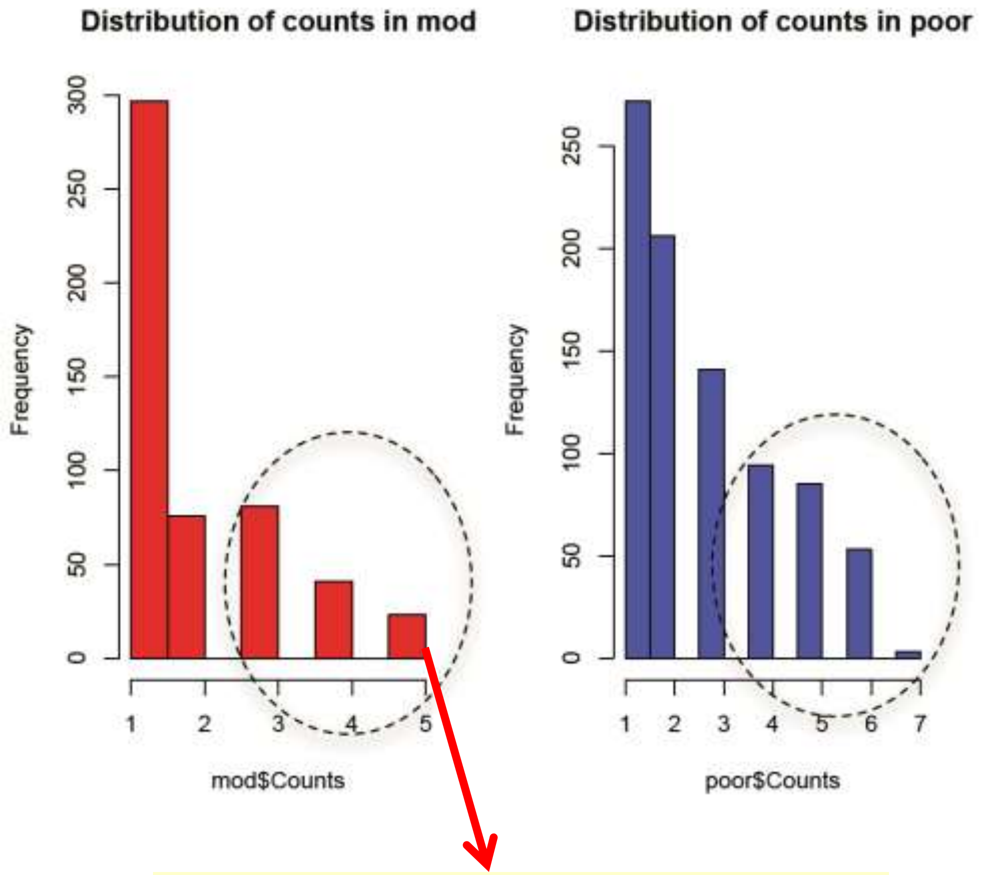
- **Some key differences**
  - Proteomic profile has much fewer features
  - Proteomic profiling study has much fewer samples

Part 2: Delivering more powerful proteomic profile analysis



- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

Typical frequency distribution of proteins detected in proteomic profiles



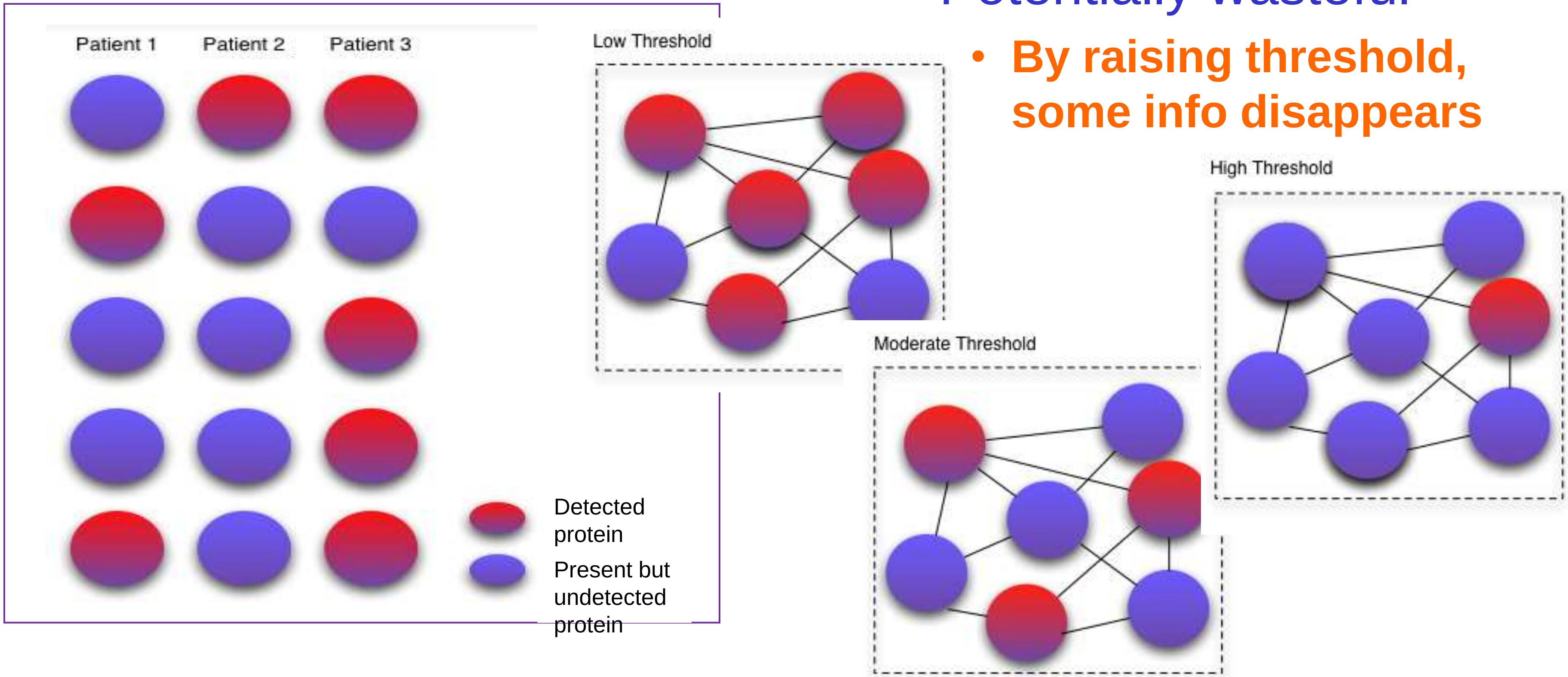
Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

Issues in Proteomic Profiling

- **Coverage**
- **Consistency**

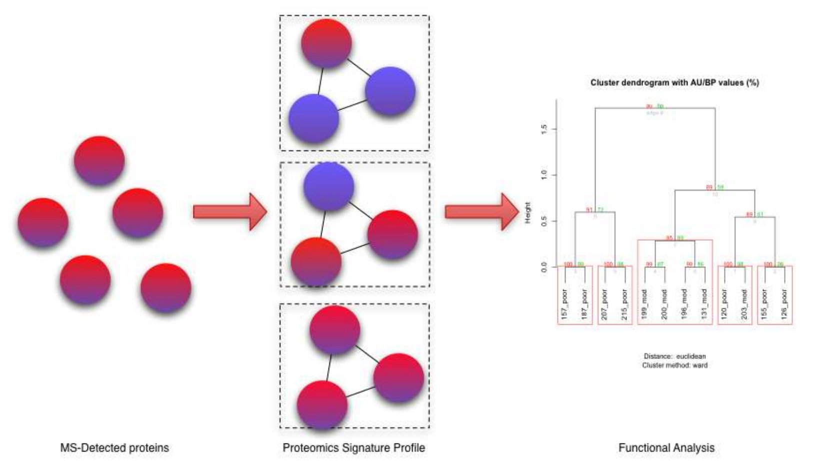
⇒ **Thresholding**

- – Somewhat arbitrary
- – Potentially wasteful
  - **By raising threshold, some info disappears**

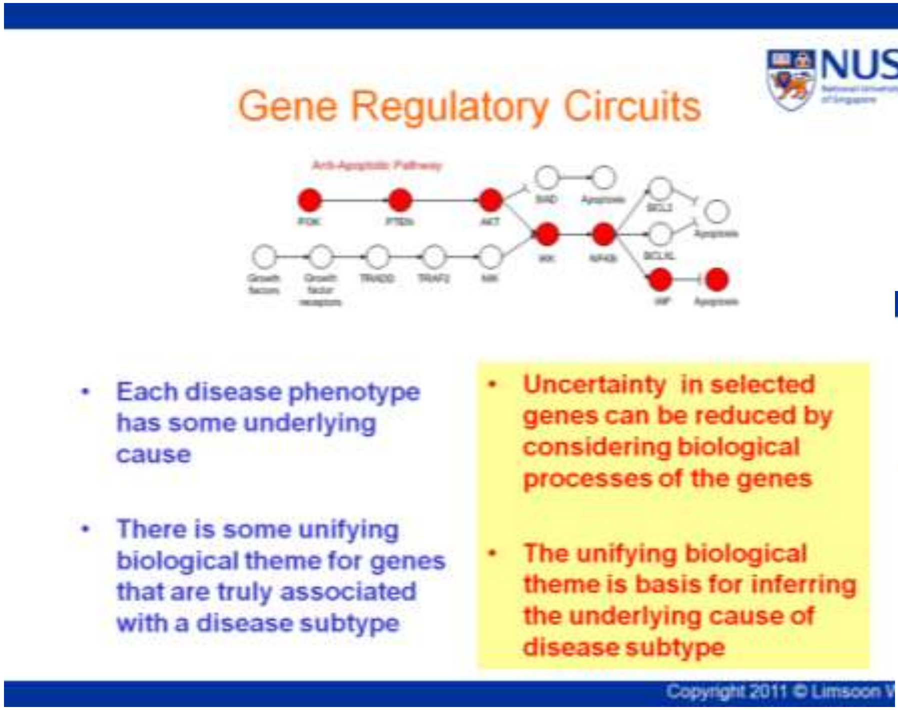


Part 2: Delivering more powerful proteomic profile analysis

- Common issues in proteomic profile analysis
- **Improving consistency**
  - PSP
  - PDS
- Improving coverage
  - CEA
  - PEP
  - Max Link



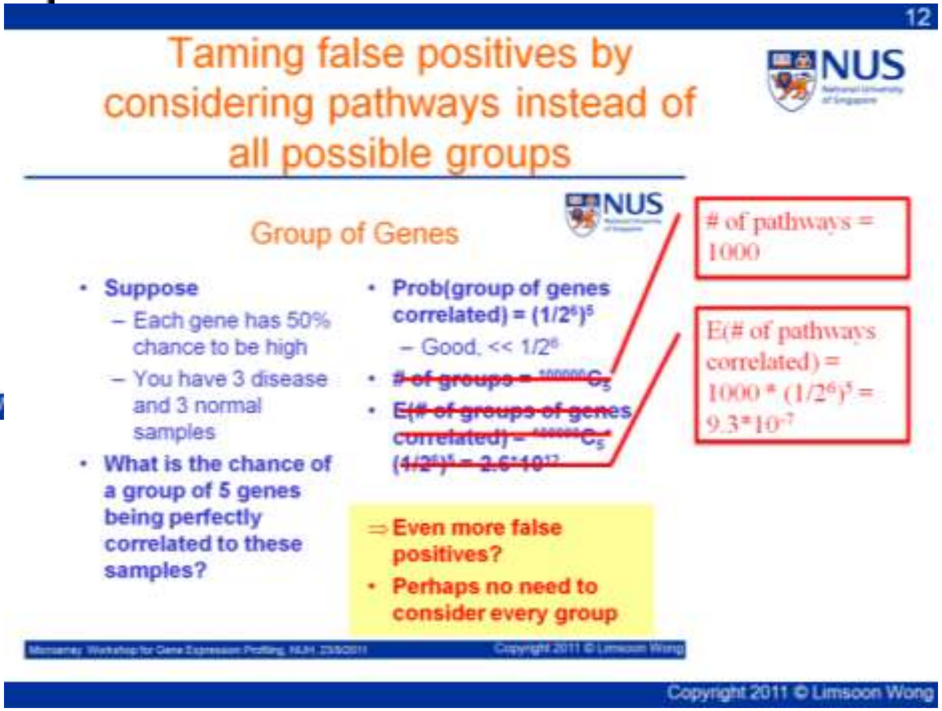
An inspiration from gene expression profile analysis



Gene Regulatory Circuits

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype

- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype



Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- What is the chance of a group of 5 genes being perfectly correlated to these samples?

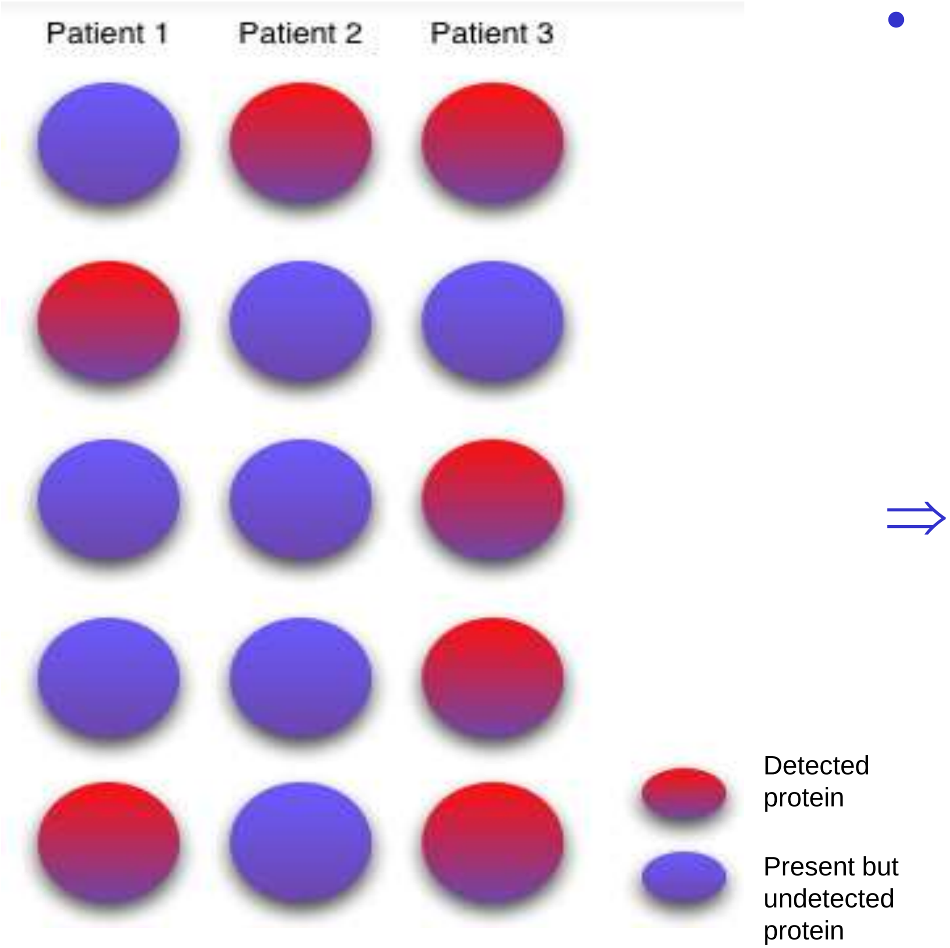
- Prob(group of genes correlated) = $(1/2^6)^5$
  - Good, << $1/2^6$
- ~~# of groups = $^{10000}C_5$~~
- ~~E(# of groups of genes correlated) = $^{10000}C_5 (1/2^6)^5 = 2.6*10^{12}$~~

⇒ Even more false positives?
- Perhaps no need to consider every group

# of pathways = 1000

E(# of pathways correlated) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

Intuitive Example



- **Suppose the failure to form a protein complex causes a disease**
  - If any component protein is missing, the complex can’t form

⇒ **Diff patients suffering from the disease can have a diff protein component missing**

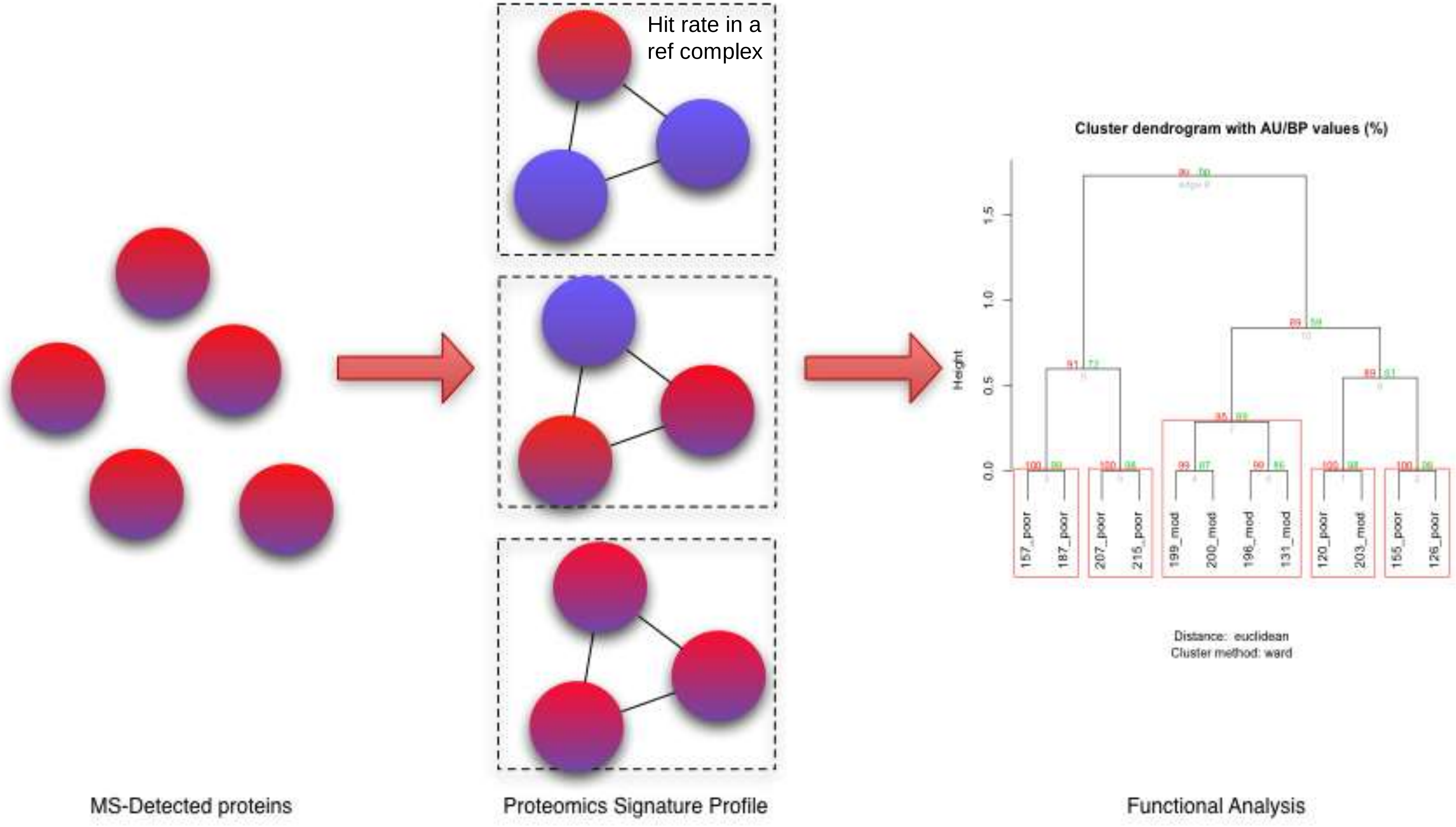
  - Construct a profile based on complexes?

We try an adaptation of SNet on proteomics profiles…

“Proteomic Signature Profiling” (PSP)

Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

“Threshold-free” Principle of PSP



Hit rate in a ref complex

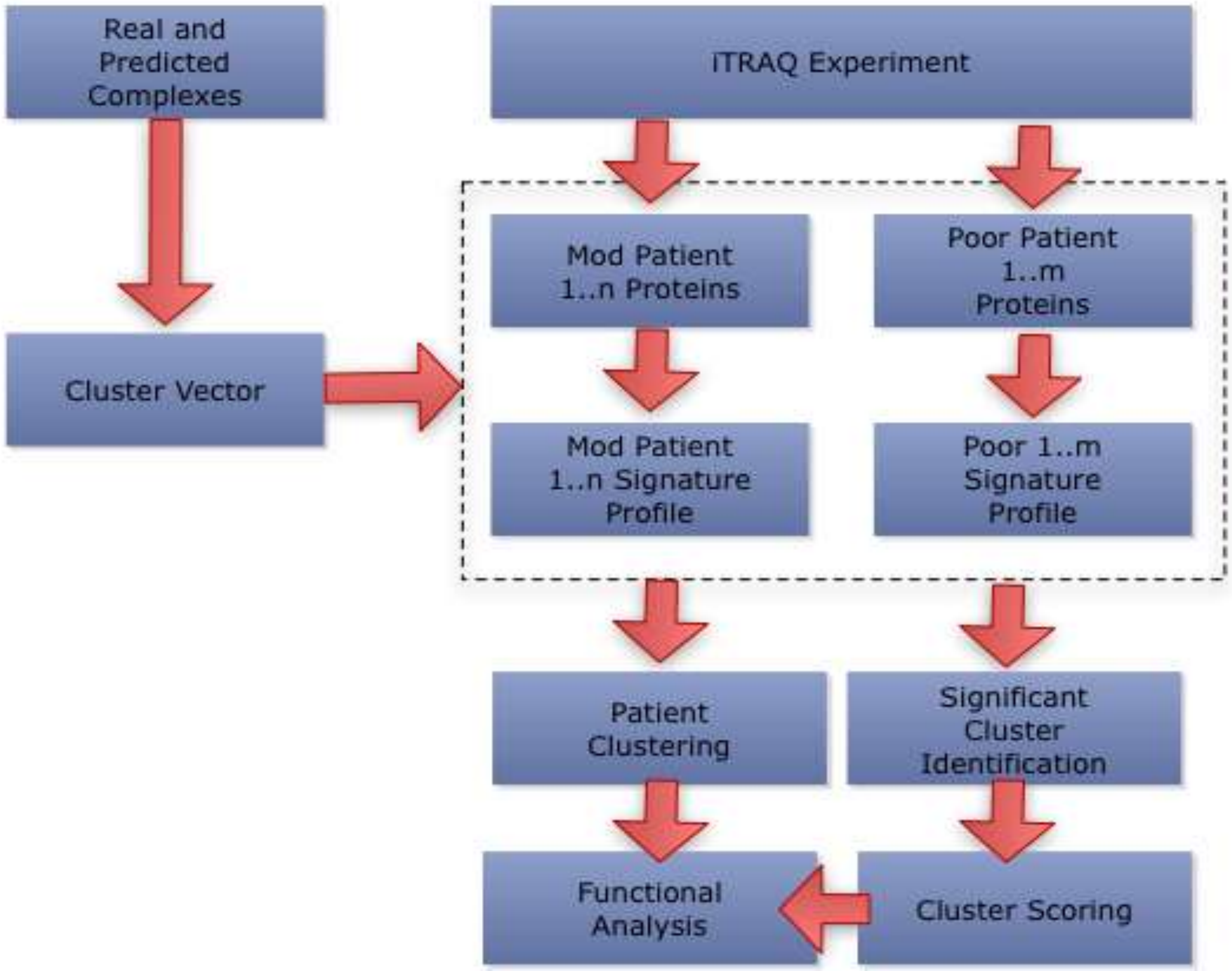
MS-Detected proteins

Proteomics Signature Profile

Functional Analysis

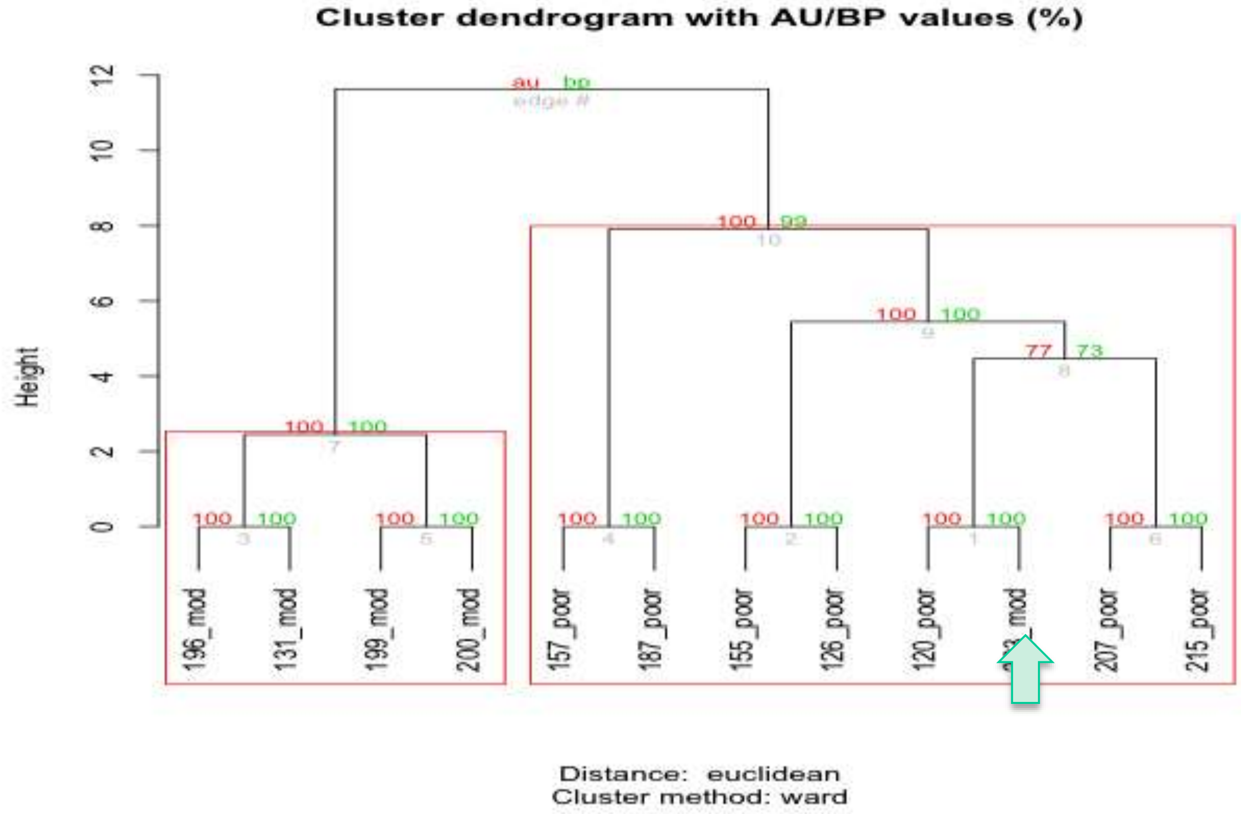
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

Applying PSP to a HCC Dataset


Real and Predicted Complexes
iTRAQ Experiment
Mod Patient 1..n Proteins
Poor Patient 1..m Proteins
Cluster Vector
Mod Patient 1..n Signature Profile
Poor 1..m Signature Profile
Patient Clustering
Significant Cluster Identification
Functional Analysis
Cluster Scoring

