Disease gene expression and proteomic profile analysis based on regulatory networks

Limsoon Wong

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

- A related 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

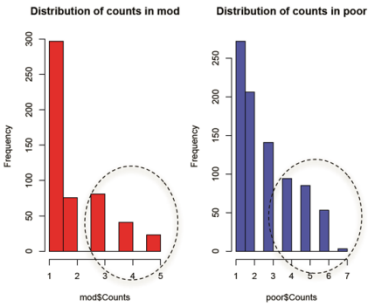
  http://www.comp.nus.edu.sg/~wongls/talks/apbc2012-tutorialnotes.pdf

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



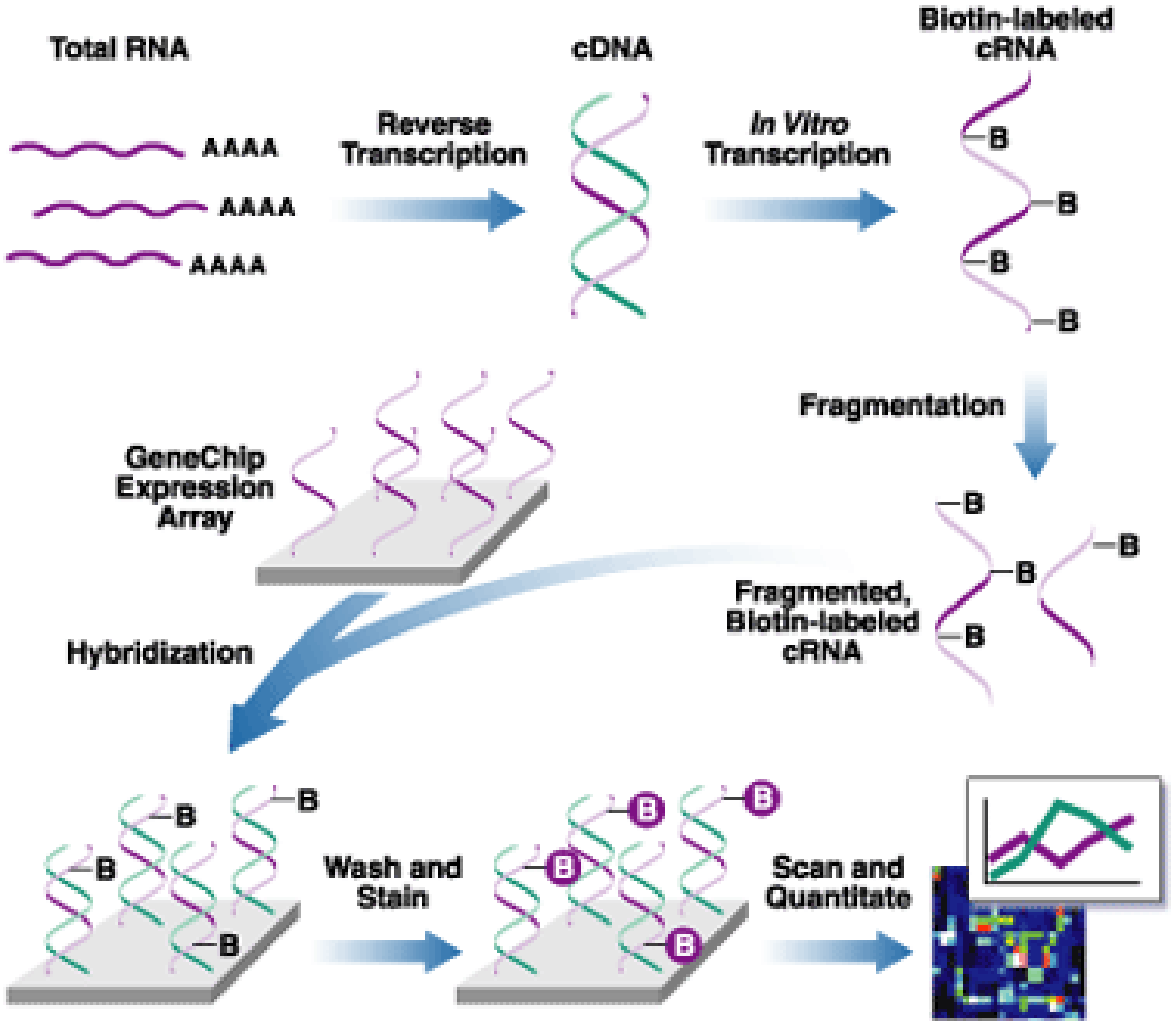
- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP
  - PDS
- **Improving coverage**
  - CEA
  - PEP
  - Max Link

Part 1:
Delivering Reproducible Gene Expression analysis

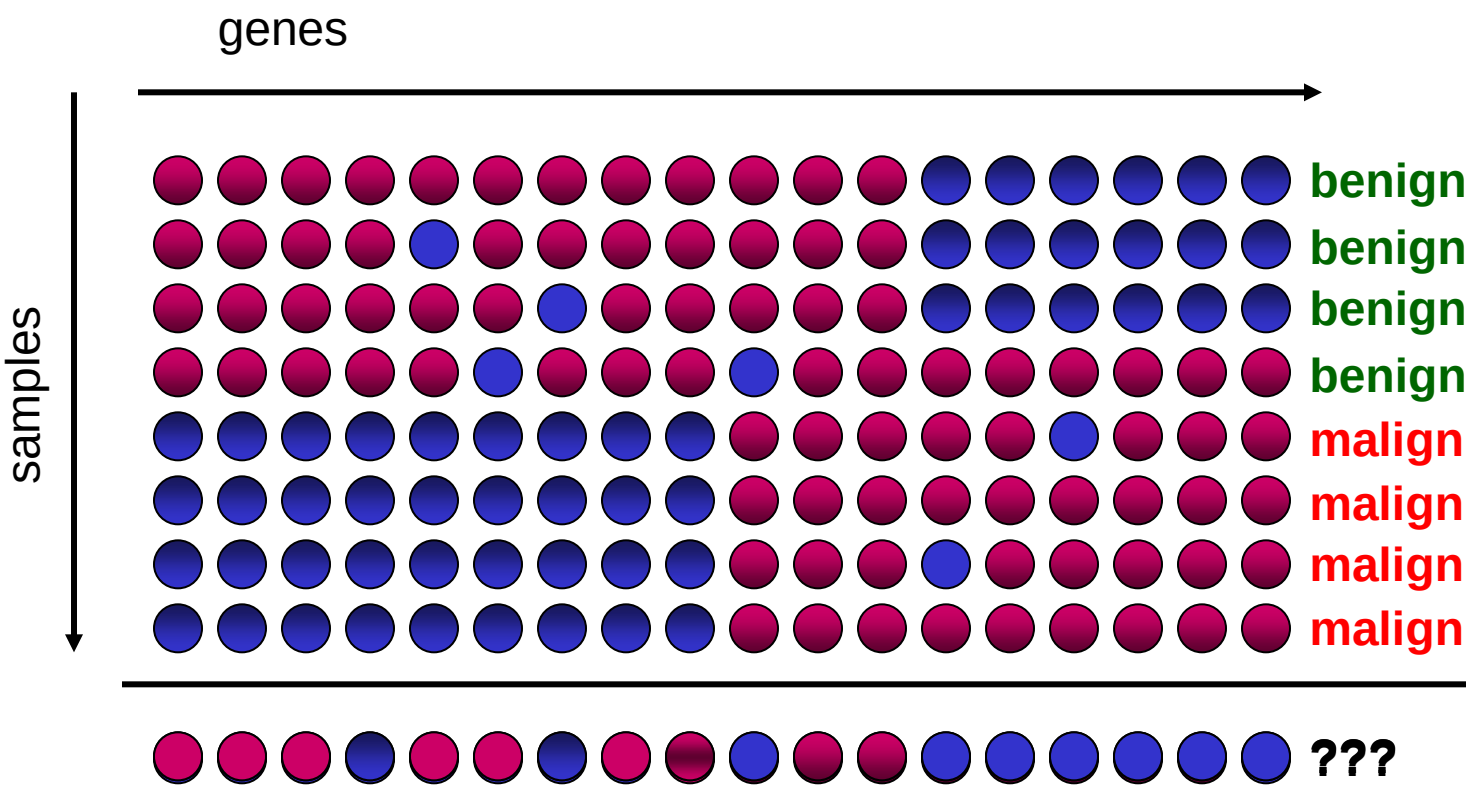
Limsoon Wong

Gene Expression Measurement by Affymetrix GeneChip Array

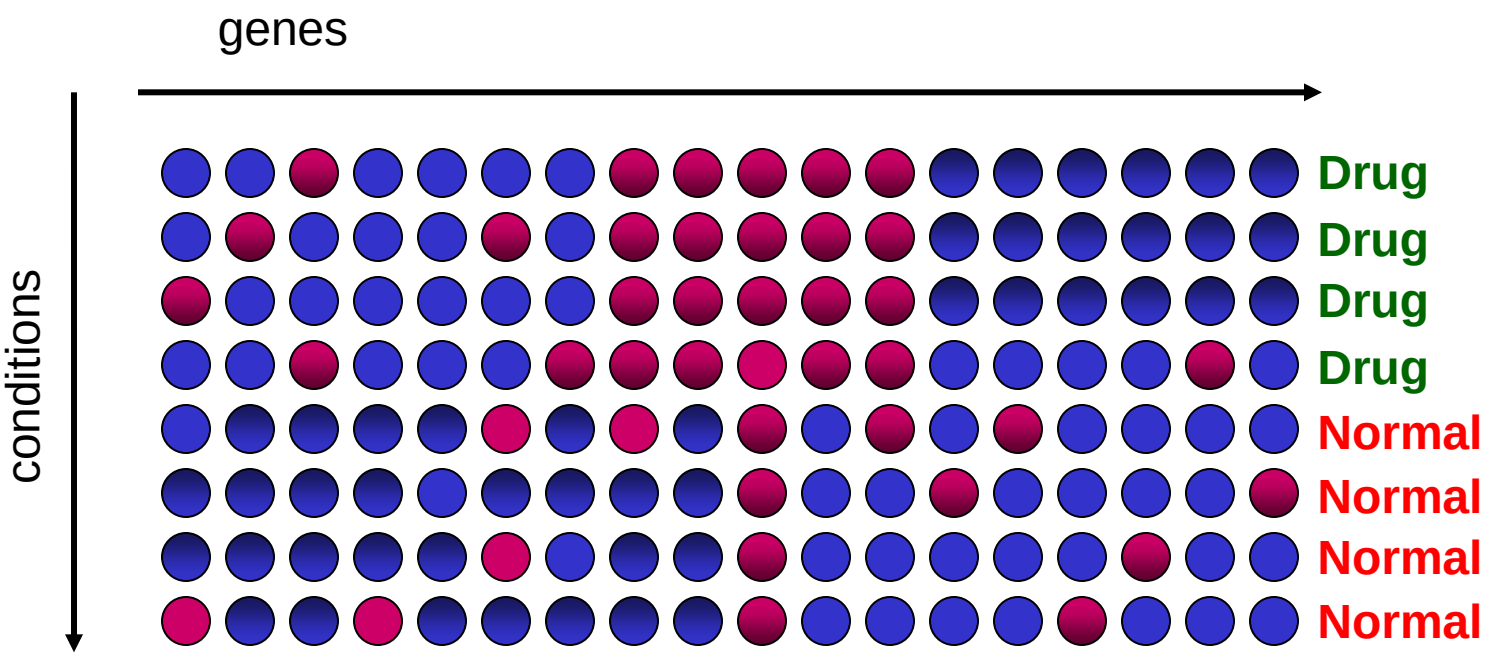
Click to watch an interesting movie explaining the working o microarray



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**

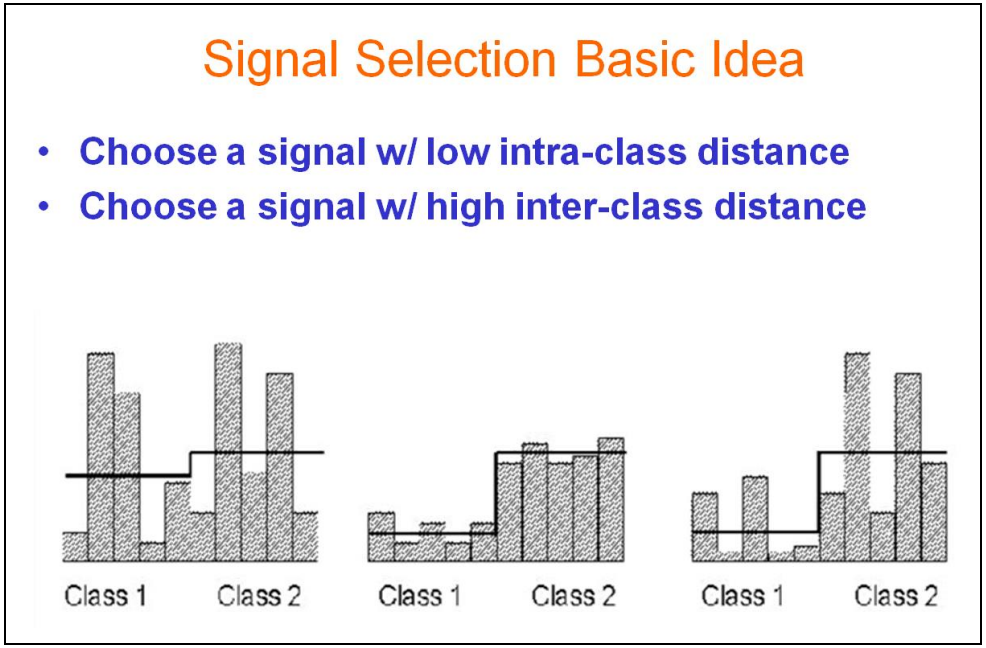
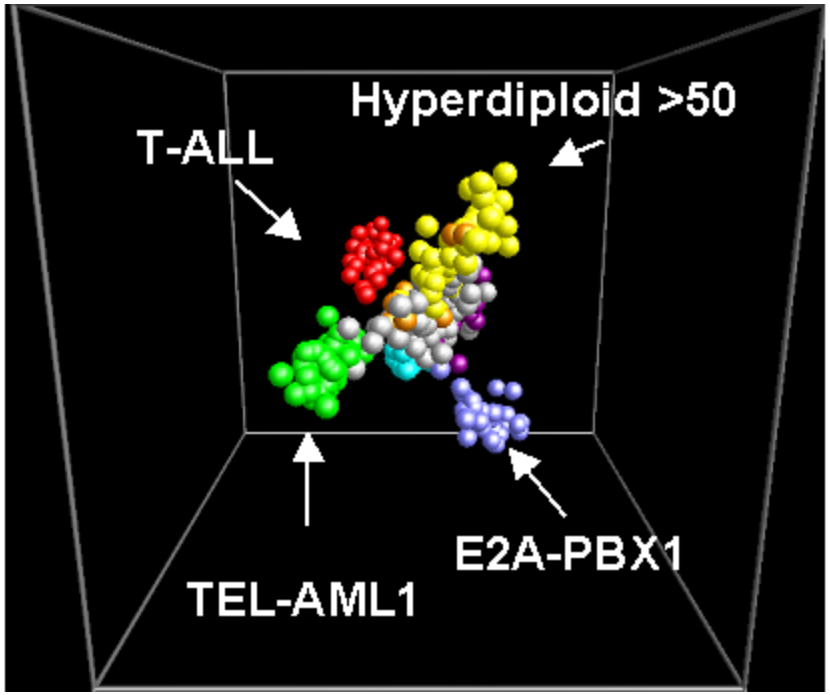


