Analysis of Gene Expression and Proteomic Profiles based on Biological Networks

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/apbc2012-tutorial.pdf

- The notes for this tutorial can be downloaded at

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

Tutorial Outline

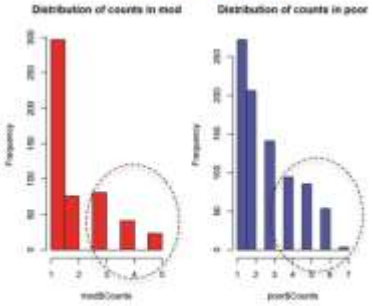
NUS

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

NUS

Part 2: Delivering more powerful proteomic profile analysis



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

NUS

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

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

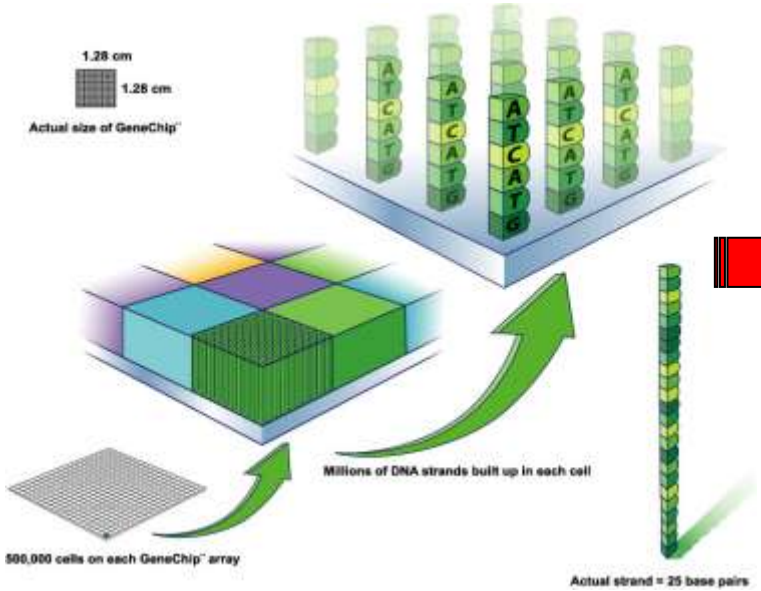
Analysis of Gene Expression and Proteomic Profiles based on Biological Networks

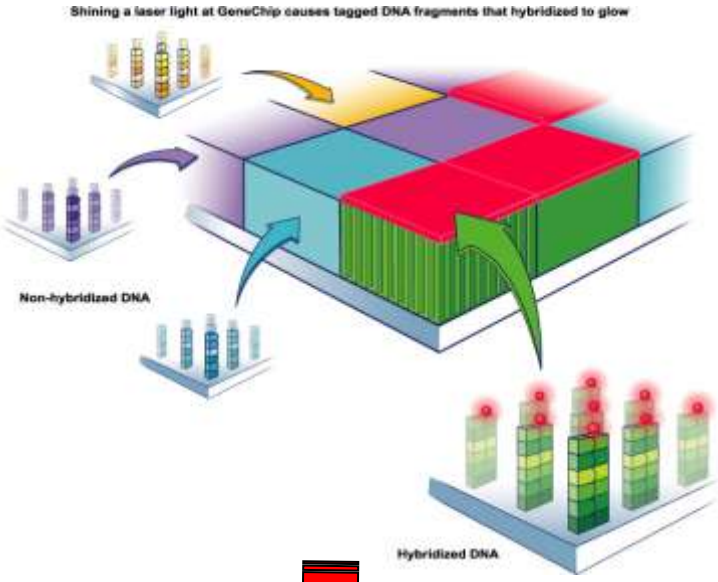
Part 1

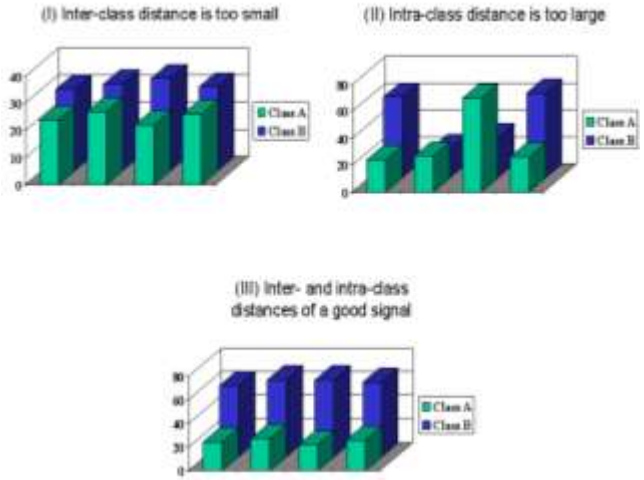
Limsoon Wong

NUS
National University of Singapore

Diagnosis Using Microarray







	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-Murl	5	2	19	297.5	A	M16762 Mouse int
AFFX-Murl	3	2	19	554.2	A	M37897 Mouse int
AFFX-Murl	4	2	19	308.6	A	M25892 Mus mus
AFFX-Murf	1	3	19	141	A	M83649 Mus mus
AFFX-BioE	13	1	19	9340.6	P	J04423 E coli bioE
AFFX-BioE	15	0	19	12862.4	P	J04423 E coli bioE
AFFX-BioE	12	0	19	8716.5	P	J04423 E coli bioE
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 Bacteriop
AFFX-CreX	20	0	20	401741.8	P	X03453 Bacteriop
AFFX-BioE	7	5	18	-483	A	J04423 E coli bioE
AFFX-BioE	5	4	18	313.7	A	J04423 E coli bioE
AFFX-BioE	7	6	20	-1016.2	A	J04423 E coli bioE

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

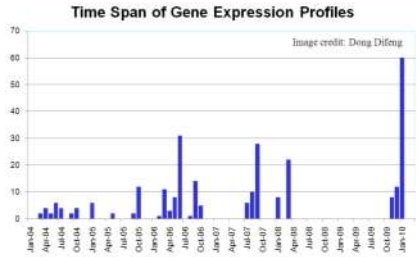
- **Natural fluctuations of gene expression in a person**

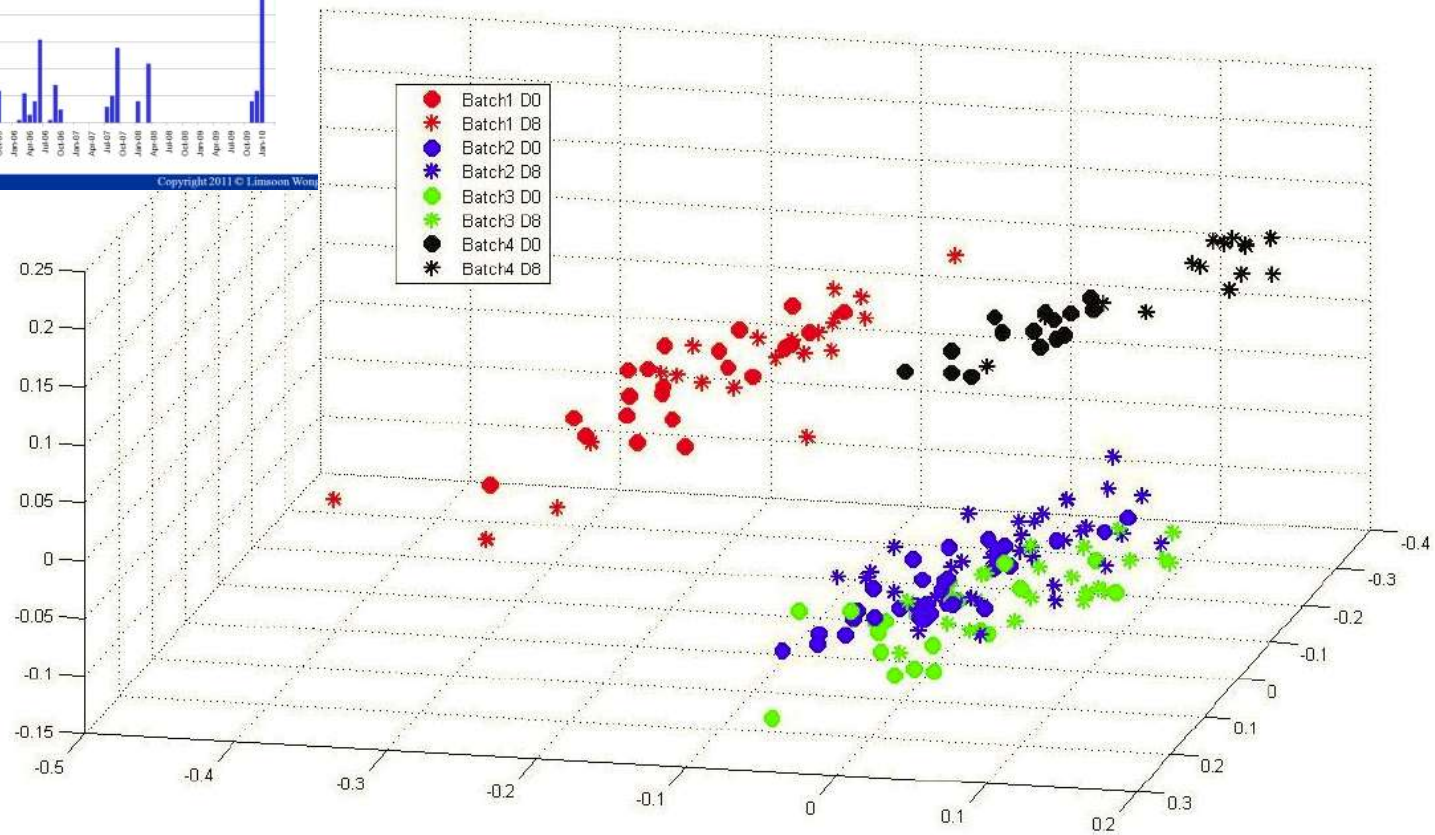
- **Noise in experimental protocols**
  - Numbers mean diff things in diff batches
  - Numbers mean diff things in data obtained from diff platforms

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

Batch Effects

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





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

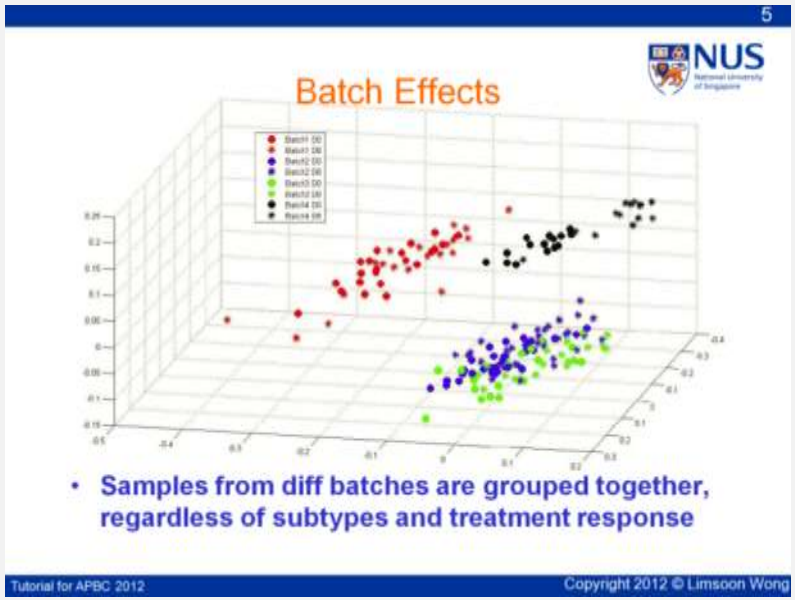
Percentage of Overlapping Genes

- **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

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

Approaches to Normalization

- **Aim of normalization:**
  Reduce variance w/o increasing bias

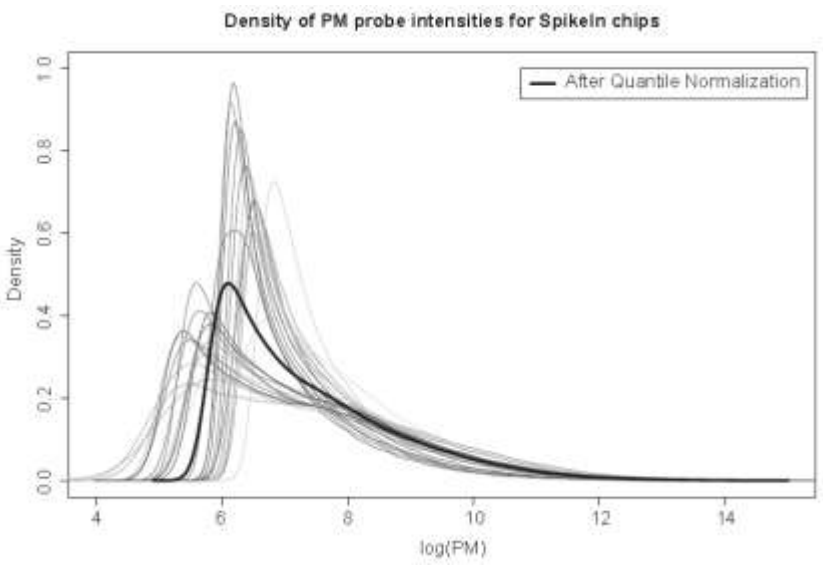
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

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

- **Quantile normalization**

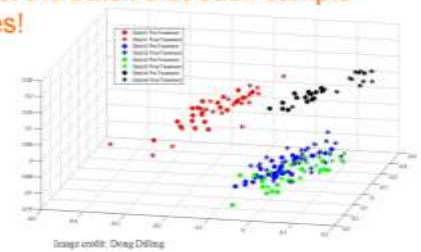
Quantile Normalization

- **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***



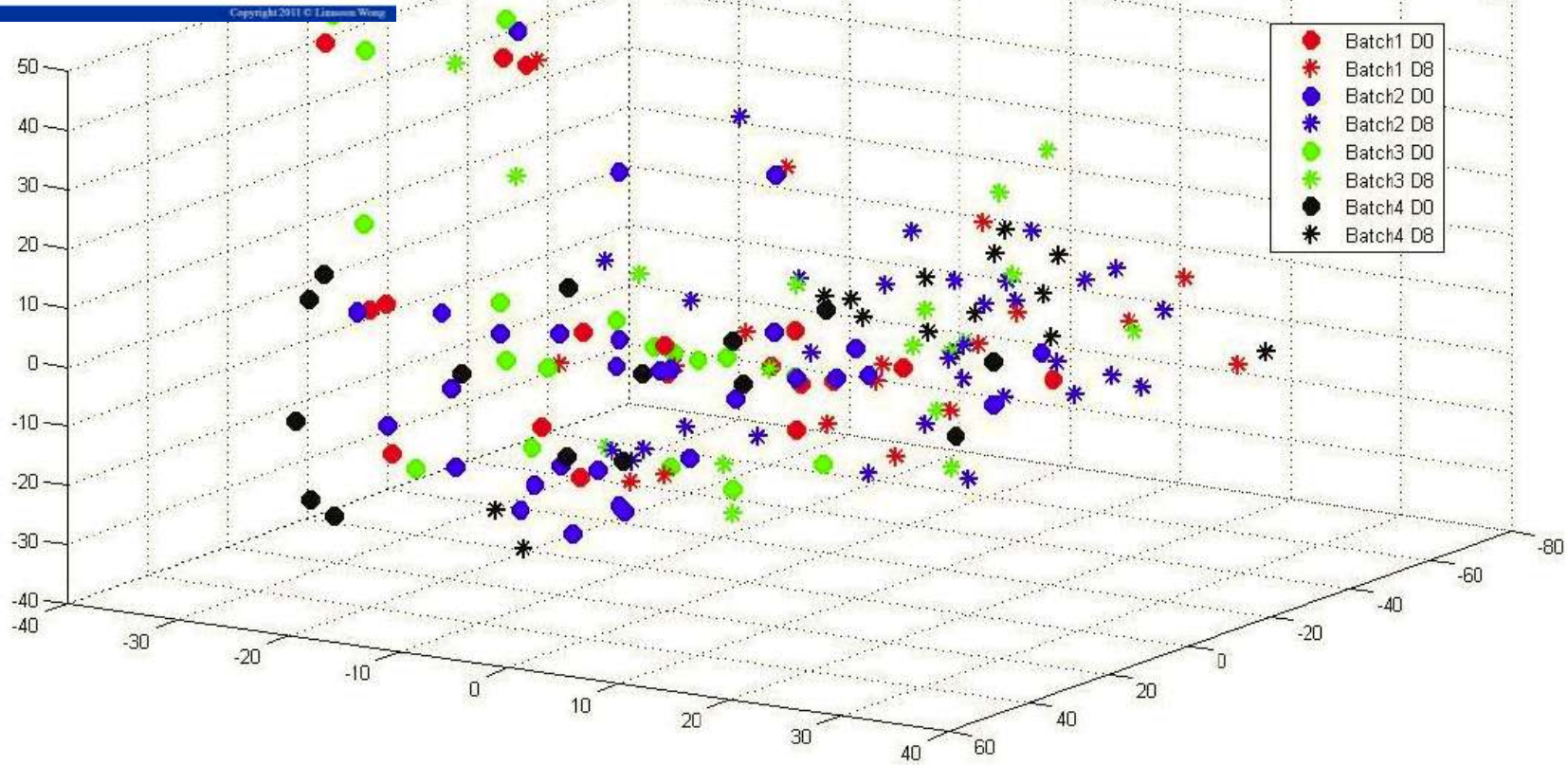
- Implemented in some microarray s/w, e.g., EXPANDER