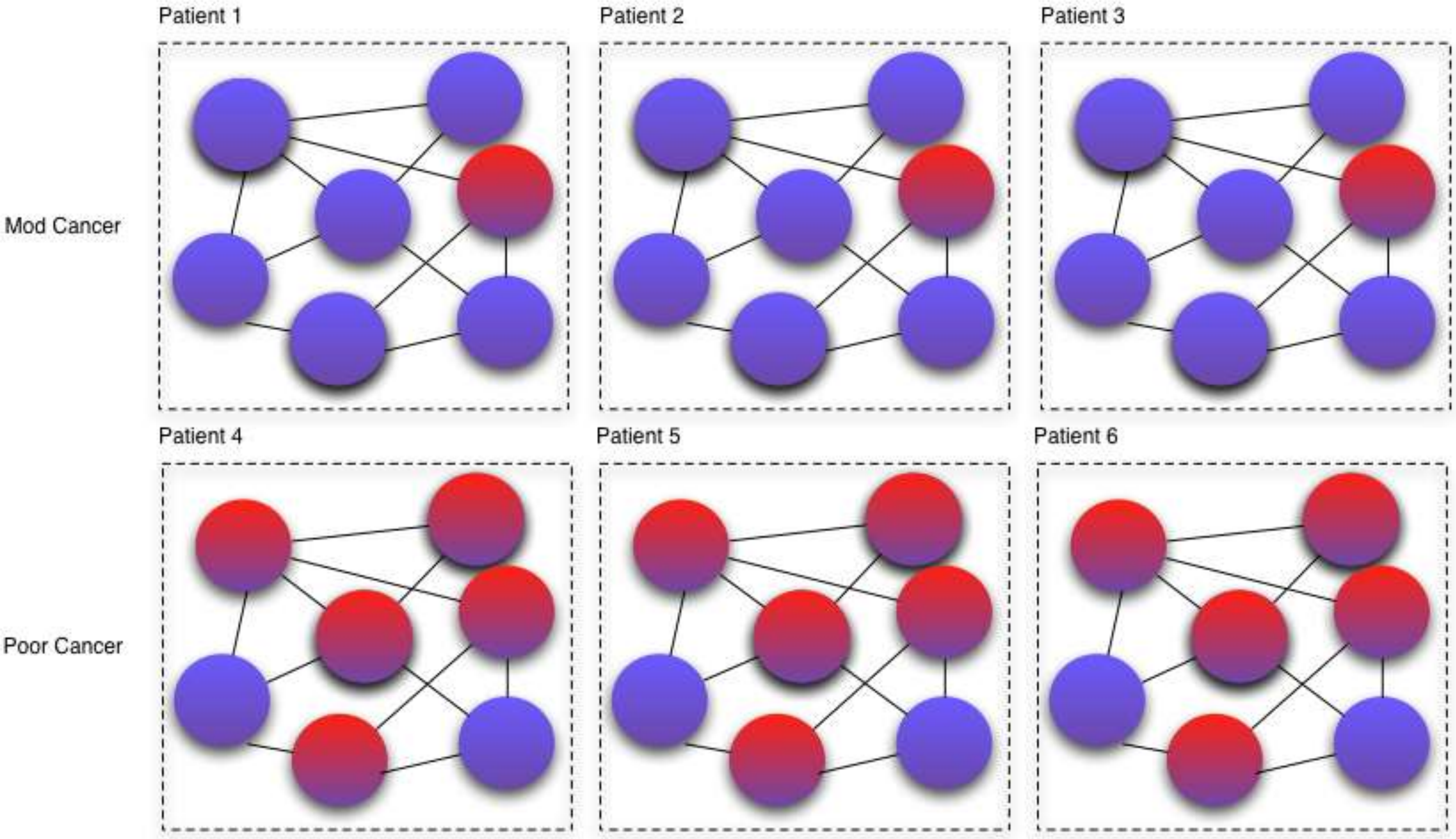
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

Consistency: Samples segregate by their classes with high confidence



Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

Feature Selection



🔴 Detected Protein

🔵 Undetected Protein

$$t\_score = \frac{\bar{HA} - \bar{HB}}{S_{HA,HB}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

$$S_{HA,HB} = \sqrt{\frac{(m-1)S_{HA}^{2} + (n-1)S_{HB}^{2}}{m+n-2}}$$

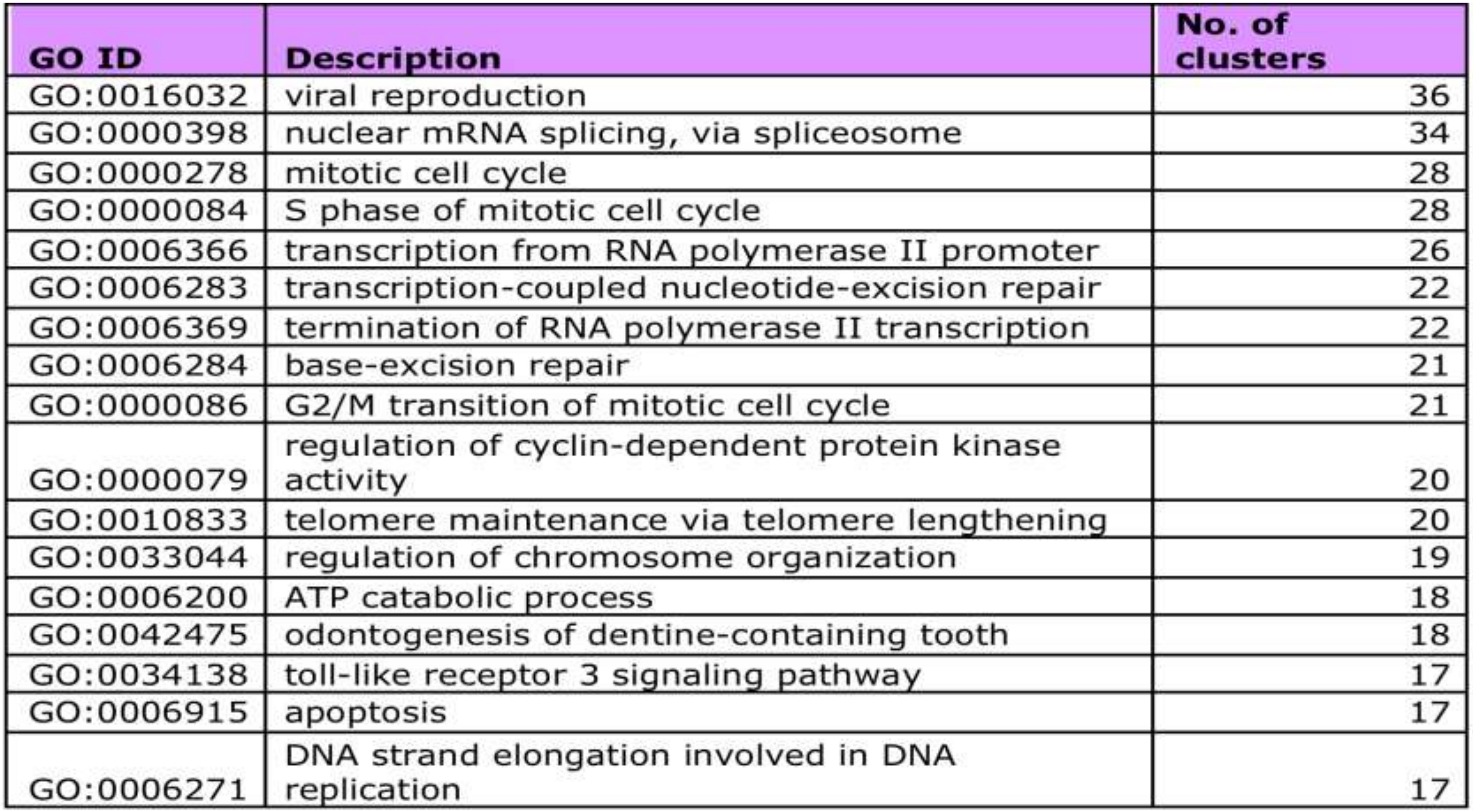
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

Top-Ranked Complexes

Cluster_ID	p_val	mod_score	poor_score	cluster_name
5179	0.000300541	0.513951977	3.159758312	NCOA6-DNA-PK-Ku-PARP1 complex
5235	0.000300541	0.513951977	3.159758312	WRN-Ku70-Ku80-PARP1 complex
1193	0.000300541	0.513951977	3.159758312	Rap1 complex
159	0	0	2.810927655	Condensin I-PARP-1-XRCC1 complex
2657	0.008815869	0	2.55616281	ESR1-CDK7-CCNH-MNAT1-MTA1-HDAC2 complex
3067	0.00911641	0	2.55616281	RNA polymerase II complex, incomplete (CDK8 complex), chromatin structure modifying
1226	0.013323983	0.715352108	2.420592827	H2AX complex I
5176	0	0.513951977	2.339059313	MGC1-DNA-PKcs-Ku complex
1189	0	0.513951977	2.339059313	DNA double-strand break end-joining complex
5251	0	0.513951977	2.339059313	Ku-ORC complex
2766	0	0.513951977	2.339059313	TERF2-RAP1 complex

Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

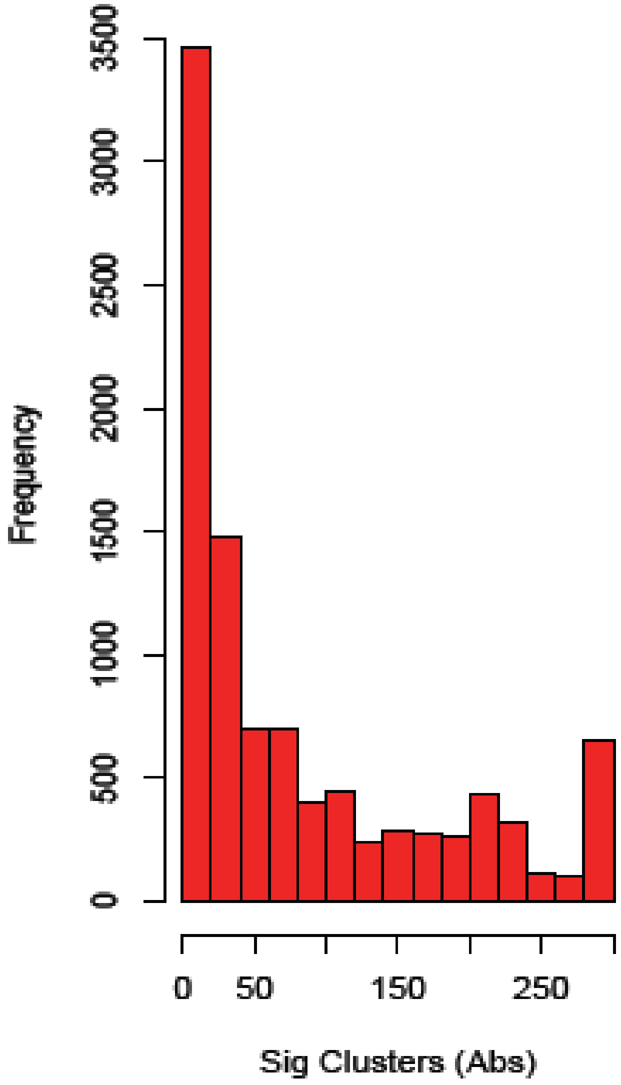
Top-Ranked GO Terms



GO ID	Description	No. of clusters
GO:0016032	viral reproduction	36
GO:0000398	nuclear mRNA splicing, via spliceosome	34
GO:0000278	mitotic cell cycle	28
GO:0000084	S phase of mitotic cell cycle	28
GO:0006366	transcription from RNA polymerase II promoter	26
GO:0006283	transcription-coupled nucleotide-excision repair	22
GO:0006369	termination of RNA polymerase II transcription	22
GO:0006284	base-excision repair	21
GO:0000086	G2/M transition of mitotic cell cycle	21
GO:0000079	regulation of cyclin-dependent protein kinase activity	20
GO:0010833	telomere maintenance via telomere lengthening	20
GO:0033044	regulation of chromosome organization	19
GO:0006200	ATP catabolic process	18
GO:0042475	odontogenesis of dentine-containing tooth	18
GO:0034138	toll-like receptor 3 signaling pathway	17
GO:0006915	apoptosis	17
GO:0006271	DNA strand elongation involved in DNA replication	17

Goh et al**. Enhancing utility of proteomics signature profiling (PSP) with pathway derived subnets (PDSs), performance analysis and specialized ontologies**. *BMC Genomcs, to appear.*

False Positive Rate Analysis



- **Divide 7 poor patients into 2 groups**
  - Significant complexes produced by PSP here are false positives
- **Repeat many times to get dull distribution**
  - Median = 40, mode = 6
- **Cf. 523 complexes in CORUM (size ≥4) used in PSP. At p ≤ 5%, 523 * 5% ≈ 27 false positives expected**

A Shortcoming of PSP

- **Protein complex databases are still relatively small & incomplete…**

⇒ Augment the set of protein complexes by protein clusters predicted from PPI networks!

- **Many protein complex prediction methods**
  - CFinder, Adamcsek et al. *Bioinformatics*, 22:1021--1023, 2006
  - CMC, Liu et al. *Bioinformatics*, 25:1891--1897, 2009
  - CFA, Habibi et al. BMC Systems Biology, 4:129, 2010
  - …

Another Shortcoming of PSP

- **Protein complexes provided a biologically-rich feature set for PSP**
  - But it is only one aspect of biological function

- **The other aspect is biological pathways**
  - But coverage issue of proteomic profiles create lots of “holes”

- **Can we extract and use subnets from pathways?**

Another adaptation of SNet on proteomics profiles…

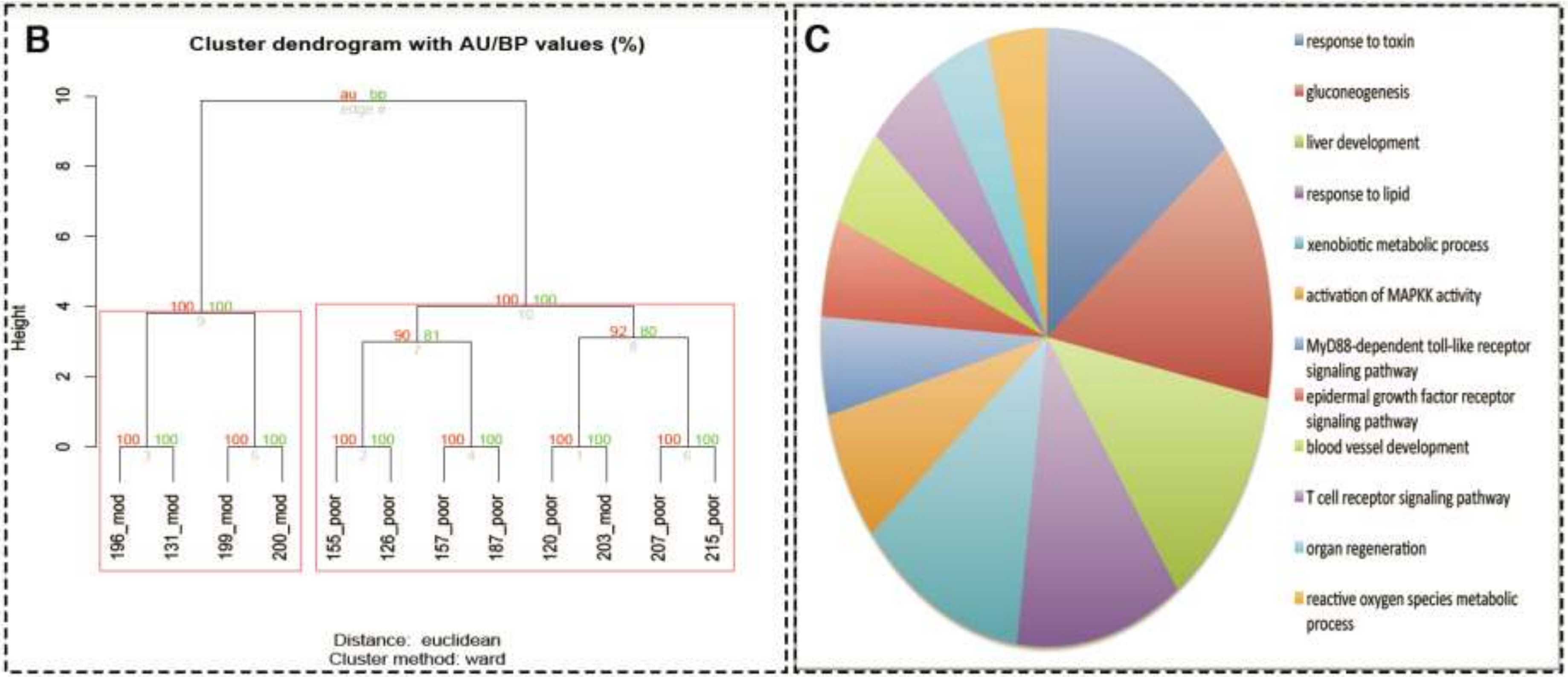
“Pathway-Derived Subnets” (PDS)

Pathway-Derived Subnets (PDS)

- **Identify the set S_i of proteins detected in more than 50% of samples having phenotype P_i**
  - Do this for each phenotype $P_1, \ldots, P_k$
- **Overlay $\cup_i S_i$ to pathways**
- **Remove nodes not covered by $\cup_i S_i$**

  $\Rightarrow$This fragments pathways into subnets
- **Use these subnets to form "proteomic signature profiles"**
  - The rest of the steps is same as PSP

PDS consistently segregates mod vs poor patients



What have we learned?

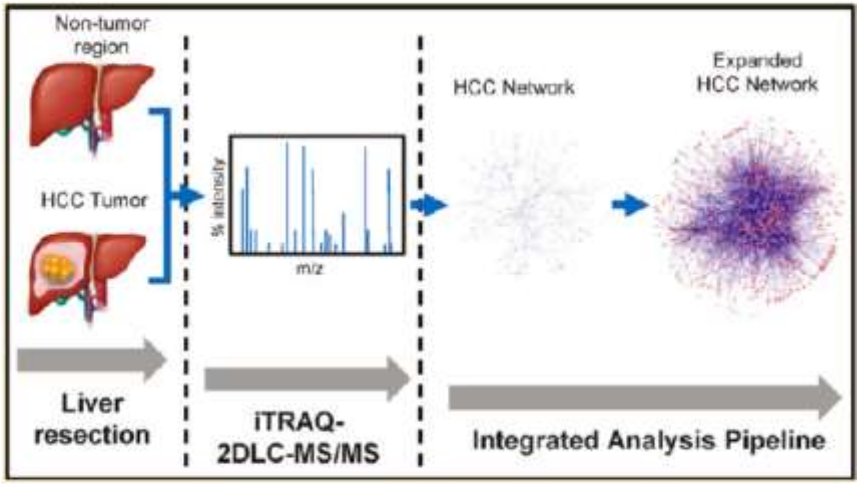
- **PSP / PDS can deal with consistency issues in proteomics**

- **GO term analysis also indicates that PSP / PDS select clusters that play integral roles in cancer**

- **PSP / PDS reveal many potential clusters and is not constrained by any prior arbitrary filtering which is a common first step in conventional analytical approaches**

Part 2: Delivering more powerful proteomic profile analysis

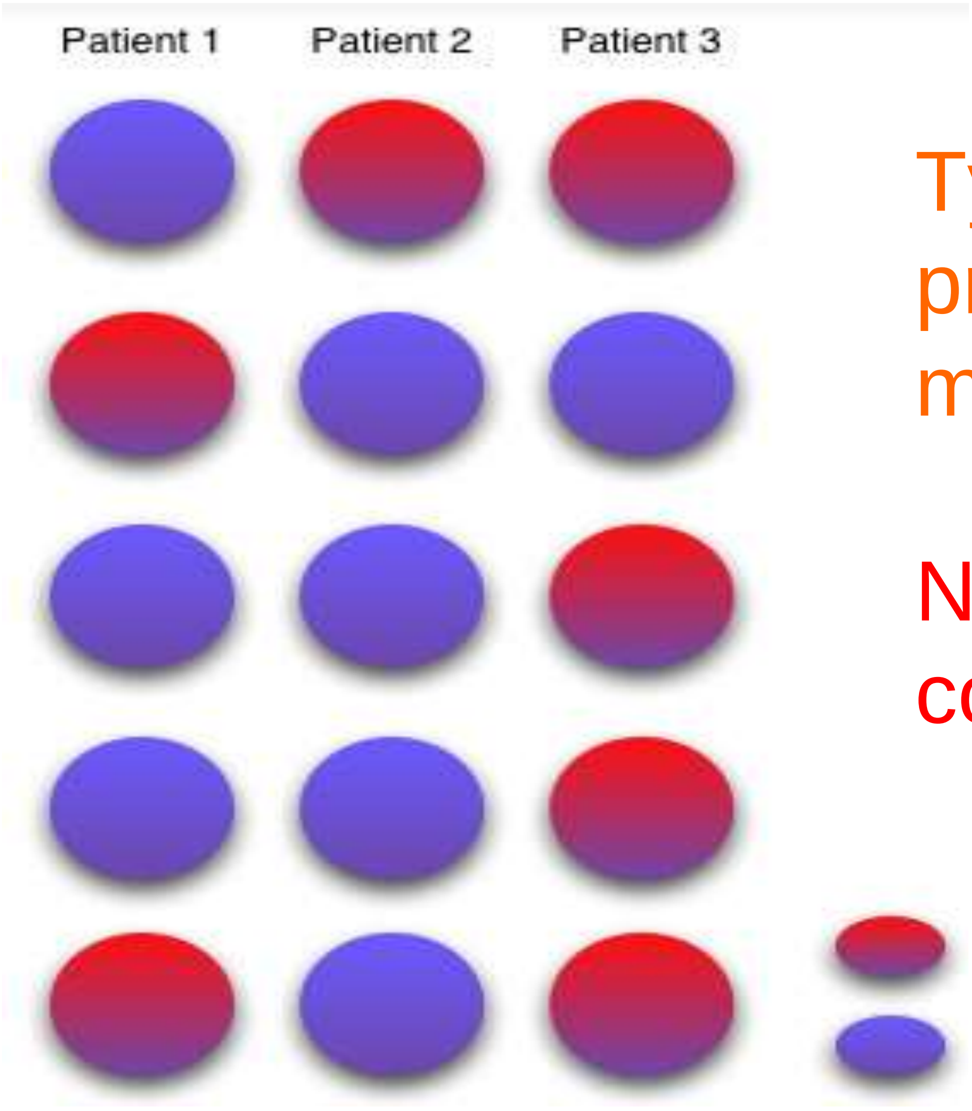
- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP, PDS
- **Improving coverage**
  - FCS,
  - CEA, PEP
  - Max Link



Peptide & protein identification by MS is still far from perfect

- **"… peptides with low scores are, nevertheless, often correct, so manual validation of such hits can often 'rescue' the identification of important proteins."**

Steen & Mann. **The ABC's and XYZ's of peptide sequencing**. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004



Detected protein

Present but undetected protein

Typical proteomic profiling misses many proteins

Need to improve coverage!

FCS

- **Rescue undetected proteins from high-scoring protein complexes**

- **Why?**

  Let A, B, C, D and E be the 5 proteins that function as a complex and thus are normally correlated in their expression. Suppose only A is not detected and all of B–E are detected. Suppose the screen has 50% reliability. Then, A’s chance of being false negative is 50%, & the chance of B–E all being false positives is $(50\%)^4$=6%. Hence, it is almost 10x more likely that A is false negative than B–E all being false positives.

- **Shortcoming: Databases of known complexes are still small**

Li et al. **Network-assisted protein identification and data interpretation in shotgun proteomics**. *Mol. Syst. Biol., 5*:303, 2009.

CEA

- **Generate cliques from PPIN**
- **Rescue undetected proteins from cliques with containing many high-confidence proteins**

- **Reason: Cliques in a PPIN often correspond to proteins at the core of complexes**

- **Shortcoming: Cliques are too strict**

⇒ **Use more power complex prediction methods**

Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *Journal of Proteome Research*, 10(5):2261--2272, May 2011

PEP

- **Map high-confidence proteins to PPIN**
- **Extract immediate neighbourhood & predict protein complexes using CFinder**
- **Rescue undetected proteins from high-ranking predicted complexes**

- **Reason: Exploit powerful protein complex prediction methods**

- **Shortcoming: Hard to predict protein complexes**
  - Do we need to know all the proteins a complex?

Goh et al. **A Network-based maximum-link approach towards MS identifies potentially important roles for undetected ARRB1/2 and ACTB in liver cancer progression.** *IJBRA*, 8(3/4):155-170, 2012

MaxLink

- **Map high-confidence proteins (“seeds”) to PPIN**
- **Identify proteins that talk to many seeds but few non-seeds**
- **Rescue these proteins**

- **Reason: Proteins interacting with many seeds are likely to be part of the same complex as these seeds**

- **Shortcoming: Likely to have more false-positives**

“Validation” of Rescued Proteins

- **Direct validation**
  - – Use the original mass spectra to verify the quality of the corresponding y- and b-ion assignments
  - – Immunological assay, etc.