Image credit: Golub et al., *Science*, 286:531–537, 1999

Terminology: DE gene = differentially expressed gene

PCA Plots



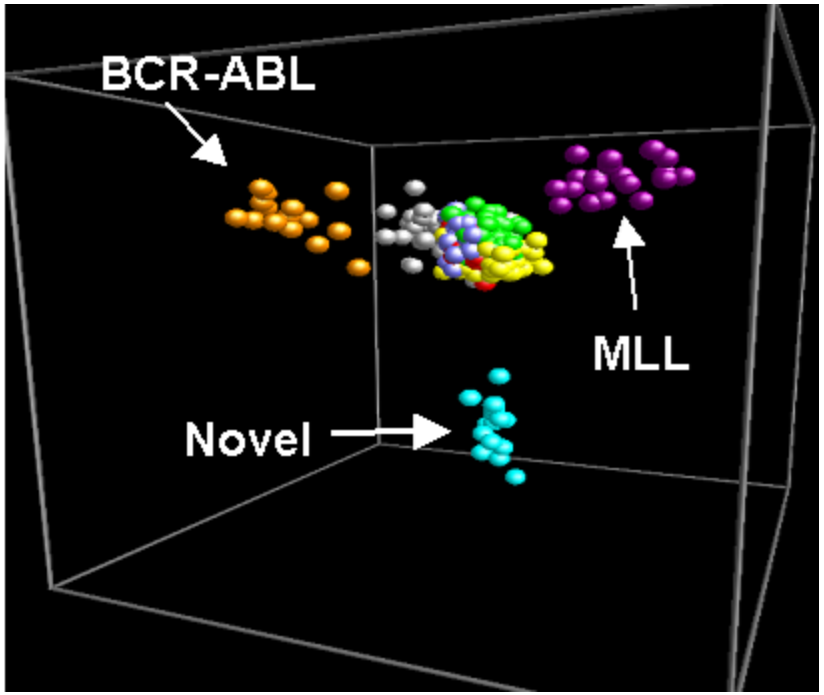


Image credit: Yeoh et al, *Cancer Cell,* 1:133-143, 2002

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

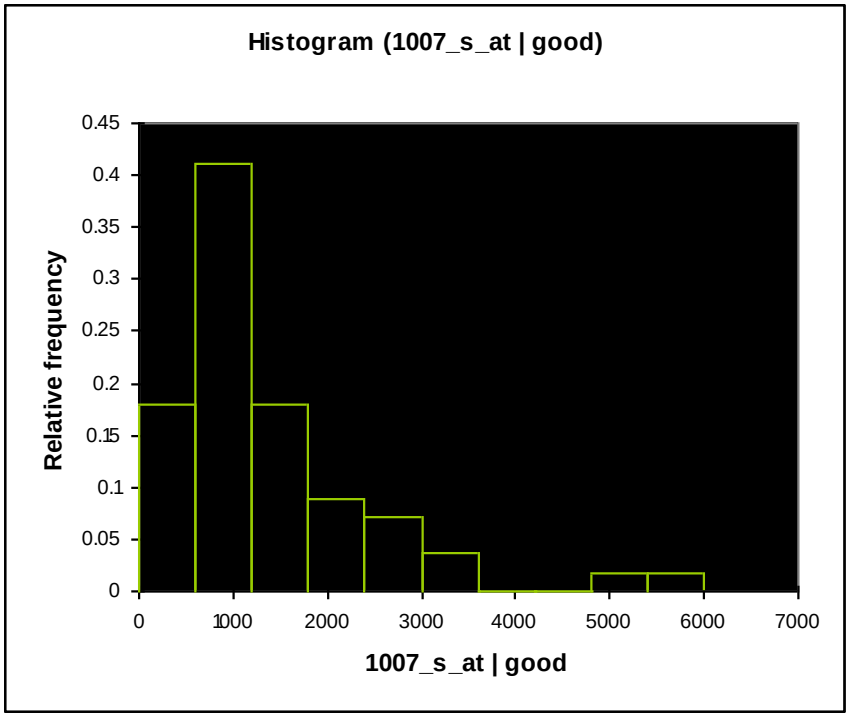
Some Headaches

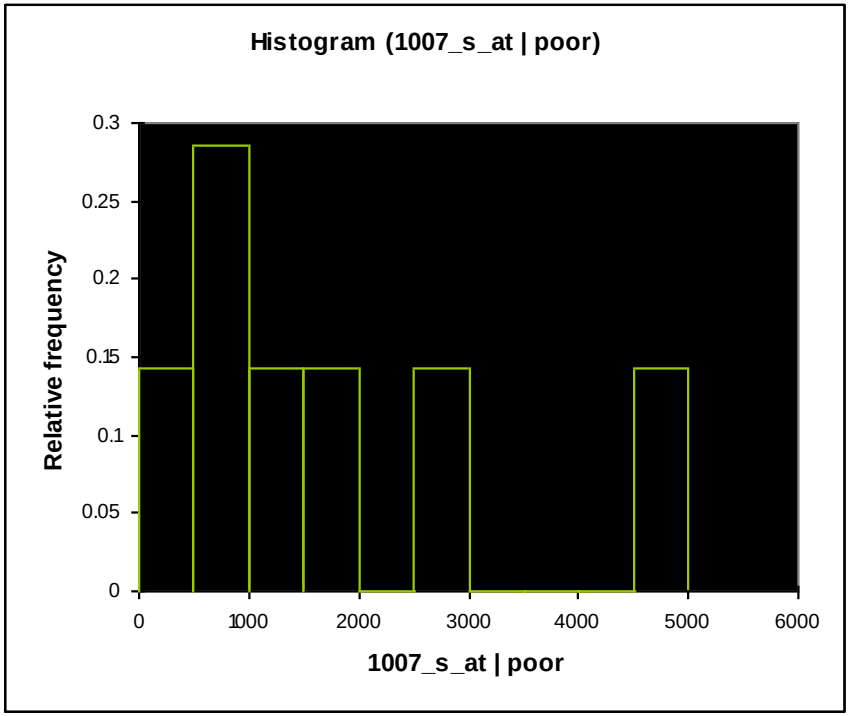
- **Natural fluctuations of gene expression in a person**

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

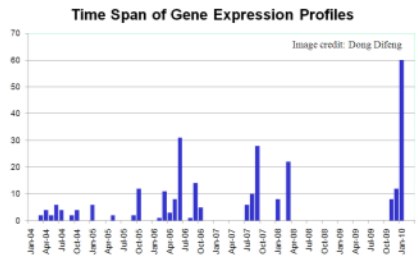
⇒ **Selected genes may not be meaningful**
  - Diff genes get selected in diff expts

Natural Fluctuations

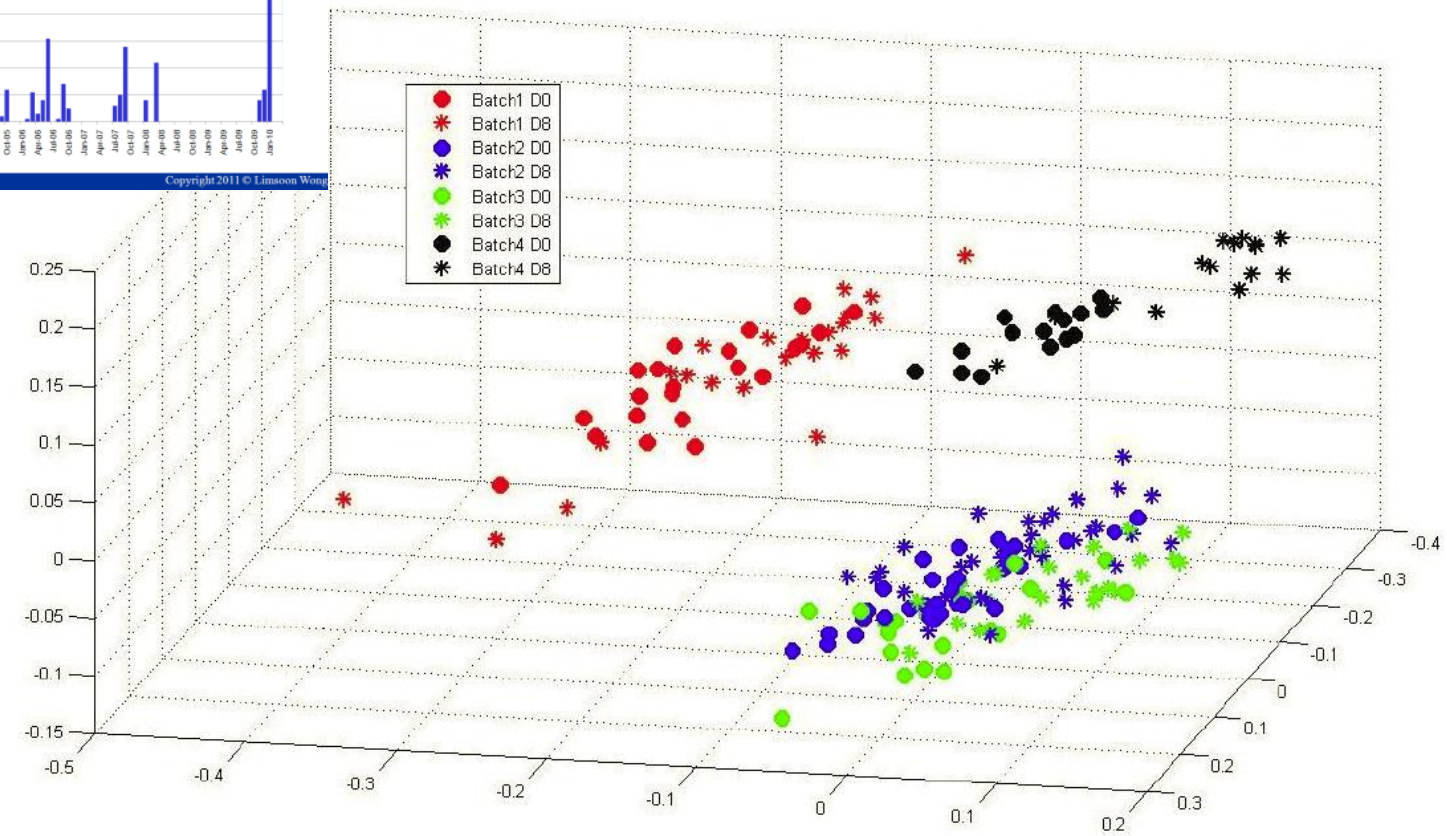




Sometimes, a gene expression study may involve batches of data collected over a long period of time…



Batch Effects



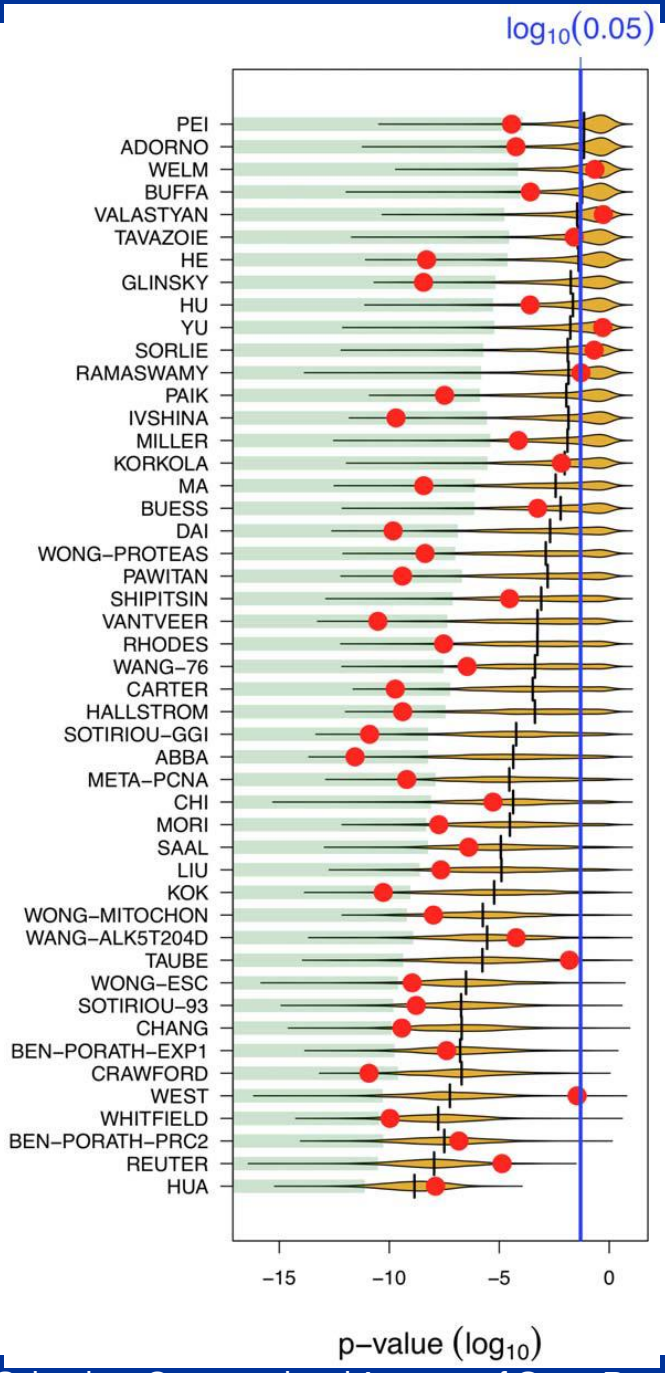
- **Samples from diff batches are grouped together, regardless of subtypes and treatment response**

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

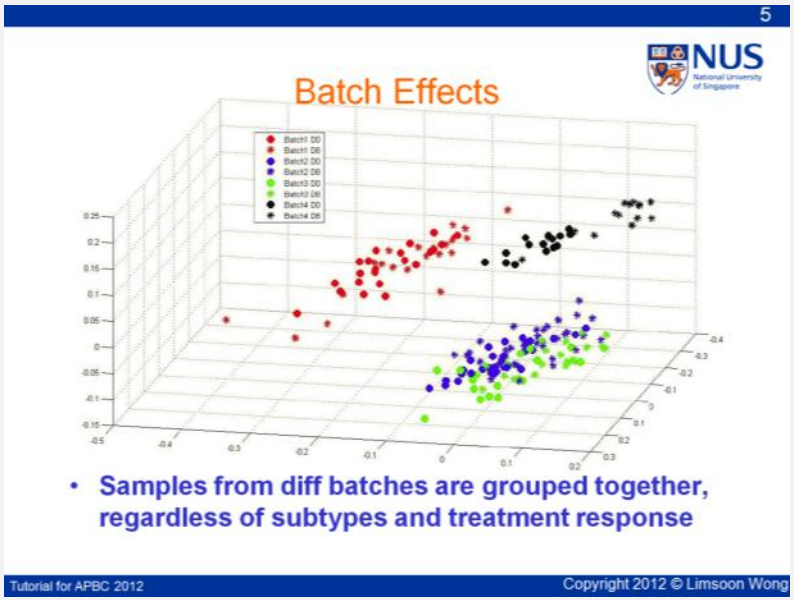
Zhang et al, Bioinformatics, 2009



"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



- 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

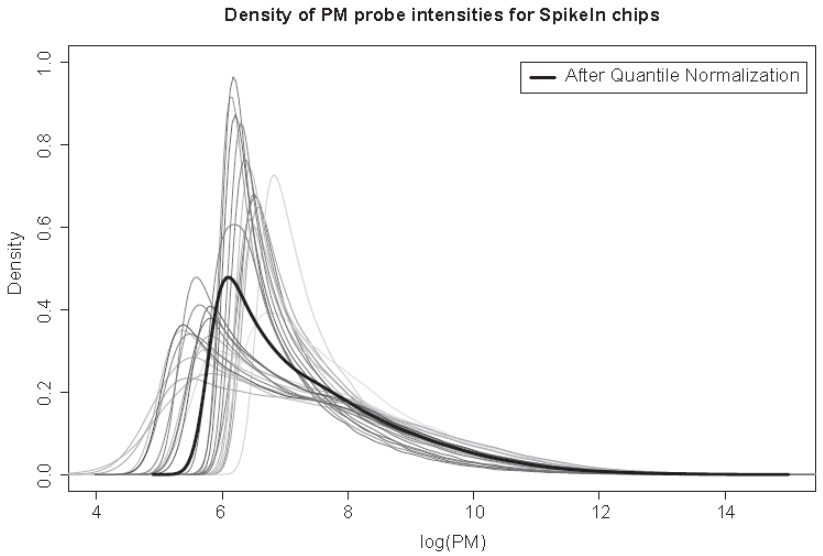
- **Scaling method**
  - Intensities are scaled so that each array has same ave value
  - E.g., Affymetrix's

- **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!