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



⇒ Need normalization to correct for batch effect

After quantile normalization

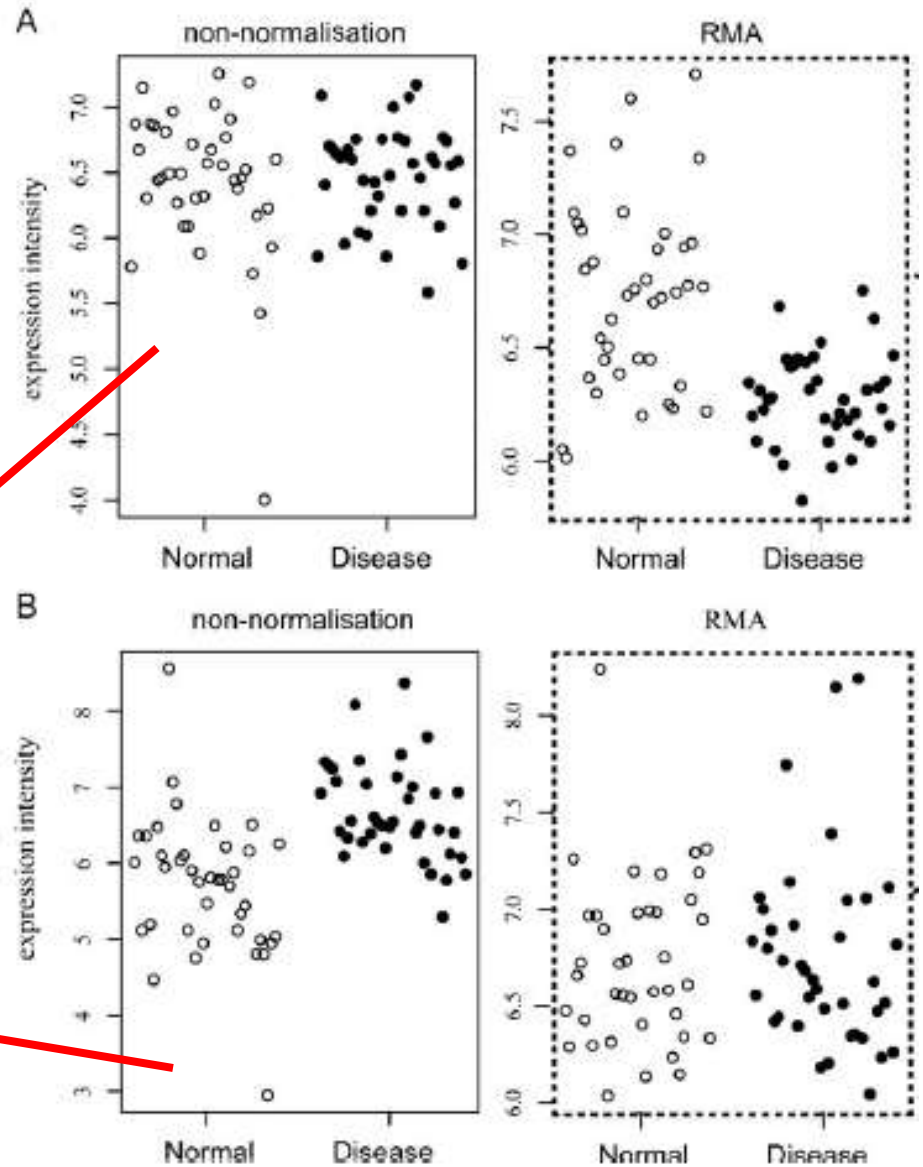


GEP after removing batch effect by quantile normalization

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

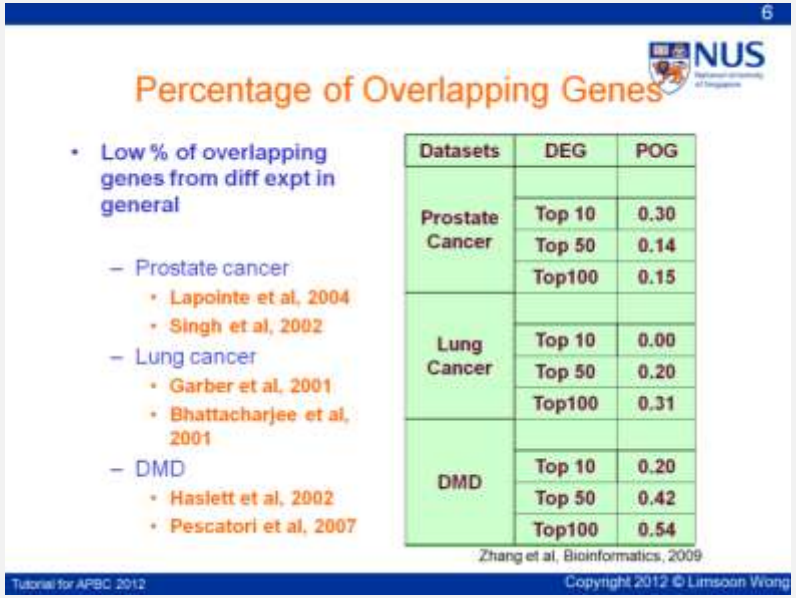
A gene was detected as an up-regulated DE gene in the non-normalized data, but was not identified as a DE gene in the quantile nornmalized data



Wang et al. *Molecular Biosystems*, in press

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**

$\Rightarrow$ **E(# of correlated genes) = 1,562**

$\Rightarrow$ **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?**

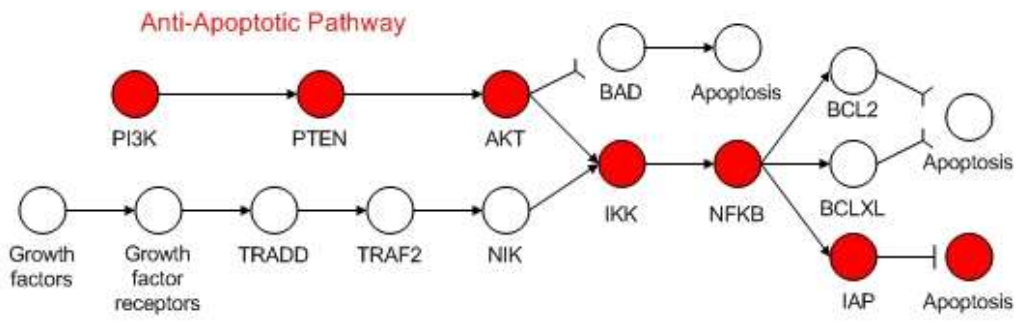
- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# of groups = ${}^{100000}C_5$**

⇒ **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**

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 = $^{100000}C_5$**~~
- ~~**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

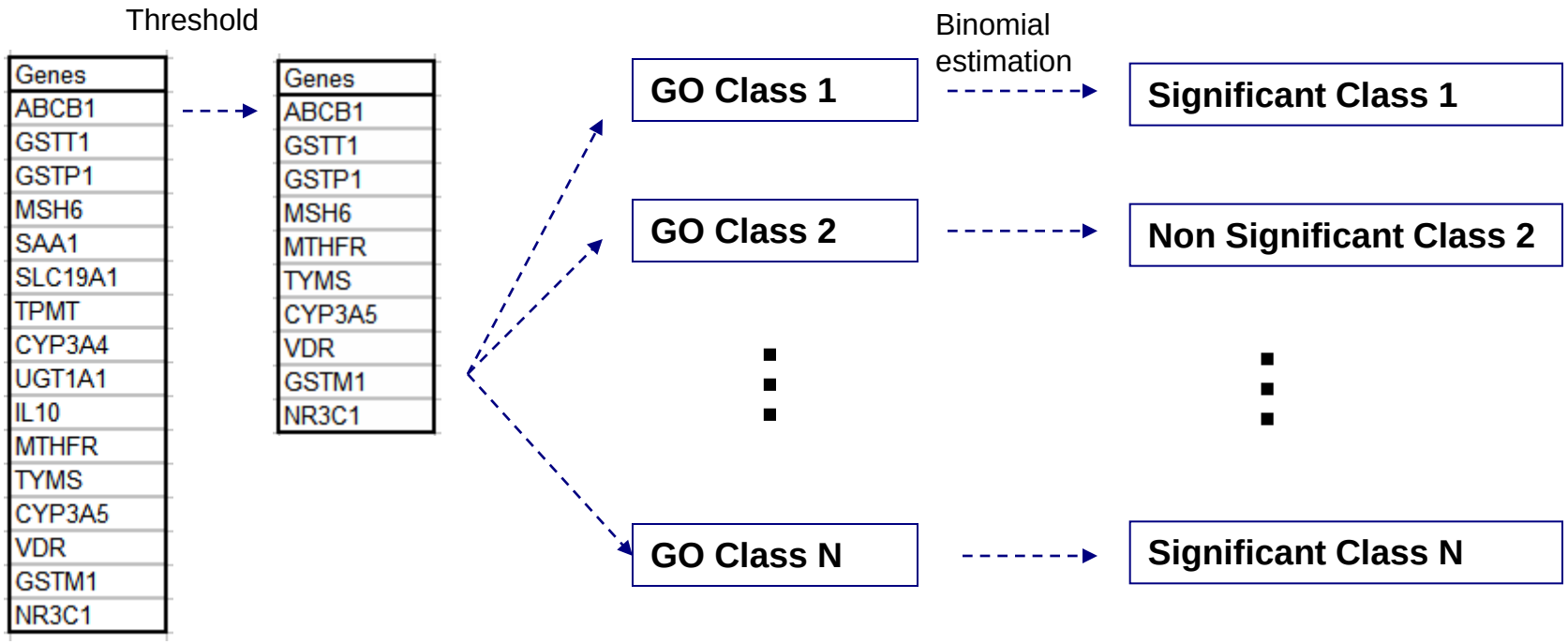
- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **SNet**
  - Soh et al
  - *BMC Genomics*, 2011

Overlap Analysis

Direct-Group Analysis

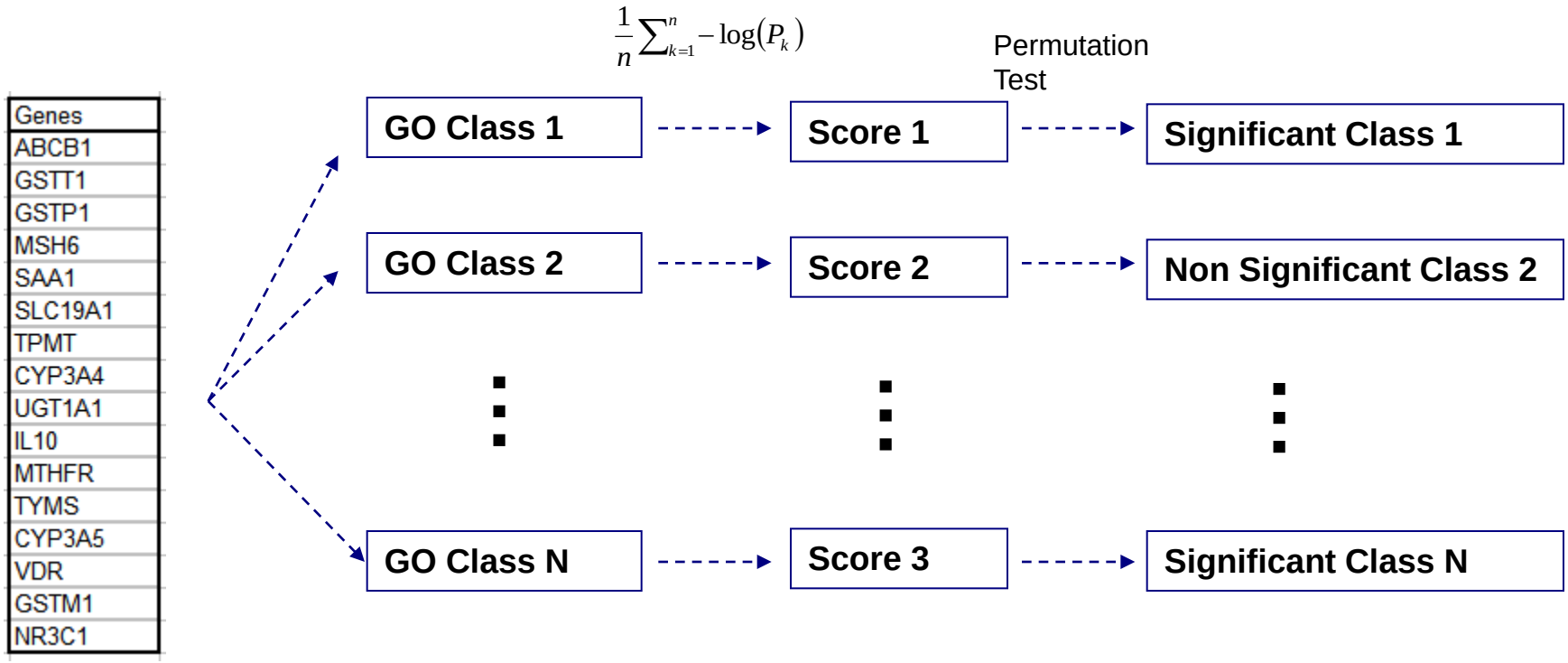
Network-Based Analysis

Overlap Analysis: ORA



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

Direct-Group Analysis: FCS

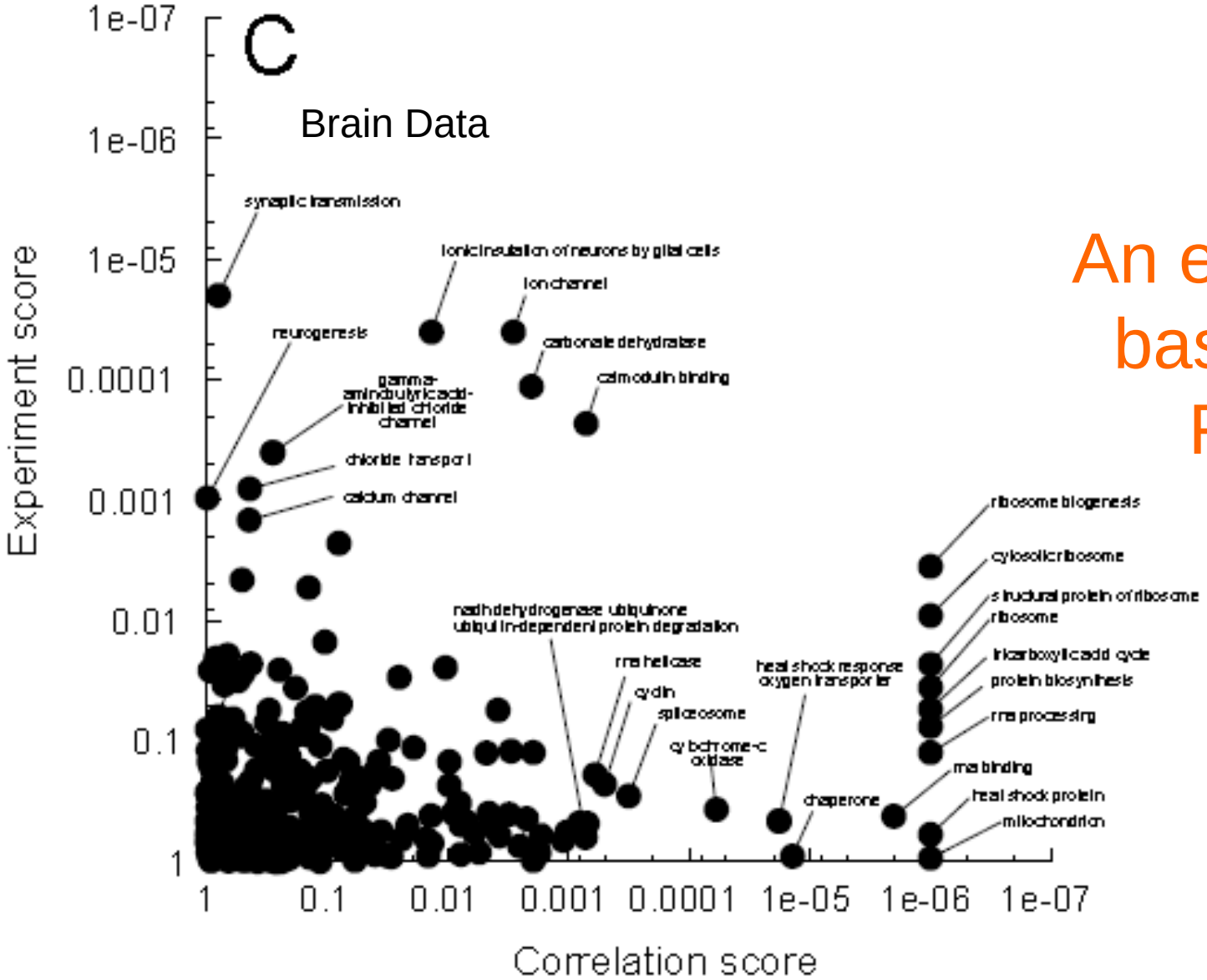


P Pavlidis et al. "Using the gene ontology for microarray data mining: A comparison of methods and application to age effects in human prefrontal cortex". *Neurochem Res*., 29(6):1213-1222, 2004.