- **Indirect validation**
  - – Check whether recovered proteins have GO terms that are enriched in the list of seeds
  - – Check whether recovered proteins show a pattern of differential expression betw disease vs normal samples that is similar to that shown by the seeds

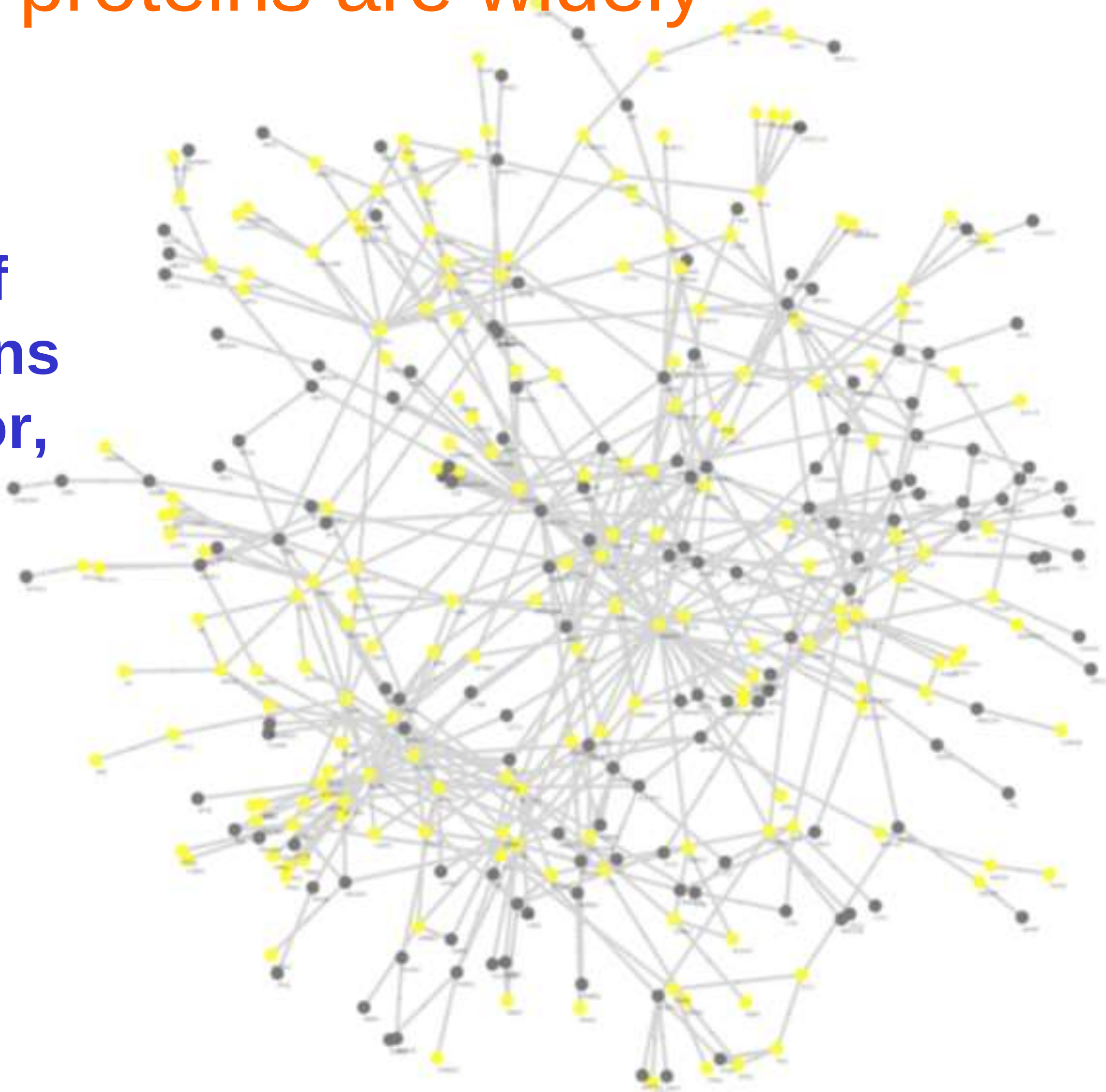
An example using the PEP approach to recover undetected proteins …

Background

- **HCC (Hepatocellular carcinoma)**
  - Classified into 3 phases: differentiated, moderately differentiated and poorly differentiated
- **Mass Spectrometry**
  - iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)
  - Coupled with 2D LC MS/MS
  - Popular because of ability to run 8 concurrent samples in one go

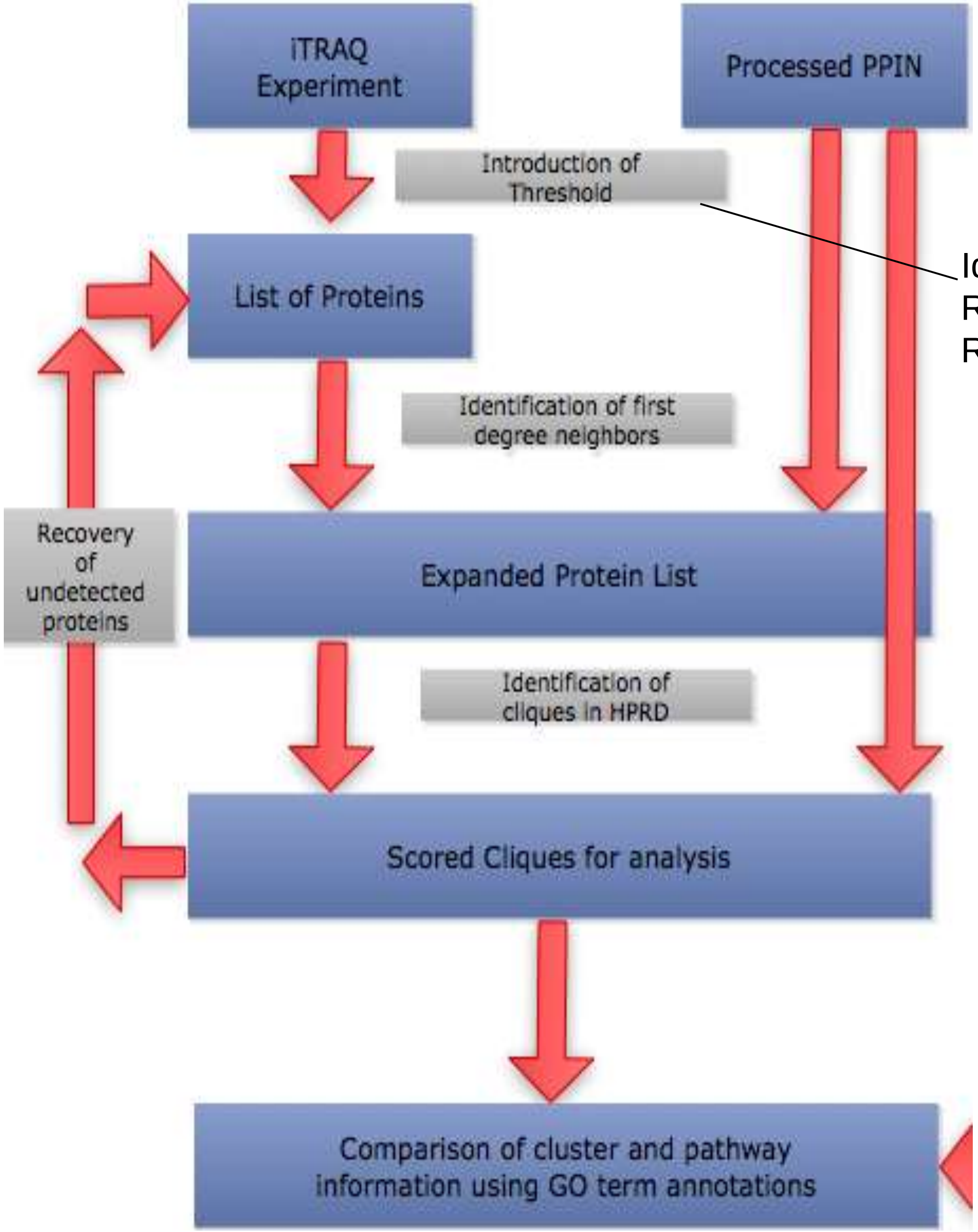
Poor and mod proteins are widely interspersed

- **In the subnet of reported proteins in mod and poor, poor and mod genes are well mixed**



● Mod and Poor

● Poor only


iTRAQ Experiment
Processed PPIN
Introduction of Threshold
List of Proteins
Identification of first degree neighbors
Recovery of undetected proteins
Expanded Protein List
Identification of cliques in HPRD
Scored Cliques for analysis
Comparison of cluster and pathway information using GO term annotations

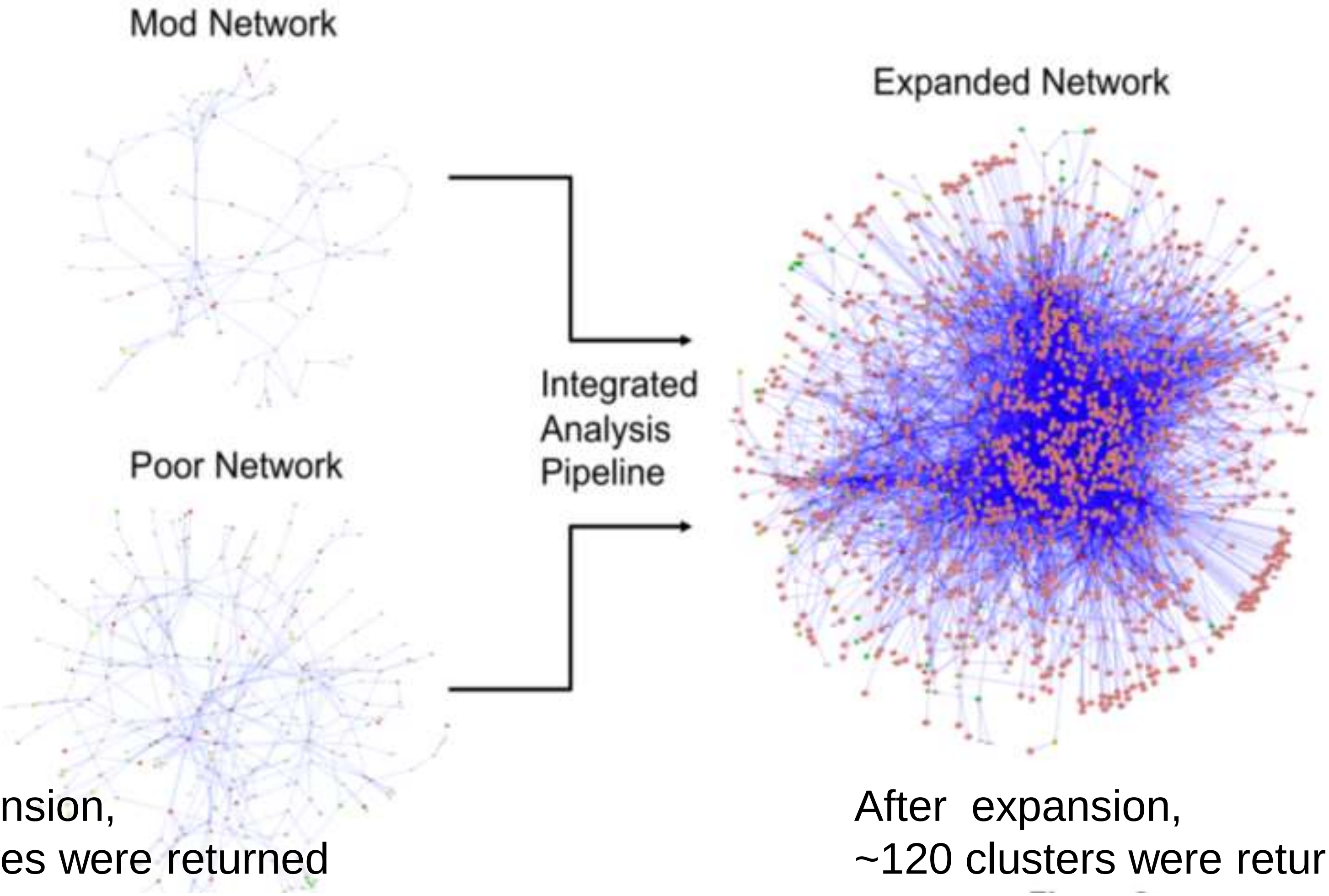
Identify the “seeds”
Ratio < 0.8 and > 1.25 for Mod (min 3 patients)
Ratio < 0.8 and > 1.25 for Poor (min 4 patients)

PEP Workflow

Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *Journal of Proteome Research*, 10(5):2261--2272, May 2011

Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Expansion to include neighbors greatly improves coverage



W/o expansion,
4 k3 cliques were returned

After expansion,
~120 clusters were returned

Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Returning to Mass Spectra

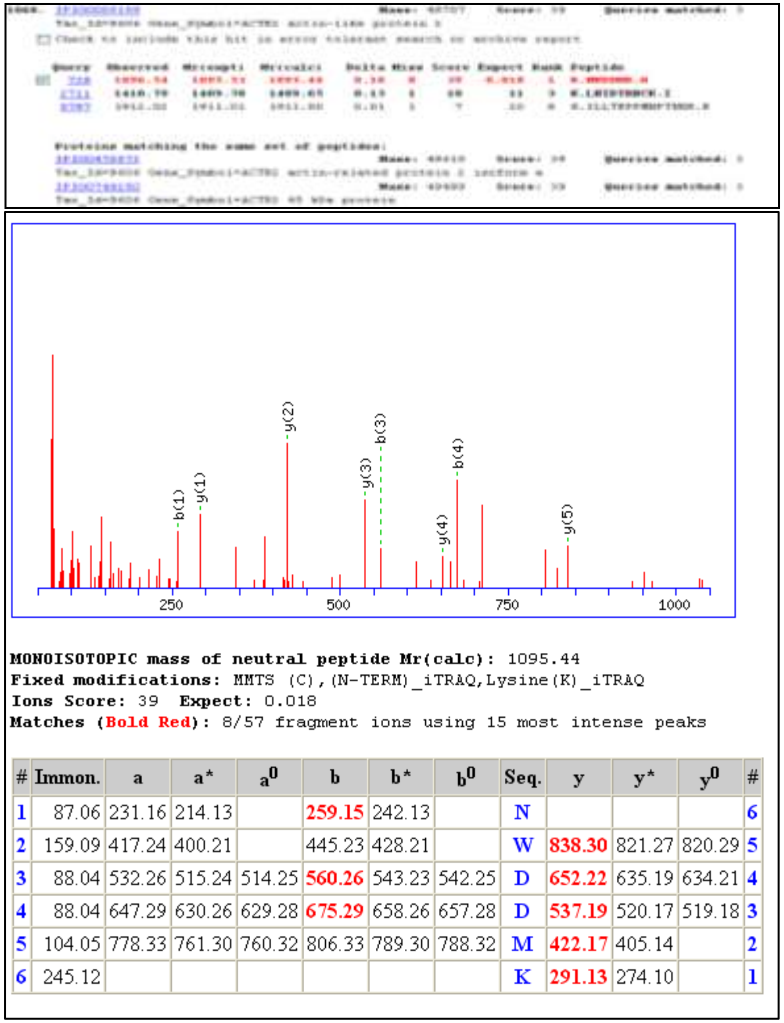
- **Test set: Several proteins (ACTR2, CDC42, GNB2L1, KIF5B, PPP2R1A, PKACA and TOP1) from top 34 clusters not detected by Paragon**
- **The test: Examine their GPS and Mascot search results and their MS/MS-to-peptide assignments**
- **Assessment of MS/MS spectra of their top ranked peptides revealed accurate y- and b-ion assignments and were of good quality ($p < 0.05$)**

⇒ In silico expansion verified

Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *Journal of Proteome Research*, 10(5):2261--2272, 2011

Successful Verification

ACTR2

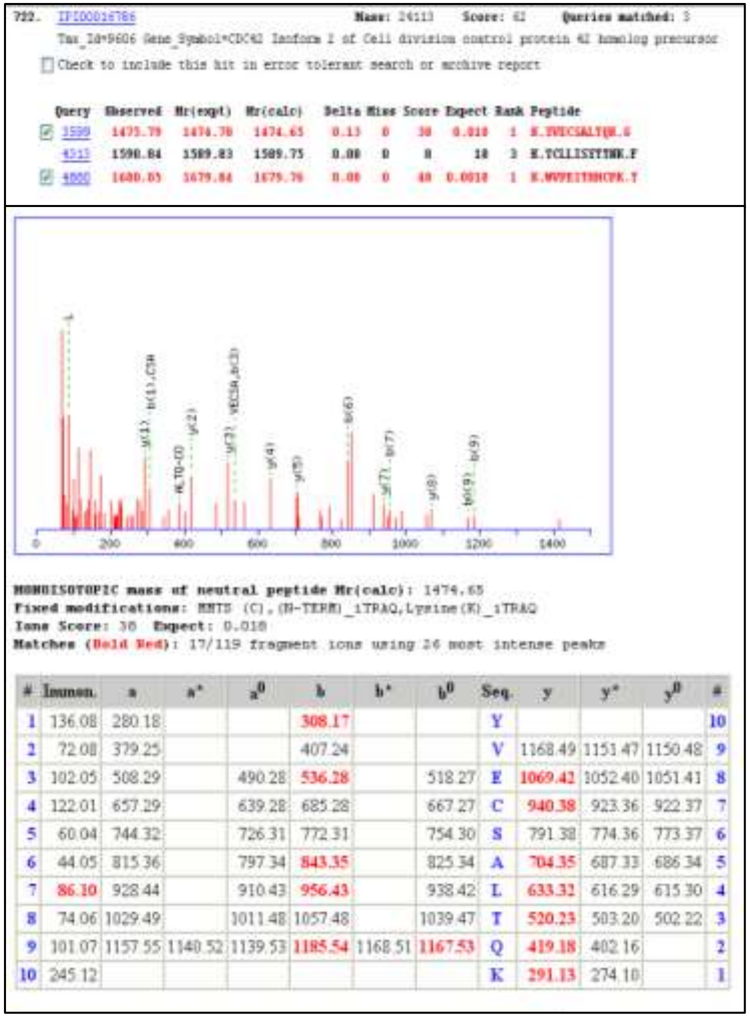


```
[illegible]   IPI[illegible]                                  Mass: [illegible]   Score: [illegible]   Queries matched: [illegible]
     Tax_Id=9606 Gene_Symbol=ACTR2 [illegible]
□ Check to include this hit in error tolerant search or archive report

     Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
☐    [illegible]
     [illegible]
     [illegible]

     Proteins matching the same set of peptides:
     IPI[illegible]                                   Mass: [illegible]   Score: [illegible]   Queries matched: [illegible]
     Tax_Id=9606 Gene_Symbol=ACTR2 actin-related protein 2 isoform a
     IPI[illegible]                                   Mass: [illegible]   Score: [illegible]   Queries matched: [illegible]
     Tax_Id=9606 Gene_Symbol=ACTR2 [illegible] protein
```

```
MONOISOTOPIC mass of neutral peptide Mr(calc): 1095.44
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 39  Expect: 0.018
Matches (Bold Red): 8/57 fragment ions using 15 most intense peaks
```

#	Immon.	a	a*	a^0	b	b*	b^0	Seq.	y	y*	y^0	#
1	87.06	231.16	214.13		**259.15**	242.13		N				6
2	159.09	417.24	400.21		445.23	428.21		W	**838.30**	821.27	820.29	5
3	88.04	532.26	515.24	514.25	**560.26**	543.23	542.25	D	**652.22**	635.19	634.21	4
4	88.04	647.29	630.26	629.28	**675.29**	658.26	657.28	D	**537.19**	520.17	519.18	3
5	104.05	778.33	761.30	760.32	806.33	789.30	788.32	M	**422.17**	405.14		2
6	245.12							K	**291.13**	274.10		1

CDC42



```
722.  IPI00016786                                 Mass: 24113    Score: 62    Queries matched: 3
      Tax_Id=9606 Gene_Symbol=CDC42 Isoform 1 of Cell division control protein 42 homolog precursor
  □ Check to include this hit in error tolerant search or archive report

      Query  Observed  Mr(expt)  Mr(calc)   Delta Miss Score Expect Rank Peptide
  ☑   3599   1475.79   1474.78   1474.65    0.13    0    38  0.018     1 K.YVECSALTQK.G
      4513   1590.84   1589.83   1589.75    0.08    0     8     18     3 K.TCLLISYTTNK.F
  ☑   4800   1680.85   1679.84   1679.76    0.08    0    48  0.0018    1 K.WVPEITHHCPK.T
```

```
MONOISOTOPIC mass of neutral peptide Mr(calc): 1474.65
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 38  Expect: 0.018
Matches (Bold Red): 17/119 fragment ions using 26 most intense peaks
```

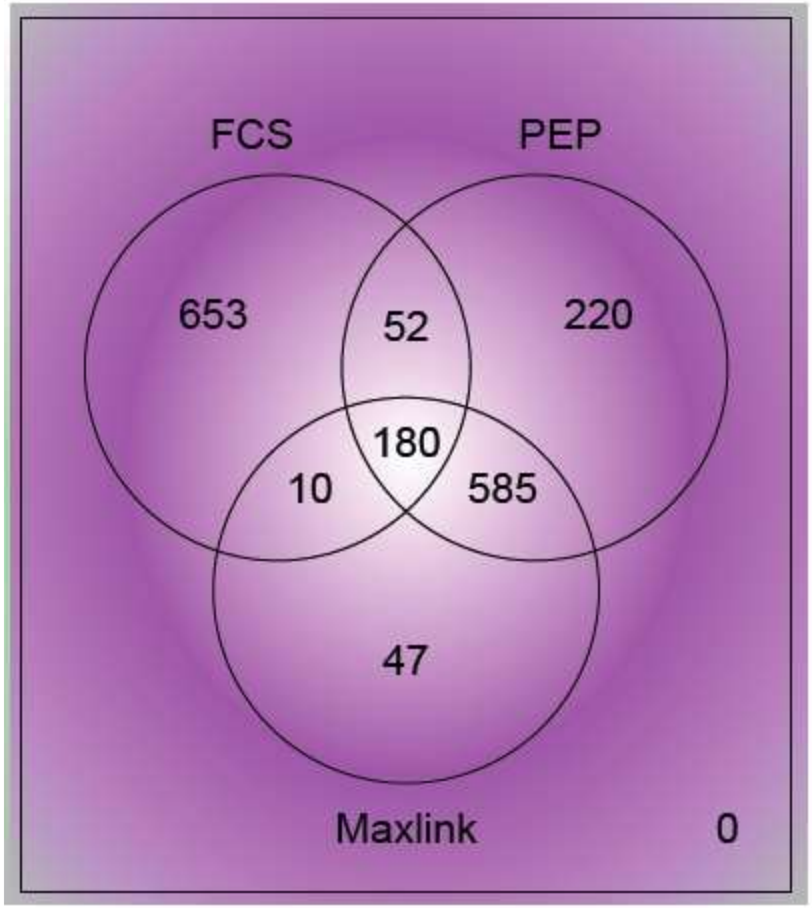
#	Immon.	a	a*	a^0	b	b*	b^0	Seq.	y	y*	y^0	#
1	136.08	280.18			**308.17**			Y				10
2	72.08	379.25			407.24			V	1168.49	1151.47	1150.48	9
3	102.05	508.29		490.28	**536.28**		518.27	E	**1069.42**	1052.40	1051.41	8
4	122.01	657.29		639.28	685.28		667.27	C	**940.38**	923.36	922.37	7
5	60.04	744.32		726.31	772.31		754.30	S	791.38	774.36	773.37	6
6	44.05	815.36		797.34	**843.35**		825.34	A	**704.35**	687.33	686.34	5
7	**86.10**	928.44		910.43	**956.43**		938.42	L	**633.32**	616.29	615.30	4
8	74.06	1029.49		1011.48	1057.48		1039.47	T	**520.23**	503.20	502.22	3
9	101.07	1157.55	1140.52	1139.53	**1185.54**	1168.51	**1167.53**	Q	**419.18**	402.16		2
10	245.12							K	**291.13**	274.10		1

Another Experiment

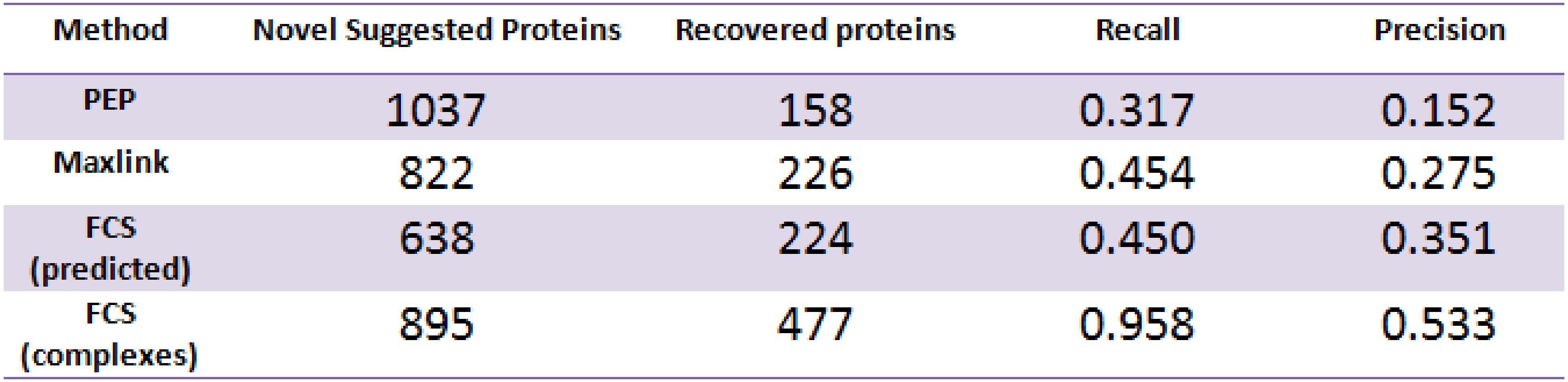
- **Valporic acid (VPA)-treated mice vs control**
  - VPA or vehicle injected every 12 hours into postnatal day-56 adult mice for 2 days
  - Role of VPA in epigenetic remodeling
- **MS was scanned against IPI rat db in round #1**
  - 291 proteins identified
- **MS was scanned against UniProtkb in round #2**
  - 498 additional proteins identified
- **All recovery methods ran on round #1 data and the recovered proteins checked against round #2**

Moderate level of agreement of reported proteins between various recovery methods

FCS (Real Complexes)

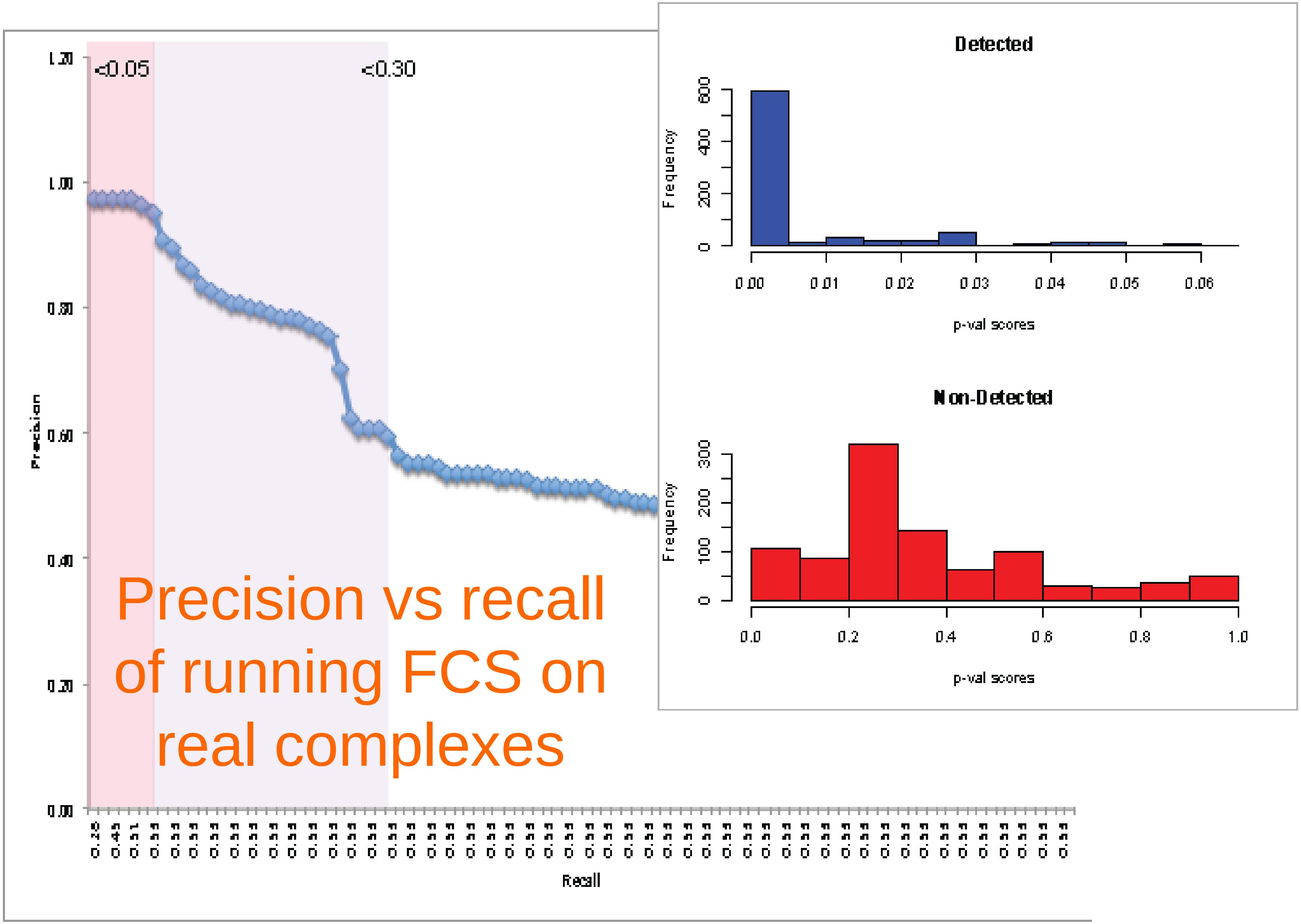


Performance Comparison



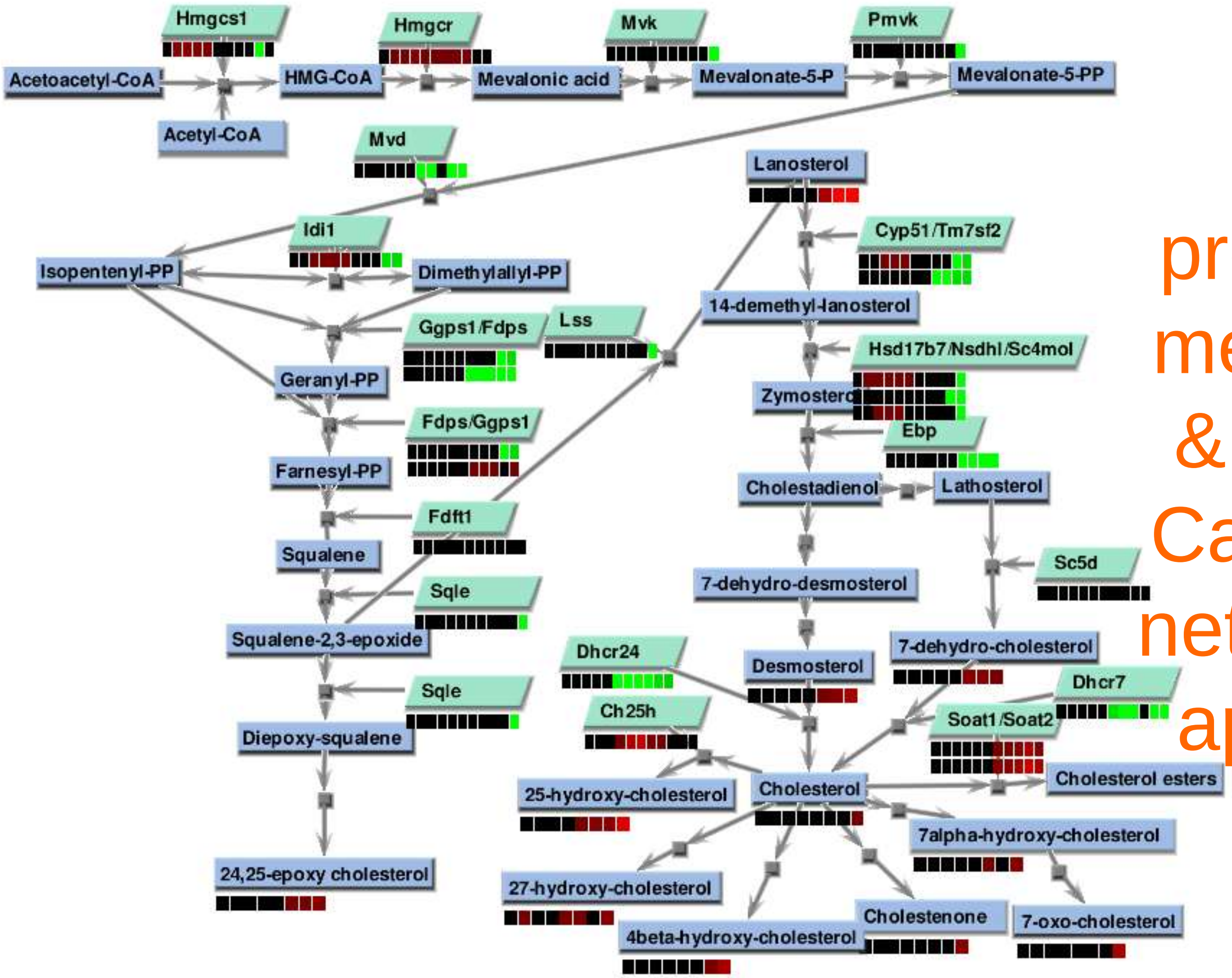
Method	Novel Suggested Proteins	Recovered proteins	Recall	Precision
PEP	1037	158	0.317	0.152
Maxlink	822	226	0.454	0.275
FCS (predicted)	638	224	0.450	0.351
FCS (complexes)	895	477	0.958	0.533

- **Looks like running FCS on real complexes is able to recover more proteins and more accurately**



Precision vs recall of running FCS on real complexes

NUS National University of Singapore



From proteomics to metabolomics & lipidomics: Can the same network-based approach be applied?