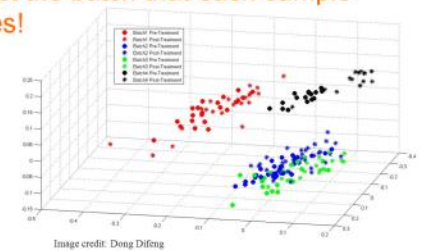
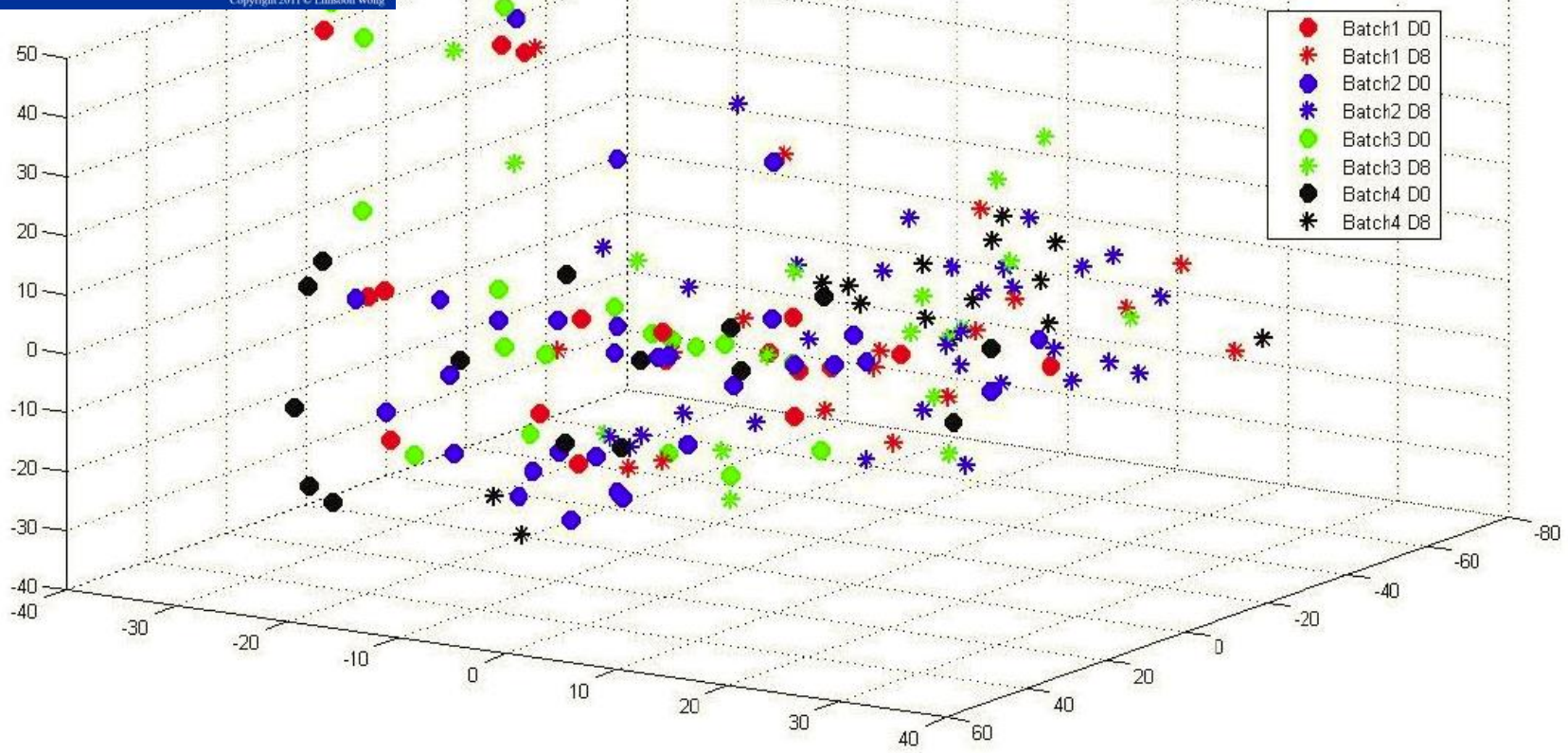


Image credit: Dong Difeng

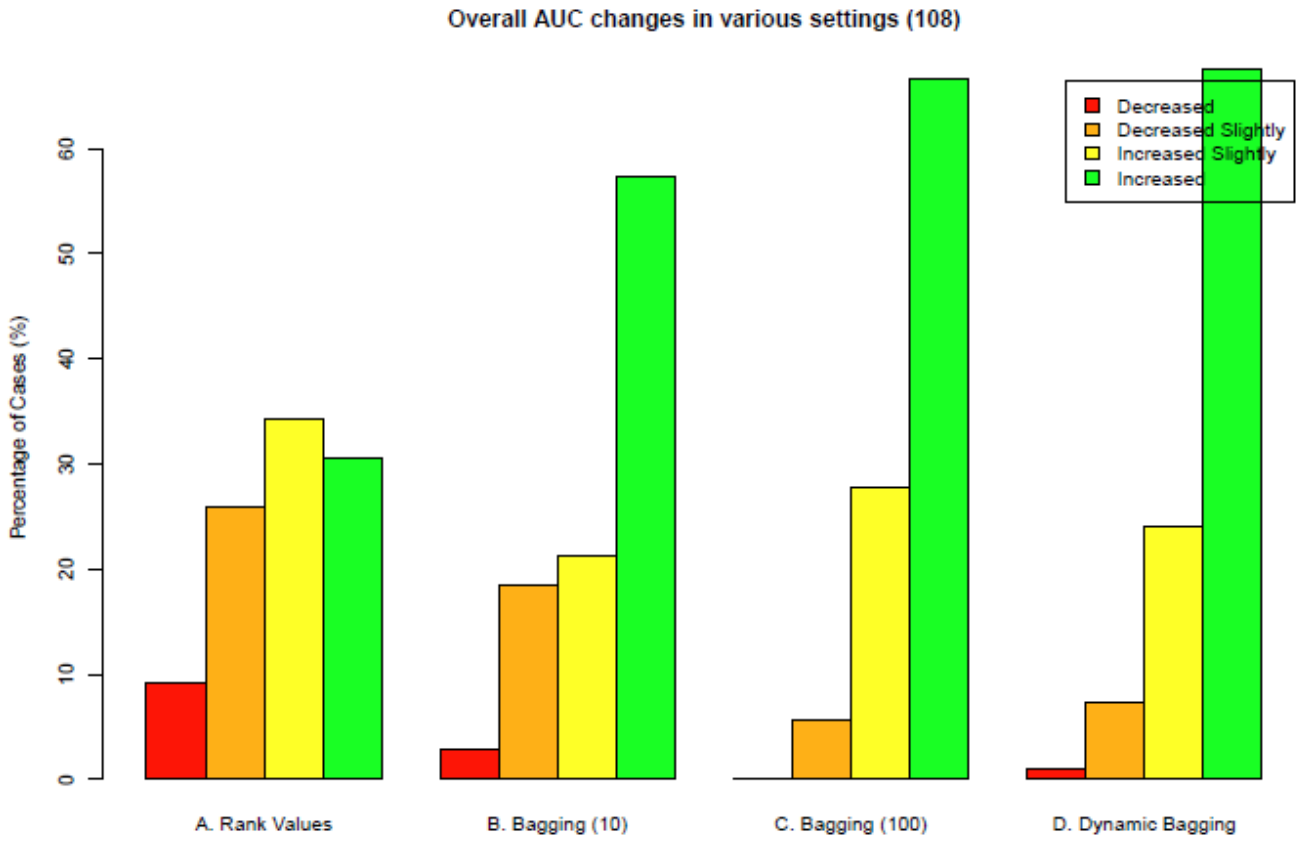
⇒ Need normalization to correct for batch effect

After quantile normalization



GEP after removing batch effect by quantile normalization

Quantile normalization improves cross-batch prediction accuracy in gene expression profile analyses



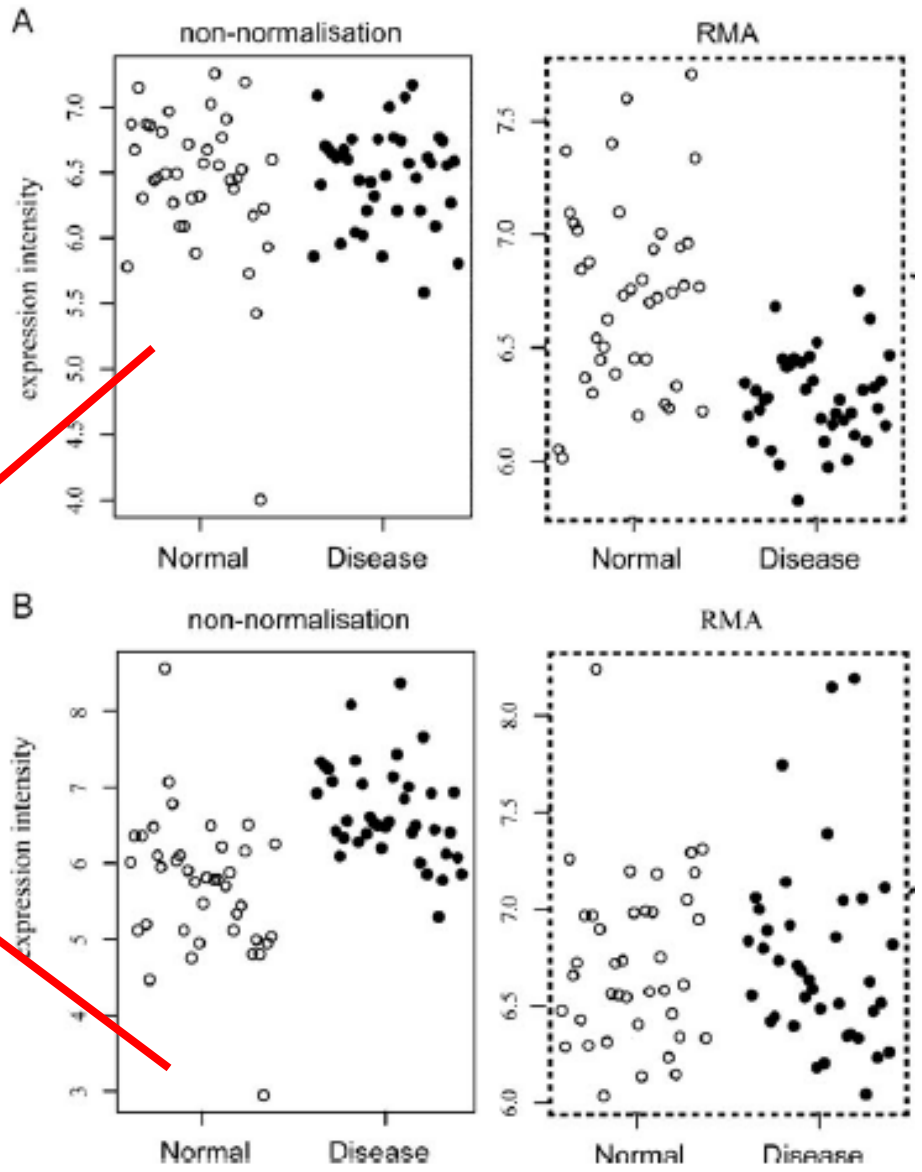
Chuan Hock Koh, Limsoon Wong. **Embracing noise to improve cross-batch prediction accuracy**. *BMC Systems Biology*, 6(Suppl 2):S3, December 2012.

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

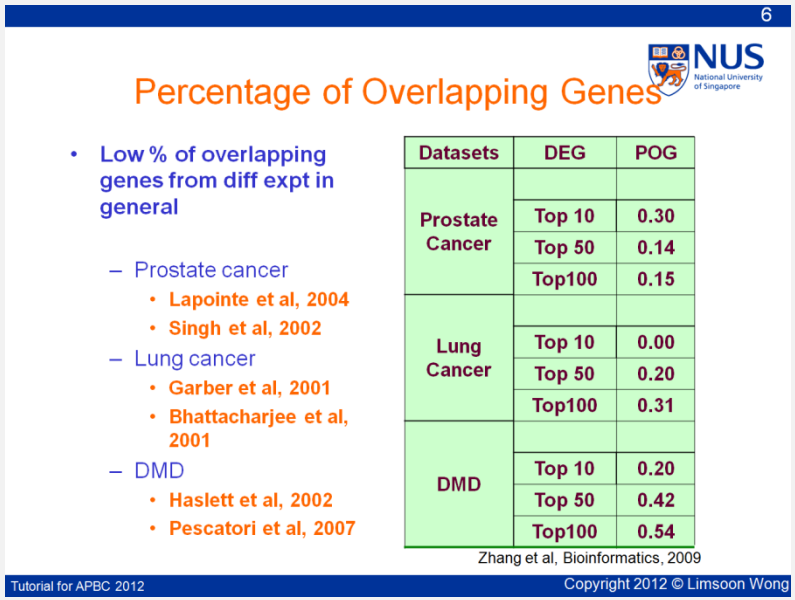
Genes are extensively upregulated in cancers. Normalizing them mislead them to be considered downregulated!