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

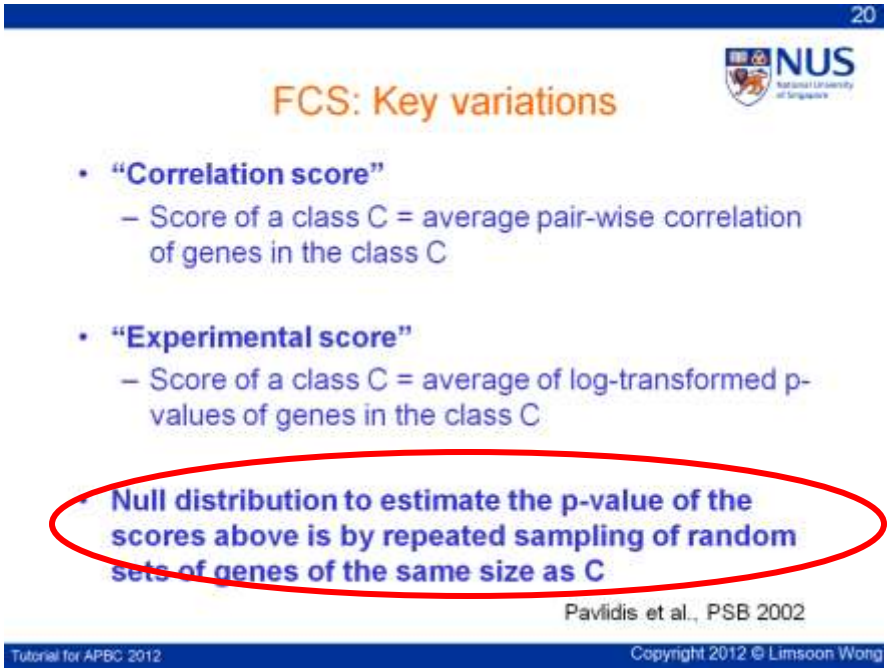


An example based on FCS

Pavlidis et al., PSB 2002

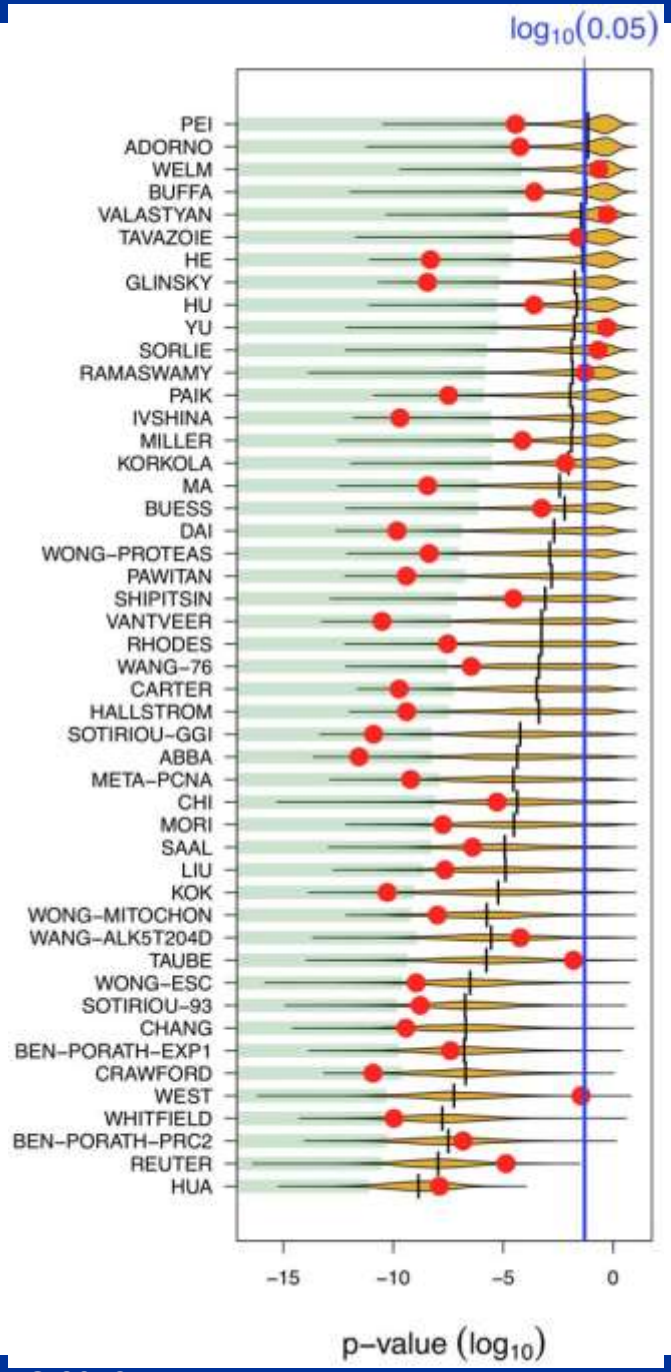
Goeman & Buhlmann. "Analyzing gene expression data in terms of gene sets: Methodological issues". *Bioinformatics*, 23(8):980-987, 2007

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



- **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.

Direct-Group Analysis: GSEA

Rank Genes

Assign score to each class based on gene rank

Permutation test

$P_{hit}(S_1, i)$

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

GO Class 1 → **Significant Class 1**

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

$P_{miss}(S_1, i)$

GO Class 2 → **Non Significant Class 2**

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

⋮

GO Class N → **Significant Class N**

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

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

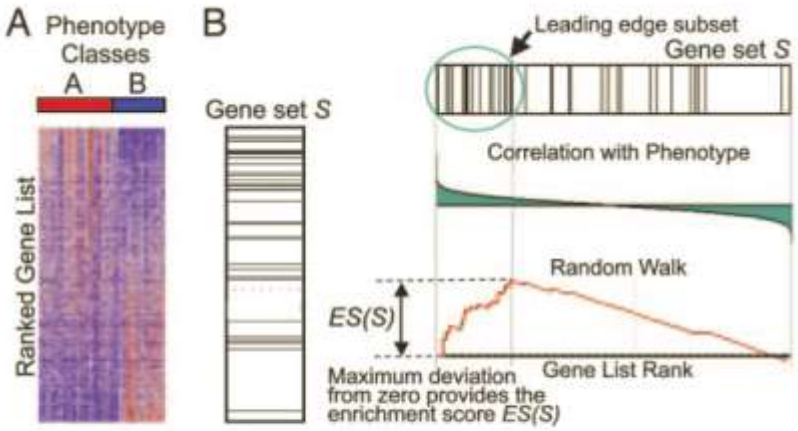


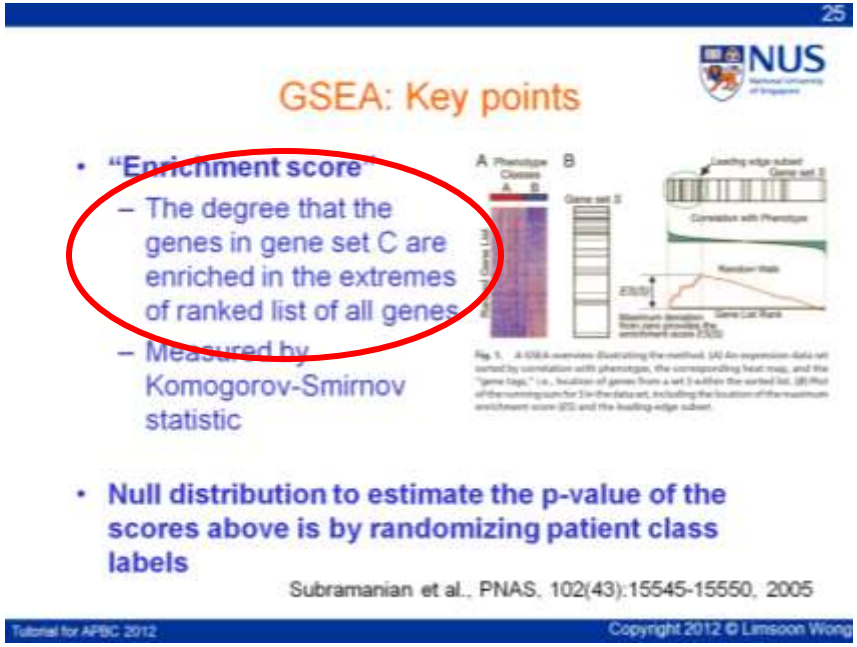
Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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



- **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 $SN^{u\_score}_{sn,d}$ with respect to phenotype d to each subnetwork sn within SN^{List} according to Equation 1.

$$SN^{u\_score}_{sn,d} = \langle SN^{i\_score}_{sn,1,d}, SN^{i\_score}_{sn,2,d}, \ldots, SN^{i\_score}_{sn,n,d} \rangle \quad (1)$$

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

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

$G^{score}_{sn,j,d}$ refers to the score of the jth 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:

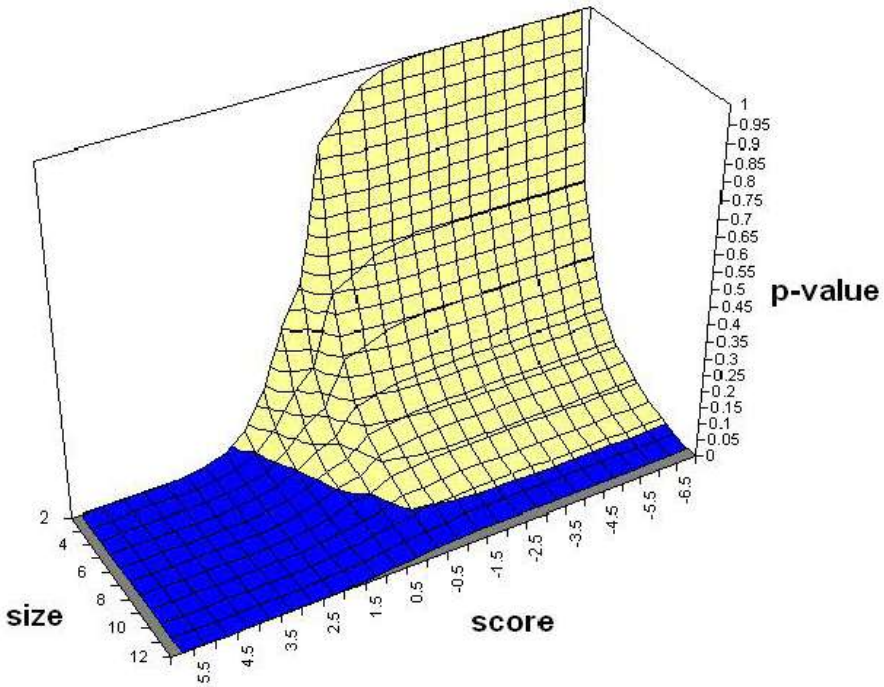
$$G^{score}_{sn,j,d} = k/n \quad (3)$$

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, $SN^{u\_score}_{sn,d}$ and $SN^{u\_score}_{sn,\neg d}$ 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}.

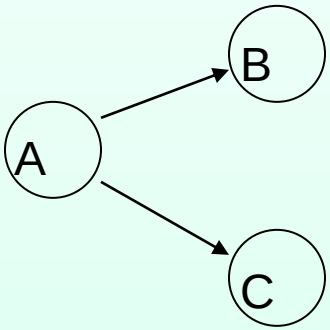
Soh et al. "Finding Consistent Disease Subnetworks Across Microarray Datasets". *BMC Bioinformatics*, 12(Suppl. 13):S15, 2011.

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



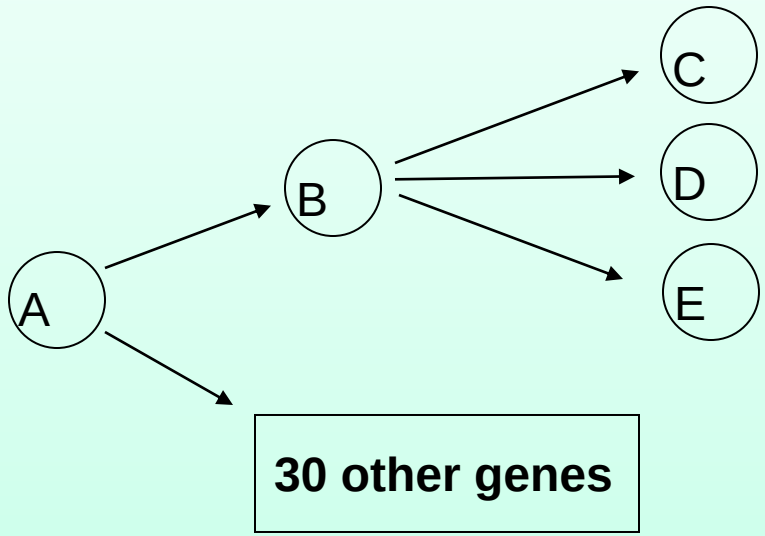
Genes A, B, C are high in phenotype *D*

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 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*

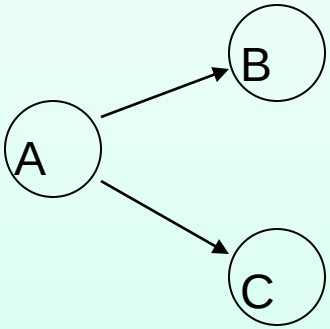
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

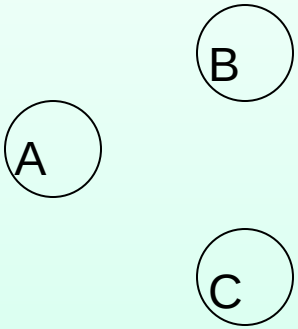
- **SNet: Able to capture the subnetwork branch within the pathway**

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.

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.