References

- Käll & Vitek. **Computational Mass Spectrometry–Based Proteomics**. *PLoS Comput Biol* , 7(12): e1002277, 2011
- Goh et al. **How advancement in biological network analysis methods empowers proteomics**. *Proteomics*, 12(4-5):550-563, 2012
- [PSP] Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *J Proteome Research*. 11(3):1571-1581, 2012
- [CEA] Li et al. **Network-assisted protein identification and data interpretation in shotgun proteomics**. *Mol. Syst. Biol., 5:*303, 2009.
- [PEP] Goh et al. **A Network-based pipeline for analyzing MS data---An application towards liver cancer**. *J Proteome Research*, 10(5):2261-2272, 2011
- [FCS] Goh et al. **Comparative network-based recovery analysis and proteomic profiling of neurological changes in valproic acid-treated mice.** *J Proteome Research*, 12(5):2116-2127, 2013

Use of Context in Gene Expression and Proteomic Profile Analysis

Part 3

Limsoon Wong

Types of Biological Networks

- **Natural biological pathways**
  - Metabolic pathway
  - Gene regulation network
  - Cell signaling network

- **Protein-protein interaction networks**

Metabolic Pathway

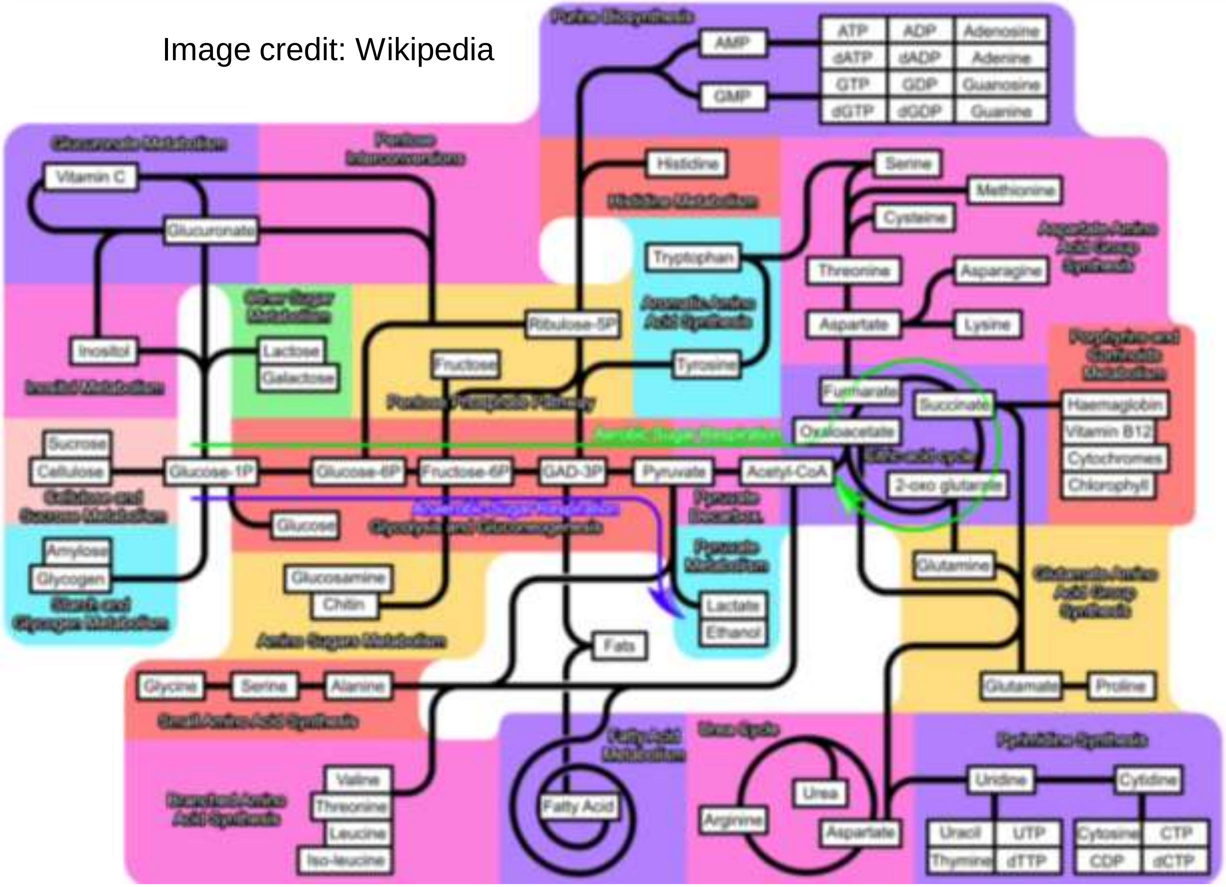


Image credit: Wikipedia

- **A series of biochem reactions in a cell**
  - Catalyzed by enzymes
  - Step-by-step modification of an initial molecule to form another product that can
    - **be used /store in the cell**
    - **initiate another metabolic pathway**

Gene Regulation Network

- **Gene regulation is the process that turns info from genes into gene products**
- **Gives a cell control over its structure & function**
  - Cell differentiation
  - Morphogenesis
  - Adaptability, …

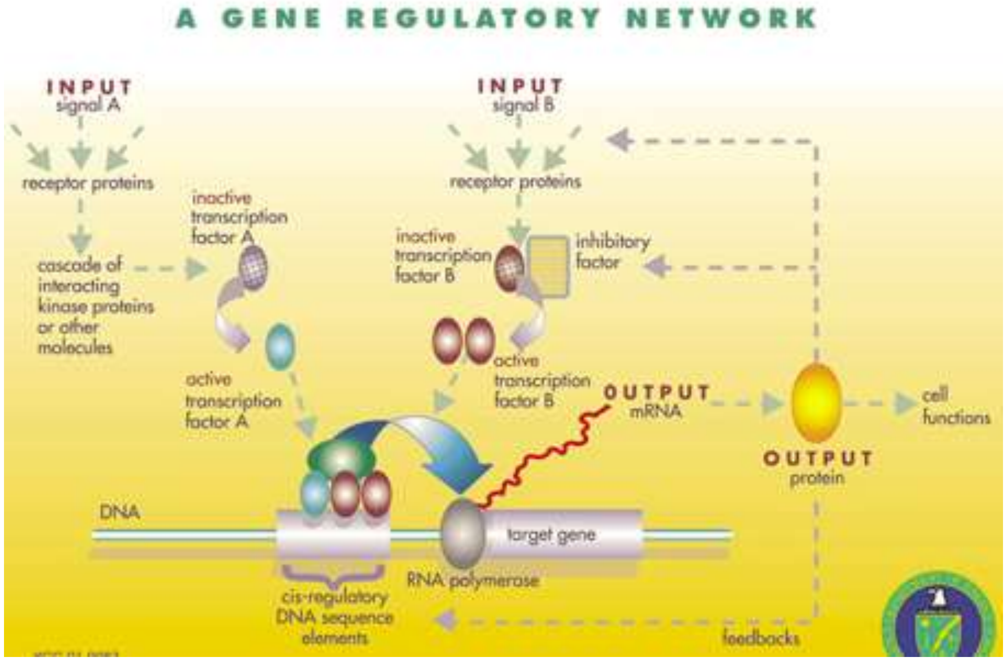

A GENE REGULATORY NETWORK
INPUT signal A
INPUT signal B
receptor proteins
receptor proteins
inactive transcription factor A
cascade of interacting kinase proteins or other molecules
inactive transcription factor B
inhibitory factor
active transcription factor A
active transcription factor B
OUTPUT mRNA
cell functions
OUTPUT protein
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks

Image credit: Genome to Life

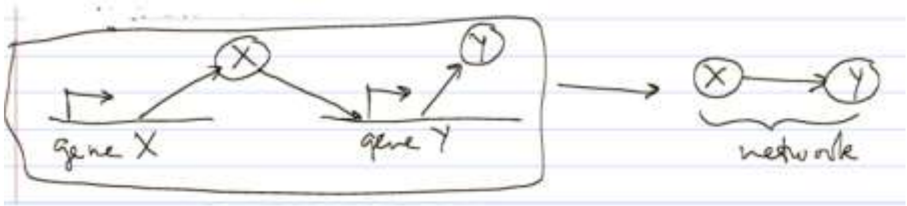


Image credit: Natasa Przulj

Cell Signaling Network

- **It is the entire set of changes induced by receptor activation**
  - Governs basic cellular activities and coordinates cell actions
- **Cells communicate with each other**
  - Direct contact (juxtacrine signaling)
  - Short distances (paracrine signaling)
  - Large distances (endocrine signaling)
- **Errors result in cancer, diabetes, ...**

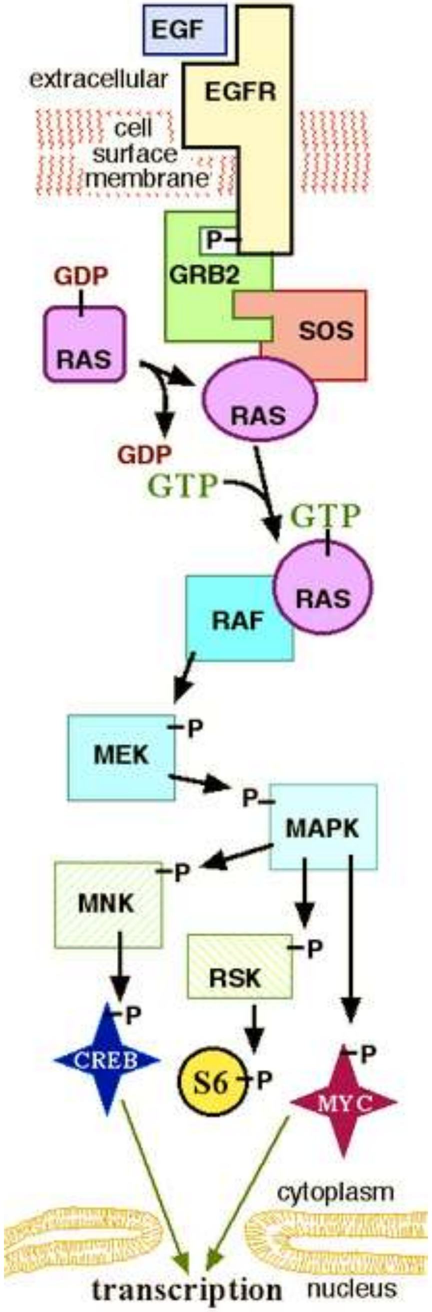
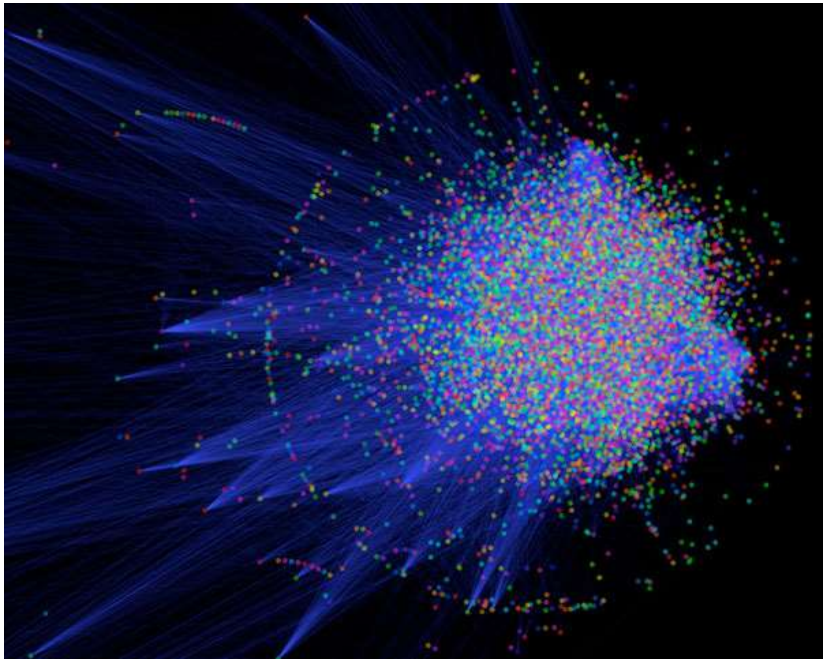


Image credit: Wikipedia

Protein Interaction Network (PPIN)

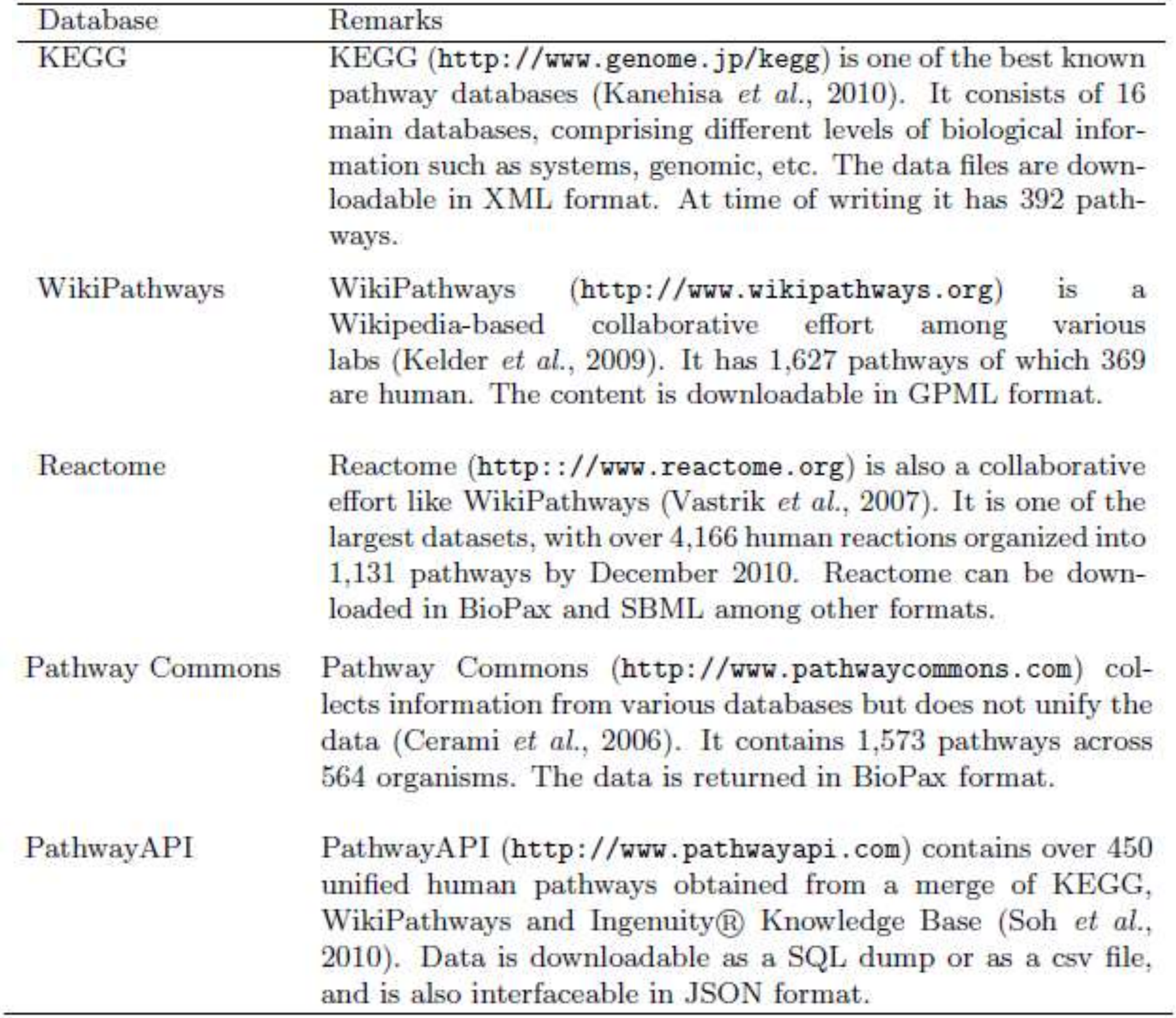
- **PPI usual refers to physical binding between proteins**
  - Stable interaction
    - **Protein complex**
    - **~70% of PPIs**
  - Transient interaction, modifying a protein for further actions
    - **Phosphorylation**
    - **Transportation**
    - **~30% of PPIs**



Visualization of the human interactome.
Image credit: Wikepedia

- **PPIN is usually a set of PPIs; it is not put into biological context**

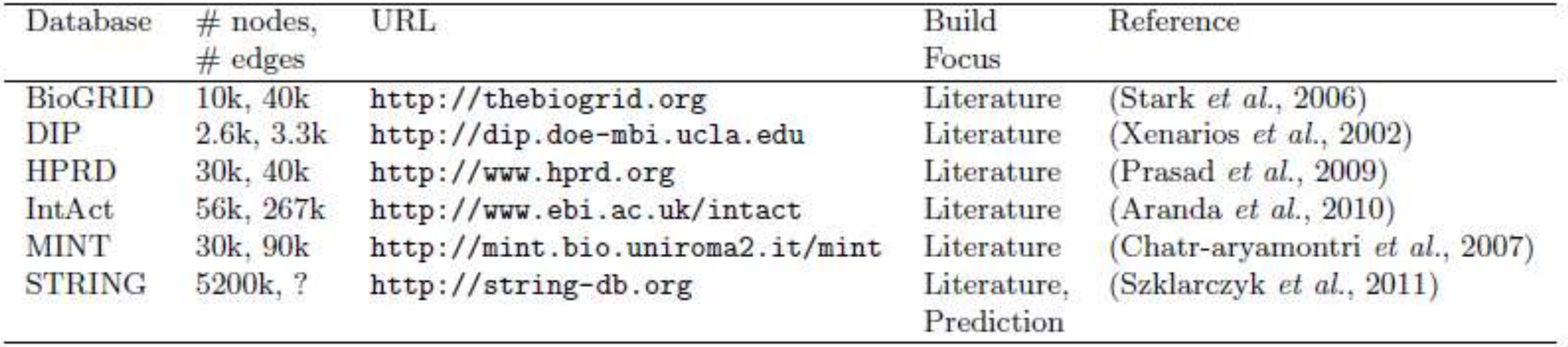
Sources of Biological Pathways



Database	Remarks
KEGG	KEGG (http://www.genome.jp/kegg) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (http://www.wikipathways.org) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (http:://www.reactome.org) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (http://www.pathwaycommons.com) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (http://www.pathwayapi.com) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

Source: Goh et al. “How advancement in biological network analysis methods empowers proteomics”. *Proteomics*, accepted.

Sources of Protein Interactions



Database	# nodes, # edges	URL	Build Focus	Reference
BioGRID	10k, 40k	http://thebiogrid.org	Literature	(Stark *et al.*, 2006)
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	(Xenarios *et al.*, 2002)
HPRD	30k, 40k	http://www.hprd.org	Literature	(Prasad *et al.*, 2009)
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	(Aranda *et al.*, 2010)
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	(Chatr-aryamontri *et al.*, 2007)
STRING	5200k, ?	http://string-db.org	Literature, Prediction	(Szklarczyk *et al.*, 2011)

Source: Goh et al. “How advancement in biological network analysis methods empowers proteomics”. *Proteomics*, accepted.

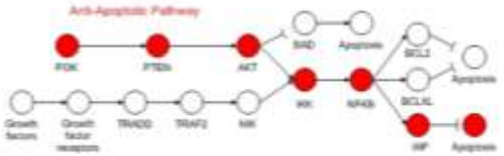
and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010

Sivachenko et al. *JBCB*, 5(2b):429-546, 2007

Gene Expression Profile Analysis

Gene Regulatory Circuits



- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype

- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype

Cop... 2011 © Limsoon W...

Contextualization!

Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- What is the chance of a group of 5 genes being perfectly correlated to these samples?

- Prob(group of genes correlated) = $(1/2^6)^5$
  - Good, << $1/2^6$
- ~~# of groups = $^{10000}C_5$~~
- ~~E(# of groups of genes correlated) = $^{10000}C_5$ * $(1/2^6)^5$ = $2.6*10^{12}$~~

⇒ Even more false positives?

- Perhaps no need to consider every group

# of pathways = 1000

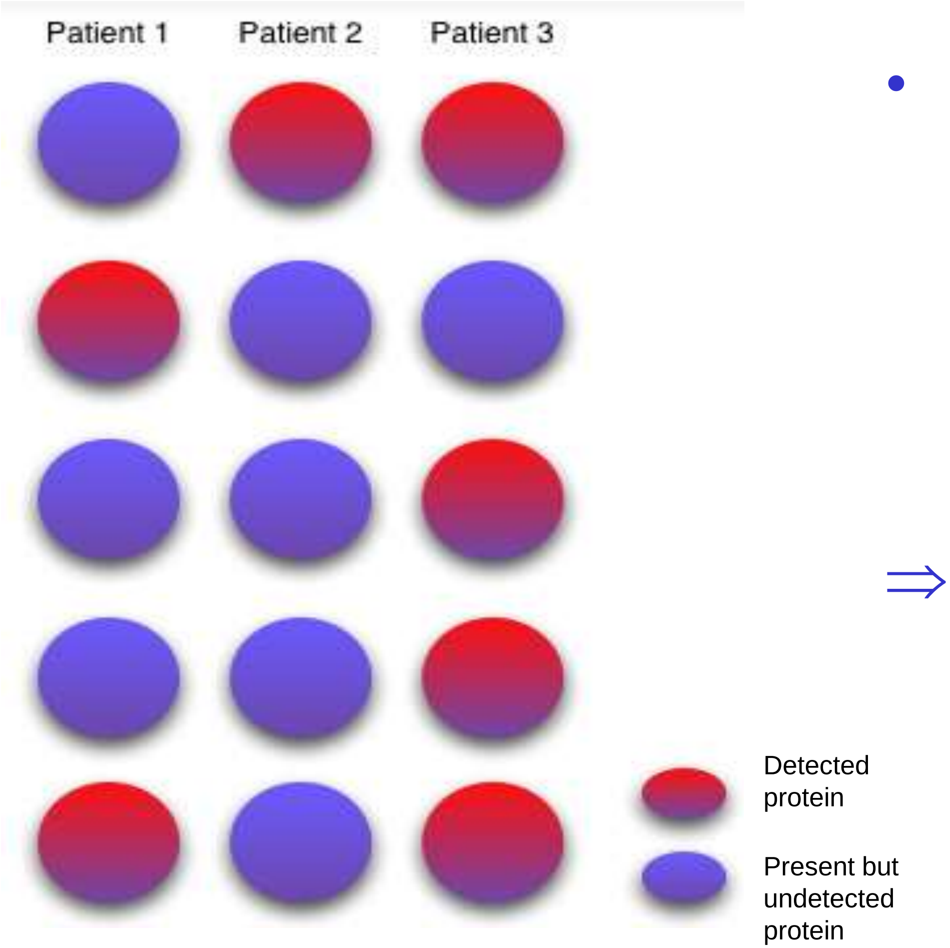
E(# of pathways correlated) = 1000 * $(1/2^6)^5$ = $9.3*10^{-7}$

Microarray Workshop for Gene Expression Profiling, NUH, 23/8/2011

Copyright 2011 © Limsoon Wong

Copyright 2011 © Limsoon Wong

Proteomic Profile Analysis



Goh et al. How advancement in biological network analysis methods empowers proteomics. *Proteomics*, in press

- **Suppose the failure to form a protein complex causes a disease**
  - If any component protein is missing, the complex can’t form

⇒ **Diff patients suffering from the disease can have a diff protein component missing**

- Construct a profile based on complexes?

Emily et al. *European Journal of Human Genetics*, 17(10):1231-1240, 2009.

Epistatic Interaction Mining

- **GWAS have linked many SNPs to diseases, but many genetic risk factors still unaccounted for**
- **Proteins coded by genes interact in cell**

⇒ **Some SNPs affect the phenotype in combination with other SNPs; i.e., epistasis**

- **Exhaustive search for epistatic effects has to test many combinations (>100,000^2) of SNPs**
  - Hard to get statistical significance
  - Take long time to run on computers

⇒ **Use biological networks to narrow the search for two-locus epistasis**

Vanunu et al. *PLoS Computational Biology*, 6(1):e1000641, 2010

Disease Causal Gene Prioritization

- **Genes causing the same or similar diseases tend to lie close to one another in PPIN**

- **Given disease Q. Look for proteins in PPIN interacting with many causal genes of diseases similar to Q**

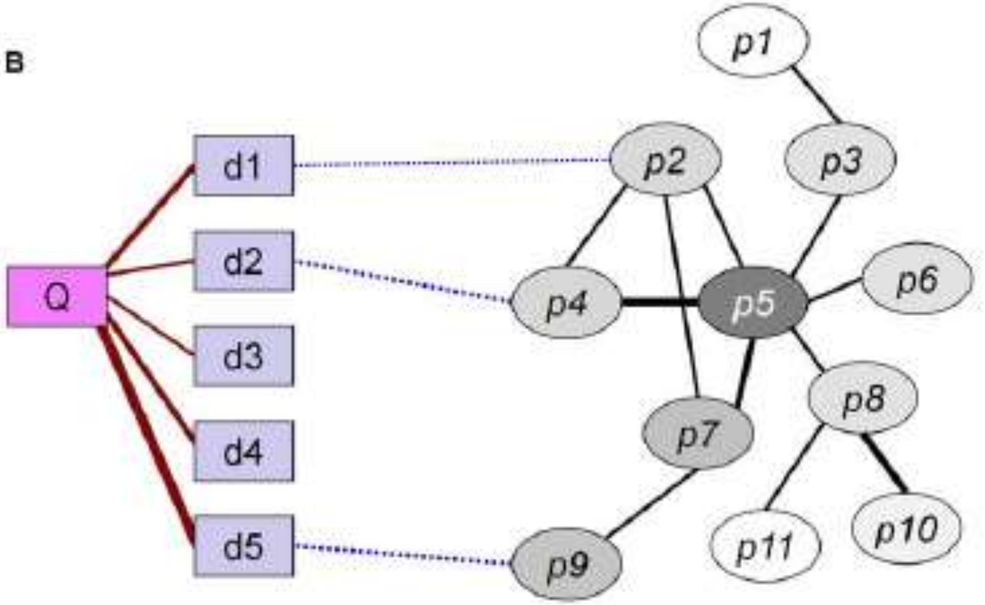
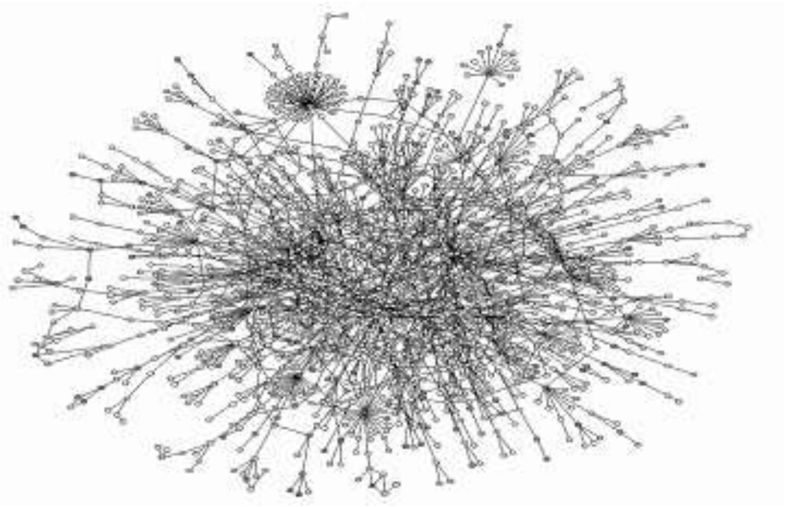


Figure 1. Illustration of the PRINCE algorithm. A query disease, denoted *Q*, has varying degrees of phenotypic similarity with other diseases, denoted *d1–d5* (marked with maroon lines, where thicker lines represent higher similarity). Known causal genes for these similar diseases are connected by dashed blue lines and used as the prior information. *p1–p11* comprise the protein set of a protein-protein interaction network, where interactions are marked with black lines and thicker lines denote edges with higher confidence. A scoring function that is smooth over the network is computed using an iterative network propagation method. At every iteration of the algorithm, each protein pumps flow to its neighbors and receives flow from them. Protein colors correspond to the flow they receive in a specific iteration, the darker the color the higher the flow. (A):

Liu et al. *Bioinformatics*, 25(15):1891-1897, 2009

Protein Complex Prediction

- **Nature of high-throughput PPI expts**
  - Proteins are taken out of their natural context!