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

Individual Genes

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

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

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

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

$\Rightarrow$ **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?**

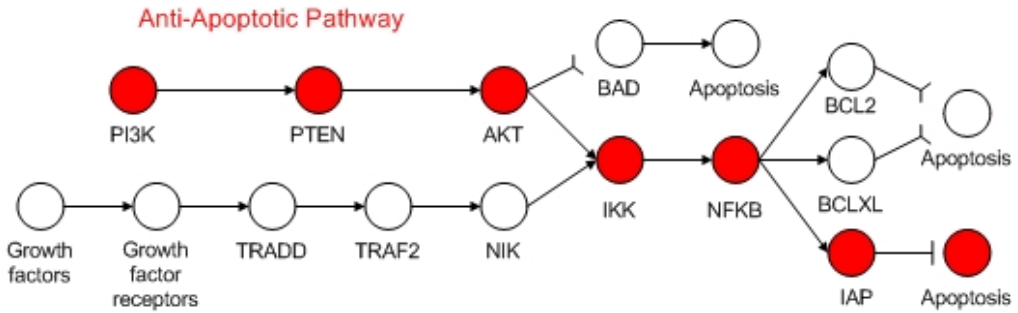
- **Prob(group of genes correlated) = $(1/2^6)^5$**
  - Good, << $1/2^6$
- **# of groups = $^{100000}C_5$**

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

$\Rightarrow$ **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

NUS National University of Singapore

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**

Towards More Meaningful Genes

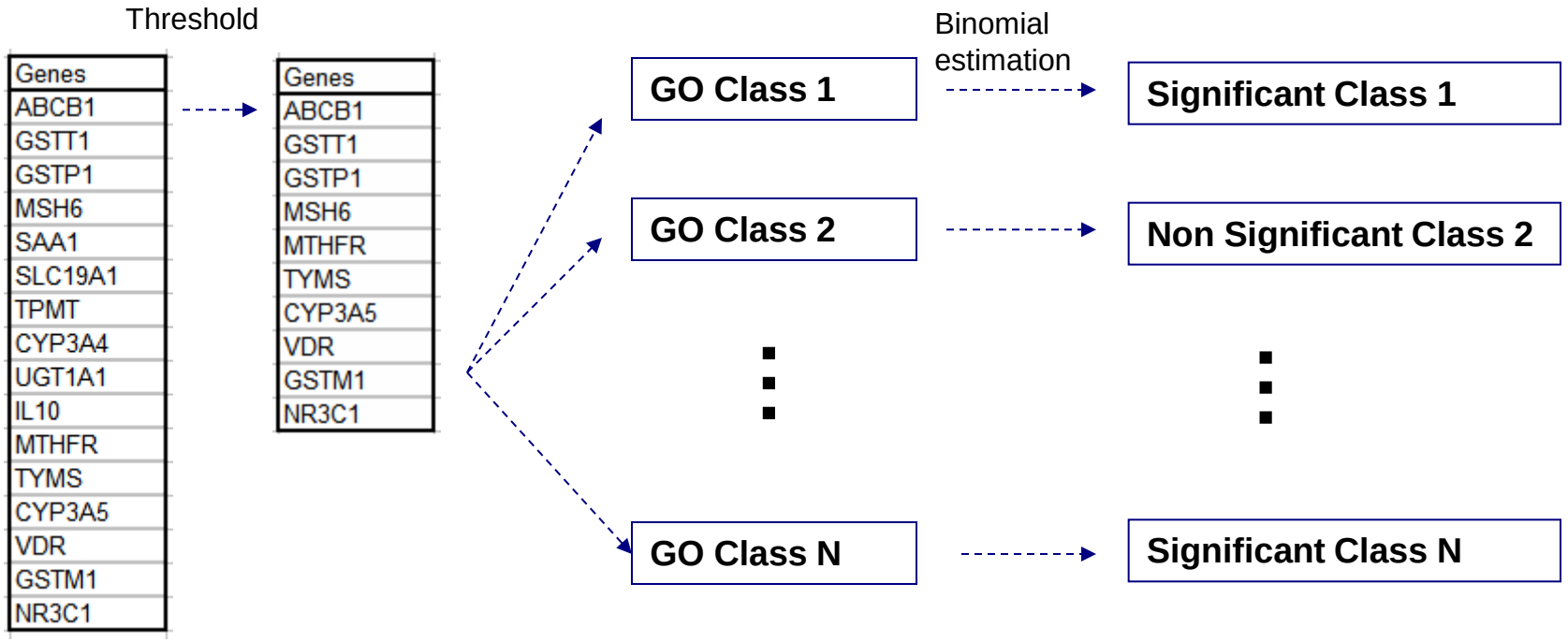
- **ORA**
  - Khatri et al
  - *Genomics*, 2002
- **FCS**
  - Pavlidis & Noble
  - PSB 2002
- **GSEA**
  - Subramanian et al
  - *PNAS*, 2005
- **SNet**
  - Soh et al
  - *BMC Genomics*, 2011

Overlap Analysis

Direct-Group Analysis

Network-Based Analysis

Overlap Analysis: ORA

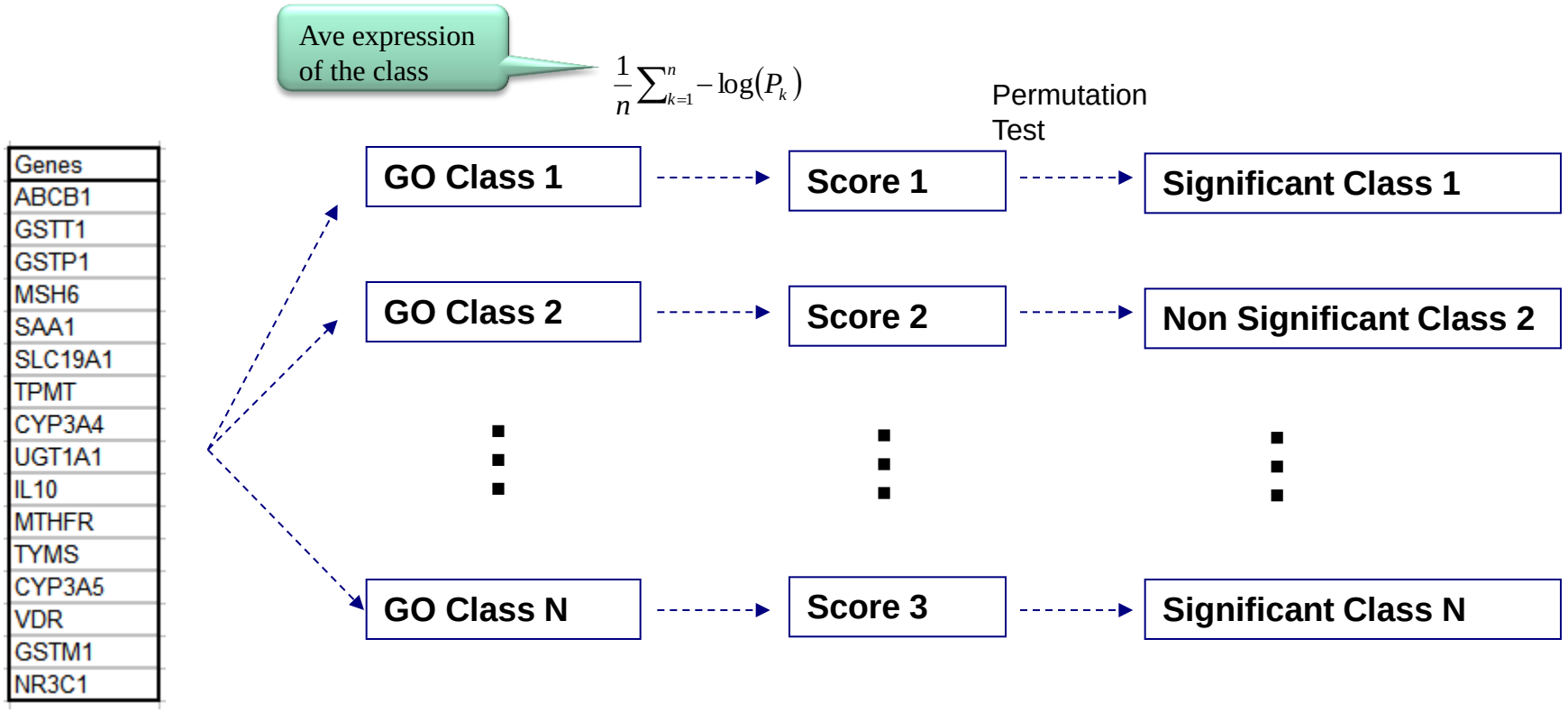


S Draghici et al. "Global functional profiling of gene expression". *Genomics*, 81(2):98-104, 2003.

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



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.

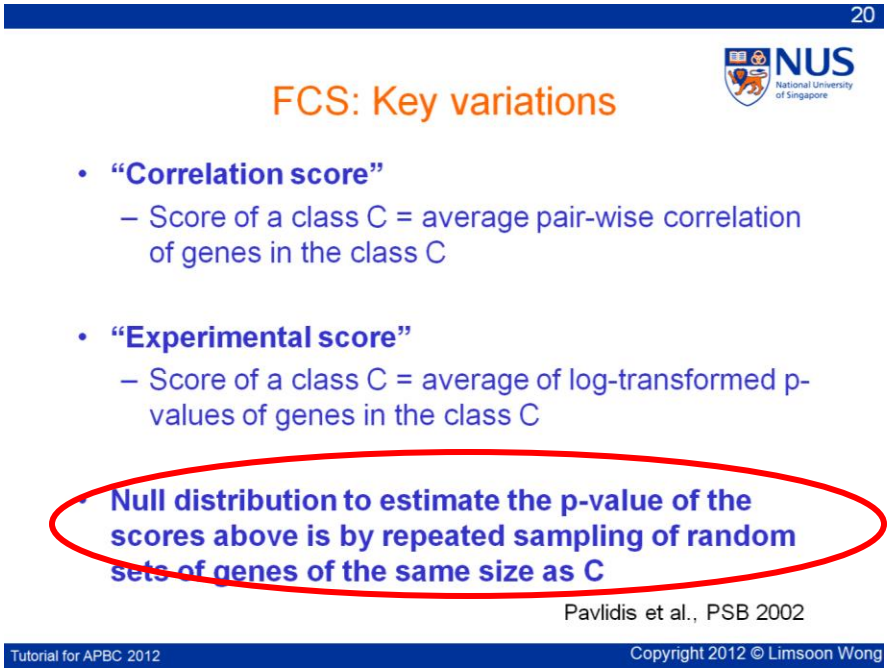
FCS: Key variations

- **"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

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**

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

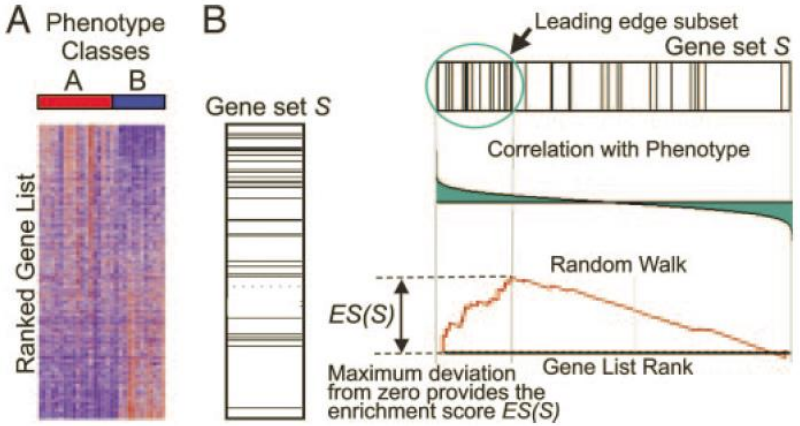


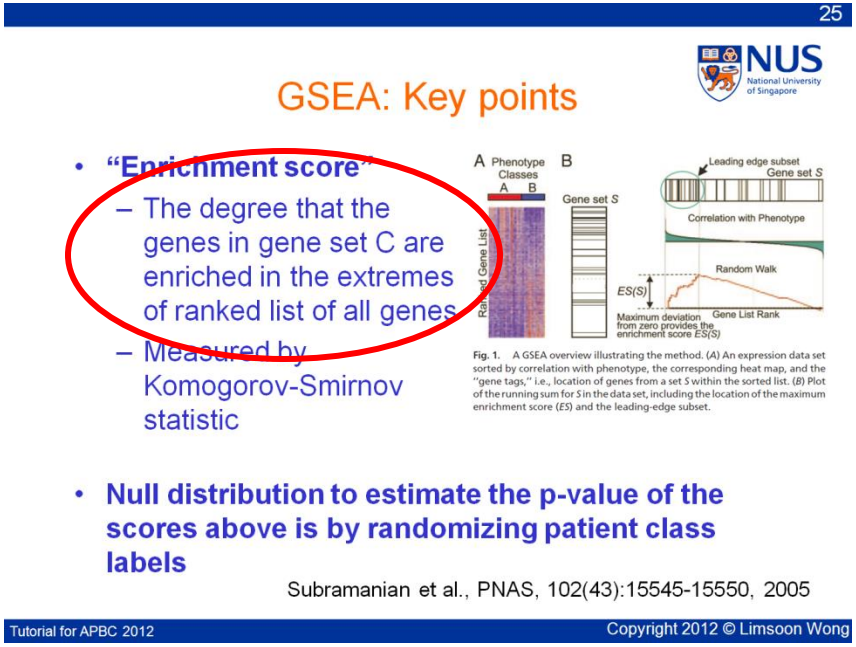
Fig. 1. A GSEA overview illustrating the method. (*A*) An expression data set sorted by correlation with phenotype, the corresponding heat map, and the "gene tags," i.e., location of genes from a set *S* within the sorted list. (*B*) Plot of the running sum for *S* in the data set, including the location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



- **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.

NUS National University of Singapore

SNet: Score Subnetworks

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

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

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:

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

$G_{sn,j,d}^{score}$ 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:

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

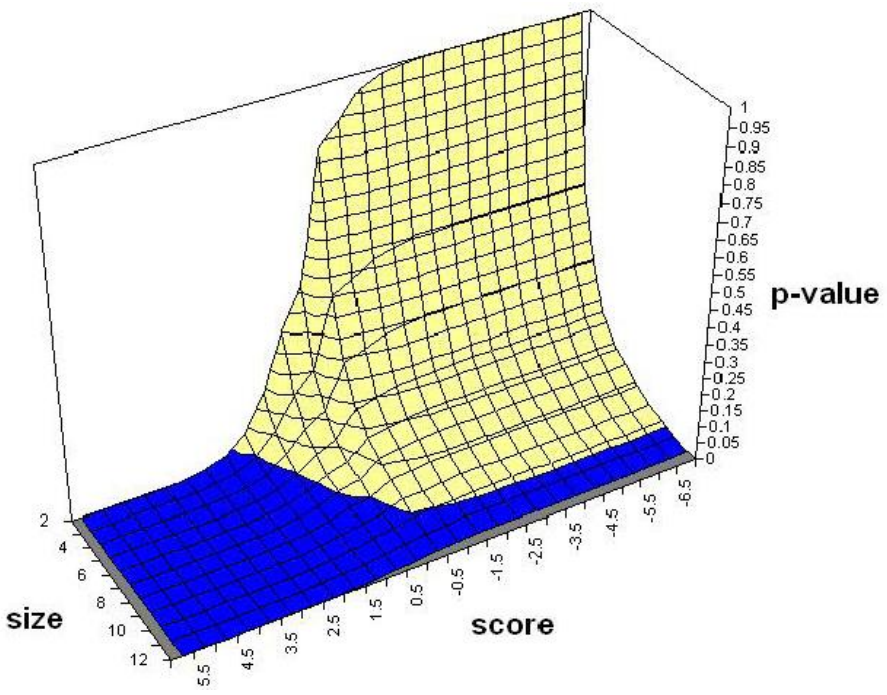
Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d. The entire Step 2 is repeated for the other disease phenotype $\neg d$, giving us the score vectors, $SN_{sn,d}^{u\_score}$ and $SN_{sn,\neg d}^{u\_score}$ 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.