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

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**
- **SNet method yields more consistent and larger disease subnetworks**

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**

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

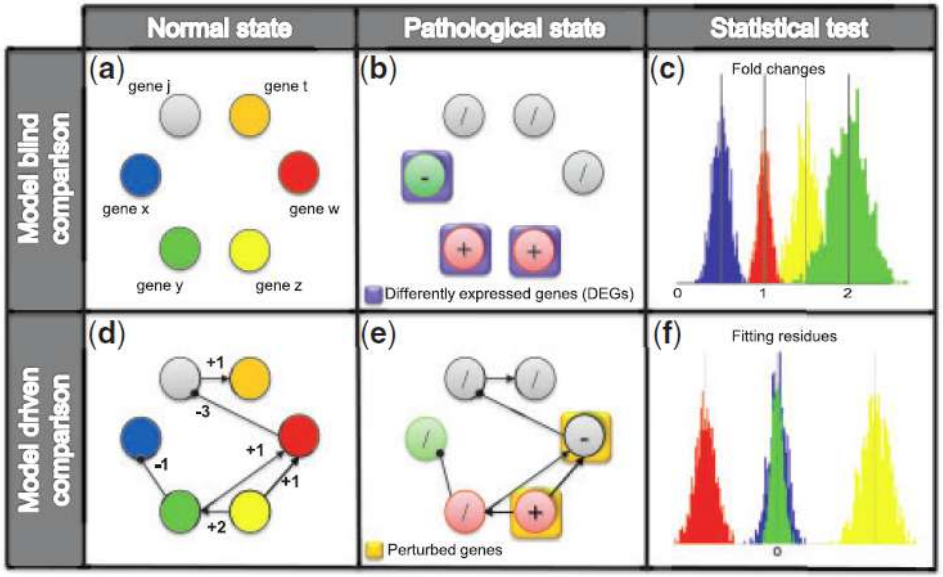


Fig. 1. System response inference: a toy genetic network consisting of six genes exemplifies the advantages of using a system-level data comparison (**a**). Standard statistical tests (i.e. *t*-test) unveil significant fold change in gene expression variations for each transcript individually (**b**), 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**).

Analysis of Gene Expression and Proteomic Profiles based on Biological Networks

Part 2

Limsoon Wong

NUS
National University of Singapore

Typical Proteomic MS Experiment

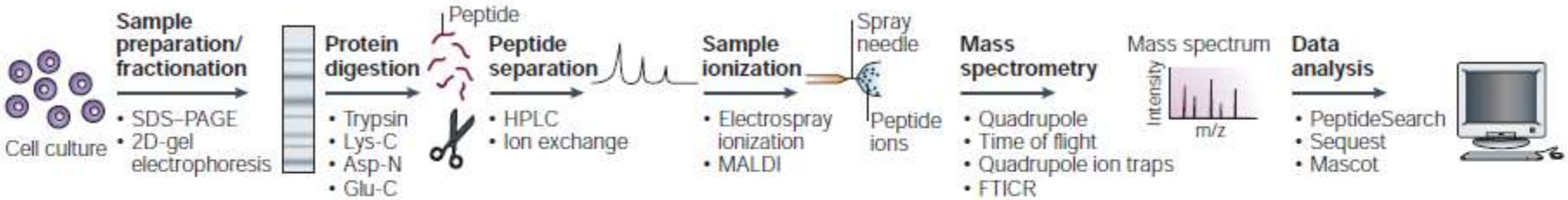
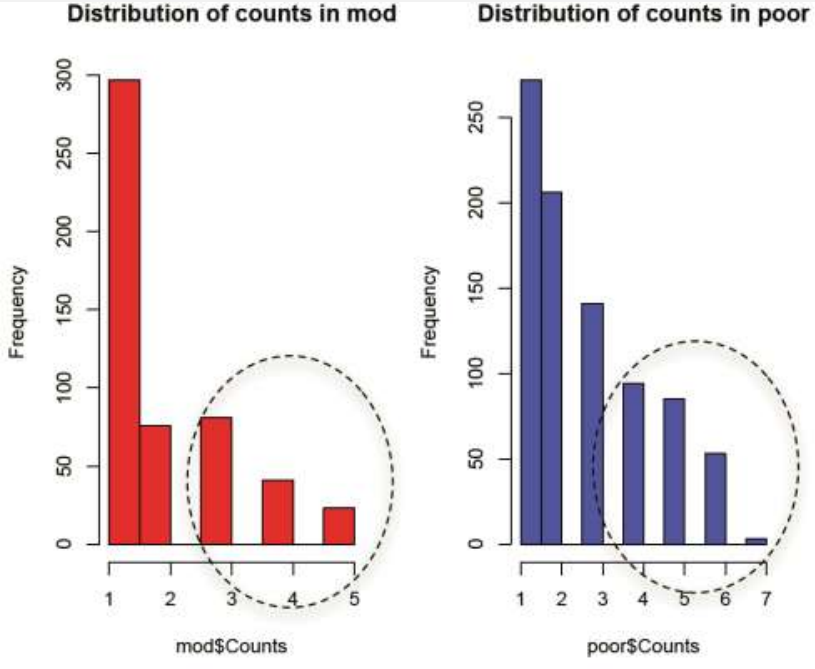


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous 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.

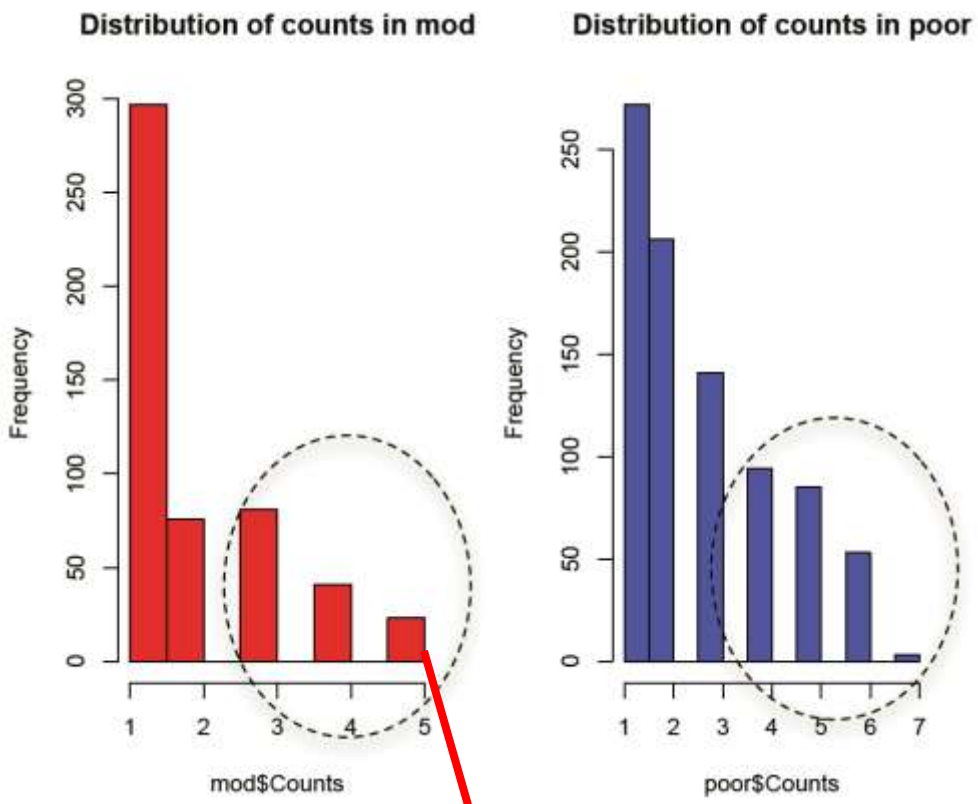
Source: Steen & Mann. The ABC's and XYZ's of peptide sequencing. *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004

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



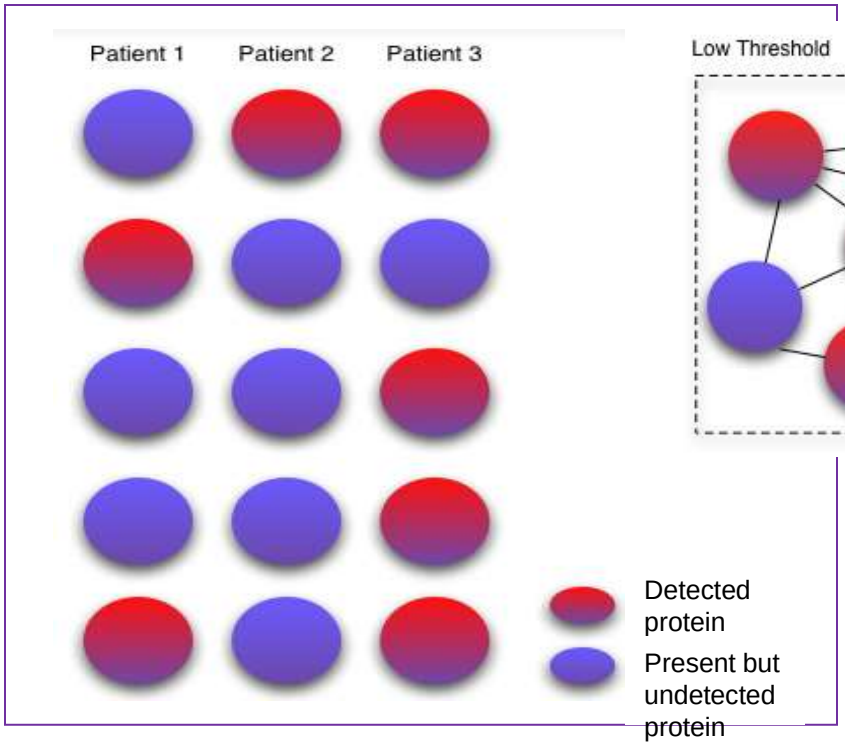
Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

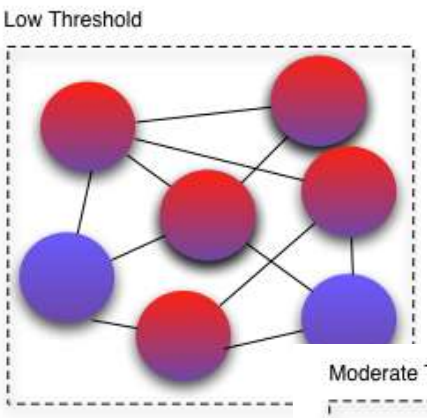
Issues in Proteomic Profiling

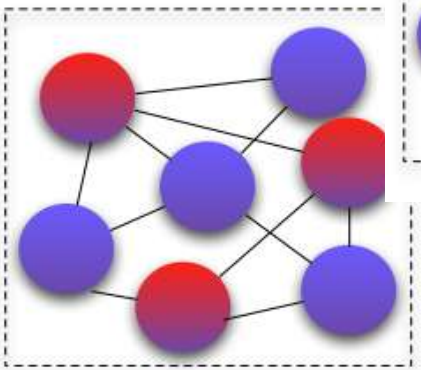
- **Coverage**
- **Consistency**

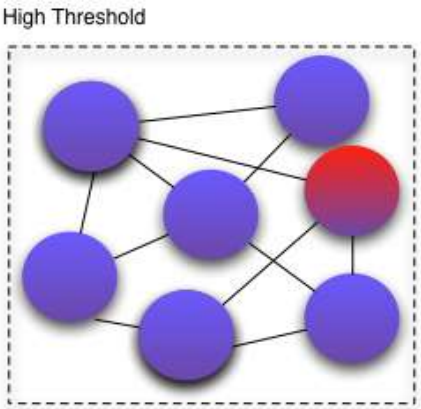
⇒ **Thresholding**

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

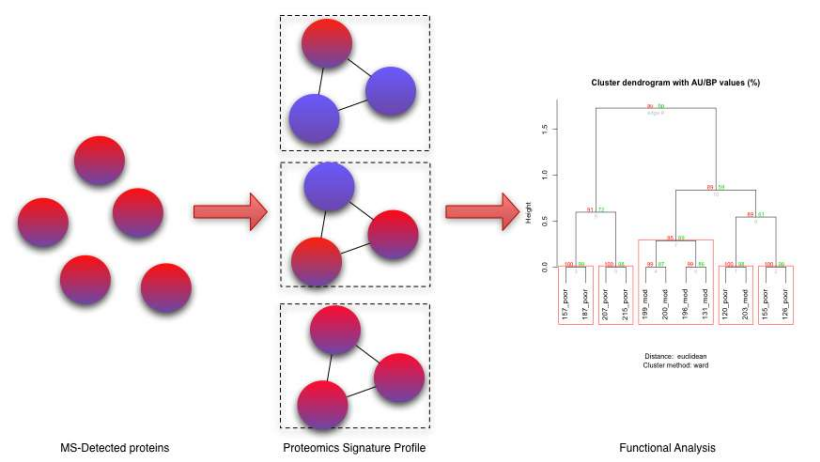








Part 2: Delivering more powerful proteomic profile analysis

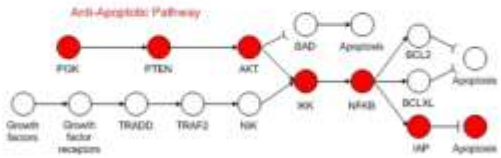


- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

An inspiration from gene expression profile analysis

11

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

12

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) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$~~

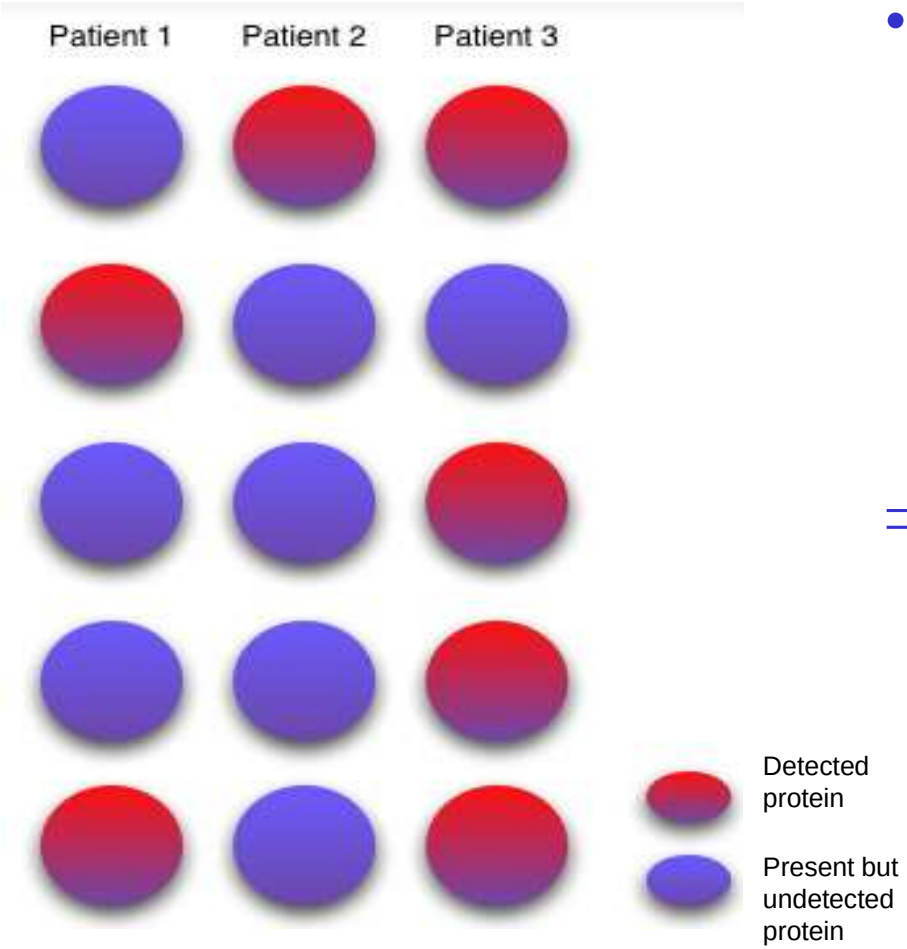
# of pathways = 1000

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

⇒ Even more false positives?