- **Can a protein interact with so many proteins simultaneously?**

- **A big "hub" and its "spokes" should probably be decomposed into subclusters**
  - Each subcluster is a set proteins that interact in the same space &time; viz., a protein complex

- **Many complexes have highly connected cores in PPIN ➔ Find complexes by clustering**
- **Issue: How to identify low edge density complexes?**

Wong. *Internet Mathematics*, 7(4):274--298, 2011

Protein Function Prediction

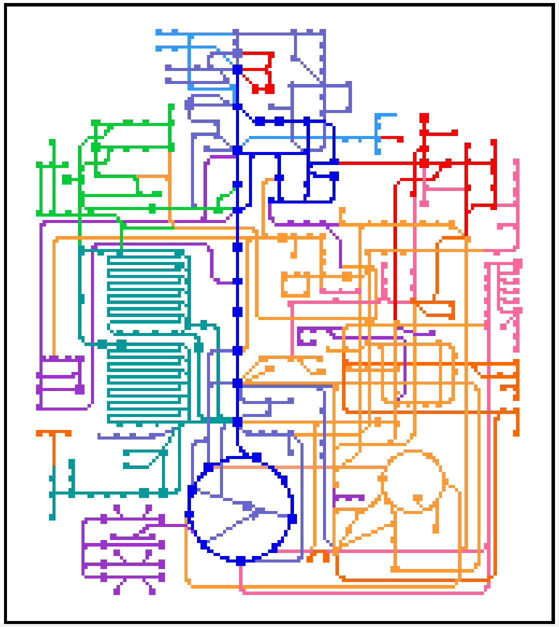
- **Proteins with similar function are topologically close in PPIN**
  - Direct functional association
  - Indirect functional association

A pair of proteins that participate in the same cellular processes or localize to the same cellular compartment are many times more likely to interact than a random pair of proteins

- **Proteins with similar function have interaction neighborhoods that are similar**

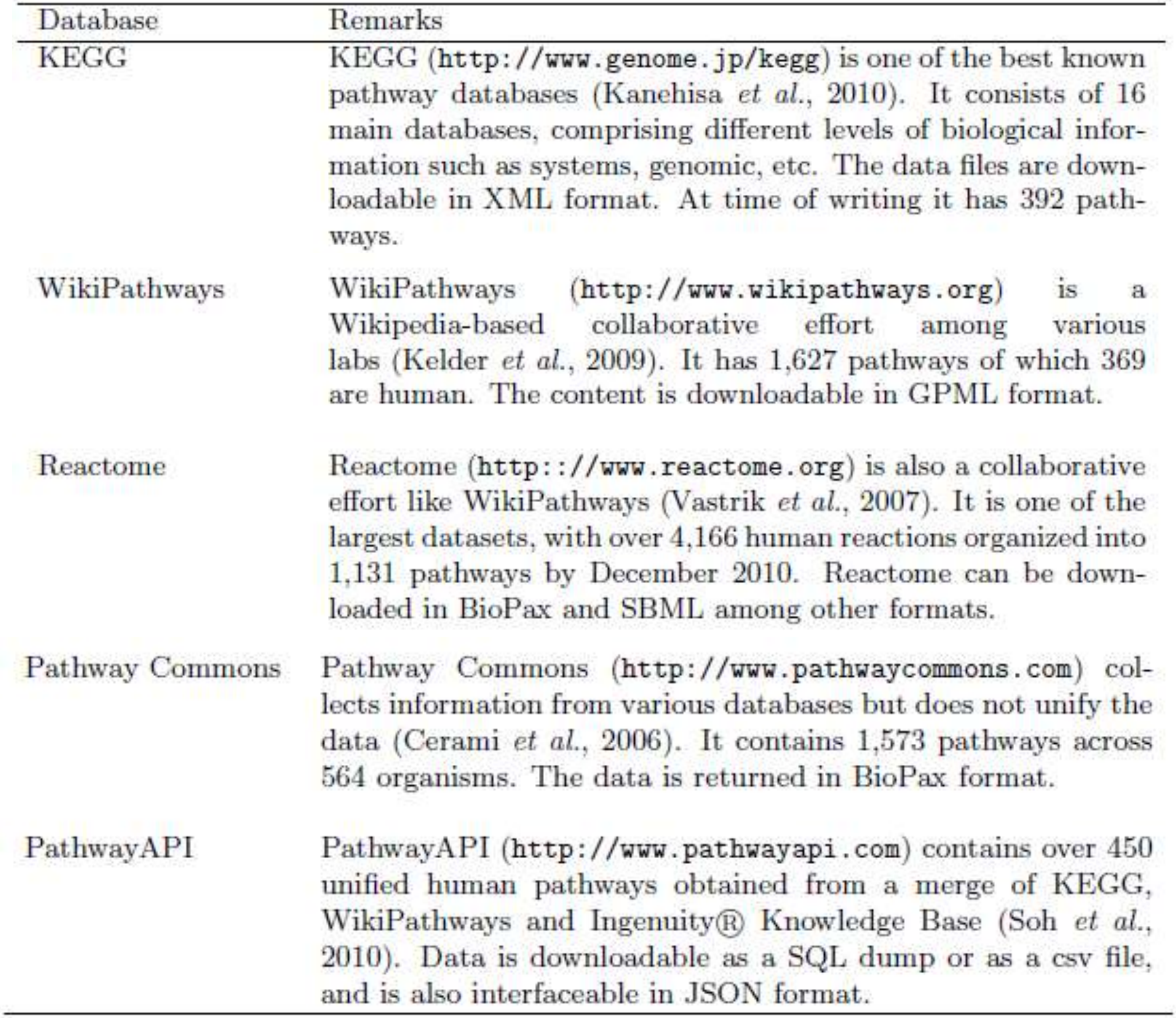
When proteins in the neighborhood of a protein X have similar functions to proteins in the neighborhood of a protein Y, then proteins X & Y likely operate in similar environment

Part 3: How good are available sources of pathway & PPI Network?



- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**

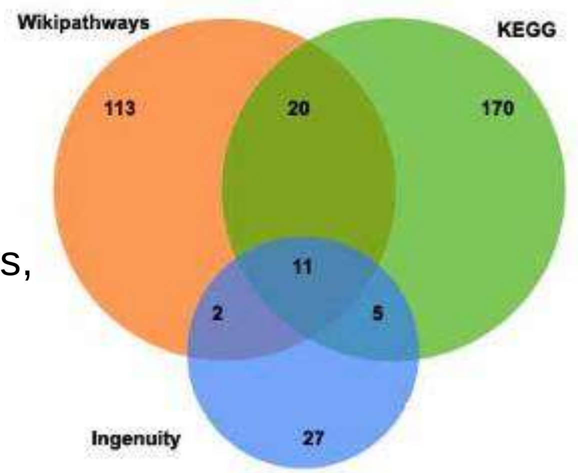
Major Sources of Biological Pathways

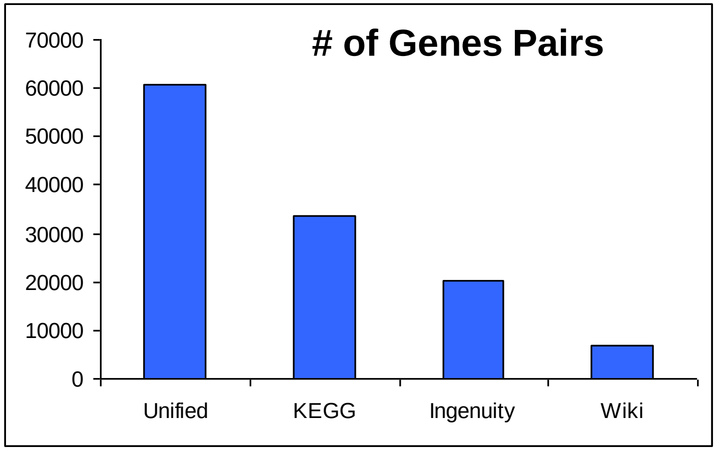


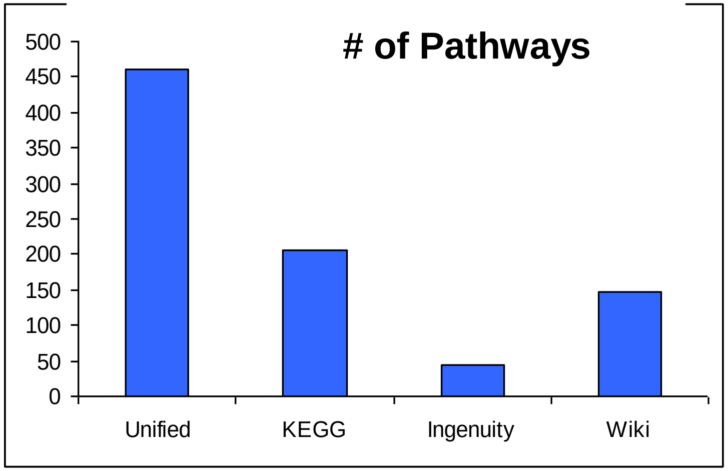
Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.
WikiPathways	WikiPathways (`http://www.wikipathways.org`) is a Wikipedia-based collaborative effort among various labs (Kelder *et al.*, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.
Reactome	Reactome (`http:://www.reactome.org`) is also a collaborative effort like WikiPathways (Vastrik *et al.*, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.
Pathway Commons	Pathway Commons (`http://www.pathwaycommons.com`) collects information from various databases but does not unify the data (Cerami *et al.*, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.
PathwayAPI	PathwayAPI (`http://www.pathwayapi.com`) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh *et al.*, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.

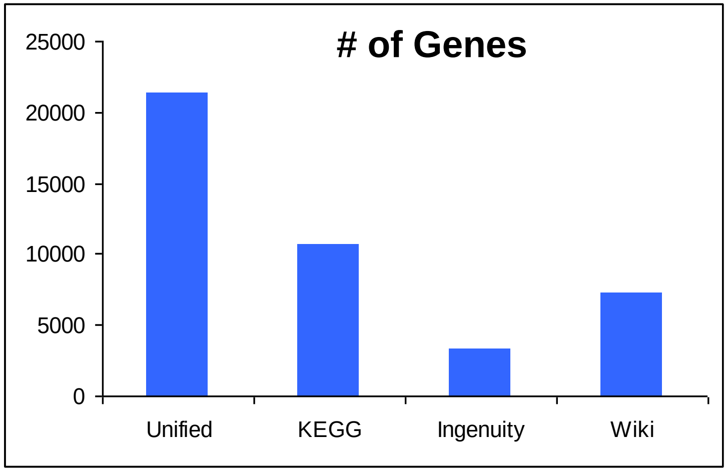
Low Comprehensiveness of Human Pathway Sources

Human pathways in Wikipathways, KEGG, & Ingenuity





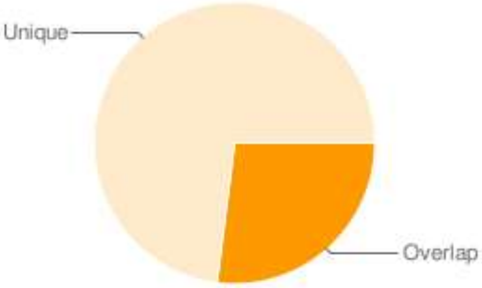




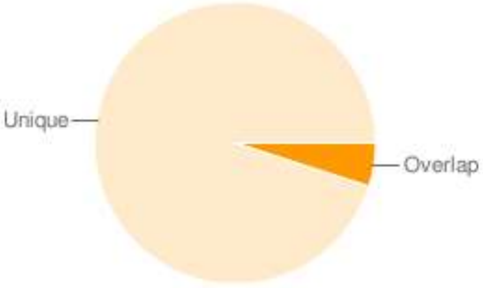
Soh et al. Consistency, Comprehensiveness, and Compatibility of Pathway Databases. *BMC Bioinformatics*, 11:449, 2010.

Low Consistency of Human Pathway Sources

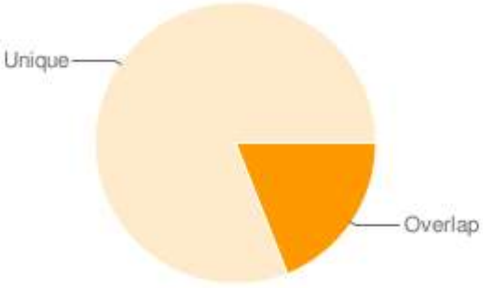
Gene Pair Overlap



Wiki vs KEGG

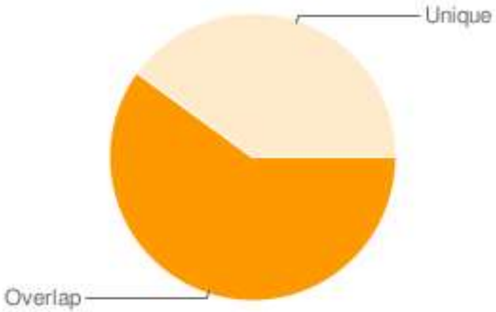


Wiki vs Ingenuity

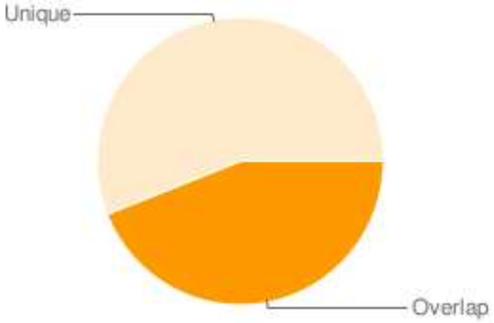


KEGG vs Ingenuity

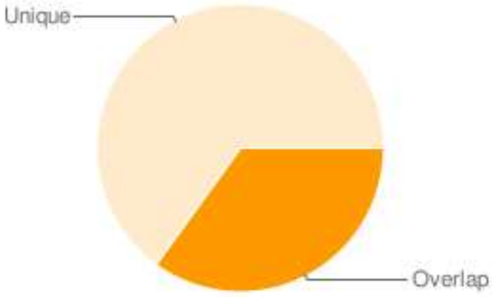
Gene Overlap



Wiki vs KEGG



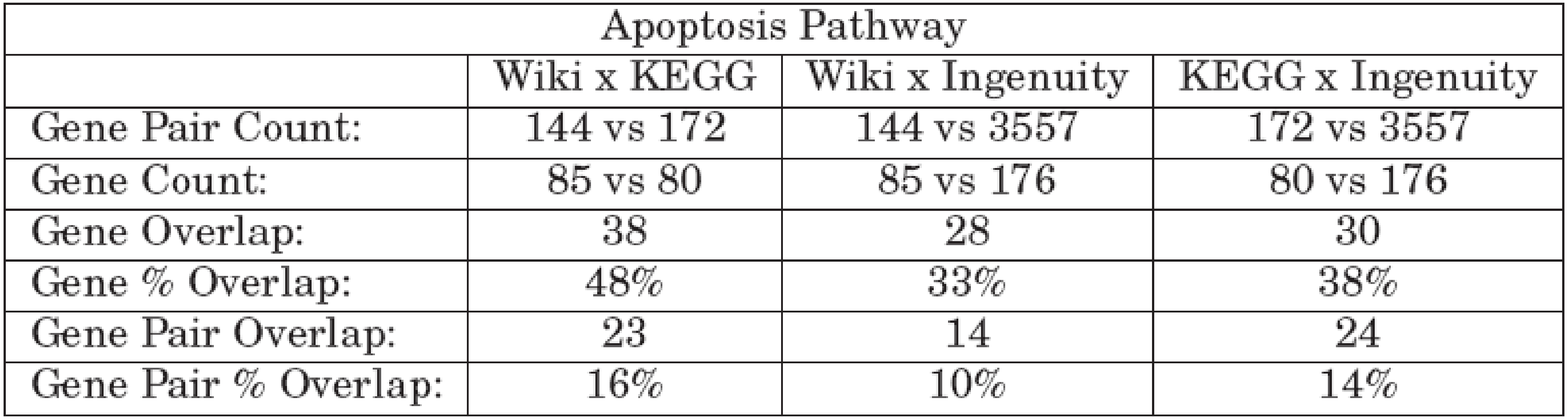
Wiki vs Ingenuity



KEGG vs Ingenuity

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

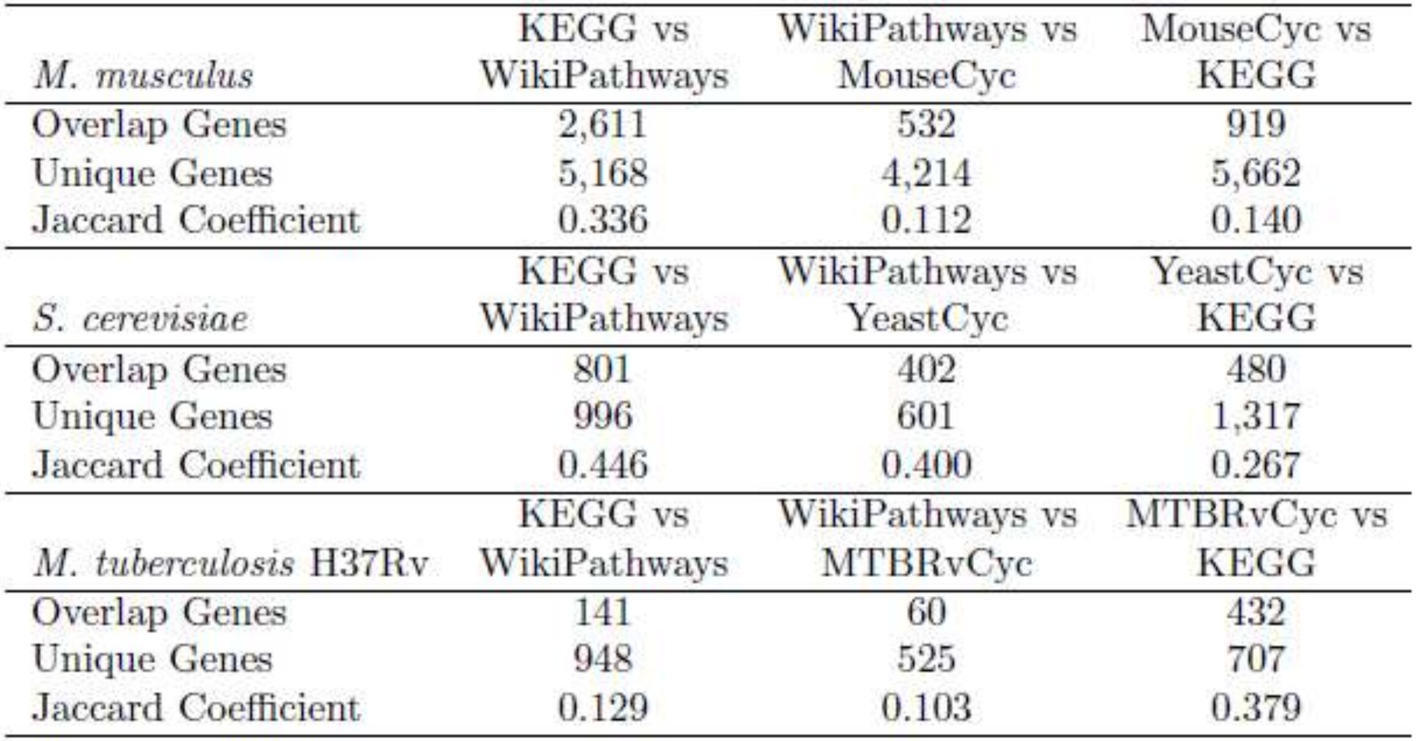
Example: Human Apoptosis Pathway



Apoptosis Pathway			
	Wiki x KEGG	Wiki x Ingenuity	KEGG x Ingenuity
Gene Pair Count:	144 vs 172	144 vs 3557	172 vs 3557
Gene Count:	85 vs 80	85 vs 176	80 vs 176
Gene Overlap:	38	28	30
Gene % Overlap:	48%	33%	38%
Gene Pair Overlap:	23	14	24
Gene Pair % Overlap:	16%	10%	14%

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

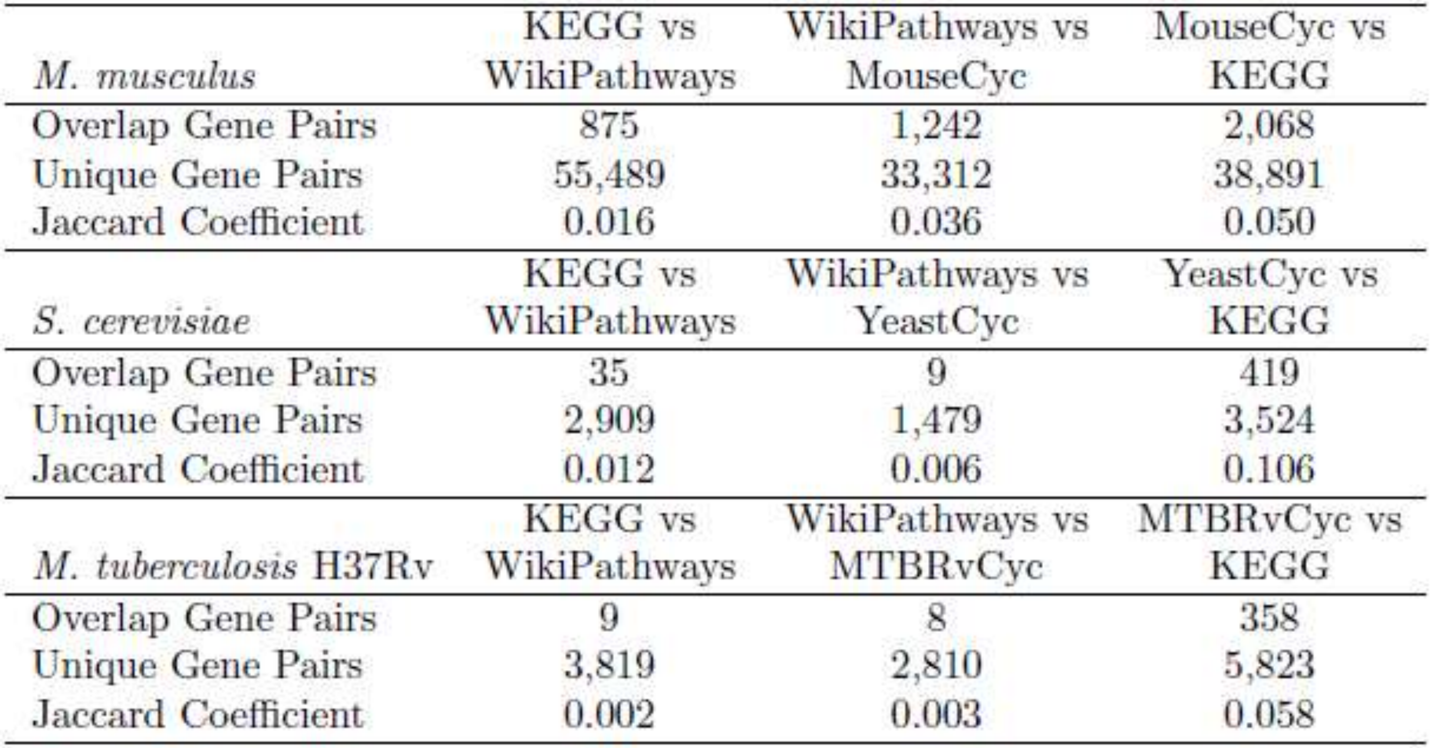
The same low inter-database consistency (in gene overlap) is observed in pathways of other organisms



M. musculus	KEGG vs WikiPathways	WikiPathways vs MouseCyc	MouseCyc vs KEGG
Overlap Genes	2,611	532	919
Unique Genes	5,168	4,214	5,662
Jaccard Coefficient	0.336	0.112	0.140
S. cerevisiae	KEGG vs WikiPathways	WikiPathways vs YeastCyc	YeastCyc vs KEGG
Overlap Genes	801	402	480
Unique Genes	996	601	1,317
Jaccard Coefficient	0.446	0.400	0.267
M. tuberculosis H37Rv	KEGG vs WikiPathways	WikiPathways vs MTBRvCyc	MTBRvCyc vs KEGG
Overlap Genes	141	60	432
Unique Genes	948	525	707
Jaccard Coefficient	0.129	0.103	0.379

Zhou et al. *BMC Systems Biology,*6(Suppl 2):S2, 2012

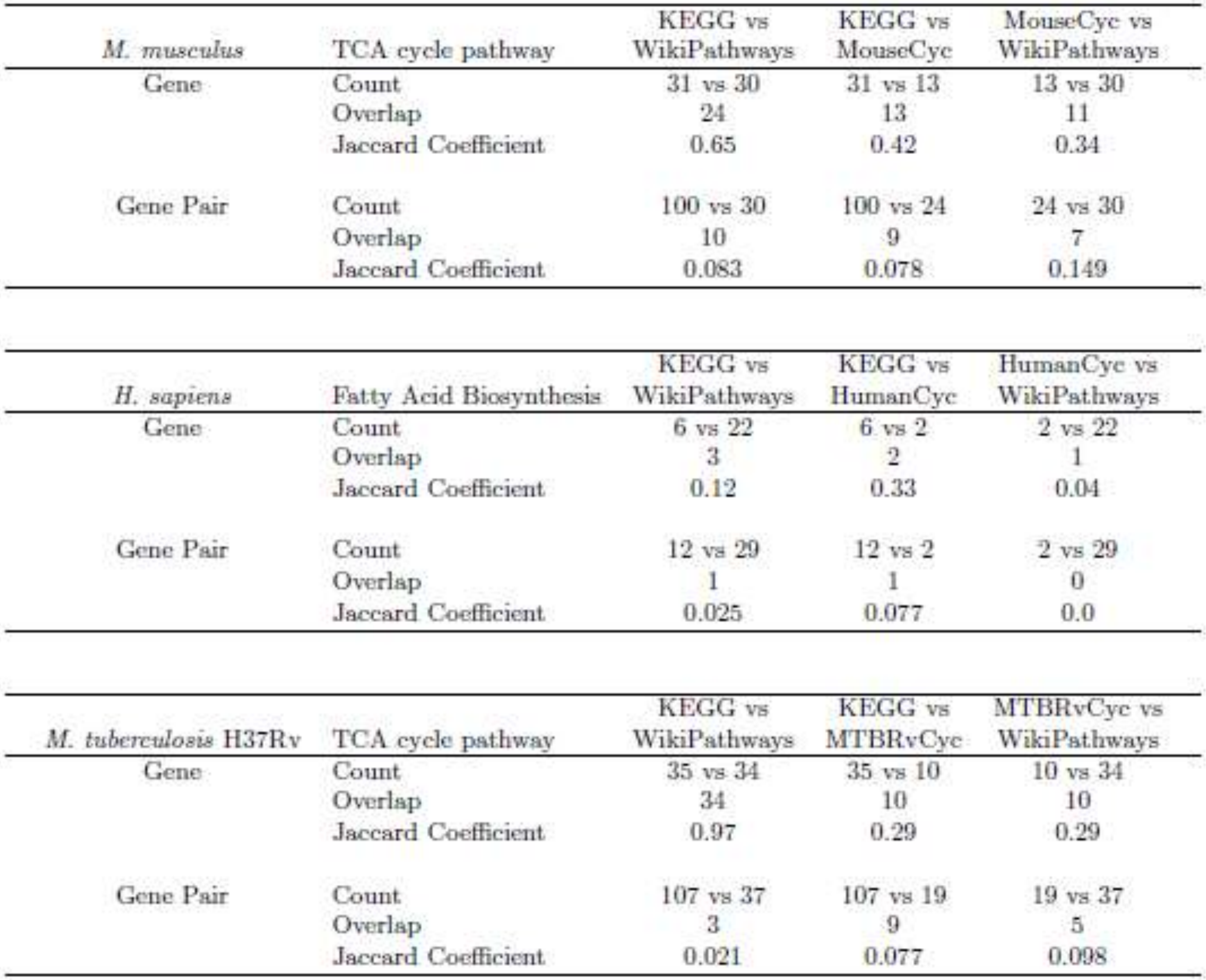
The same low inter-database consistency (in gene pair overlap) is observed in pathways of other organisms