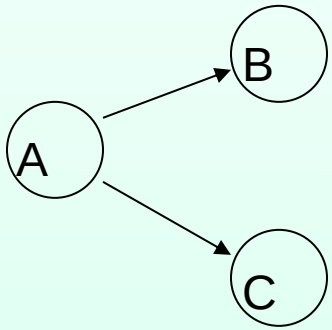
NUS National University of Singapore

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



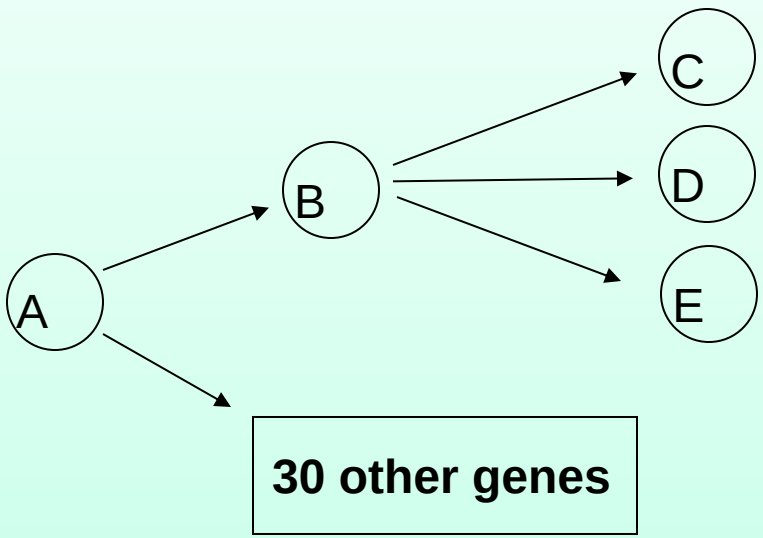
Genes A, B, C are high in phenotype *D*

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*

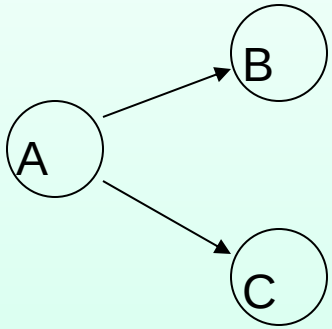
30 other genes not diff expressed

Conventional techniques: Entire network is likely to be missed

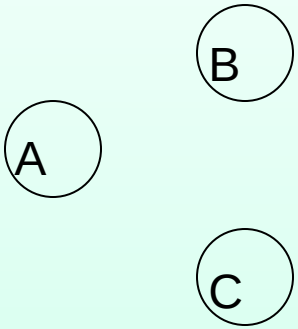
- **SNet: Able to capture the subnetwork branch within the pathway**

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.

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.

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%

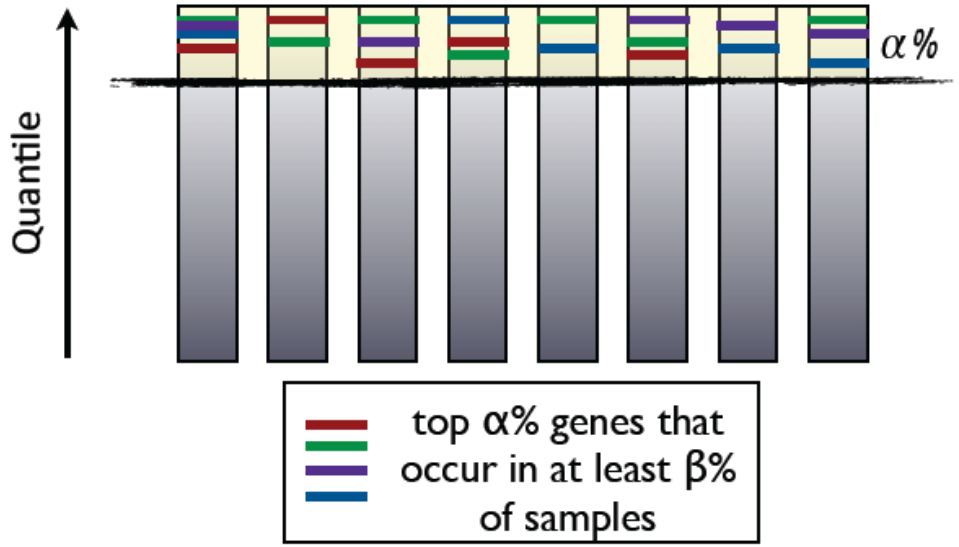
- **For each disease, take significant subnetworks extracted independently from both datasets and see how much their genes overlap**

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 represented in color bands are in at least $\beta\%$ of the samples are then taken to induce subnetworks.

- What if the real important genes are close to, but not in, the top $\alpha\%$ most highly expressed genes?
- Blindly increasing α does not help, as this will bring in lots of false-positive genes

Issue #2 with SNet

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

$G^{score}_{sn,j,d}$ 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:

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

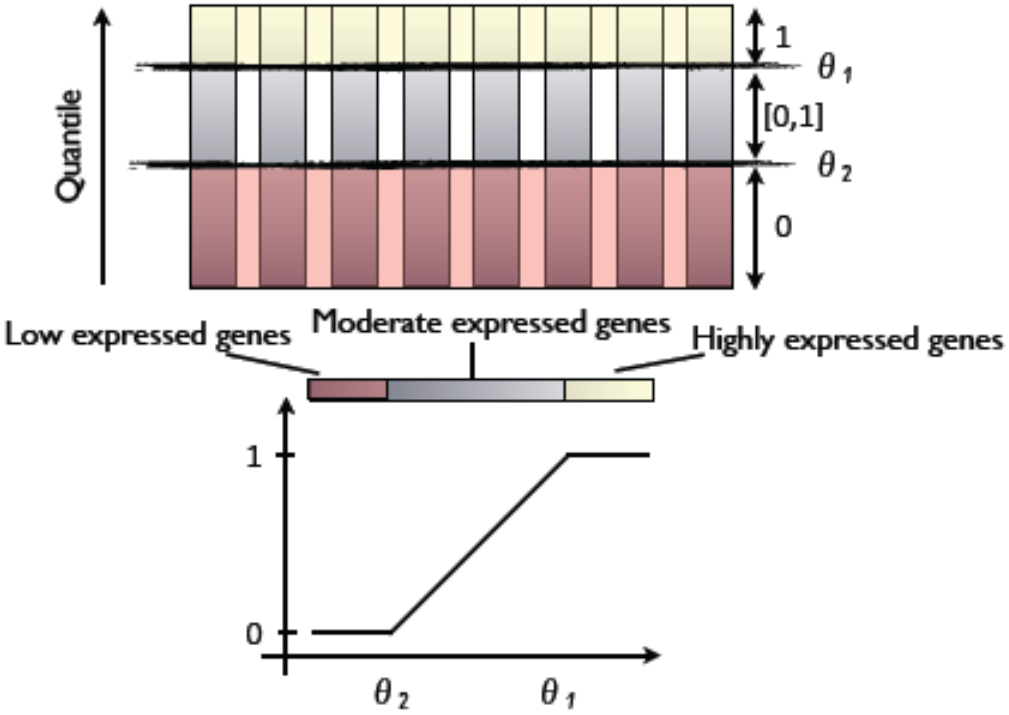
Where k is the number of patients of phenotype d who has gene x highly expressed (top $\alpha\%$) and n is the total number of patients of phenotype d.

- **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**

$\Rightarrow$ **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:

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

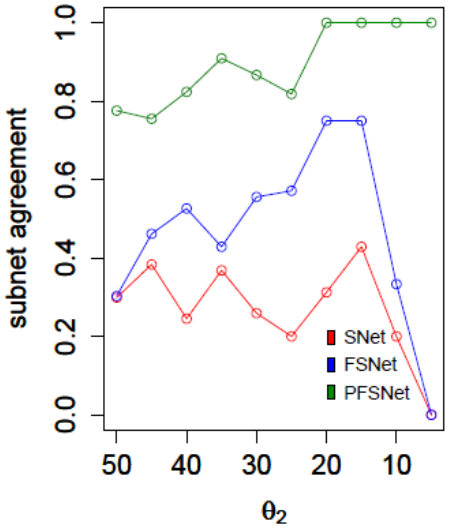
$$Score_2^{p_k}(S) = \sum_{g_i \in S} fs(e_{g_i,p_k}) * \beta_2^*(g_i) \qquad (6)$$

Paired T-Test

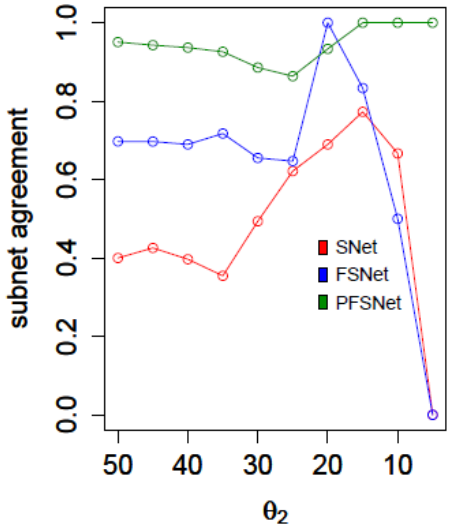
- **$Score^{Pk}_1(S)$ and $Score^{Pk}_2(S)$ are computed for the same sample Pk and subnetwork S**

$\Rightarrow$ **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

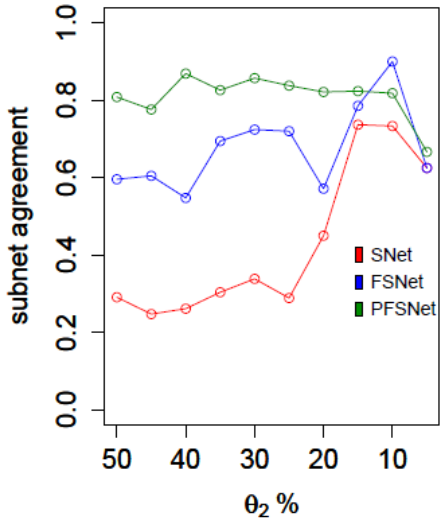


upregulated in ALL

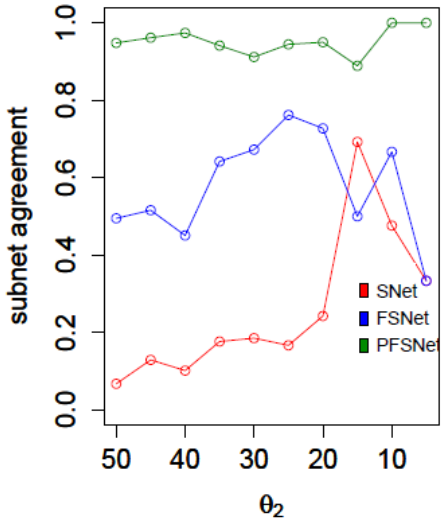


upregulated in AML

Fig. 4: Consistency of subnetworks in Leukemia dataset



upregulated in DMD

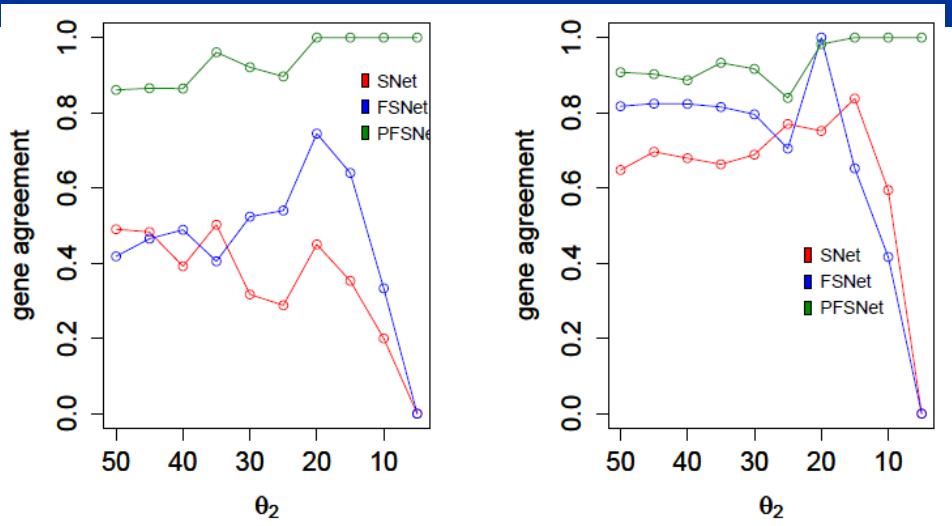


upregulated in NORM

Fig. 6: Consistency of subnetworks in DMD dataset

PSFNet vs SNet: Subnet Agreement

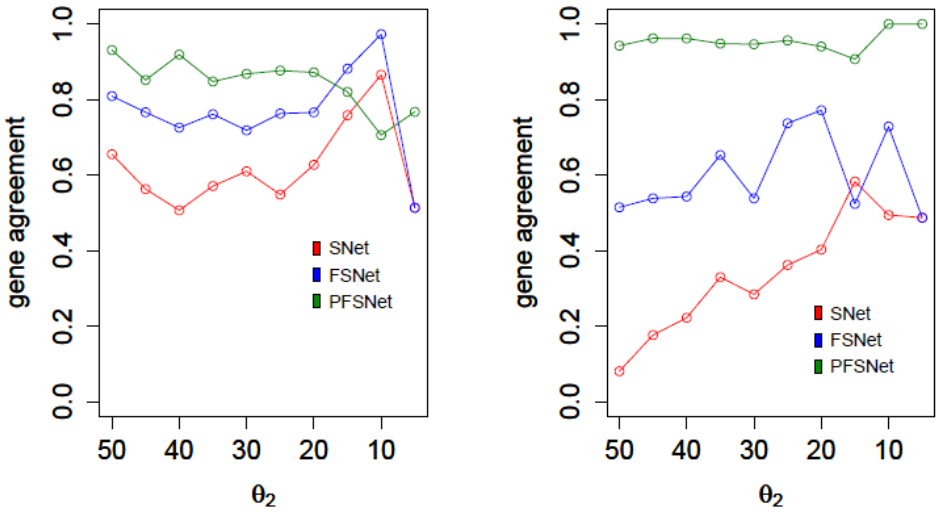
Overlap = |A∩B| / |A∪B|



upregulated in ALL upregulated in AML

Fig. 7: Consistency of genes in Leukemia dataset

PSFNet vs SNet: Gene Agreement



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

PFSNet vs T-Test: Gene Agreement

Table 2. Comparing gene-level agreement of PFSNet, FSNet, SNet, GSEA, SAM, t-test.

Dataset	PFSNet		FSNet		SNet		GSEA		SAM(5% sig)		SAM(top 100)		t-test(5% sig)		t-test(top 100)	
	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D	D	¬D
Leukemia	1.00	0.81	0.64	0.42	0.35	0.58	0.12	0.20	0.50	0.47	0.01	0.01	0.35	0.29	0.19	0.07
ALL (subtype)	0.54	0.70	0.38	0.41	0.29	0.57	0.04	0.04	0.19	0.27	0.12	0.21	0.08	0.10	0.01	0.00
DMD	0.82	0.72	0.88	0.75	0.76	0.54	0.44	0.20	0.34	0.08	0.27	0.19	0.41	0.19	0.11	0.25

For PFSNet and FSNet, threshold values of $\theta_1 = 5\%, \theta_2 = 15\%$ are used. D represents subnetworks enriched in phenotype D and $\neg D$ represents subnetworks enriched in phenotype $\neg D$. For GSEA, the "leading edge genes" were used. For SAM and t-test, we took genes at 5% significance level and also the top n genes indicated in brackets.

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

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

Table 4. Top 5 subnetworks that have biological significance.