- Perhaps no need to consider every group

Intuitive Example



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

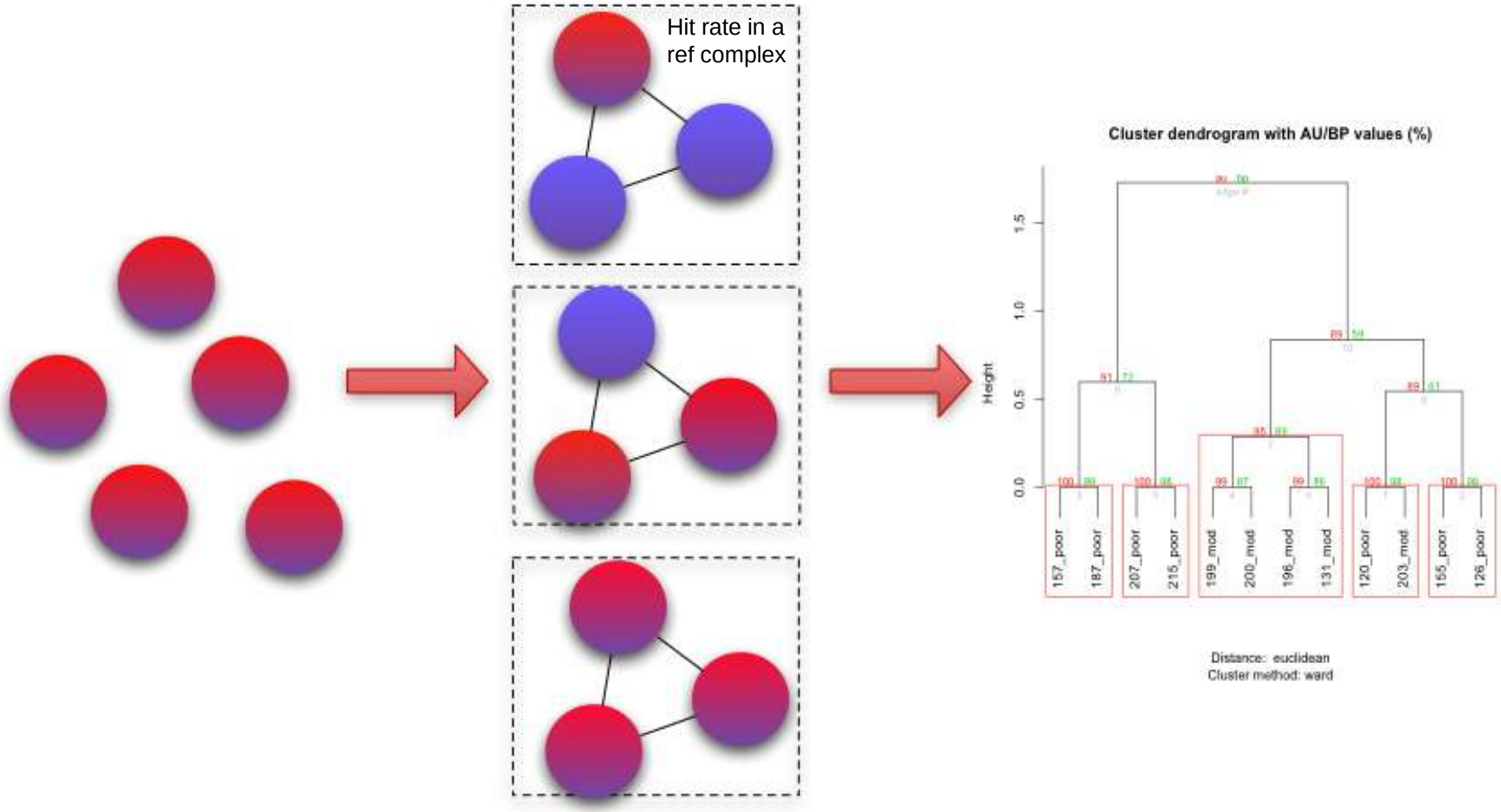
⇒ **Diff patients suffering from the disease can have a diff protein component missing**
  - Construct a profile based on complexes?

We try an adaptation of SNet on proteomics profiles…

“Proteomic Signature Profiling” (PSP)

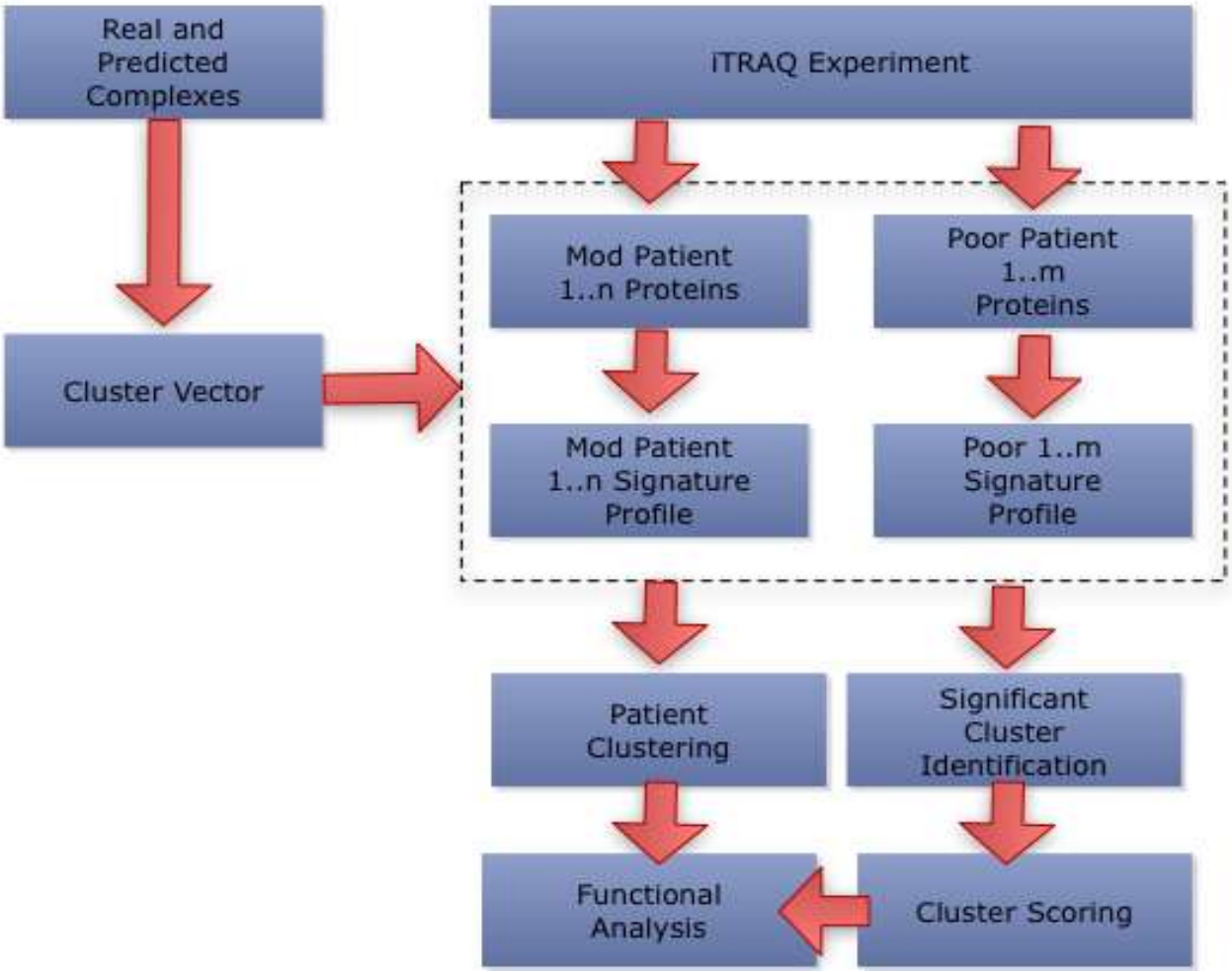
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

"Threshold-free" Principle of PSP



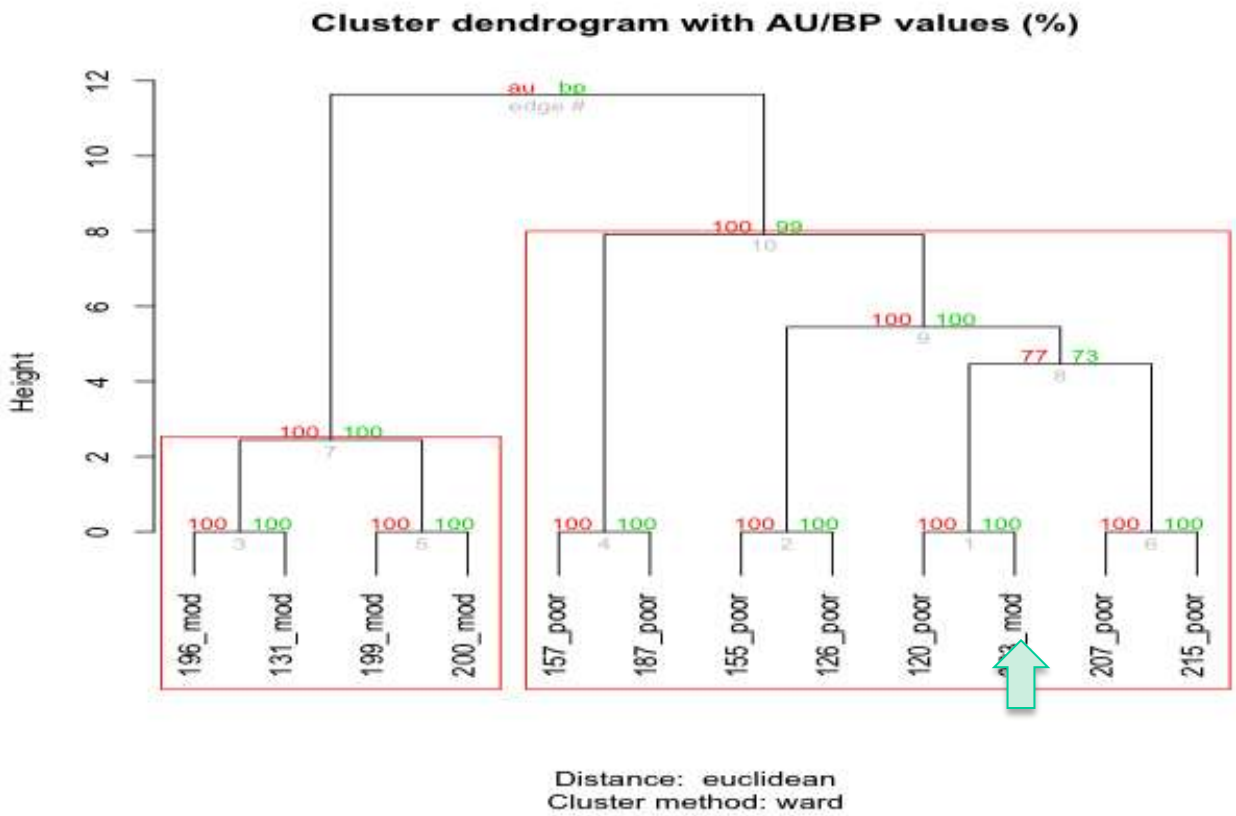
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

Applying PSP to a HCC Dataset



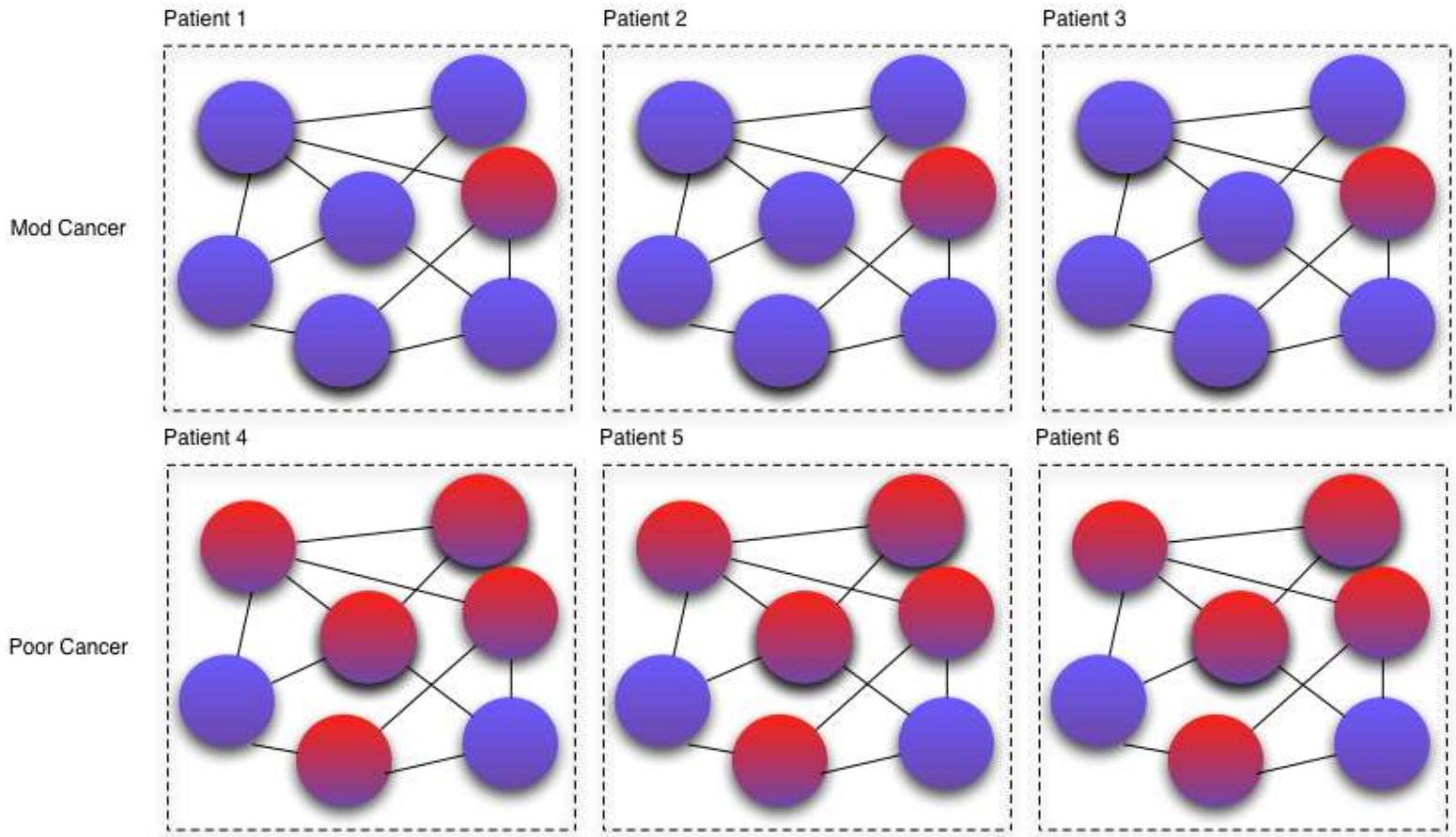
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

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

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…

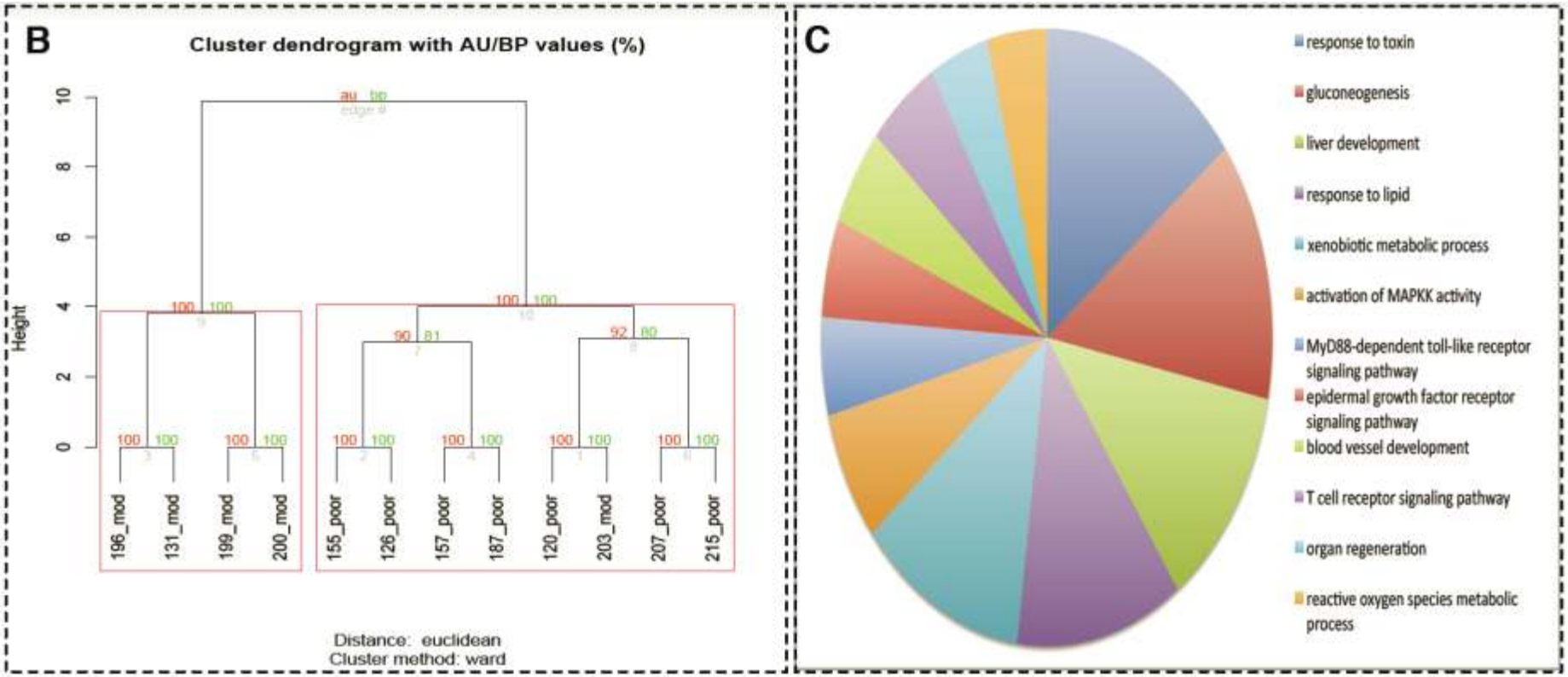
“Pathway-Derived Subnets” (PDS)

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

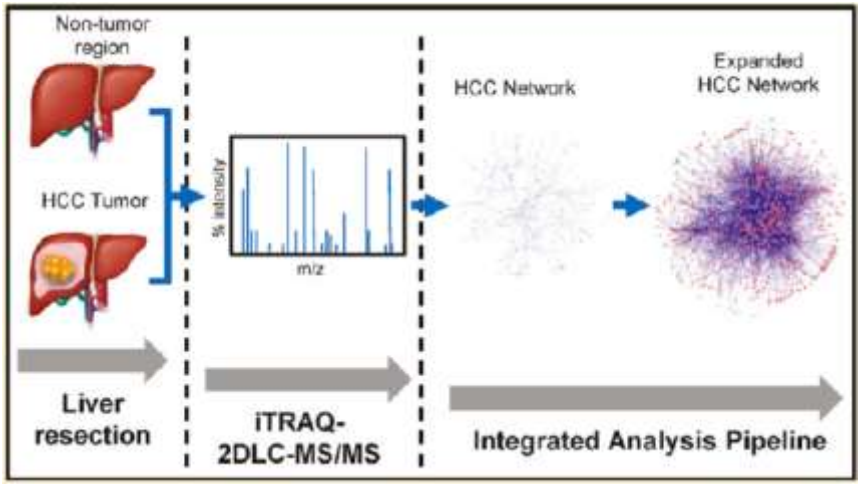
PDS consistently segregates mod vs poor patients



What have we learned?