M. musculus	KEGG vs WikiPathways	WikiPathways vs MouseCyc	MouseCyc vs KEGG
Overlap Gene Pairs	875	1,242	2,068
Unique Gene Pairs	55,489	33,312	38,891
Jaccard Coefficient	0.016	0.036	0.050
S. cerevisiae	KEGG vs WikiPathways	WikiPathways vs YeastCyc	YeastCyc vs KEGG
Overlap Gene Pairs	35	9	419
Unique Gene Pairs	2,909	1,479	3,524
Jaccard Coefficient	0.012	0.006	0.106
M. tuberculosis H37Rv	KEGG vs WikiPathways	WikiPathways vs MTBRvCyc	MTBRvCyc vs KEGG
Overlap Gene Pairs	9	8	358
Unique Gene Pairs	3,819	2,810	5,823
Jaccard Coefficient	0.002	0.003	0.058

Zhou et al. *BMC Systems Biology,*6(Suppl 2):S2, 2012

Example: TCA Cycle Pathway



		KEGG vs	KEGG vs	MouseCyc vs
M. musculus	TCA cycle pathway	WikiPathways	MouseCyc	WikiPathways
Gene	Count	31 vs 30	31 vs 13	13 vs 30
	Overlap	24	13	11
	Jaccard Coefficient	0.65	0.42	0.34
Gene Pair	Count	100 vs 30	100 vs 24	24 vs 30
	Overlap	10	9	7
	Jaccard Coefficient	0.083	0.078	0.149

		KEGG vs	KEGG vs	HumanCyc vs
H. sapiens	Fatty Acid Biosynthesis	WikiPathways	HumanCyc	WikiPathways
Gene	Count	6 vs 22	6 vs 2	2 vs 22
	Overlap	3	2	1
	Jaccard Coefficient	0.12	0.33	0.04
Gene Pair	Count	12 vs 29	12 vs 2	2 vs 29
	Overlap	1	1	0
	Jaccard Coefficient	0.025	0.077	0.0

		KEGG vs	KEGG vs	MTBRvCyc vs
M. tuberculosis H37Rv	TCA cycle pathway	WikiPathways	MTBRvCyc	WikiPathways
Gene	Count	35 vs 34	35 vs 10	10 vs 34
	Overlap	34	10	10
	Jaccard Coefficient	0.97	0.29	0.29
Gene Pair	Count	107 vs 37	107 vs 19	19 vs 37
	Overlap	3	9	5
	Jaccard Coefficient	0.021	0.077	0.098

Zhou et al. *BMC Systems Biology,*6(Suppl 2):S2, 2012

Incompatibility Issues

Pathway sources are curated. They are incomplete; but they have few errors. ➔ Makes sense to combine them. But…

- Data extraction method variations
- Format variations
- Data differences
- Gene/GeneID name differences
- Pathway name differences

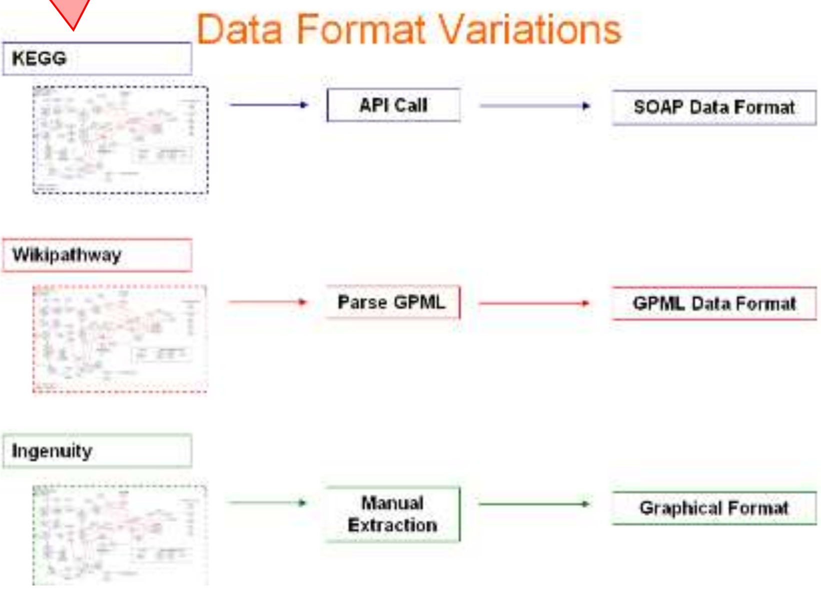
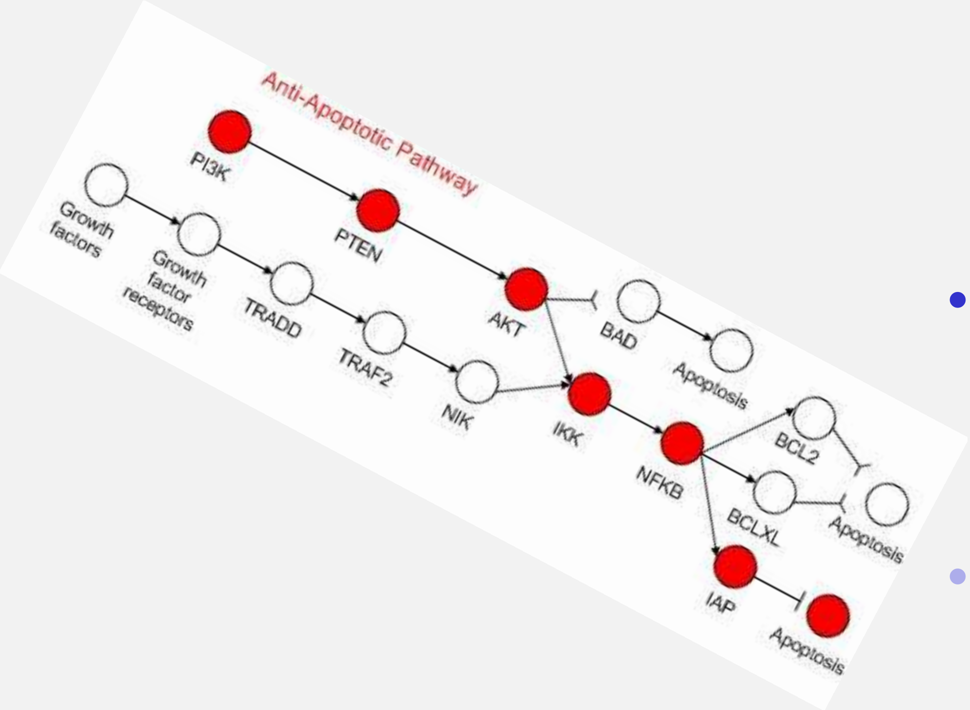


Image credit: Donny Soh's PhD dissertation, 2009

Part 2: How good are available sources of pathway information?



- **Sources of pathway info**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**

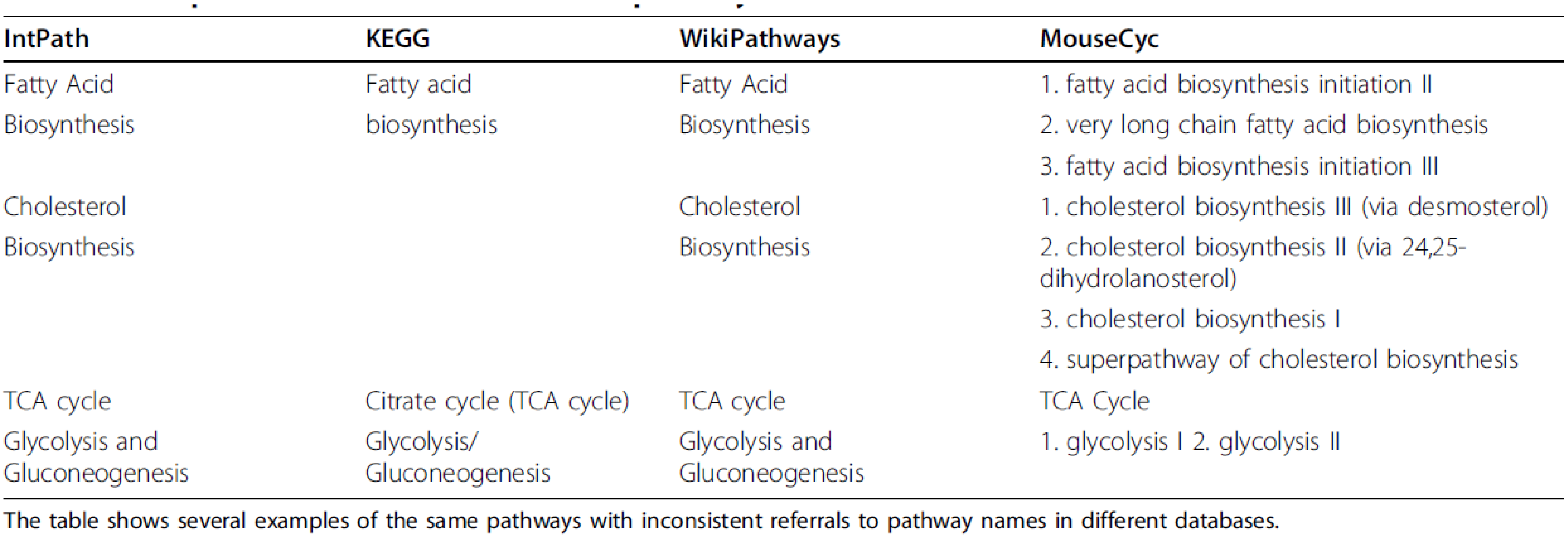
Things to deal with

- **Any integration of incompatible pathway databases must deal with**
  - Data extraction method variations
  - Format variations
  - Data differences
  - Gene name / gene id differences
  - Pathway name differences
- **We discuss only pathway name differences**
- **For other issues, consult**
  - Zhou et al. IntPath---an integrated pathway gene relationship database for model organisms and important pathogens, *BMC Bioinformatics*, 6(Suppl 2):S2, 2012

The same pathways in the different sources are often given different names.

So how do we even know two pathways are the same and should be compared / merged?

Example of Pathway Name Differences



IntPath	KEGG	WikiPathways	MouseCyc
Fatty Acid Biosynthesis	Fatty acid biosynthesis	Fatty Acid Biosynthesis	1. fatty acid biosynthesis initiation II 2. very long chain fatty acid biosynthesis 3. fatty acid biosynthesis initiation III
Cholesterol Biosynthesis		Cholesterol Biosynthesis	1. cholesterol biosynthesis III (via desmosterol) 2. cholesterol biosynthesis II (via 24,25-dihydrolanosterol) 3. cholesterol biosynthesis I 4. superpathway of cholesterol biosynthesis
TCA cycle	Citrate cycle (TCA cycle)	TCA cycle	TCA Cycle
Glycolysis and Gluconeogenesis	Glycolysis/ Gluconeogenesis	Glycolysis and Gluconeogenesis	1. glycolysis I 2. glycolysis II

The table shows several examples of the same pathways with inconsistent referrals to pathway names in different databases.

Zhou et al. *BMC Systems Biology,*6(Suppl 2):S2, 2012

Possible Ways to Match Pathways

- **Match based on name (LCS)**
  - Pathways w/ similar name should be the same pathway
  - But annotations are very noisy

  ⇒Likely to mismatch pathways?

  ⇒Likely to match too many pathways?

- **Are the followings good alternative approaches?**
  - Match based on overlap of genes
  - Match based on overlap of gene pairs

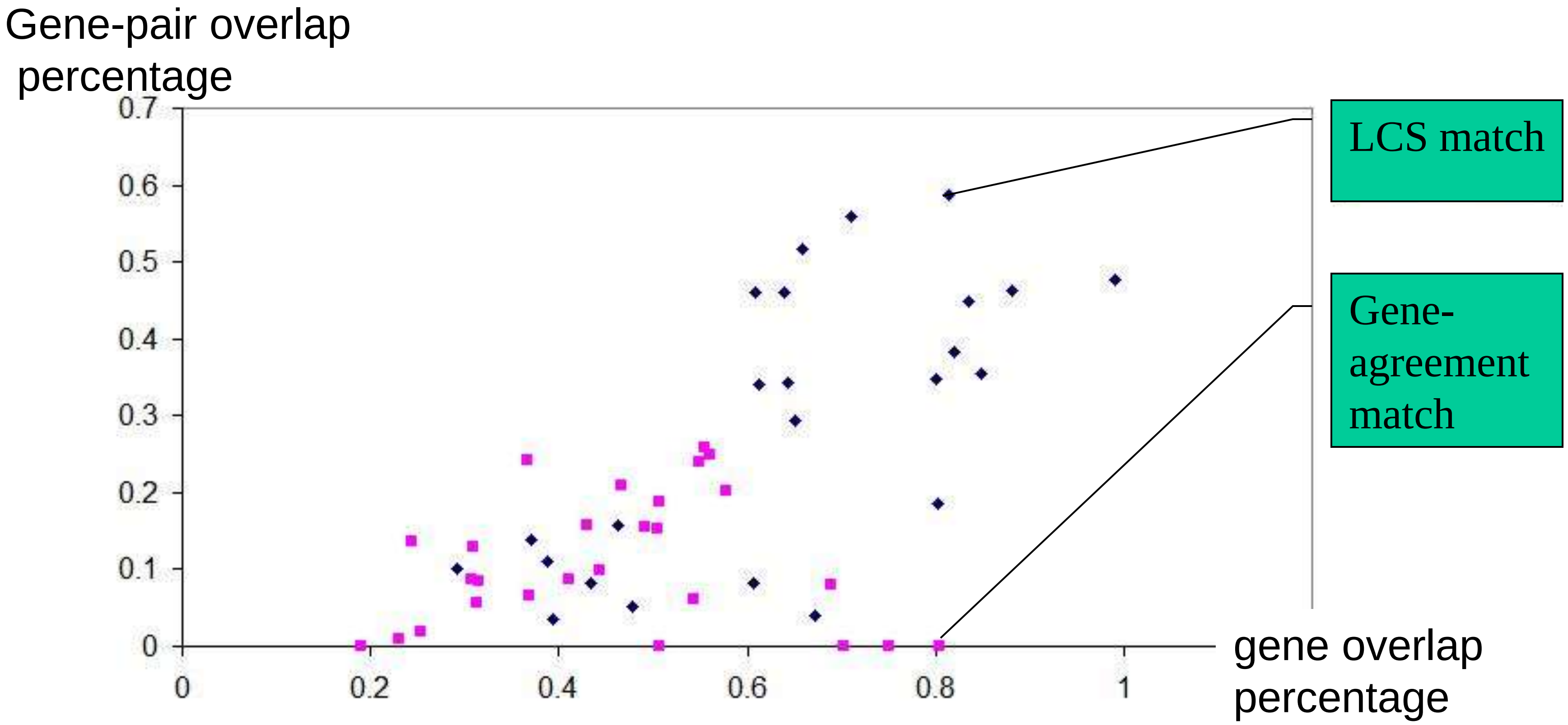
LCS vs Gene-Agreement Matching

- **Accuracy**
  - 94% of LCS matches are in top 3 gene agreement matches
  - 6% of LCS matches not in top 3 of gene agreement matches; but their gene-pair agreement levels are higher
- **Completeness**
  - Let Pi be pathway in db A that LCS cannot find match in db B
  - Let Qi be pathway in db B with highest gene agreement to Pi
  - Gene-pair agreement of Pi-Qi is much lower than pathway pairs matched by LCS

LCS is better than gene-agreement based matching!

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

LCS vs Gene-Agreement Matching

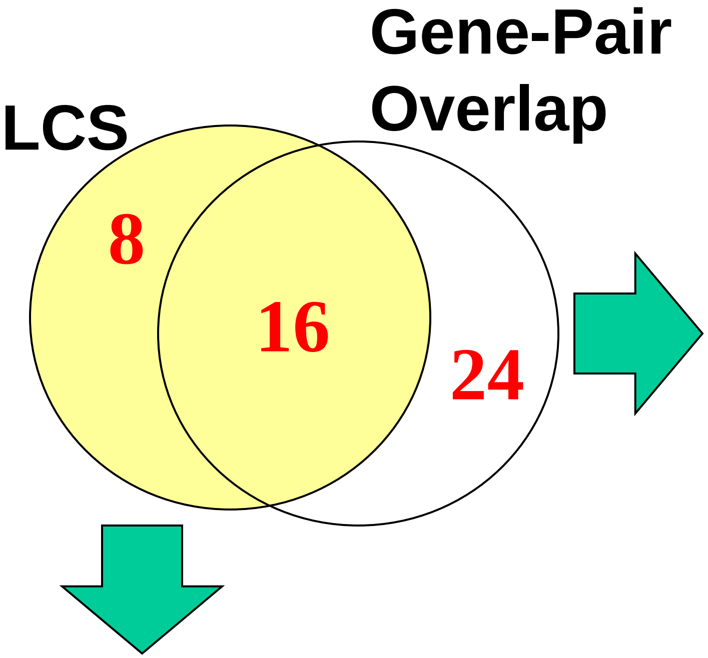


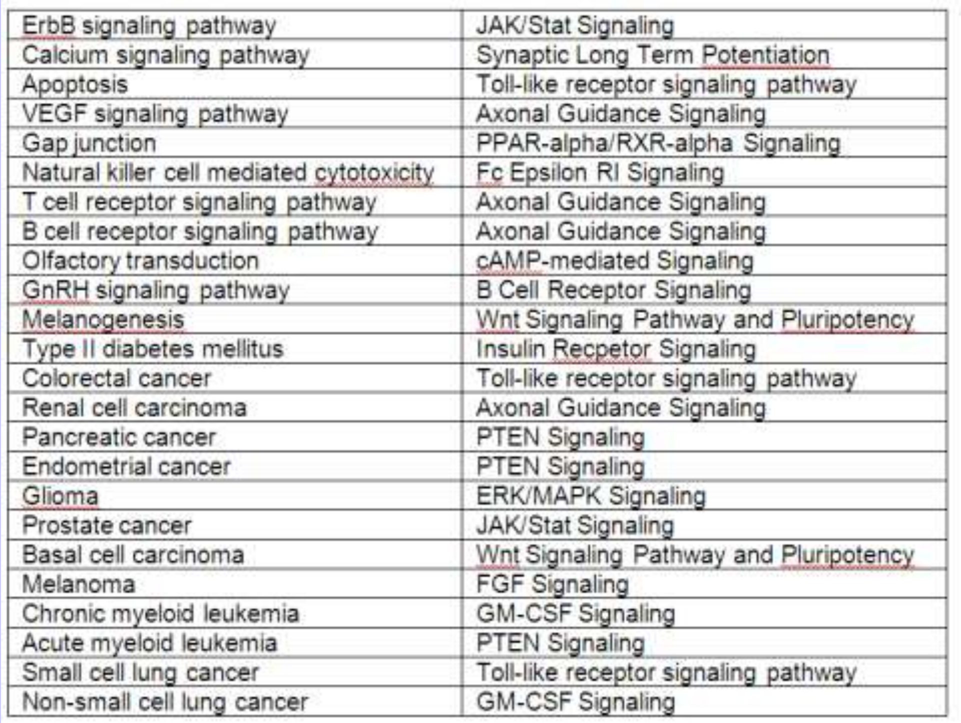
- **LCS consistently has higher gene-pair agreement**

⇒ LCS is better than gene-agreement based matching!

Soh et al. *BMC Bioinformatics*, 11:449, 2010.

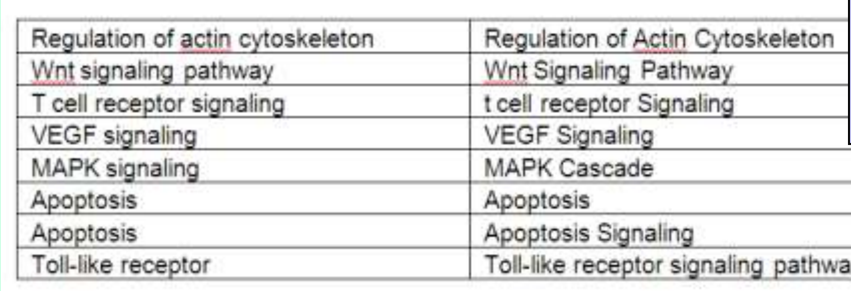
LCS vs Gene-Pair Agreement Matching





ErbB signaling pathway	JAK/Stat Signaling
Calcium signaling pathway	Synaptic Long Term Potentiation
Apoptosis	Toll-like receptor signaling pathway
VEGF signaling pathway	Axonal Guidance Signaling
Gap junction	PPAR-alpha/RXR-alpha Signaling
Natural killer cell mediated cytotoxicity	Fc Epsilon RI Signaling
T cell receptor signaling pathway	Axonal Guidance Signaling
B cell receptor signaling pathway	Axonal Guidance Signaling
Olfactory transduction	cAMP-mediated Signaling
GnRH signaling pathway	B Cell Receptor Signaling
Melanogenesis	Wnt Signaling Pathway and Pluripotency
Type II diabetes mellitus	Insulin Recpetor Signaling
Colorectal cancer	Toll-like receptor signaling pathway
Renal cell carcinoma	Axonal Guidance Signaling
Pancreatic cancer	PTEN Signaling
Endometrial cancer	PTEN Signaling
Glioma	ERK/MAPK Signaling
Prostate cancer	JAK/Stat Signaling
Basal cell carcinoma	Wnt Signaling Pathway and Pluripotency
Melanoma	FGF Signaling
Chronic myeloid leukemia	GM-CSF Signaling
Acute myeloid leukemia	PTEN Signaling
Small cell lung cancer	Toll-like receptor signaling pathway
Non-small cell lung cancer	GM-CSF Signaling

The 24 pathway pairs singled out by maximal gene-pair overlap



Regulation of actin cytoskeleton	Regulation of Actin Cytoskeleton
Wnt signaling pathway	Wnt Signaling Pathway
T cell receptor signaling	t cell receptor Signaling
VEGF signaling	VEGF Signaling
MAPK signaling	MAPK Cascade
Apoptosis	Apoptosis
Apoptosis	Apoptosis Signaling
Toll-like receptor	Toll-like receptor signaling pathway

The 8 pathway pairs singled out by LCS

Note: We consider only pathway pairs that have at least 20 reaction overlap.

LCS vs Gene-Pair Agreement Matching

- **Gene-pair agreement match will miss when**
  - Pathway P in db A has few overlap with pathway P in db B due to incompleteness of db, even if pathway name matches perfectly!
  - Example: wnt signaling pathway, VEGF signaling pathway, MAPK signaling pathway, etc. in KEGG don’t have largest gene-pair overlap w/ corresponding pathways in Wikipathways & Ingenuity

⇒ **Bad for getting a more complete unified pathway P**

LCS vs Gene-Pair Agreement Matching

- **Pathways having large gene-pair overlap are not necessarily the same pathways**

- **Examples**
  – "Synaptic Long Term Potentiation" in Ingenuity vs "calcium signalling" in KEGG
  – "PPAR-alpha/RXR-alpha Signaling" in Ingenuity vs "TGF-beta signaling pathway" in KEGG

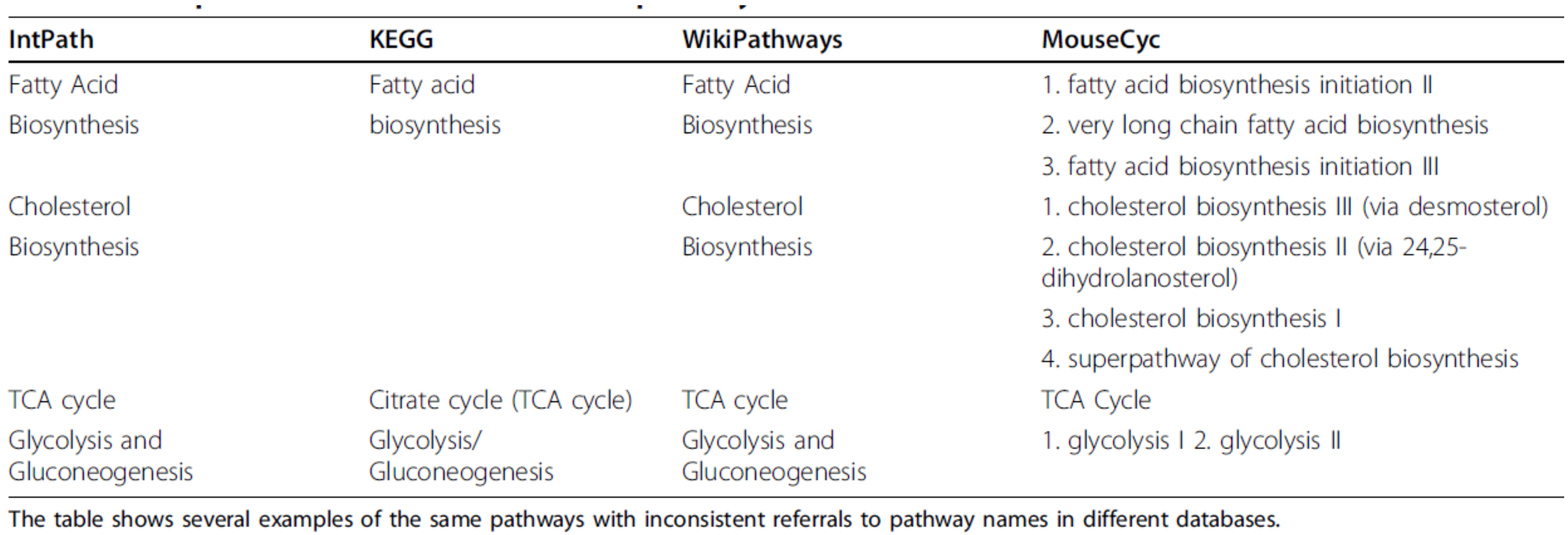
⇒ **Difficult to set correct gene-pair overlap threshold to balance against false positive matches**

Further Improvement to LCS

- **Please read the reference below (esp. page 10) for some of the improvements made to LCS**
  - Zhou et al. IntPath---an integrated pathway gene relationship database for model organisms and important pathogens, *BMC Bioinformatics*, 6(Suppl 2):S2, 2012.

An Interesting Question

- **If two pathways are merged, how do you choose the name of the resulting merged pathway?**
  - – Pick the longer of the two original names?
  - – Pick the shorter?
  - – Pick randomly?