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

* indicates subnetworks that were not found in SNet and # indicates pathways that were missed by GSEA

DMD: Striated Muscle Contraction

For DMD, the subnetwork responsible for striated muscle contraction is shown in figure 5a (supplementary material). The cause of Duchenne muscular dystrophy is well known to stem from the gene Dystrophin, which codes for a protein attached to the cell membrane (sacrolemma) of striated muscle cells (Goldstein and McNally, 2010). When its expression is perturbed, the cell membrane becomes fragile and permits an amplification in calcium signals into the muscle cell causing a cascade of signals to induce cell death. Our subnetwork is generated around the Dystrophin gene and implicates other genes belonging to the Myosin (MYBPC1,MYBPC2) and Troponin (TNNI1,TNNI2) family. The Myosin and Troponin genes are responsible for controlling muscle contractions. The down-regulation of Troponin in DMD patients might help explain muscle contracture, a condition in which the muscle shortens. This is because with lower abundance of Troponin, Myosin is able to bind to Actin. This mechanism together with the amplification of calcium causes the muscle to constantly contract, shortening over time (Goldstein and McNally, 2010; Krans, 2010).

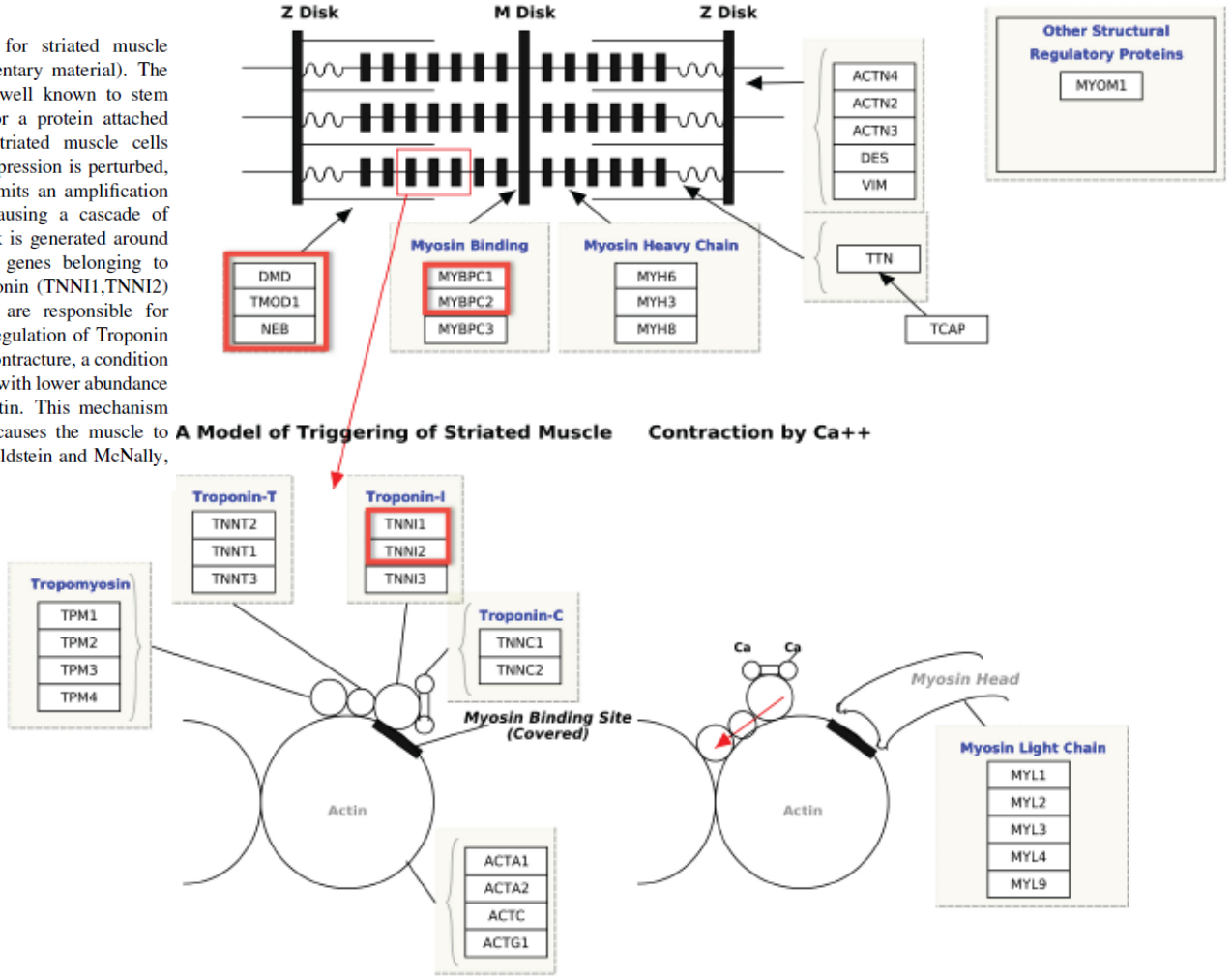
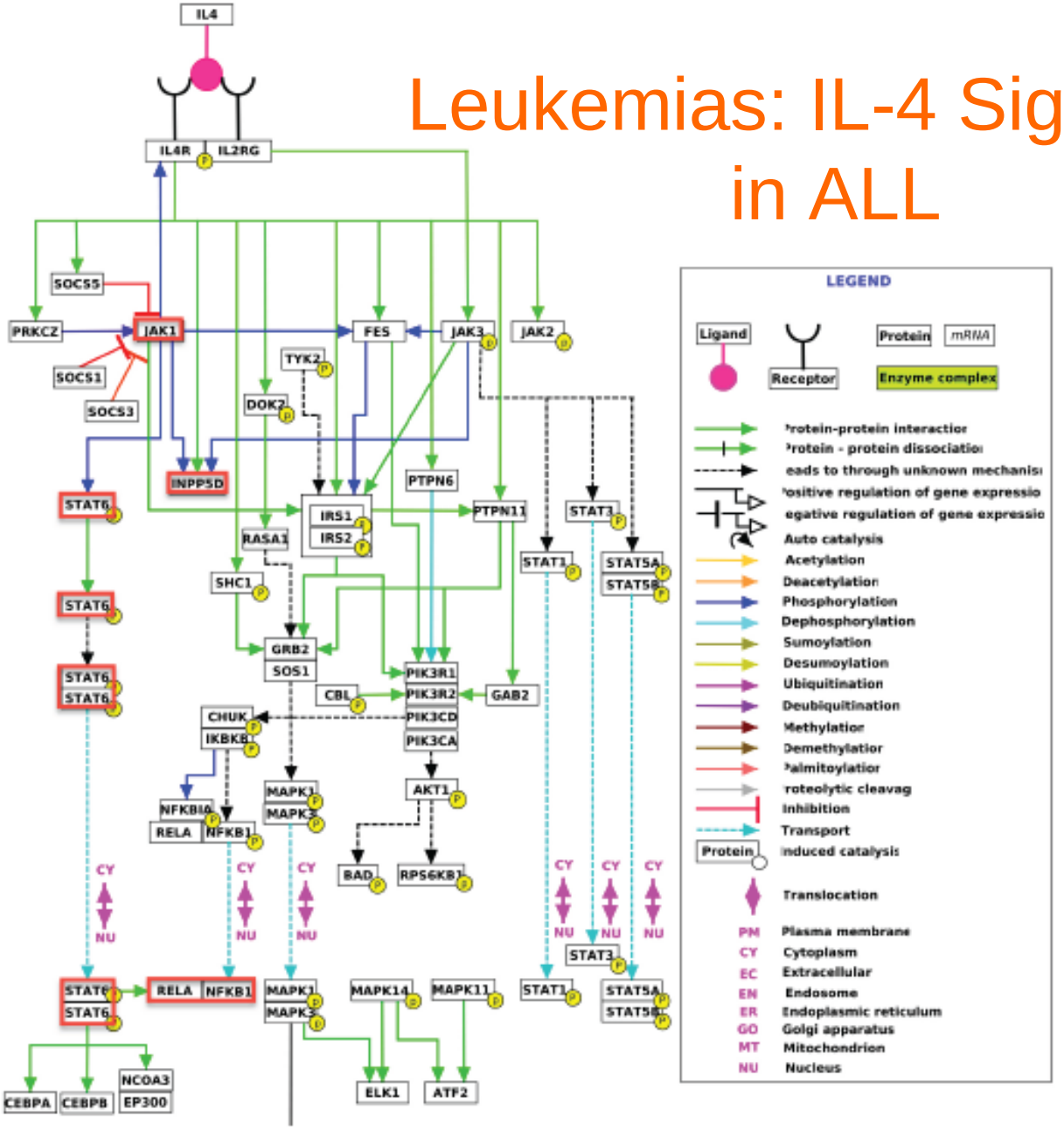


Fig. 5. An example of a biologically relevant pathway for DMD. The nodes from the induced subnetwork identified by PFSNet is highlighted with red boxes.

Leukemias: IL-4 Signaling in ALL



For the Leukemia dataset (in which patients are either classified to have acute lymphoblastic leukemia or acute myeloid leukemia), one of the significant subnetworks that is biologically relevant is part of the Interleukin-4 signaling pathway; see figure 6b (supplementary material). The binding of Interleukin-4 to its receptor (Cardoso *et al.*, 2008) causes a cascade of protein activation involving JAK1 and STAT6 phosphorylation. STAT6 dimerizes upon activation and is transported to the nucleus and interacts with the RELA/NFKB1 transcription factors, known to promote the proliferation of T-cells (Rayet and Gelinas, 1999). In contrast, acute myeloid leukemia does not have genes in this subnetwork up-regulated and are known to be unrelated to lymphocytes.

NUS National University of Singapore

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**

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
- Koh & Wong. **Embracing noise to improve cross-batch prediction accuracy**. *BMC Systems Biology*, 6(Suppl 2):S3, 2012
- [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**

Part 2
Delivering More Powerful Proteomic Profile Analysis

Limsoon Wong

Typical Proteomic MS Experiment

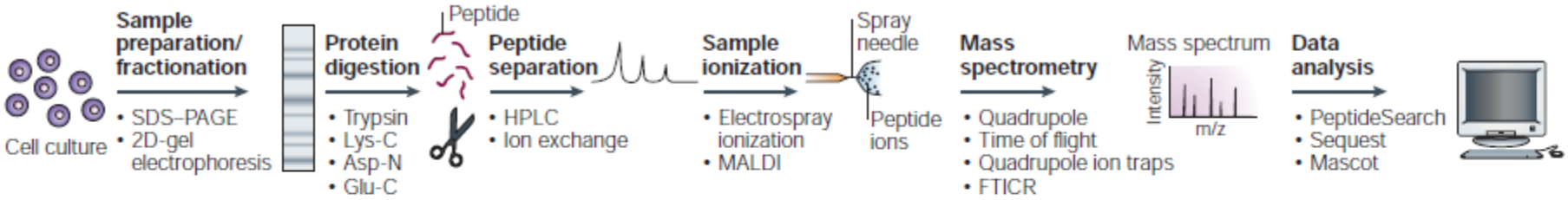
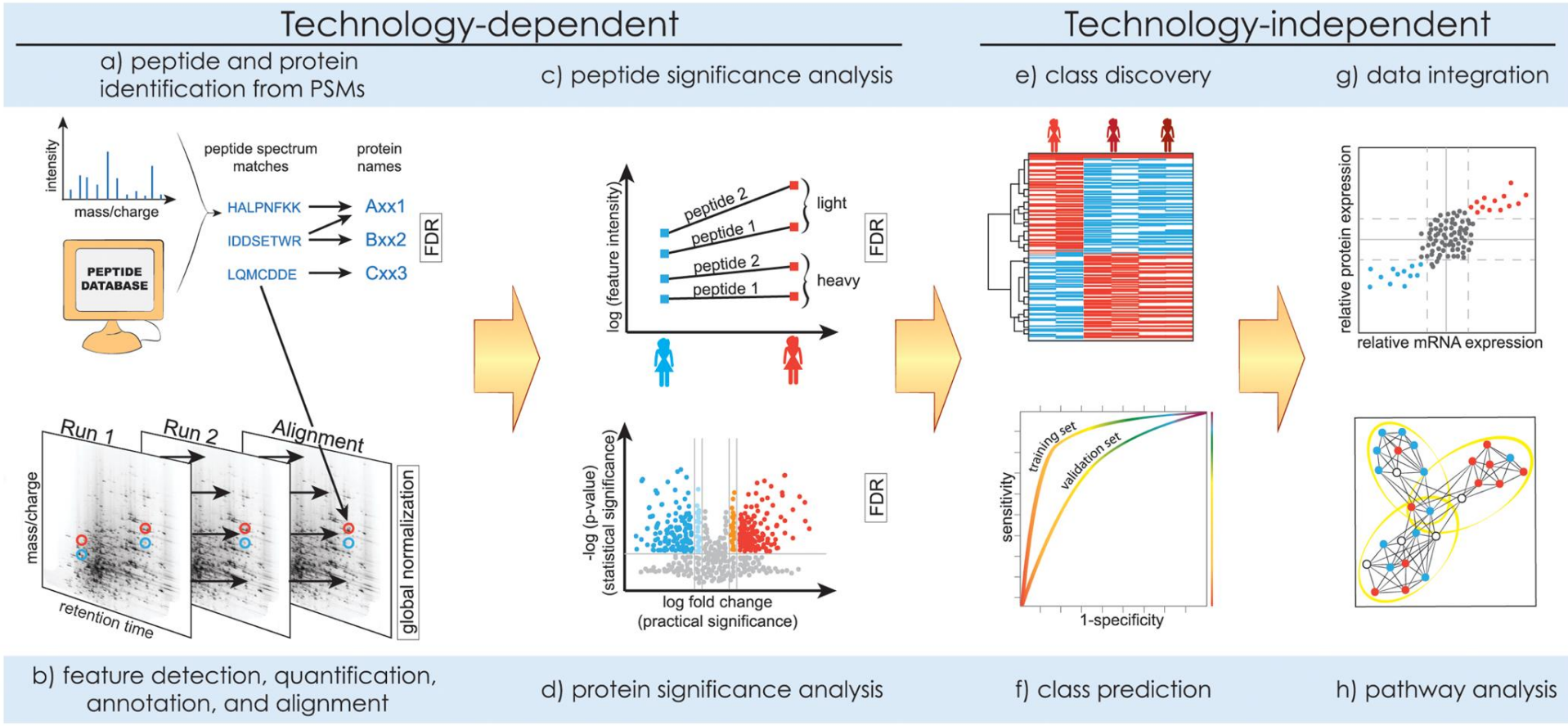


Figure 1 | **The mass-spectrometry/proteomic experiment.** A protein population is prepared from a biological source — for example, a cell culture — and the last step in protein purification is often SDS–PAGE. The gel lane that is obtained is cut into several slices, which are then in-gel digested. Numerous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