- **PSP can deal with consistency issues in proteomics**
- **GO term analysis also indicates that PSP selects clusters that play integral roles in cancer**
- **PSP reveals 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

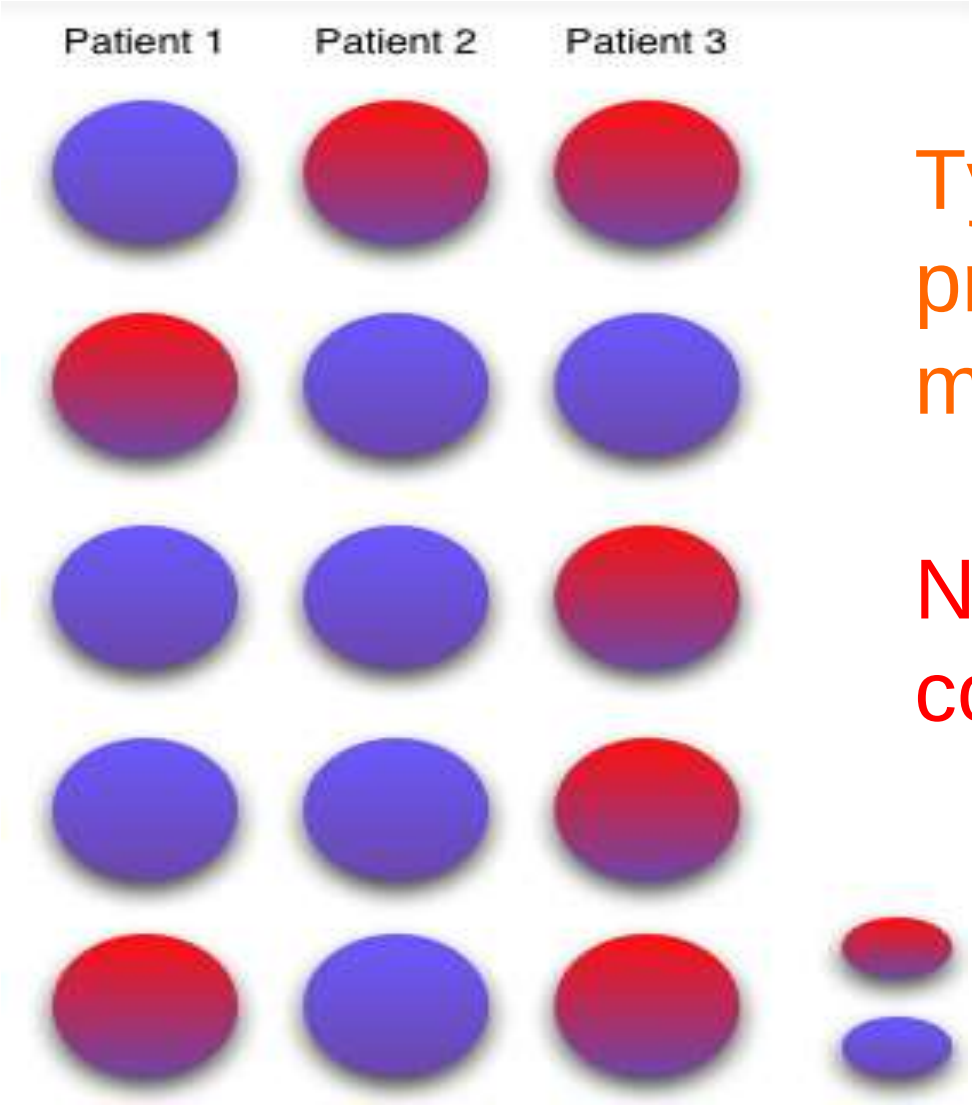


- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

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



Typical proteomic profiling misses many proteins

Need to improve coverage!

Basic Approach

- **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**

$\Rightarrow$ **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.** APBC 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 …

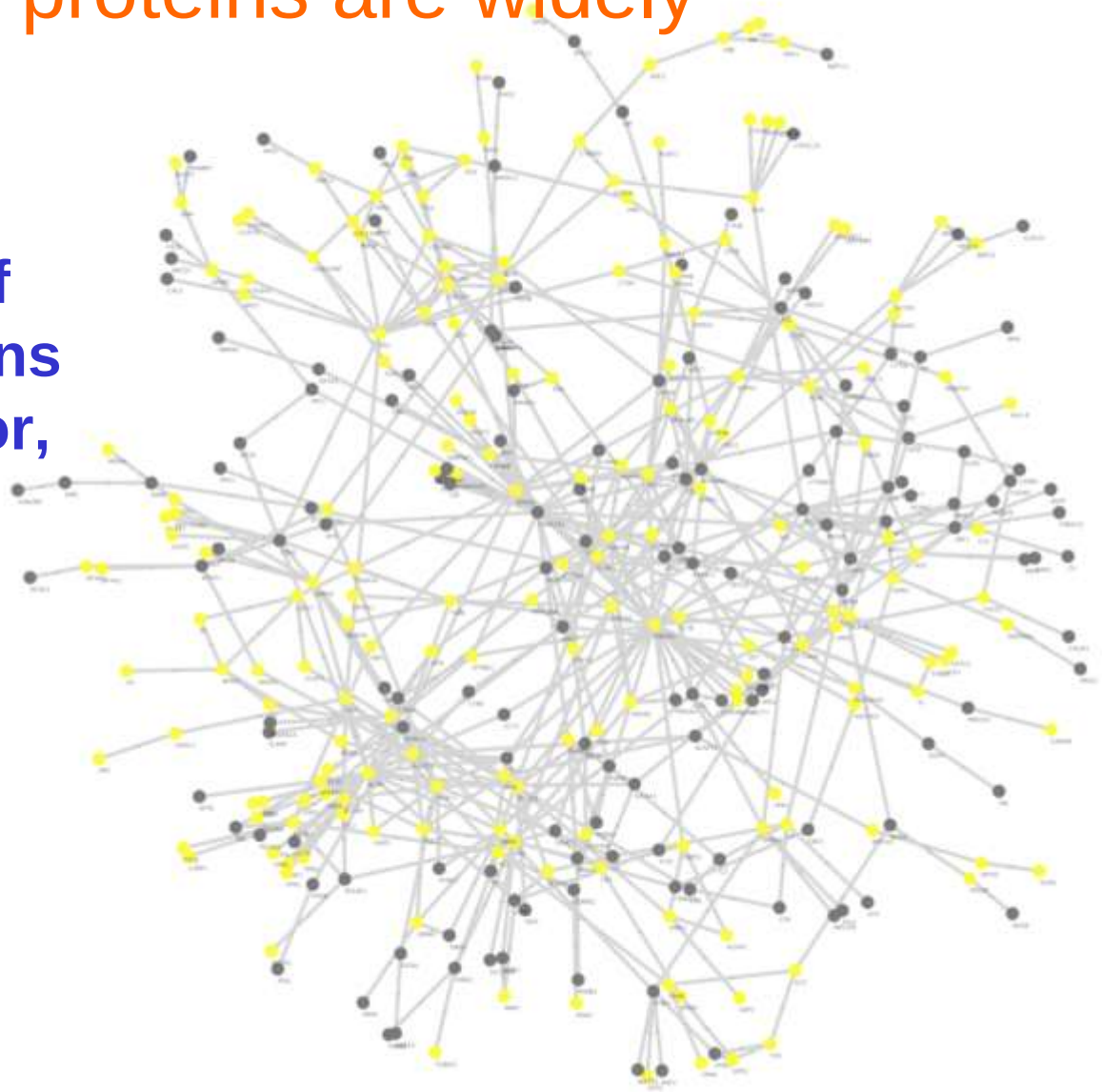
Background

- **HCC (Hepatocellular carcinoma)**
  - Classified into 3 phases: differentiated, moderately differentiated and poorly differentiated

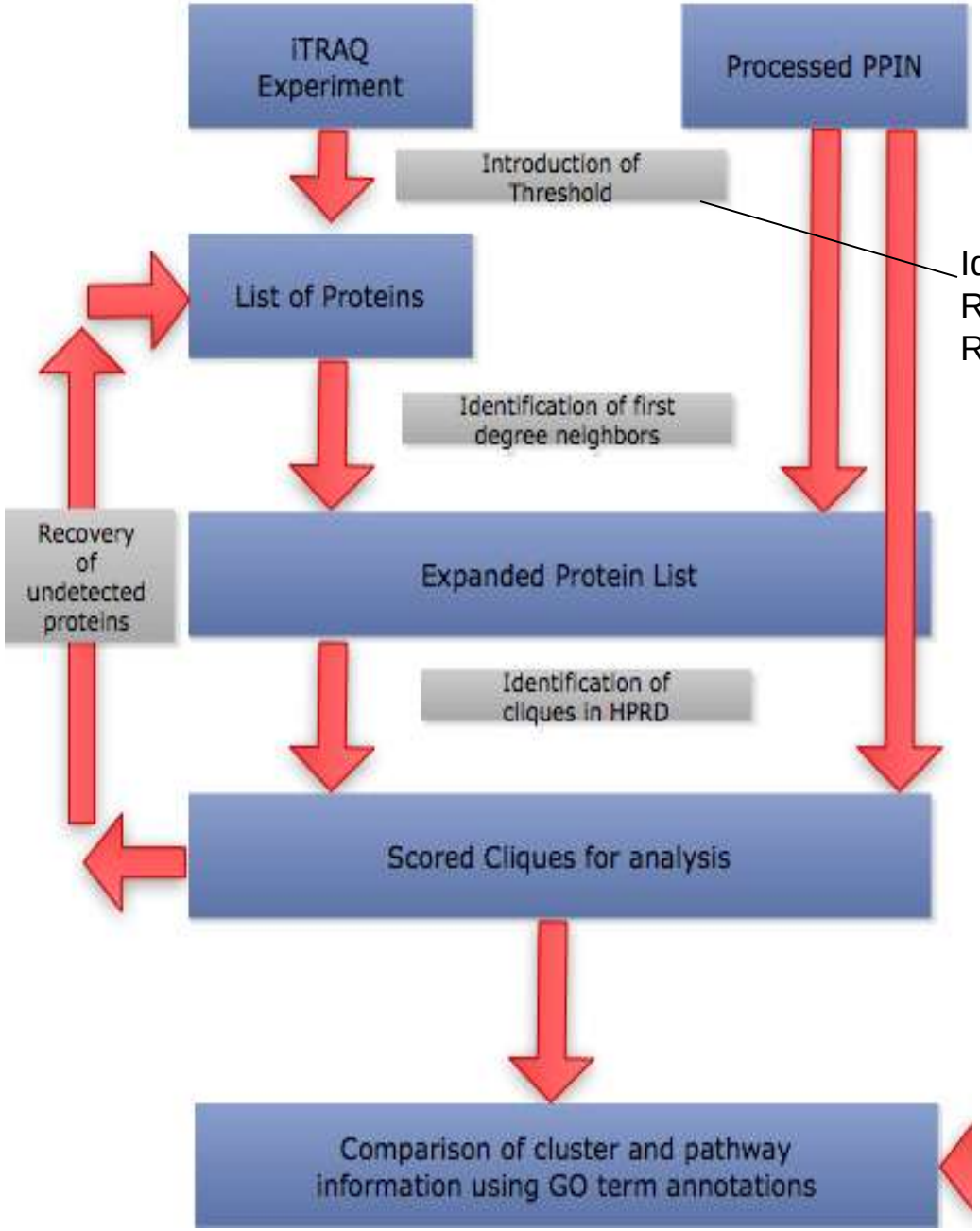
- **Mass Spectrometry**
  - iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)
  - Coupled with 2D LC MS/MS
  - Popular because of ability to run 8 concurrent samples in one go

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



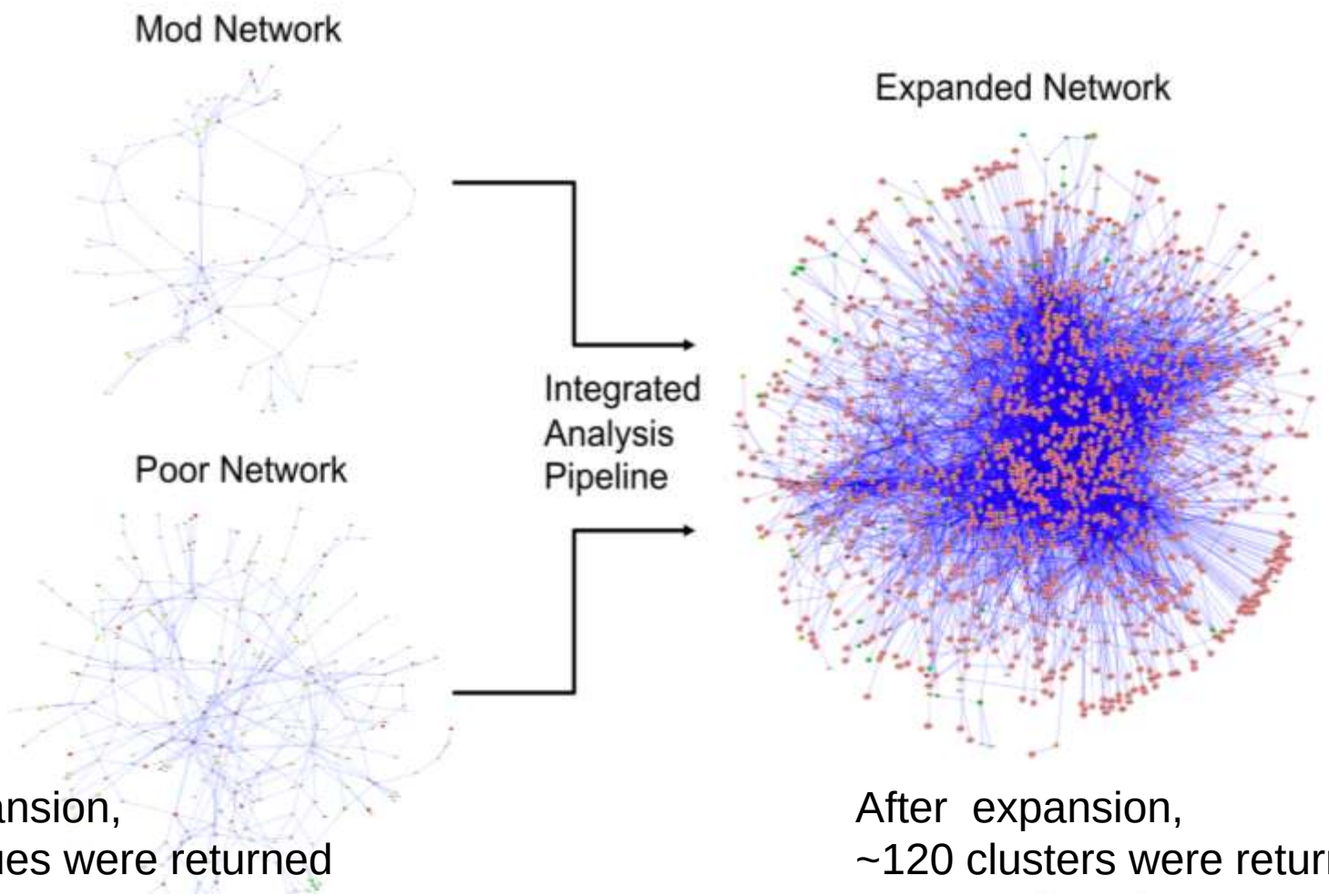
Identify the "seeds"
Ratio < 0.8 and > 1.25 for Mod (min 3 patients)
Ratio < 0.8 and > 1.25 for Poor (min 4 patients)