IntPath	**KEGG**	**WikiPathways**	**MouseCyc**
Fatty Acid Biosynthesis	Fatty acid biosynthesis	Fatty Acid Biosynthesis	1. fatty acid biosynthesis initiation II 2. very long chain fatty acid biosynthesis 3. fatty acid biosynthesis initiation III
Cholesterol Biosynthesis		Cholesterol Biosynthesis	1. cholesterol biosynthesis III (via desmosterol) 2. cholesterol biosynthesis II (via 24,25-dihydrolanosterol) 3. cholesterol biosynthesis I 4. superpathway of cholesterol biosynthesis
TCA cycle	Citrate cycle (TCA cycle)	TCA cycle	TCA Cycle
Glycolysis and Gluconeogenesis	Glycolysis/ Gluconeogenesis	Glycolysis and Gluconeogenesis	1. glycolysis I 2. glycolysis II

The table shows several examples of the same pathways with inconsistent referrals to pathway names in different databases.

Zhou et al. *BMC Systems Biology,*6(Suppl 2):S2, 2012

The Answer

- **The general pathway name is chosen as the shortest pathway names from among the identified related pathways**

⇒ This usually works well as the name of the integrated pathway

- **But in some cases, the shortest name contains “suffix” or “prefix”---like “I”, “II”---that causes the integrated pathway name to give the wrong idea of describing only a specific aspect of the integrated pathway**

⇒ Remove such suffixes and prefixes when generating integrated pathway names

- **In a small number of cases, several similar pathways are included in one pathway name. In these cases, the shortest name is not appropriate as the name of the integrated pathway**

⇒ Replace the keyword of the integrated pathway name to cover more pathway information

Zhou et al. *BMC Systems Biology,*6(Suppl 2):S2, 2012

- **Having found a good way to match up pathways in different datasources, we proceeded to build a big unified pathway db….**

PathwayAPI
= KEGG
+ Wikipathways
+ Ingenuity

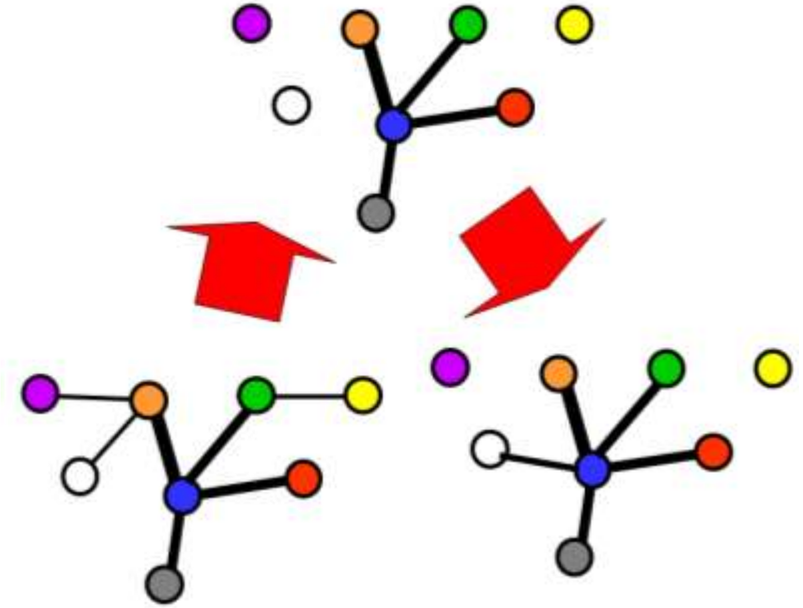
Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *BMC Bioinformatics*, 11:449, September 2010.

What have we learned?

- **Significant lack of concordance betw db's**
  - Level of consistency for genes is 0% to 88%
  - Level of consistency for genes pairs is 0%-61%
  - Most db contains less than half of the pathways in other db's

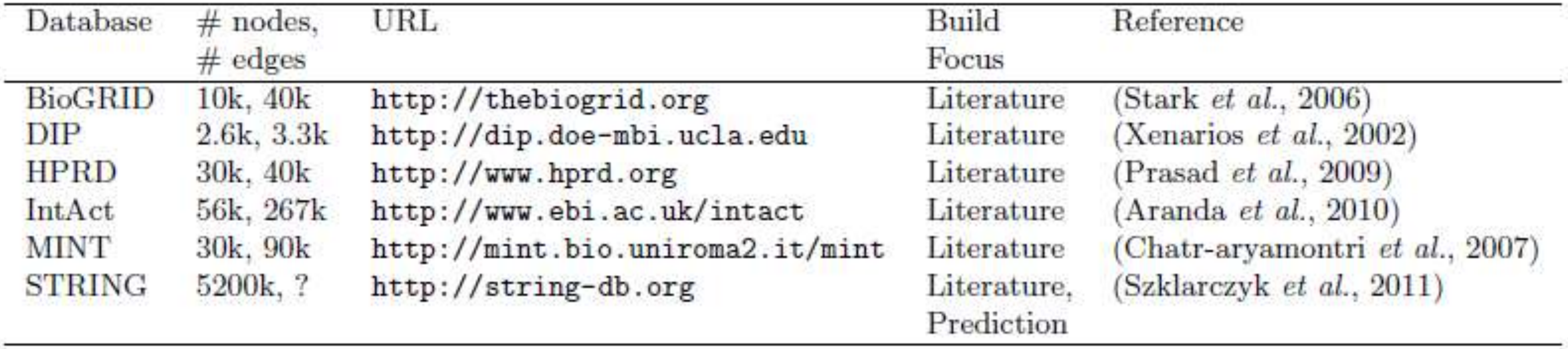
- **Matching pathways by name is better than matching by gene overlap or gene-pair overlap**

Part 3: How good are available sources of pathway & PPI Network?



- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**

Sources of Protein Interactions



Database	# nodes, # edges	URL	Build Focus	Reference
BioGRID	10k, 40k	http://thebiogrid.org	Literature	(Stark *et al.*, 2006)
DIP	2.6k, 3.3k	http://dip.doe-mbi.ucla.edu	Literature	(Xenarios *et al.*, 2002)
HPRD	30k, 40k	http://www.hprd.org	Literature	(Prasad *et al.*, 2009)
IntAct	56k, 267k	http://www.ebi.ac.uk/intact	Literature	(Aranda *et al.*, 2010)
MINT	30k, 90k	http://mint.bio.uniroma2.it/mint	Literature	(Chatr-aryamontri *et al.*, 2007)
STRING	5200k, ?	http://string-db.org	Literature, Prediction	(Szklarczyk *et al.*, 2011)

Source: Goh et al. “How advancement in biological network analysis methods empowers proteomics”. *Proteomics*, accepted.

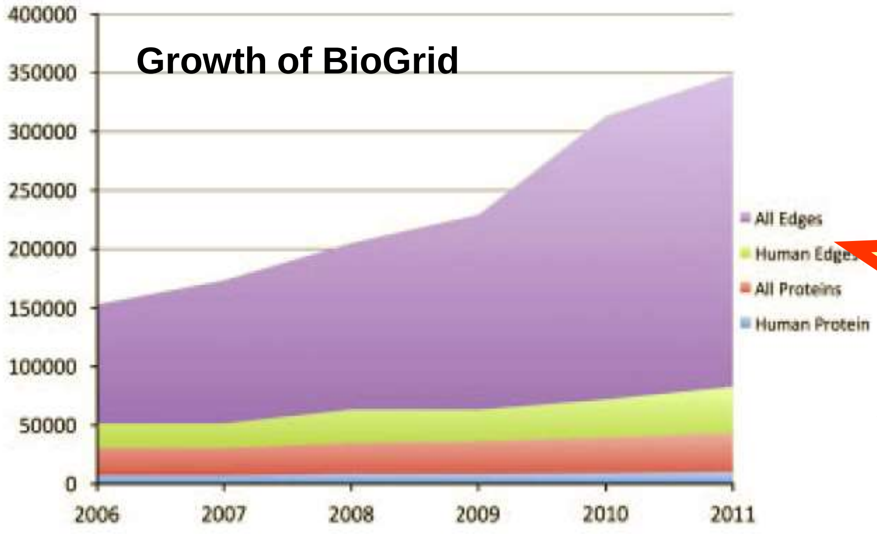
and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010

PPI Detection Assays

- **Many high-throughput assays for PPIs**
  - Y2H
  - TAP
  - Synthetic lethality

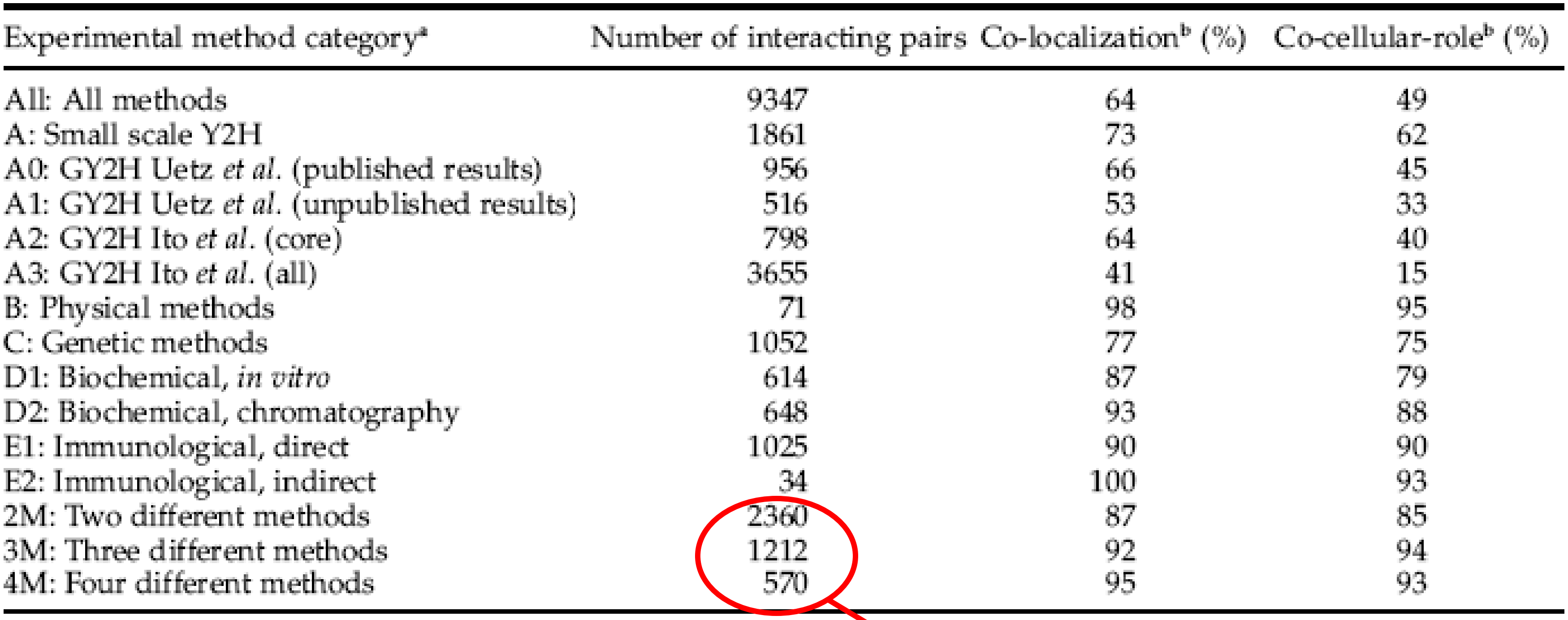
Generating *large amounts* of expt data on PPIs can be done with ease



- **But …**

High-throughput approaches sacrifice quality for **quantity**: (a) limited or biased coverage: ***false negatives***, & (b) high error rates: ***false positives***

Noise in PPI Networks



Experimental method category[a]	Number of interacting pairs	Co-localization[b] (%)	Co-cellular-role[b] (%)
All: All methods	9347	64	49
A: Small scale Y2H	1861	73	62
A0: GY2H Uetz *et al.* (published results)	956	66	45
A1: GY2H Uetz *et al.* (unpublished results)	516	53	33
A2: GY2H Ito *et al.* (core)	798	64	40
A3: GY2H Ito *et al.* (all)	3655	41	15
B: Physical methods	71	98	95
C: Genetic methods	1052	77	75
D1: Biochemical, *in vitro*	614	87	79
D2: Biochemical, chromatography	648	93	88
E1: Immunological, direct	1025	90	90
E2: Immunological, indirect	34	100	93
2M: Two different methods	2360	87	85
3M: Three different methods	1212	92	94
4M: Four different methods	570	95	93

Sprinzak et al., *JMB*, 327:919-923, 2003

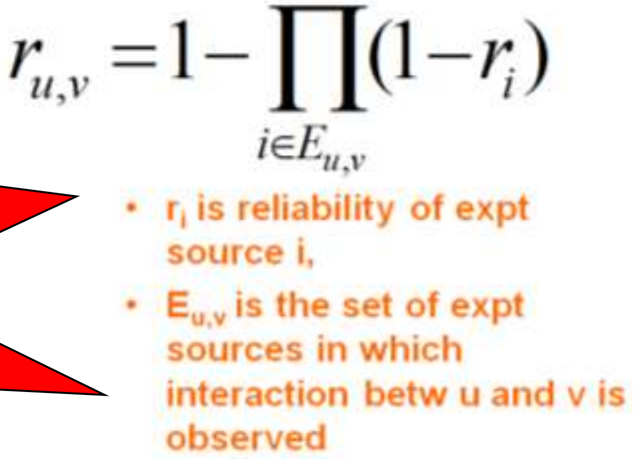
Large disagreement betw methods

- **High level of noise**

⇒ **Need to clean up before making inference on PPI networks**

Dealing with noise in PPIN using Reproducibility

- **If a PPI is reported in a few independent expts, it is more reliable than those reported in only one expt**



$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)$$

- r_i is reliability of expt source i,
- $E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed

Good idea. But you need to do more expts
➔ More time & more $ has to be spent

Dealing with noise in PPIN using Functional Homogeneity

- **If two proteins in a PPI participate in the same function or pathway, it is more reliable than those whose proteins do not share function & pathway**

Good idea. But the two proteins in the PPI you are looking at may not have functional annotation

Exercise

- What fraction of yeast PPIs in BioGrid share function?
- What fraction of yeast protein pairs share function?

Dealing with noise in PPIN using Localization Coherence

- **Two proteins should be in the same place to interact. Agree?**

Good idea. But the two proteins in the PPI you are looking at may not have localization annotation

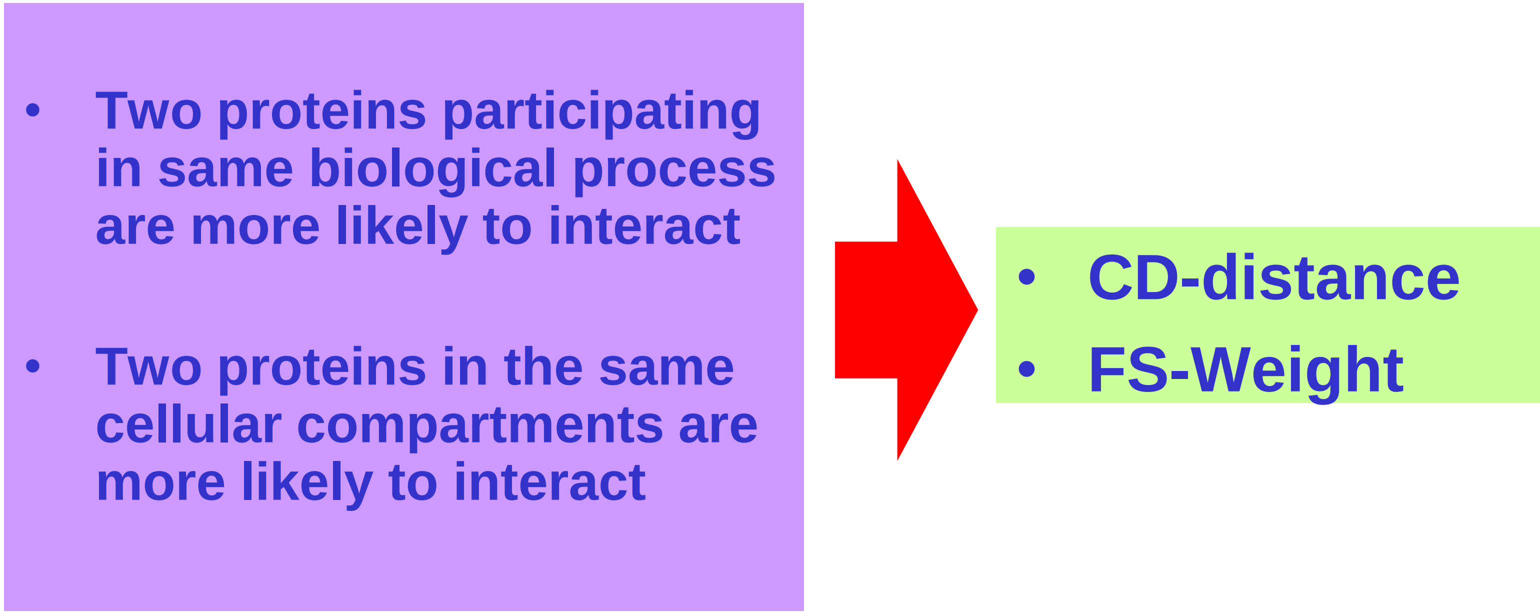
Exercise

– What fraction of yeast PPIs in BioGrid are in the same cellular compartment?

– What fraction of yeast protein pairs are in the same cellular compartment?

Chua & Wong. Increasing the Reliability of Protein Interactomes. *Drug Discovery Today*, 13(15/16):652--658, 2008

Dealing with noise in PPIN using local topology around a PPI edge

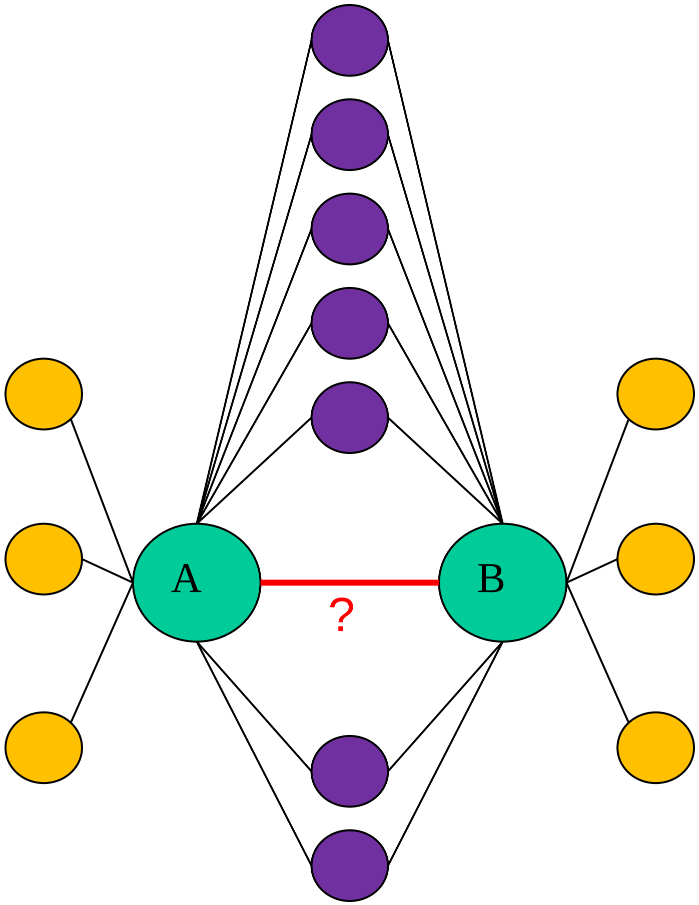


- **Two proteins participating in same biological process are more likely to interact**
- **Two proteins in the same cellular compartments are more likely to interact**

- **CD-distance**
- **FS-Weight**

CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment

Topology of neighbourhood of real PPIs



- Suppose 20% of putative PPIs are noise

⇒ ≥ 3 purple proteins are real partners of both A and B

⇒ A and B are likely localized to the same cellular compartment (Why?)

- Fact: Proteins in the same cellular compartment are 10x more likely to interact than other proteins

⇒ A and B are likely to interact

Brun, et al. *Genome Biology*, 5(1):R6, 2003

Czekanowski-Dice Distance

- **Given a pair of proteins (u, v) in a PPI network**
  - N_u = the set of neighbors of u
  - N_v = the set of neighbors of v
- **CD(u,v) =** $$\frac{2|N_u \cap N_v|}{|N_u| + |N_v|}$$
- **Consider relative intersection size of the two neighbor sets, not absolute intersection size**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, CD(u,v)=1
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, CD(u,v)=1

Liu et al. *GIW2008*, pp. 138-149

Adjusted CD-Distance

- **Variant of CD-distance that penalizes proteins with few neighbors**

$$\textbf{wL(u,v)} = \frac{2|N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$

$$\lambda_u = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_u|\}, \lambda_v = \max\{0, \frac{\sum_{x \in G} |N_x|}{|V|} - |N_v|\}$$

- **Suppose average degree is 4, then**
  - Case 1: $|N_u| = 1$, $|N_v| = 1$, $|N_u \cap N_v| = 1$, wL(u,v)=0.25
  - Case 2: $|N_u| = 10$, $|N_v| = 10$, $|N_u \cap N_v| = 10$, wL(u,v)=1

A thought…

wL(u,v) = $$\frac{2|N_u \cap N_v|}{|N_u|+\lambda_u+|N_v|+\lambda_v}$$

- **Weight of interaction reflects its reliability**

⇒ Can we get better results if we use this weight to re-calculate the score of other interactions?

Liu et al. *GIW2008*, pp. 138-149

Iterated CD-Distance

- **$wL^0(u,v) = 1$ if $(u,v) \in G$, otherwise $wL^0(u,v)=0$**
- **$wL^1(u,v) =$** $$\frac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v}$$
- **$wL^k(u,v) =$** $$\frac{\sum_{x \in N_u \cap N_v} w_L^{k-1}(u,x) + \sum_{x \in N_u \cap N_v} w_L^{k-1}(v,x)}{\sum_{x \in N_u} w_L^{k-1}(u,x) + \lambda^k{}_u + \sum_{x \in N_v} w_L^{k-1}(v,x) + \lambda^k{}_v}$$
- **$\lambda^k_u = \max\{0,$** $$\frac{\sum_{x \in V} \sum_{y \in N_x} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in N_u} w_L^{k-1}(u,x)$$ **$\}$**
- **$\lambda^k_v = \max\{0,$** $$\frac{\sum_{x \in V} \sum_{y \in N_x} w_L^{k-1}(x,y)}{|V|} - \sum_{x \in N_v} w_L^{k-1}(v,x)$$ **$\}$**

Validation

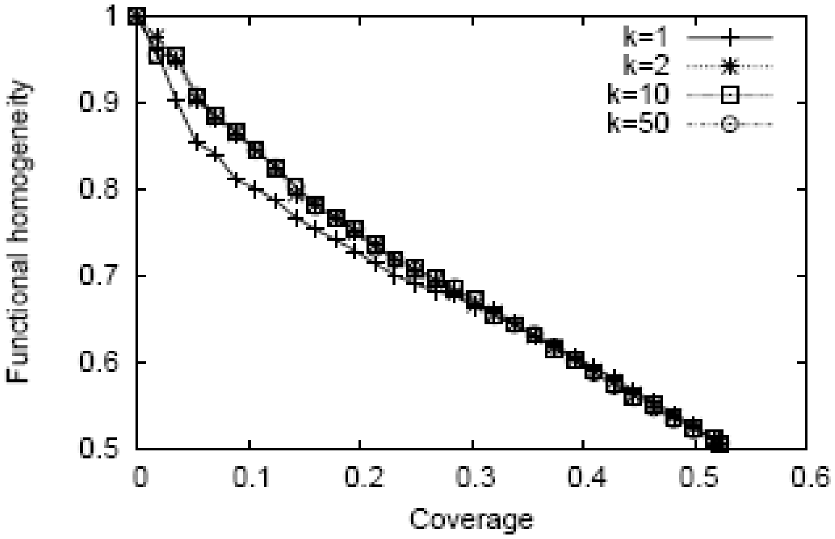
- **DIP yeast dataset**
  - Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  - Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs

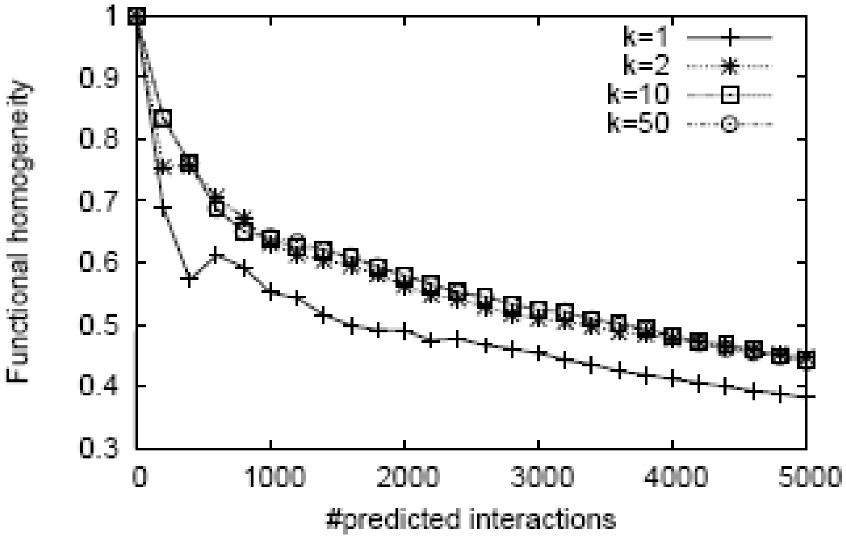
- **Let's see how much better iterated CD-distance is over the baseline above, as well as over the original CD-distance/FS-weight**

Liu et al. *GIW2008*, pp. 138-149

How many iteration is enough?

Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%

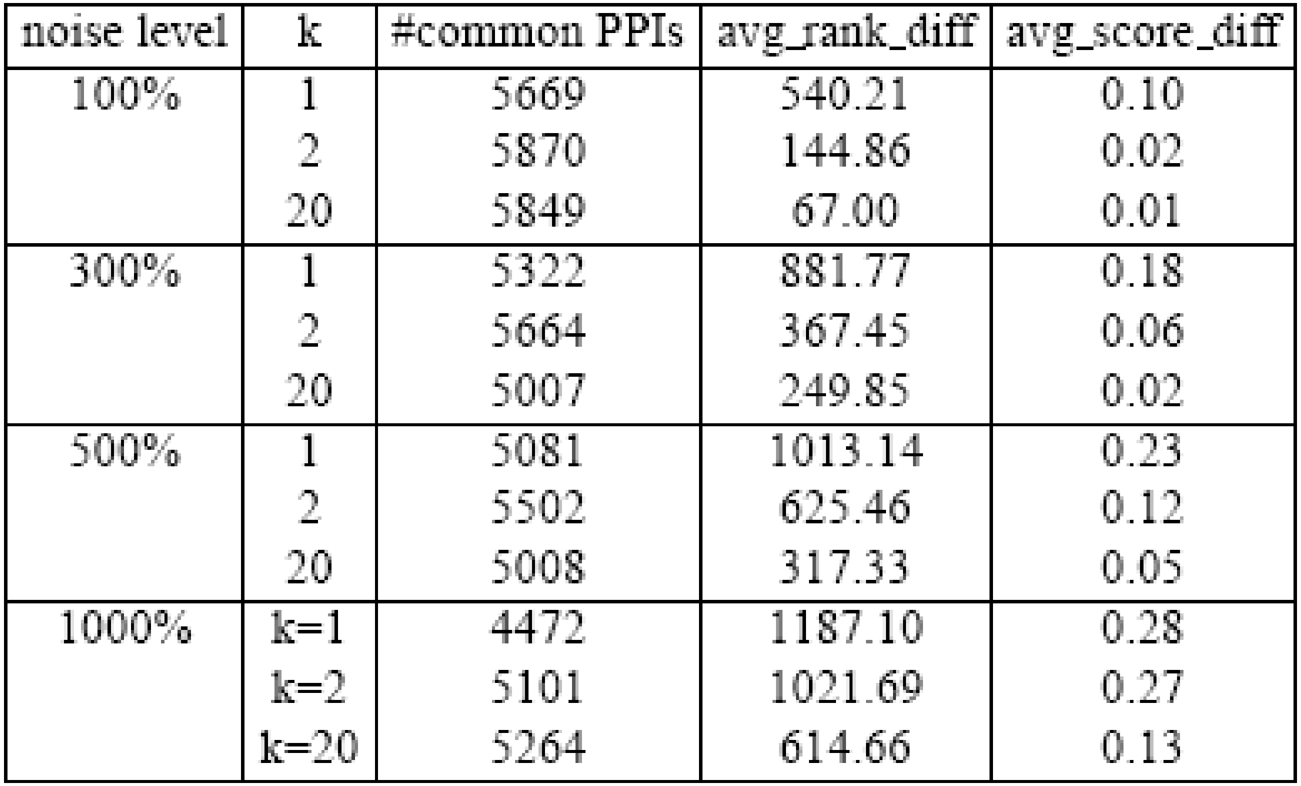




- **Iterated CD-distance achieves best performance wrt functional homogeneity at k=2**
- **Ditto wrt localization coherence (not shown)**

Liu et al. *GIW2008*, pp. 138-149

How many iteration is enough?