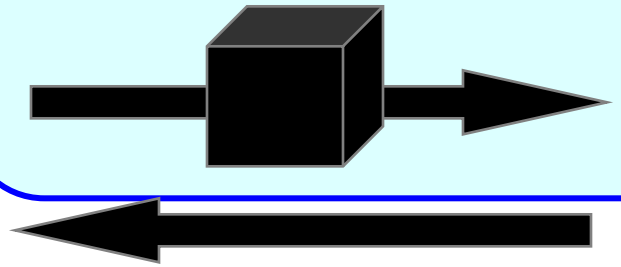
Kall and Vitek, 2011

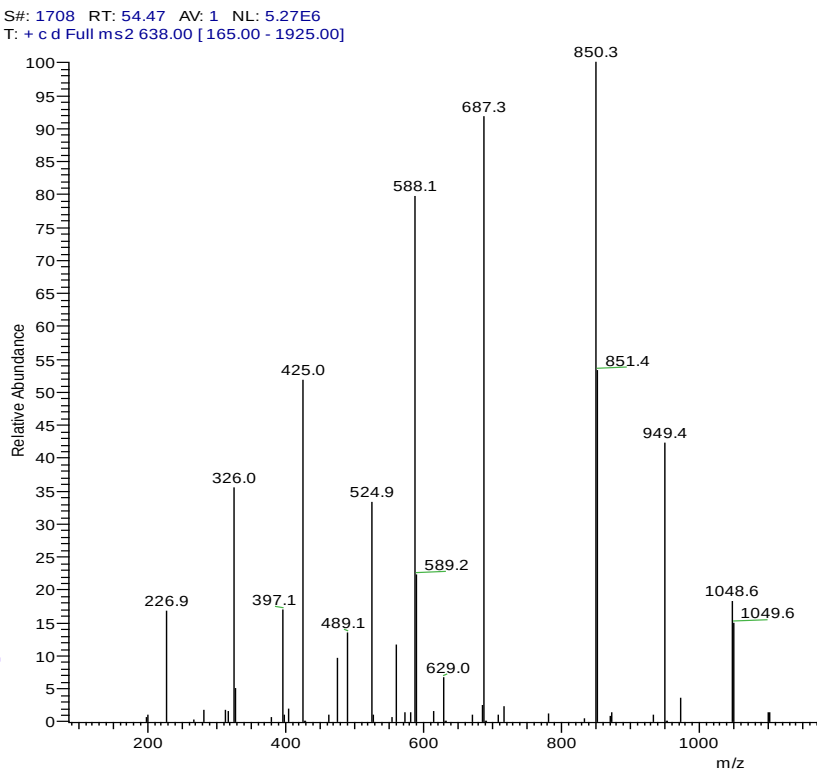
Protein Identification by Mass Spec

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

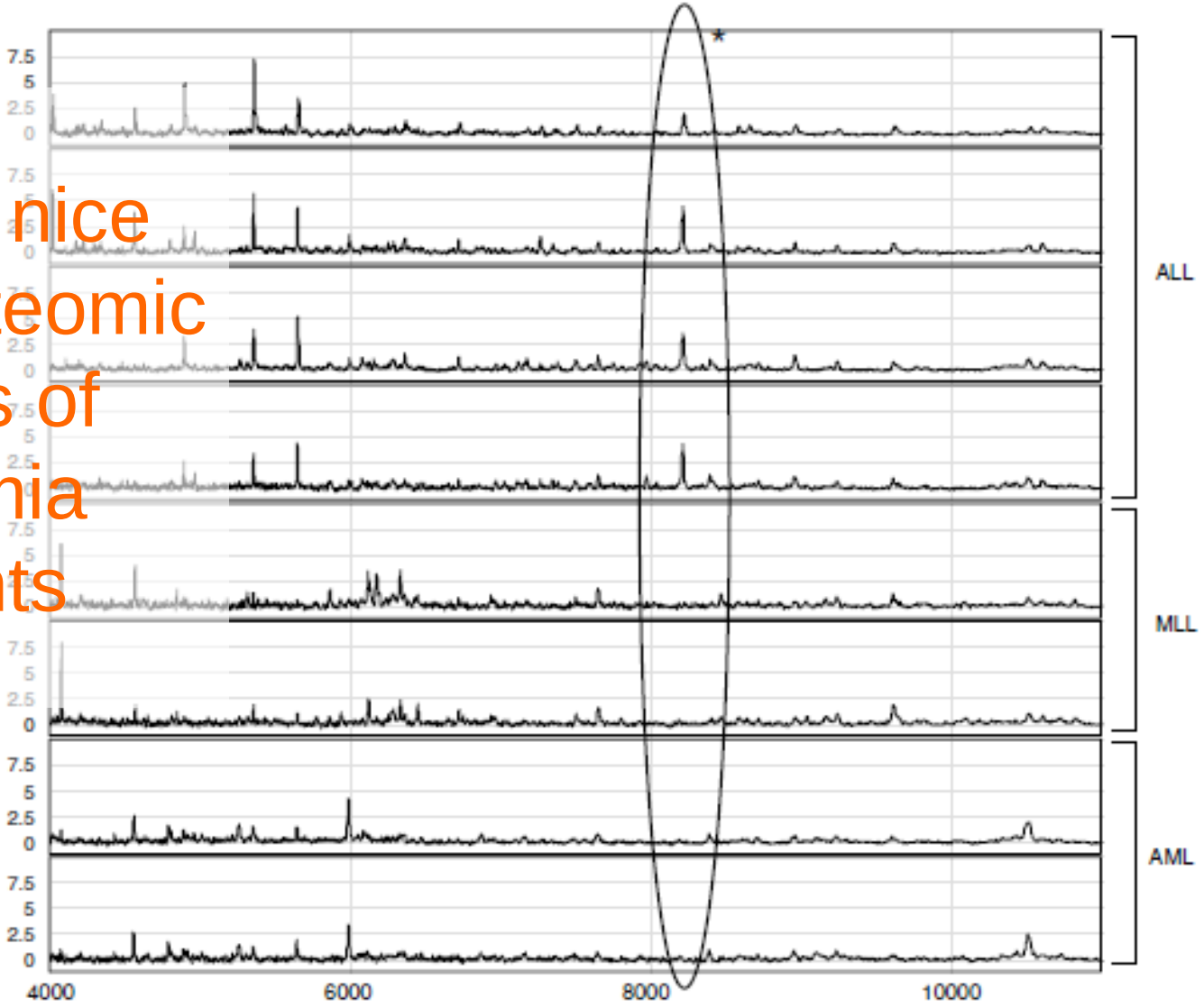


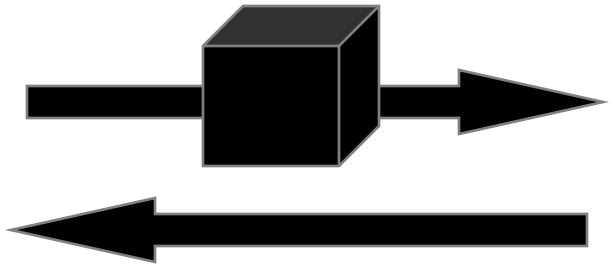
Figure 1 Spectra from SELDI-TOF MS analysis of REH, 697, MV4;11, and Kasumi cell lines. Protein (4 μg) from each cell type was analyzed on SAX2 ProteinChip® Arrays. ALL cell lines shown are REH and 697, the MLL cell line is MV4;11, and the AML cell line is Kasumi. The asterisk indicates the differentially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

Sequence

MS/MS instrument



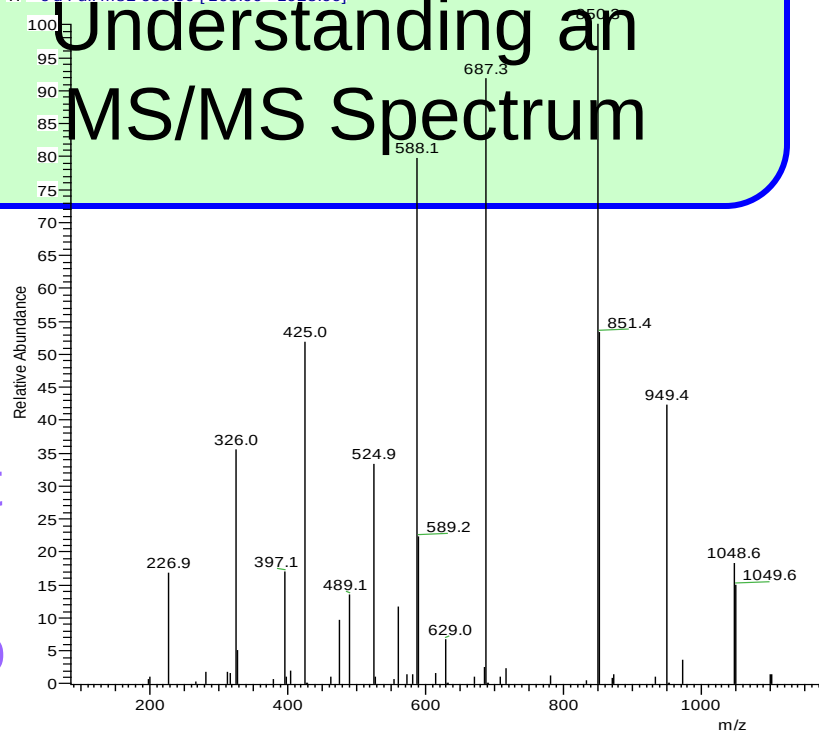
Database search

- Sequest, Mascot, InSpect

de Novo interpretation

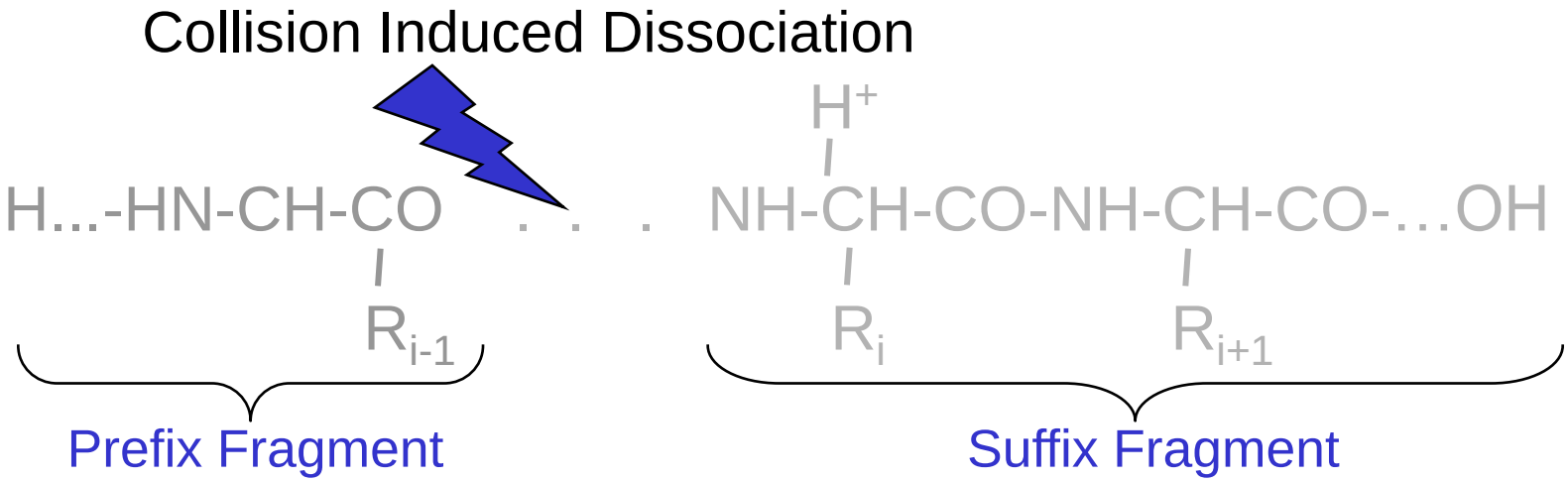
- Lutefisk, Peaks, PepNovo

Step 2: Understanding an MS/MS Spectrum



Source: Leong Hon Wai

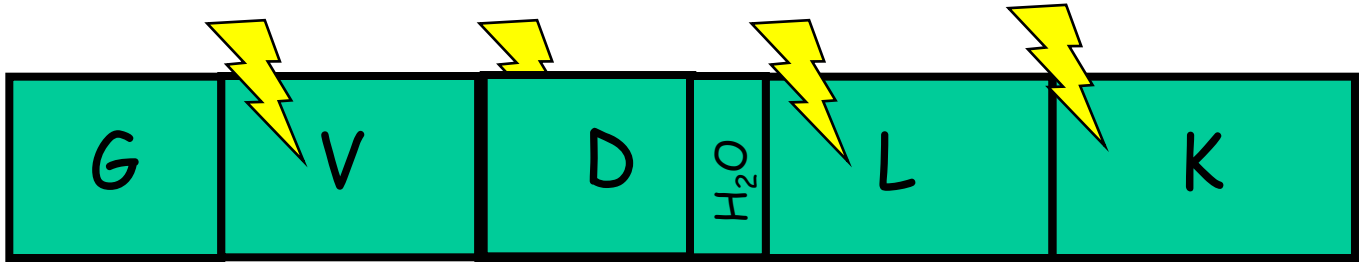
Peptide Fragmentation

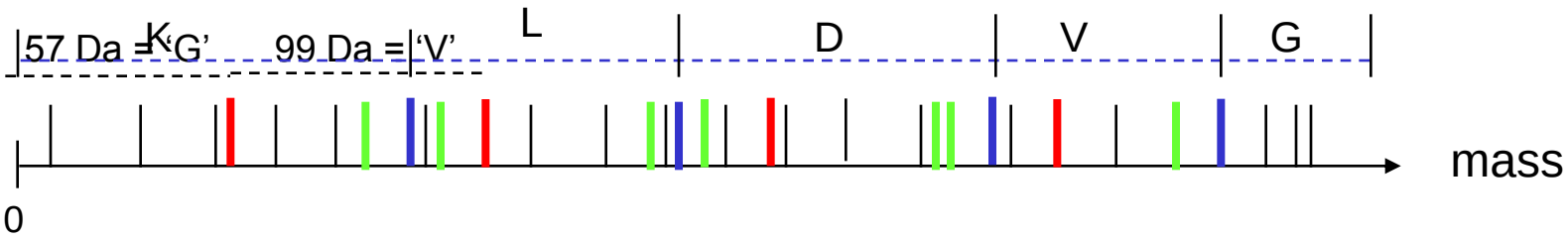


- **Peptides tend to fragment along the backbone**
- **Fragments can also loose neutral chemical groups like NH_3 and H_2O**