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**

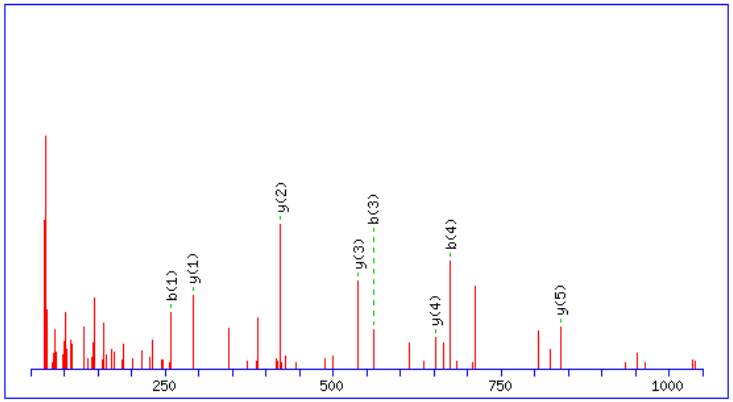
- **Assessment of MS/MS spectra of their top ranked peptides revealed accurate y- and b-ion assignments and were of good quality ($p < 0.05$)**

$\Rightarrow$ **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

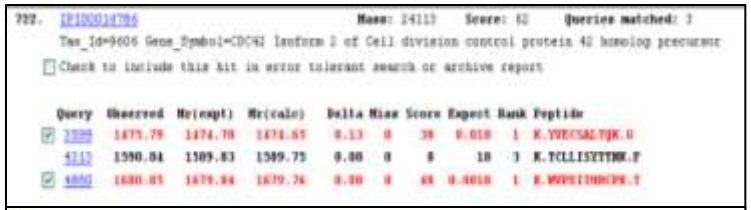
ACTR2

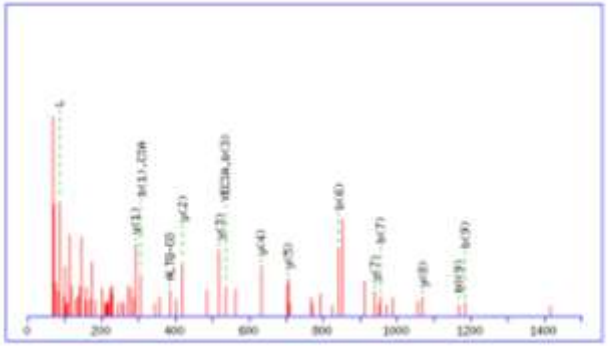


MONOISOTOPIC mass of neutral peptide Mr(calc): 1095.44
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 39 **Expect:** 0.018
Matches (Bold Red): 8/57 fragment ions using 15 most intense peaks

#	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



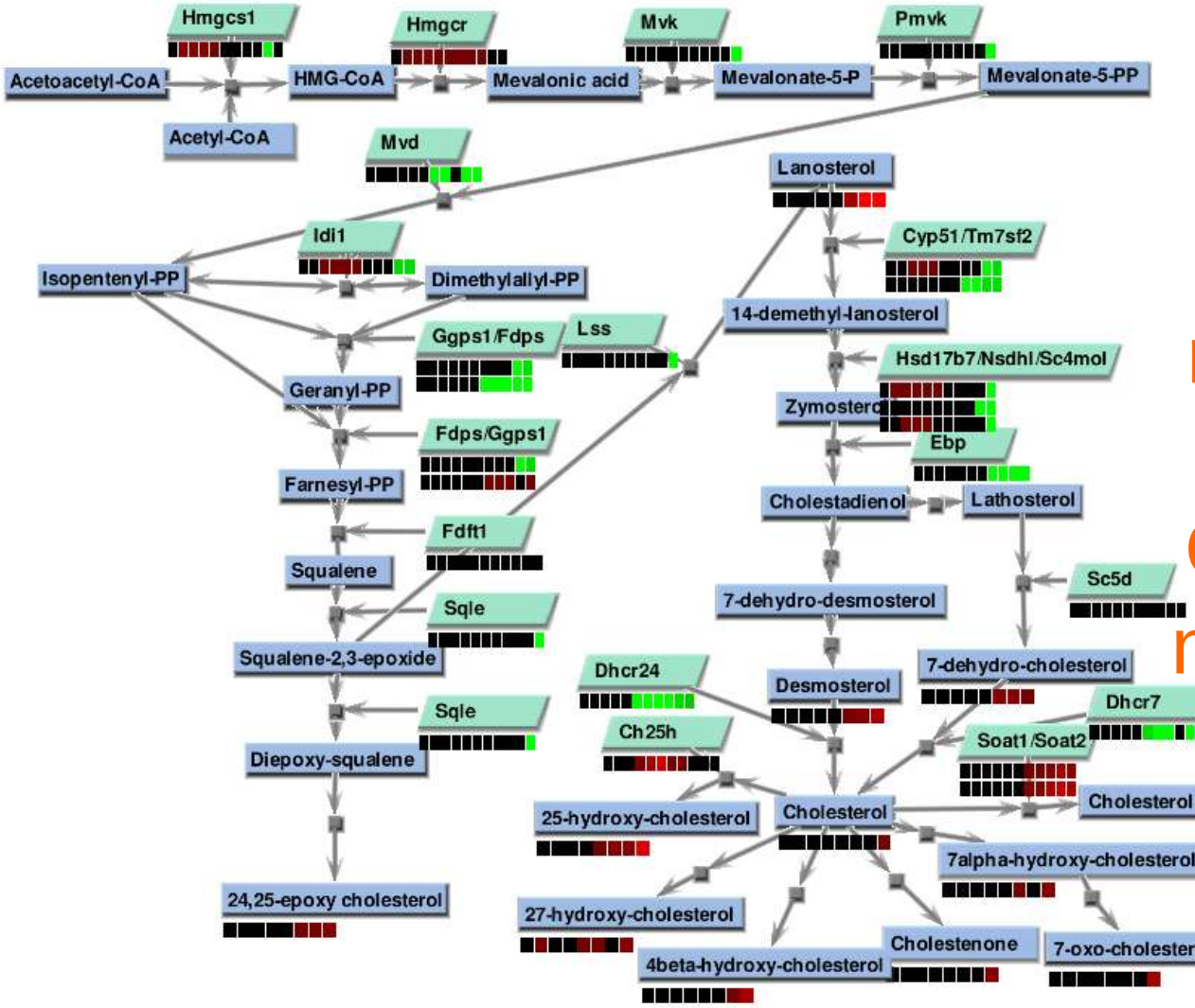


MONOISOTOPIC mass of neutral peptide Mr(calc): 1474.65
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 30 **Expect:** 0.018
Matches (Bold Red): 17/119 fragment ions using 26 most intense peaks

#	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

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*, in press
- [PSP] Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted
- [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
- [MaxLink] Goh et al. **A Network-based maximum-link approach towards MS**. APBC 2012



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

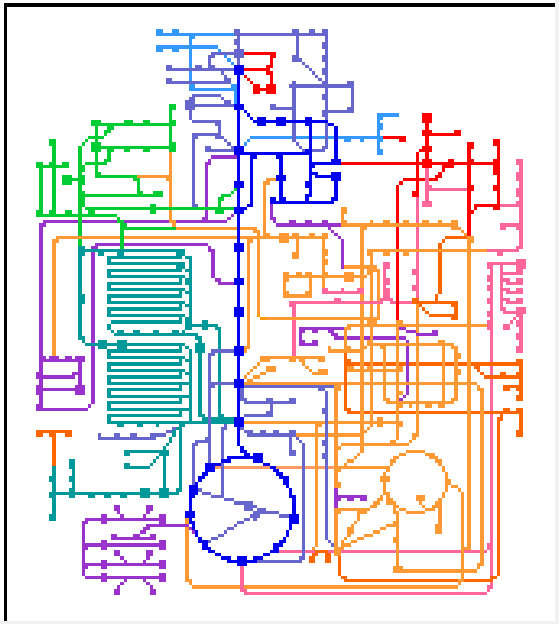
Analysis of Gene Expression and Proteomic Profiles based on Biological Networks

Part 3

Limsoon Wong

NUS National University of Singapore

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)

and Protein Complexes

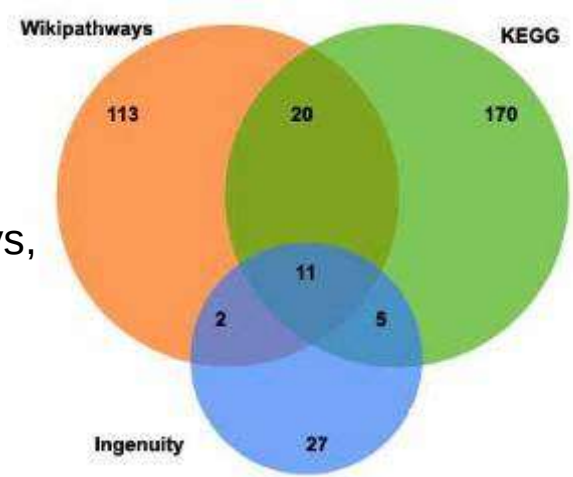
- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, NAR, 2010

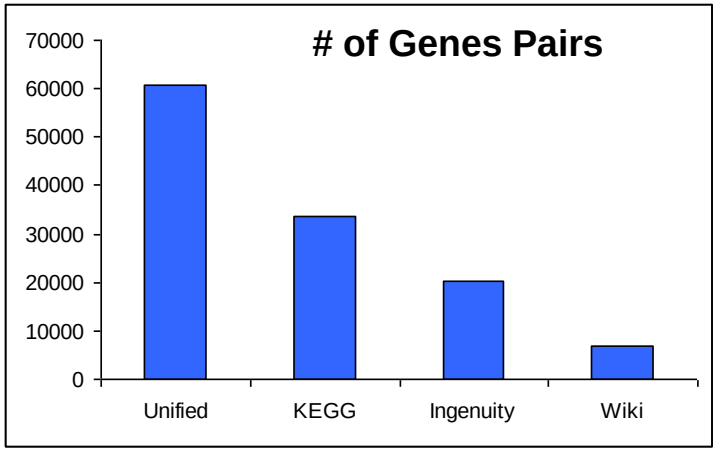
Sources of Biological Pathways

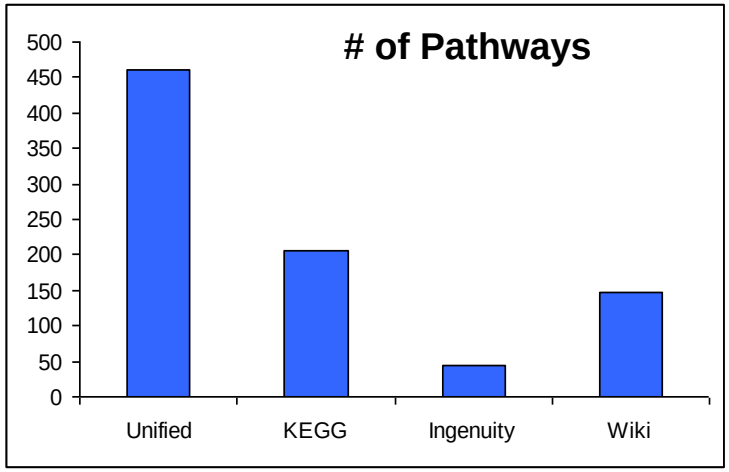
Database	Remarks
KEGG	KEGG (`http://www.genome.jp/kegg`) is one of the best known pathway databases (Kanehisa *et al.*, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time 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.

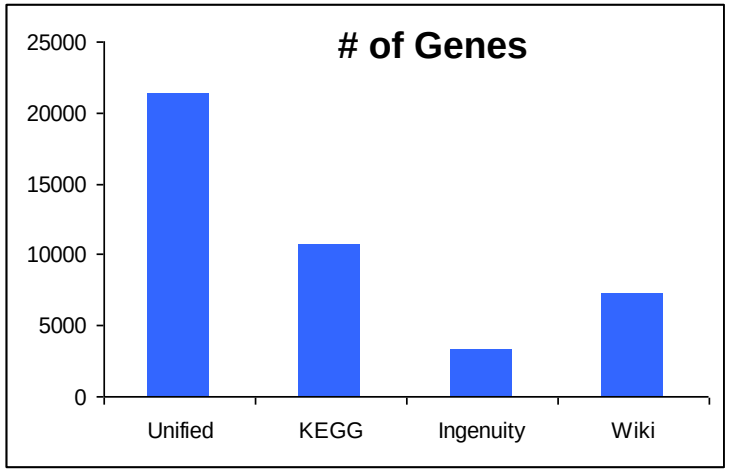
Low Comprehensiveness of Pathway Sources