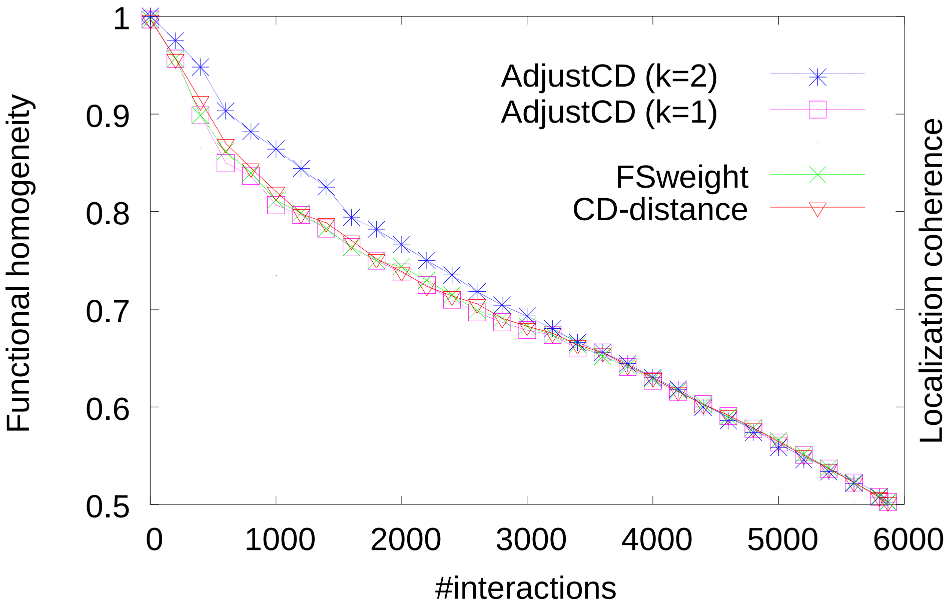
noise level	k	#common PPIs	avg_rank_diff	avg_score_diff
100%	1	5669	540.21	0.10
	2	5870	144.86	0.02
	20	5849	67.00	0.01
300%	1	5322	881.77	0.18
	2	5664	367.45	0.06
	20	5007	249.85	0.02
500%	1	5081	1013.14	0.23
	2	5502	625.46	0.12
	20	5008	317.33	0.05
1000%	k=1	4472	1187.10	0.28
	k=2	5101	1021.69	0.27
	k=20	5264	614.66	0.13

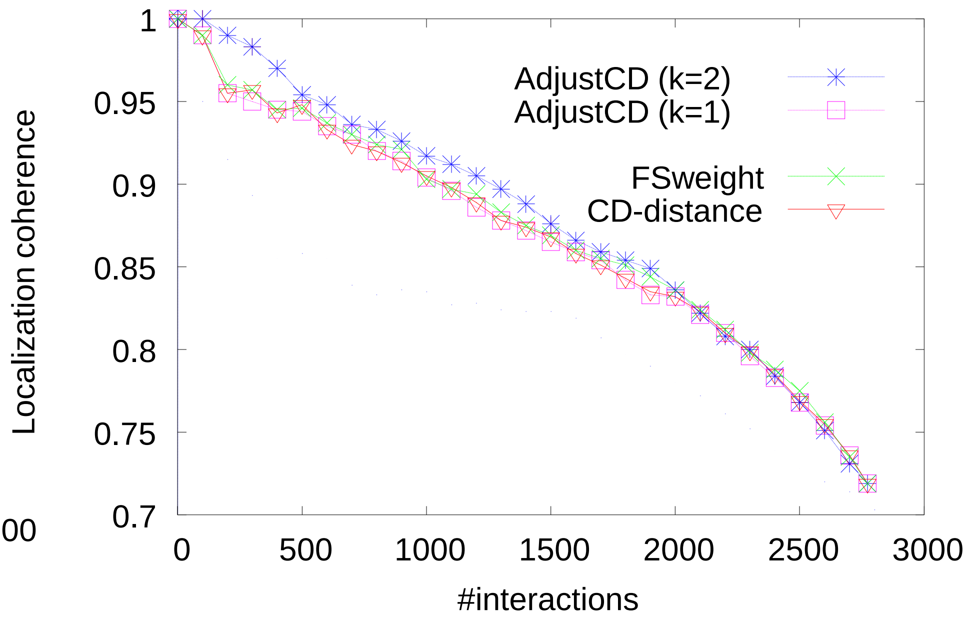
- **Iterative CD-distance at diff k values on noisy network**

⇒ # of iterations depends on amt of noise

Identifying False Positive PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



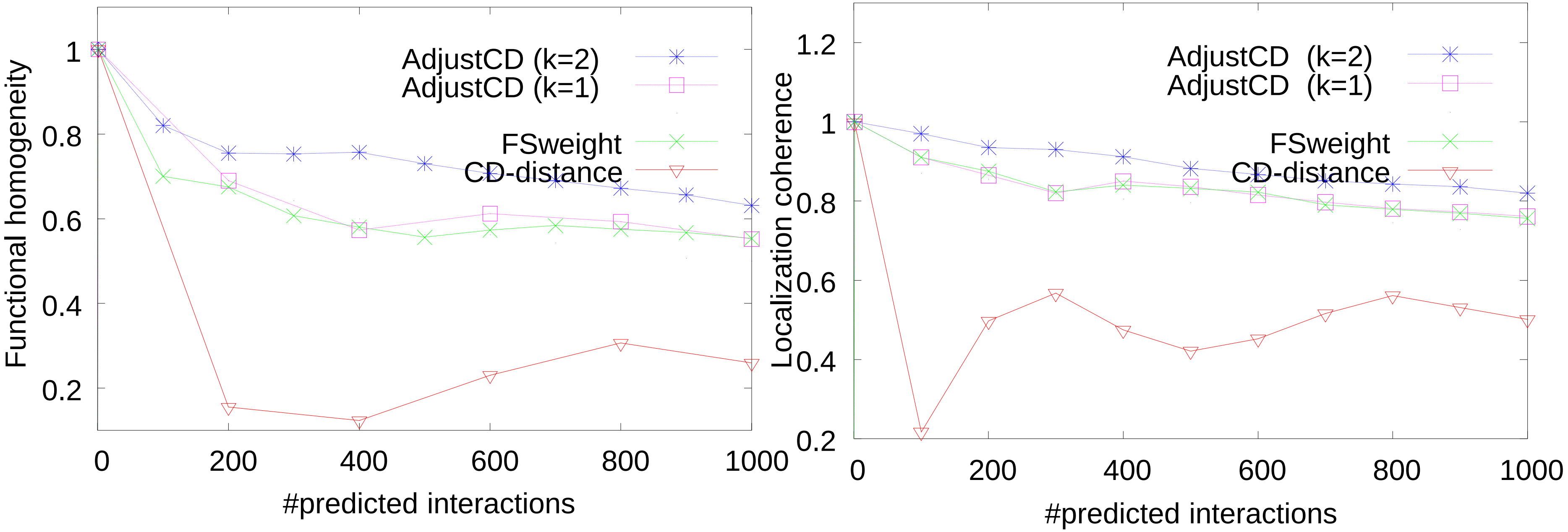


- **Iterated CD-distance is an improvement over previous measures for assessing PPI reliability**

Liu et al**.** *GIW2008,* pp. 138-149

Identifying False Negative PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



- **Iterated CD-distance is an improvement over previous measures for predicting new PPIs**

Liu et al. *GIW2008*, pp. 138-149

Combining multiple types of info to predict whether a PPI edge is real

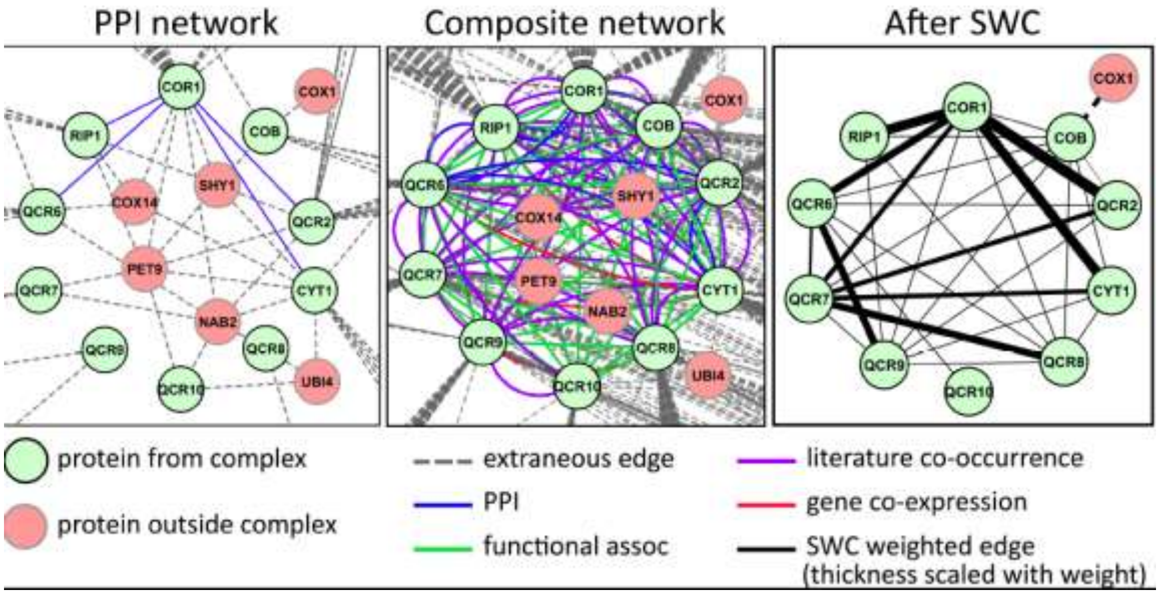
- **Sometimes you do have additional independent info available**
  - Several PPI expts
  - Functional annotations
  - Localization information

- **You can combine these pieces of info in the following standard way:**

$$r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i)$$

  - r_i is reliability of expt source i,
  - $E_{u,v}$ is the set of expt sources in which interaction betw u and v is observed

Another way to combine more types of info to predict if a PPI is real



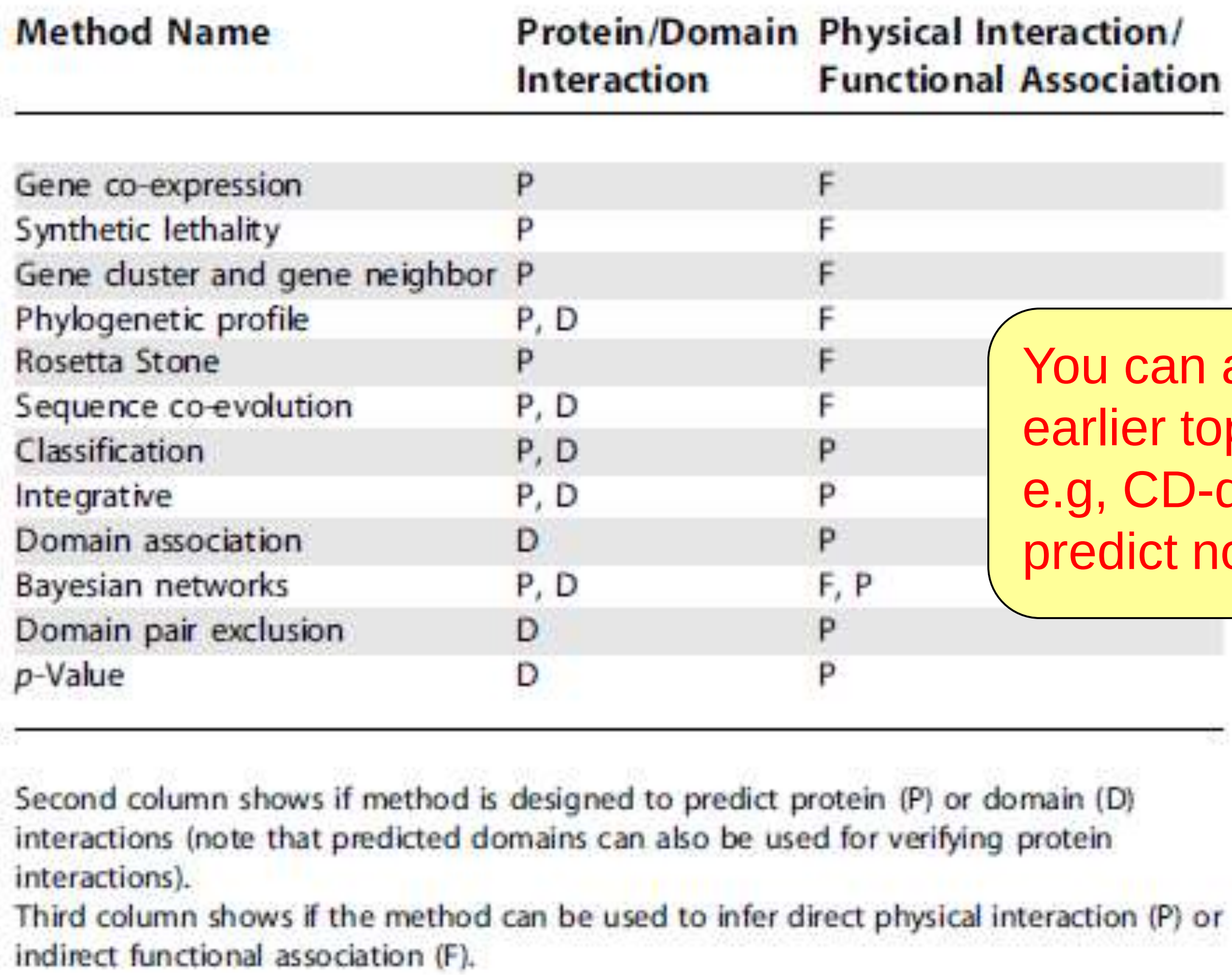
- **Overlay literature co-occurrence, gene co-expression, etc. on PPIN**
- **Machine learning to learn characteristic of real PPI**

$$
\begin{aligned}
& weight_{raw}(e) \\
= \; & P(e \; is \; comp | F_1 = f_1, F_2 = f_2, \ldots) \\
= \; & \frac{P(F_1 = f_1, F_2 = f_2, \ldots | e \; is \; comp) P(e \; is \; comp)}{P(F_1 = f_1, F_2 = f_2, \ldots)} \\
= \; & \frac{\prod_i P(F_i = f_i | e \; is \; comp) P(e \; is \; comp)}{\prod_i P(F_i = f_i)}
\end{aligned}
$$

Yong, et al. “Supervised maximum-likelihood weighting of composite protein networks for complex prediction”. *BMC Systems Biology*, 6(Suppl 2):S13, 2012

Shoemaker & Panchenko. *PLoS Computational Biology*, 3(4):e43, 2007

PPI Prediction Methods



Method Name	Protein/Domain Interaction	Physical Interaction/ Functional Association
Gene co-expression	P	F
Synthetic lethality	P	F
Gene cluster and gene neighbor	P	F
Phylogenetic profile	P, D	F
Rosetta Stone	P	F
Sequence co-evolution	P, D	F
Classification	P, D	P
Integrative	P, D	P
Domain association	D	P
Bayesian networks	P, D	F, P
Domain pair exclusion	D	P
p-Value	D	P

Second column shows if method is designed to predict protein (P) or domain (D) interactions (note that predicted domains can also be used for verifying protein interactions).
Third column shows if the method can be used to infer direct physical interaction (P) or indirect functional association (F).

You can also use our earlier topology scores, e.g, CD-distance to predict novel PPIs

Dandekar et al. *Trends Biochem Sci*, 23:324–328, 1998

PPI Prediction by Gene Clusters

- **Gene clusters or operons encoding co-regulated genes are usually conserved, despite shuffling effects of evolution**

⇒ Find conserved gene clusters

- Predict the genes to interact & form operons

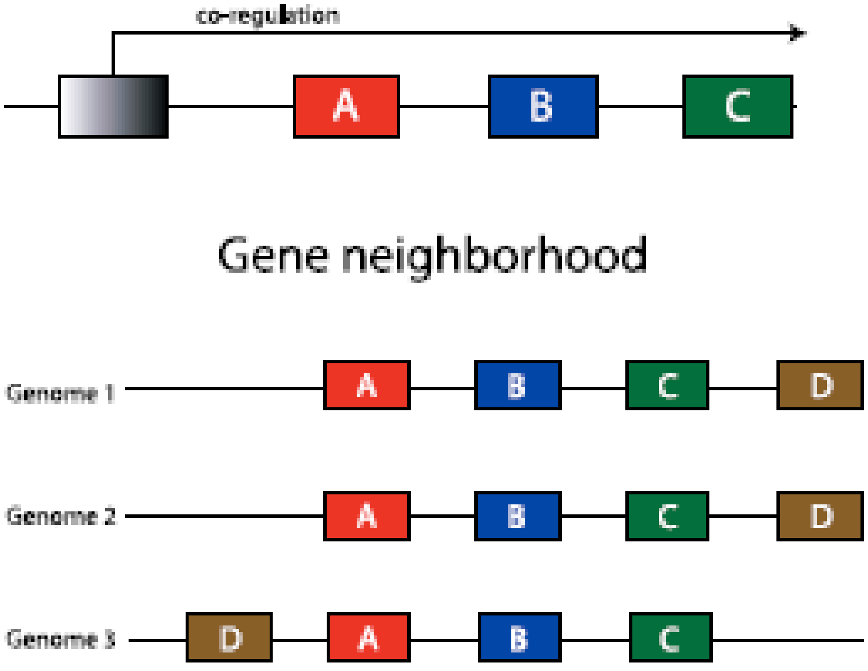


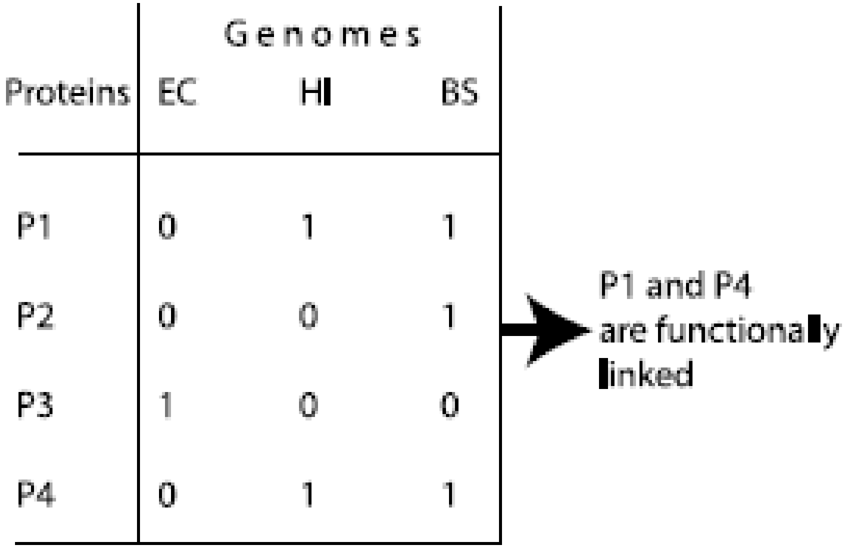
Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

Pellegrini et al. *PNAS*, 96:4285–4288, 1999

PPI Prediction by Phylogenetic Profiling

- **Components of complexes and pathways should be present simultaneously in order to perform their functions**

- Functionally linked and interacting proteins co-evolve and have orthologs in the same subset of fully sequenced organisms



	Genomes		
Proteins	EC	HI	BS
P1	0	1	1
P2	0	0	1
P3	1	0	0
P4	0	1	1

→ P1 and P4 are functionally linked

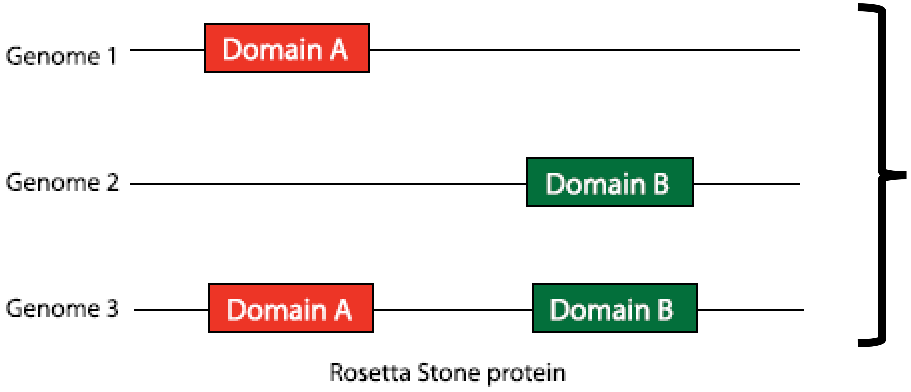
Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

Enright et al. *Nature,* 402:86–90, 1999

PPI Prediction by Rosetta Stone

- **Some interacting proteins have homologs in other genomes that are fused into one protein chain, a so-called Rosetta Stone protein**

- Gene fusion occurs to optimize co-expression of genes encoding for interacting proteins



Predict A & B interact

Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

See [Juan et al, *PNAS*, 105(3):934-939, 2008] for an impt further development to this idea

PPI Prediction by Seq Co-Evolution

- **Interacting proteins co-evolve**
  - – Changes in one protein leading to loss of function are compensated by correlated changes in another protein

- Co-evolution is quantified by correlation of distance matrices used to construct the trees

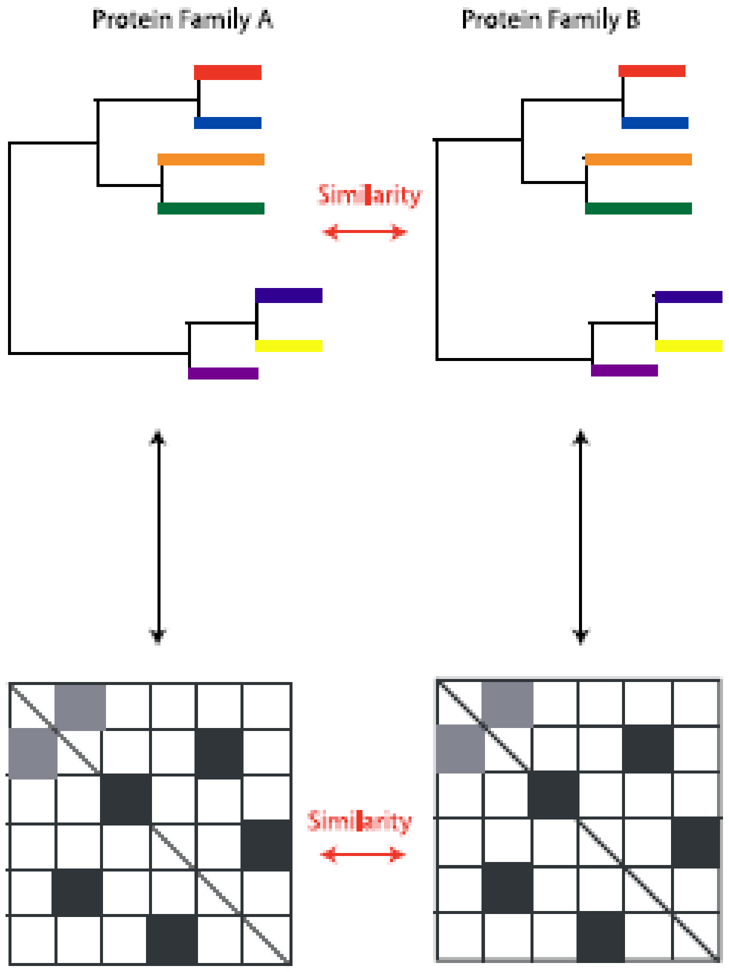
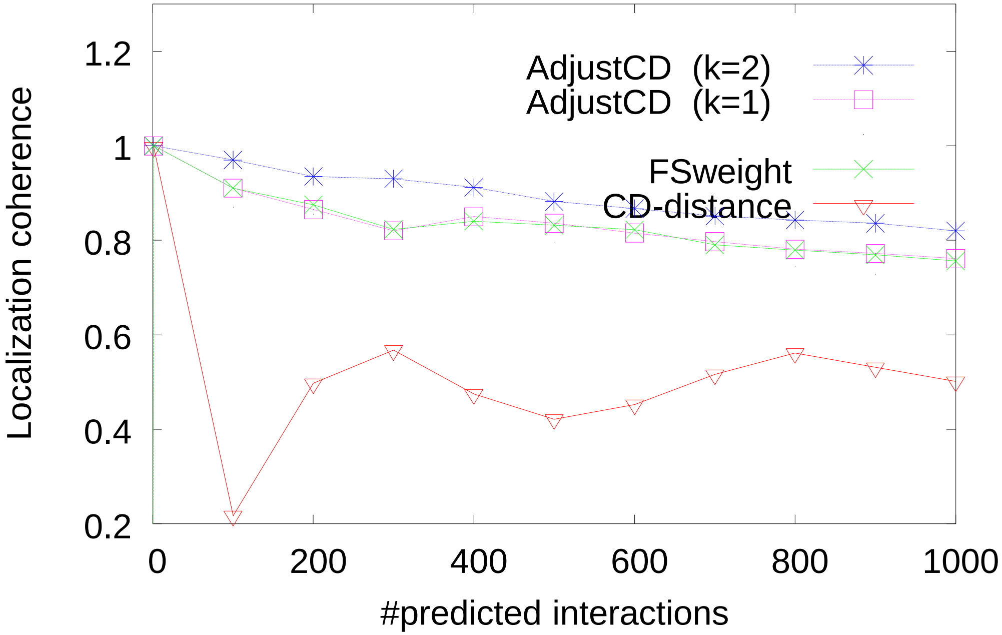


Image credit: Shoemaker & Panchenko. *PLoS Comp Biol*, 3(4):e43, 2007

Liu et al. *GIW2008*, pp. 138-149

PPI Prediction by Iterated CD-Distance

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%



$$wL^k(u,v) = \frac{\sum_{x \in Nu \cap Nv} w_L^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} w_L^{k-1}(v,x)}{\sum_{x \in Nu} w_L^{k-1}(u,x) + \lambda^k_u + \sum_{x \in Nv} w_L^{k-1}(v,x) + \lambda^k_v}$$

- Predict (u,v) interact if $wL^k(u,v)$ is large

What have we learned?

- **It is possible to predict PPIs using a variety of information and methods**
  - Gene cluster, gene fusion, phylogenetic profile, sequence co-evolution, …

For those who are interested to go further:

- **How do you predict cross-species PPI's between a host and a pathogen?**

Must Read

- Soh et al. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases.** *BMC Bioinformatics*, 11:449, 2010
- Zhou et al. **IntPath---an integrated pathway gene relationship database for model organisms and important pathogens,** *BMC Systems Biology*, 6(Suppl 2):S2, 2012
- Ng & Tan. **Discovering protein-protein interactions.** *JBCB*, 1(4):711-741, 2004
- Chua & Wong. **Increasing the Reliability of Protein Interactomes.** *Drug Discovery Today*, 13(15/16):652-658, 2008
- Shoemaker & Panchenko. **Deciphering protein-protein Interactions. Part II. Computational methods to predict protein and domain interaction partners.** *PLoS Computational Biology*, 3(4):e43, 2007

Acknowledgements

- **Kenny Chua**
- **Difeng Dong**
- **Wilson Goh**
- **Kevin Lim**
- **Guimei Liu**
- **Donny Soh**
- **Chern Han Yong**
- **Hufeng Zhou**

- **Singapore funding agencies**
  - A*STAR
  - Ministry of Education
  - National Research Foundation

- **UK funding agencies**
  - Wellcome Trust scholarship