Human pathways in Wikipathways, KEGG, & Ingenuity



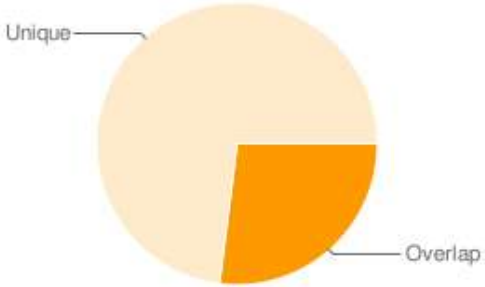




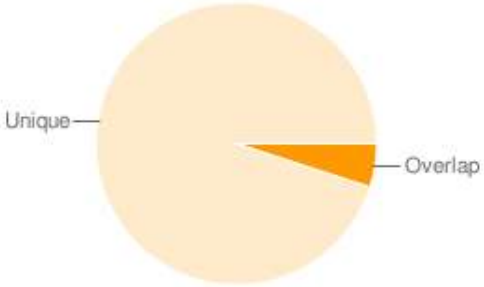


Low Consistency of Pathway Sources

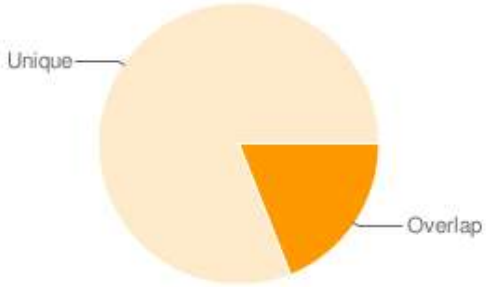
Gene Pair Overlap



Wiki vs KEGG

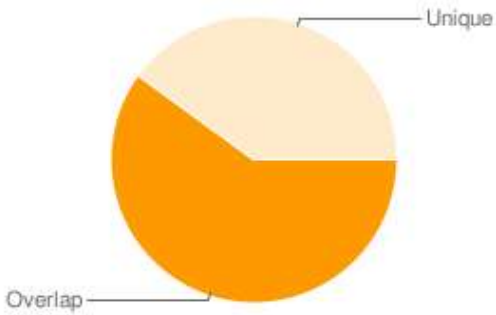


Wiki vs Ingenuity

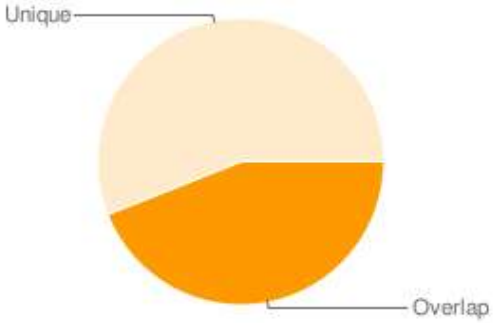


KEGG vs Ingenuity

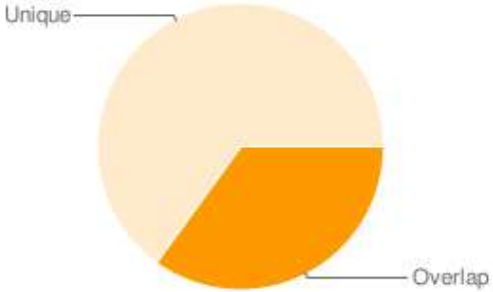
Gene Overlap



Wiki vs KEGG



Wiki vs Ingenuity



KEGG vs Ingenuity

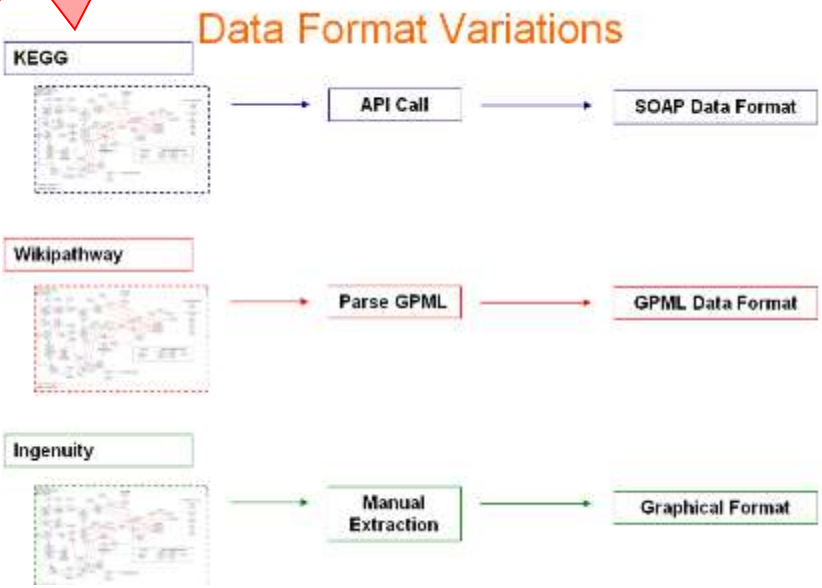
Example: Apoptosis Pathway

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%

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

Incompatibility Issues

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

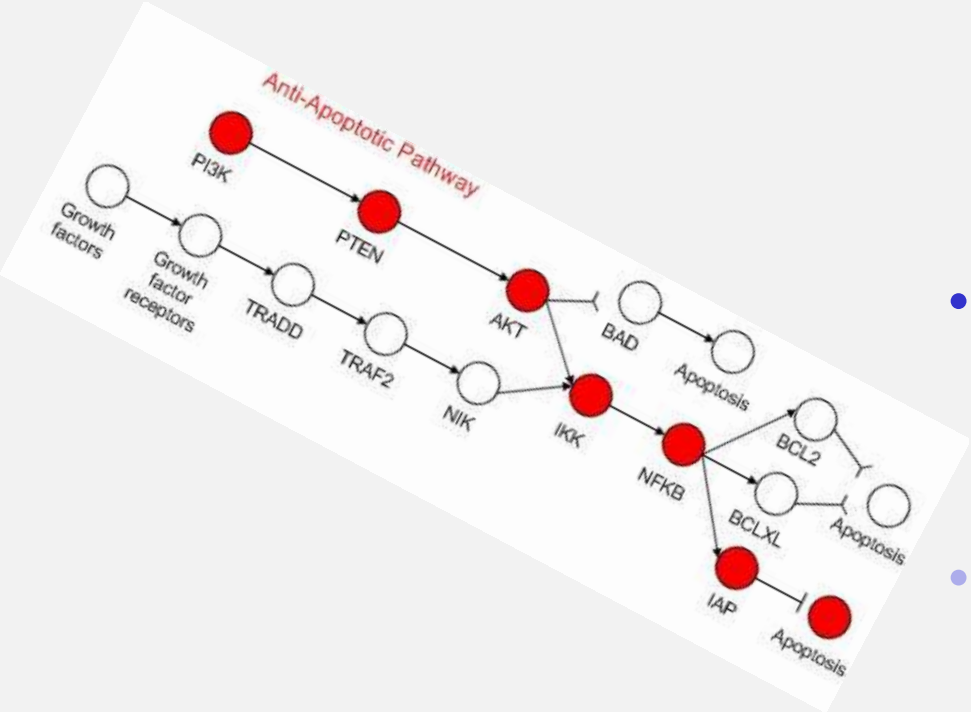


The preceding analyses hide an intricate issue…

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?

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



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

Possible Ways to Match Pathways

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

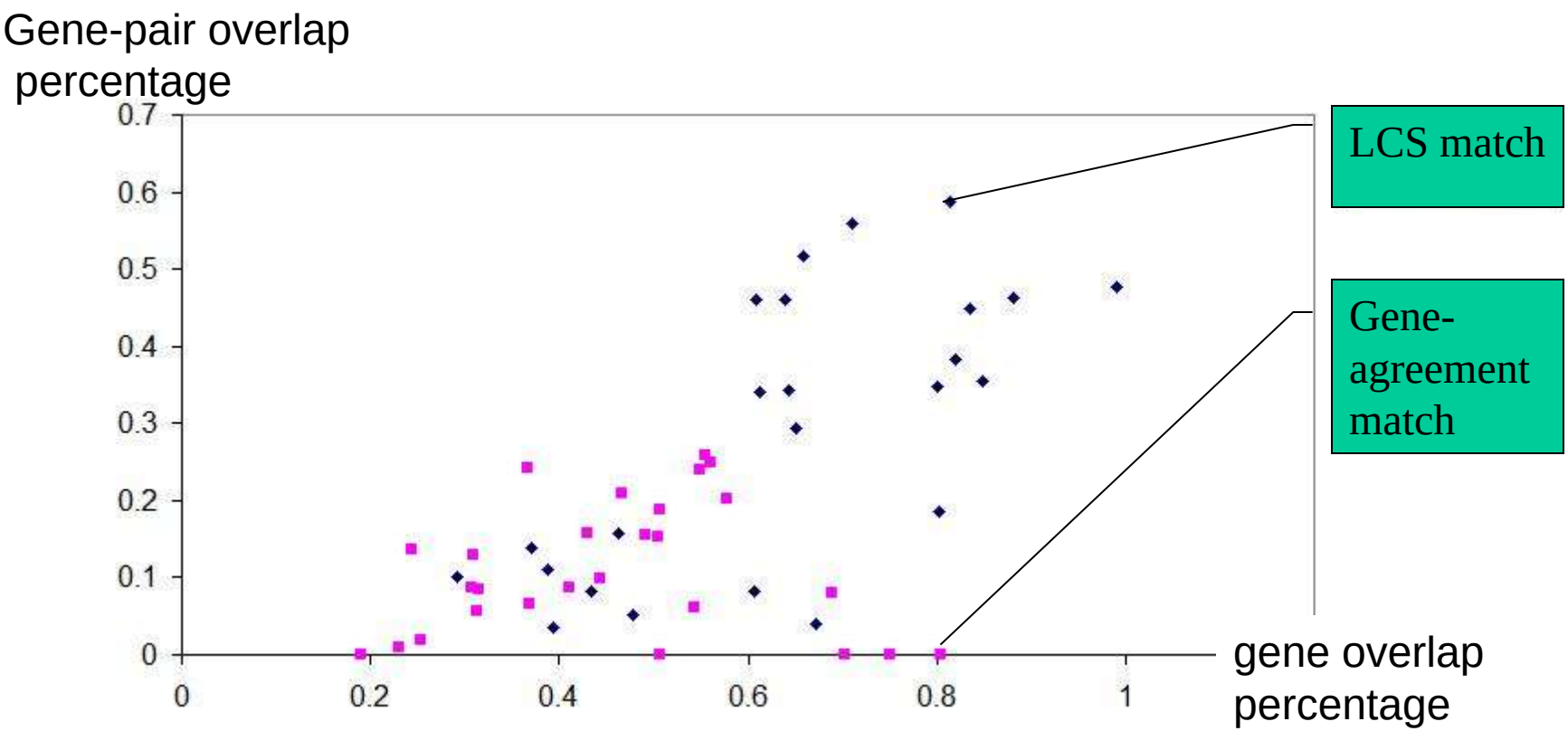
  ⇒Likely to mismatch pathways?

  ⇒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

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

LCS vs Gene-Agreement Matching

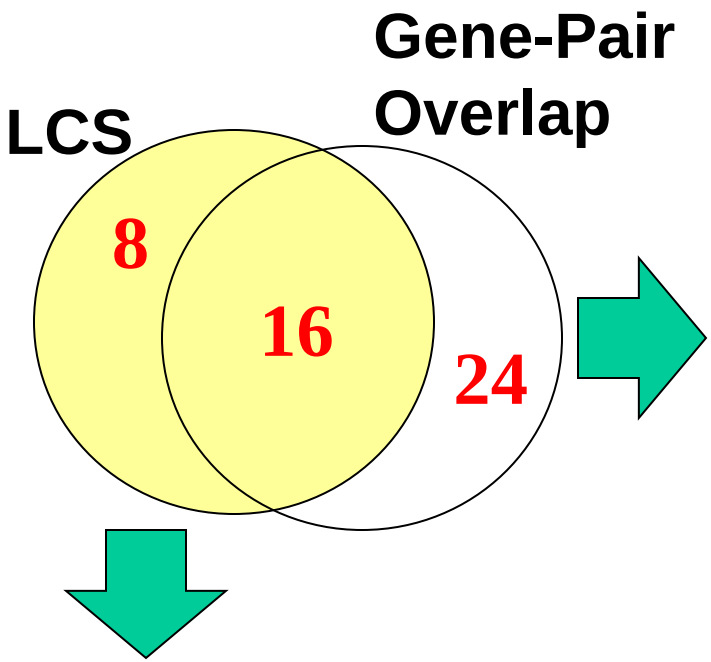


- **LCS consistently has higher gene-pair agreement**

⇒ LCS is better than gene-agreement based matching!

Soh et al. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *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.

- **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

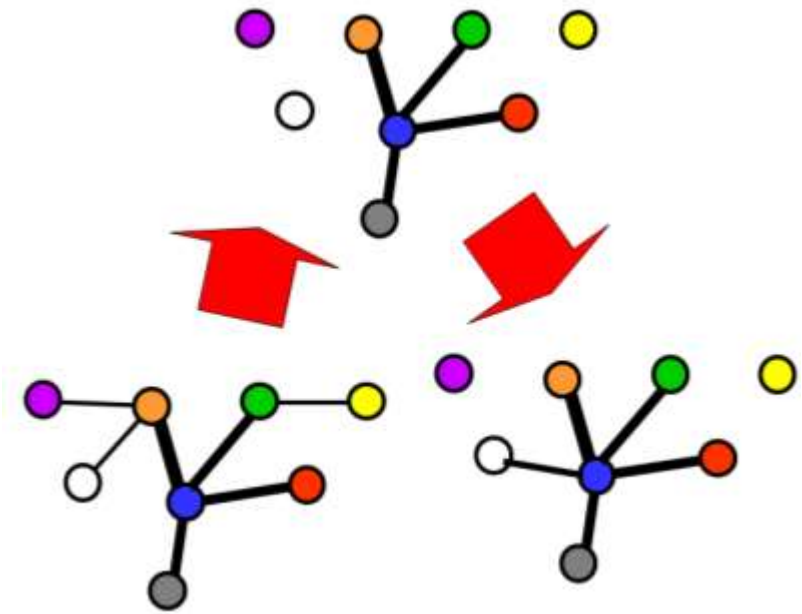
Donny Soh, Difeng Dong, Yike Guo, Limsoon Wong. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *BMC Bioinformatics*, 11:449, September 2010.

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?

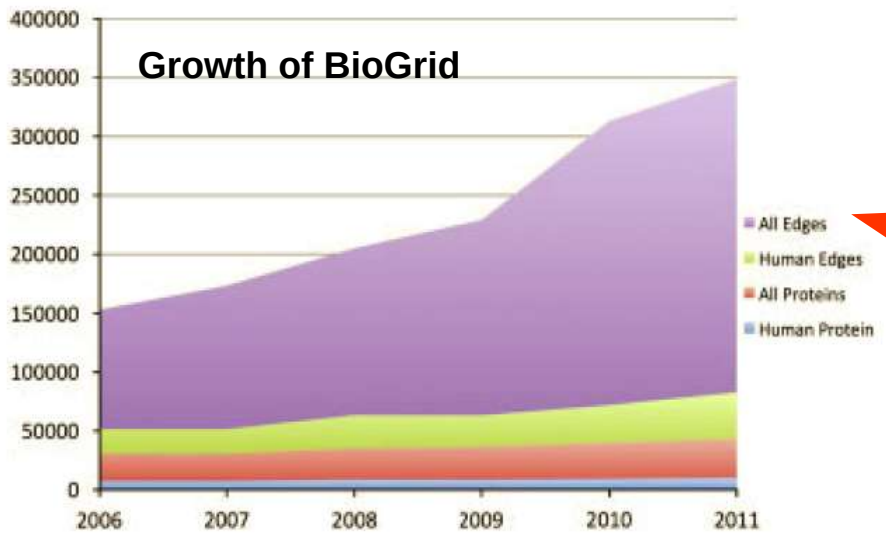


- **Sources of pathway & PPIN**
  - Comprehensiveness
  - Consistency
  - Compatibility
- **Integration**
  - Pathway matching
- **PPIN cleansing**

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**

$\Rightarrow$ **Need to clean up before making inference on PPI networks**

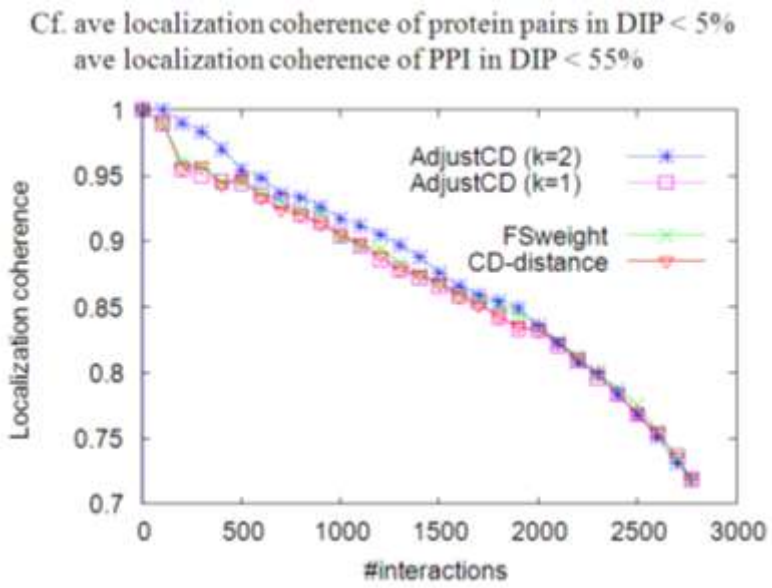
Chua & Wong. **Increasing the Reliability of Protein Interactomes**.
Drug Discovery Today, 13(15/16):652--658, 2008

Dealing with noise in PPIN

- **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



References

- A Ruepp et al. **CORUM: The comprehensive resource of mammalian protein complexes---2009**. *Nucleic Acids Research*, 38:D497-D501, 2010
- M Kanehisa et al. **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Research*, 38:D355-D360, 2010
- I Vastrik et al. **Reactome: A knowledge base of biologic pathways and processes**. *Genome Biology*, 8:R39, 2007
- EG Cerami et al. **Pathway Commons, a web resource for biological pathway data**. *Nucleic Acids Research*, 39:D685-D690, 2011
- D Soh et al. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases**. *BMC Bioinformatics*, 11:449, 2010
- Chua & Wong. **Increasing the Reliability of Protein Interactomes.** *Drug Discovery Today*, 13(15/16):652--658, 2008

NUS National University of Singapore

Acknowledgements

Donny Soh

Difeng Dong

Wilson Goh

- A*STAR AIP scholarship
- A*STAR SERC PSF grant
- NRF CRP grant
- Wellcome Trust scholarship

Agency for Science, Technology and Research SINGAPORE

wellcome trust

NATIONAL RESEARCH FOUNDATION
Prime Minister's Office, Republic of Singapore