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

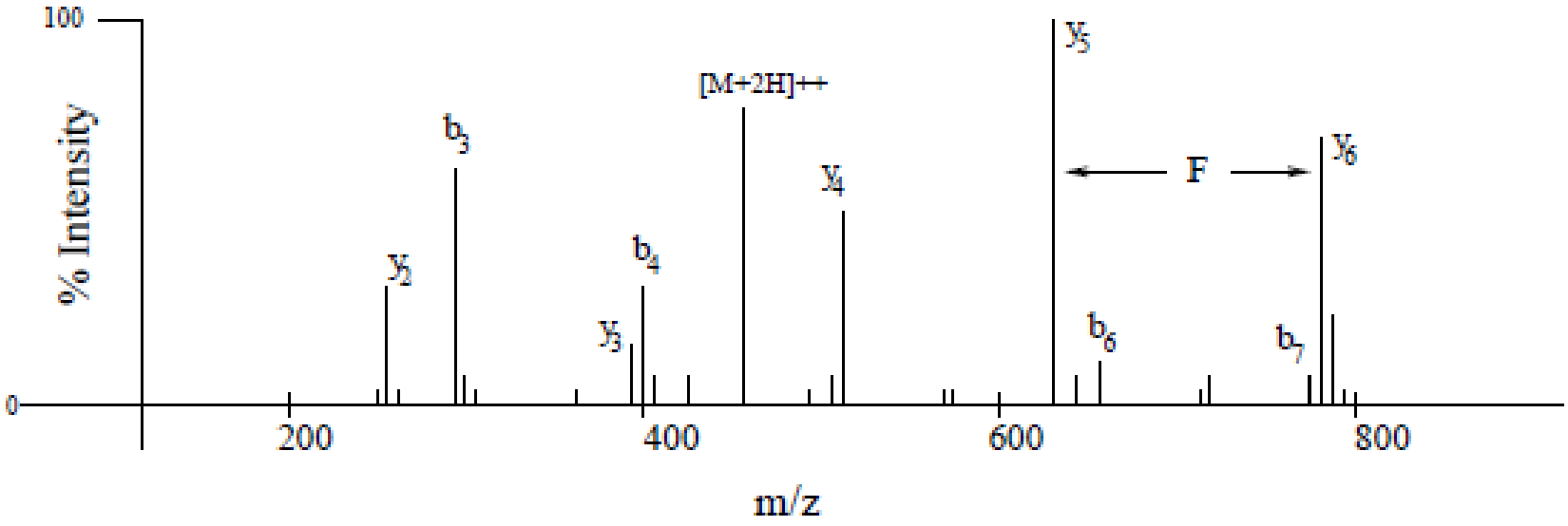
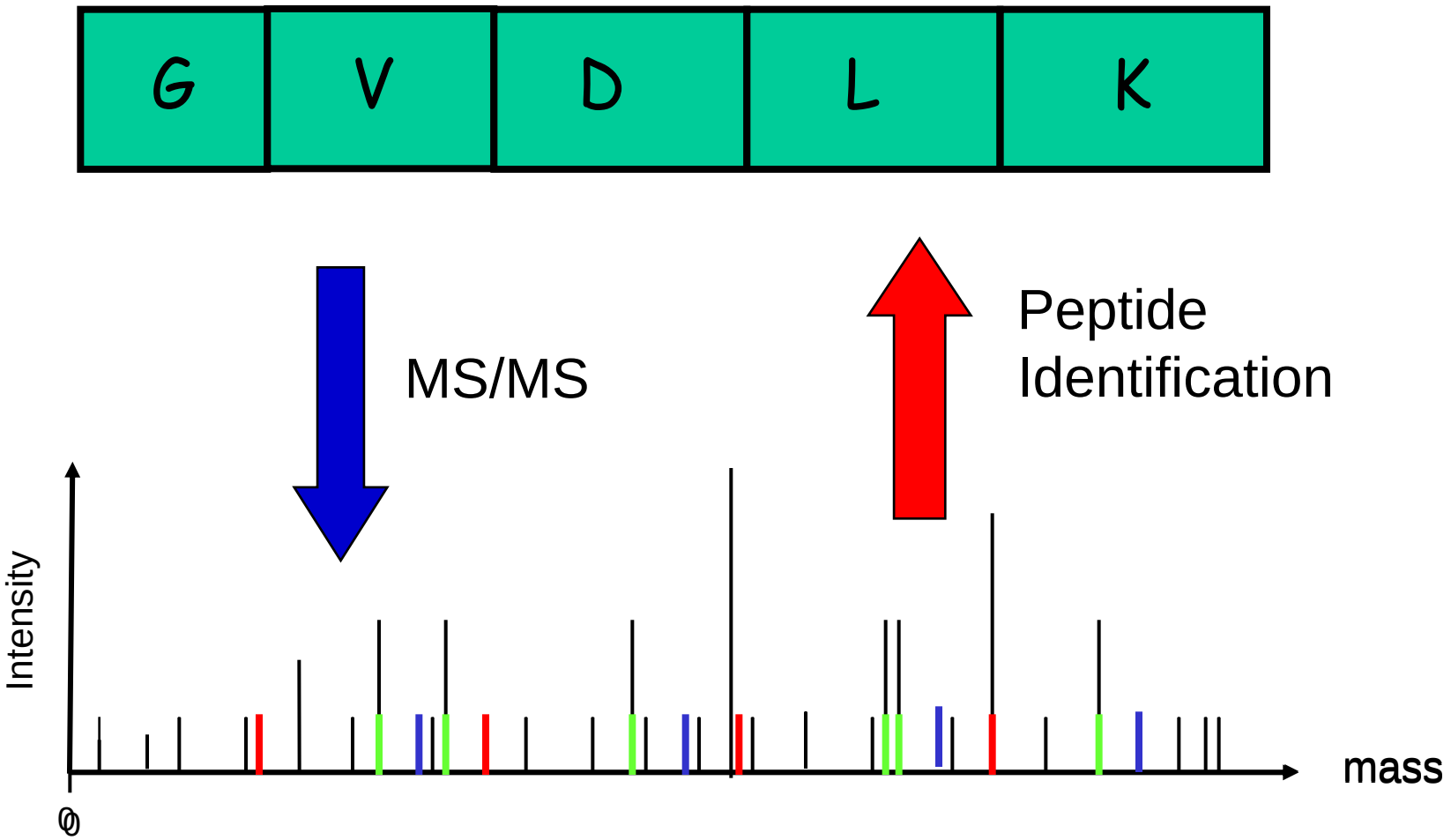


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

Protein Identification with MS/MS

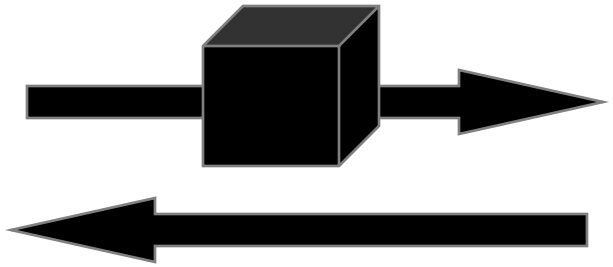


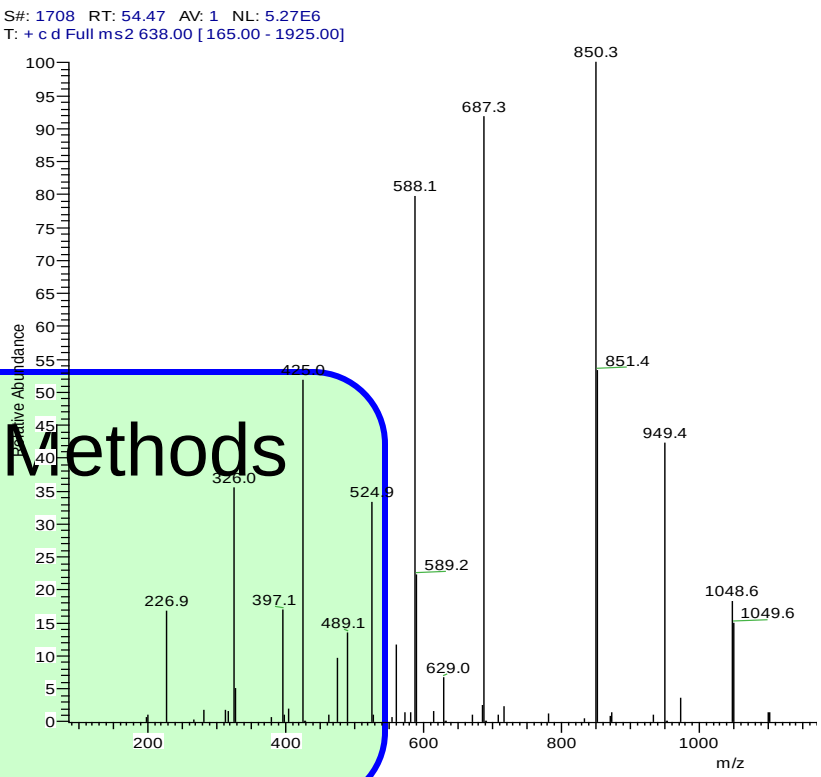
Source: Leong Hon Wai

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

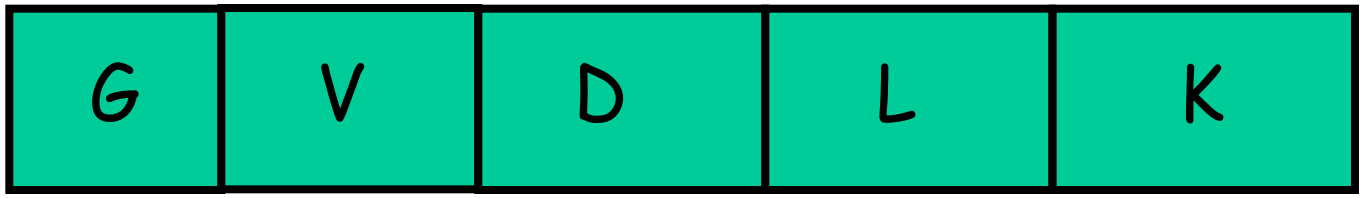
- **Database search**
  - Used for spectrum from known peptides
  - Rely on completeness of database

- **General Approach**
  - Match given spectrum with known peptide
  - Enhanced with advanced statistical analysis and complex scoring functions

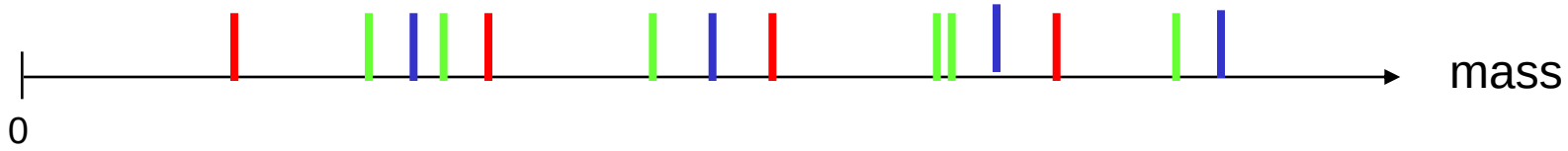
- **Methods**
  - SEQUEST, MASCOT, InsPecT, Paragon

Theoretical Spectrum for a Peptide

- **Given this peptide**



- **Its theoretical spectrum is**

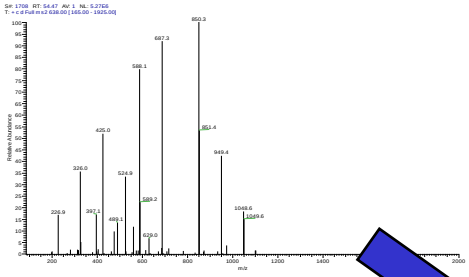


- **Theoretical spectrum is dependent on**
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

Source: Leong Hon Wai

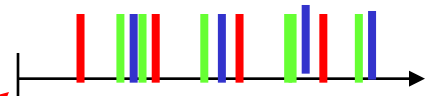
Database Search Algorithm

Database Search



Database of known peptides

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



Theoretical spectrum

Match

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

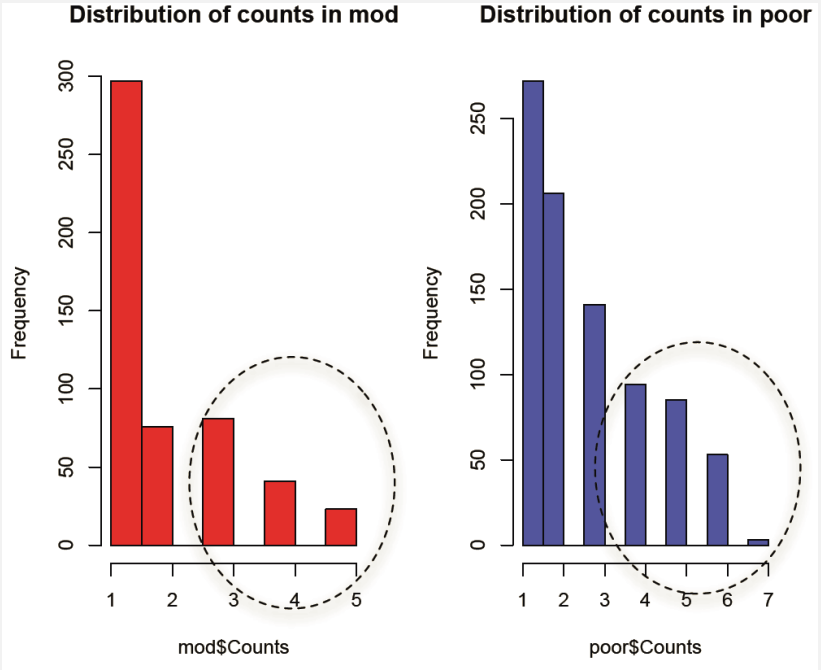
- **After all the peptides have been identified, they are grouped into protein identifications**
- **Peptide scores are added up to yield protein scores**
- **Confidence of a particular peptide identification increases if other peptides identify the same protein and decreases if no other peptides do 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

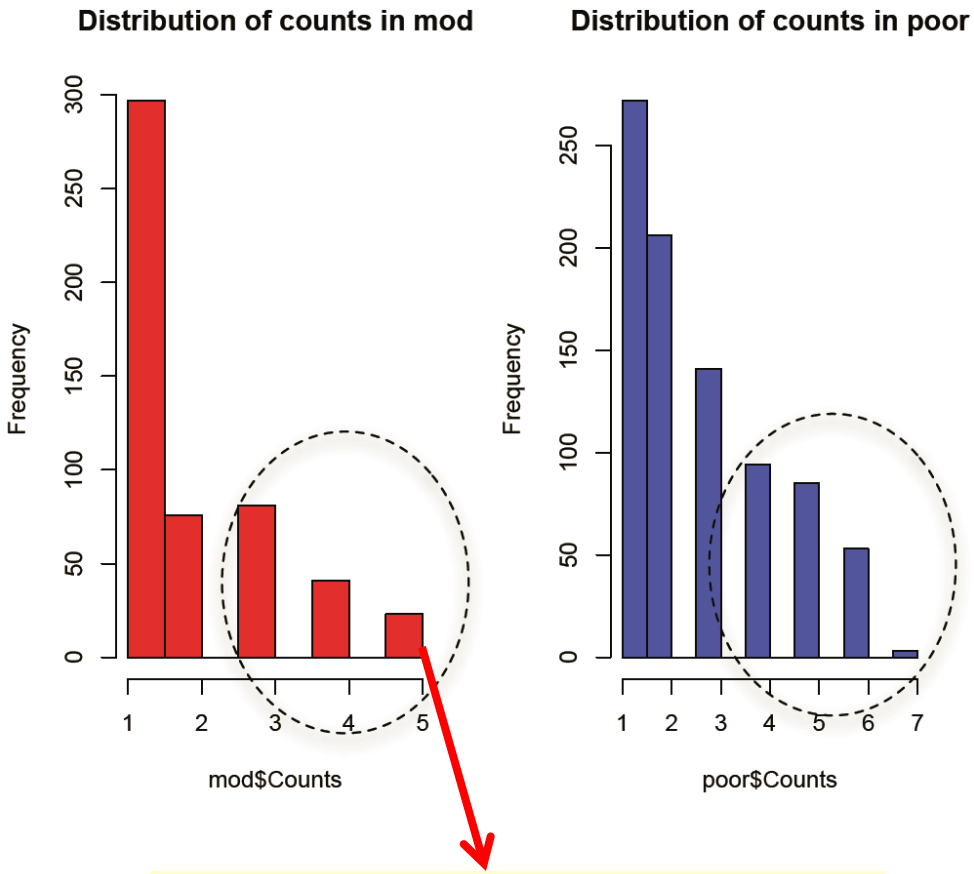
- **Once the proteins are identified, the proteomic profile of a sample can be constructed**
  - I.e., which protein is found in the sample and how abundant it is
- **Similar to gene expression profile. So gene expression profile analysis techs can be applied**
- **Some key differences**
  - Proteomic 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



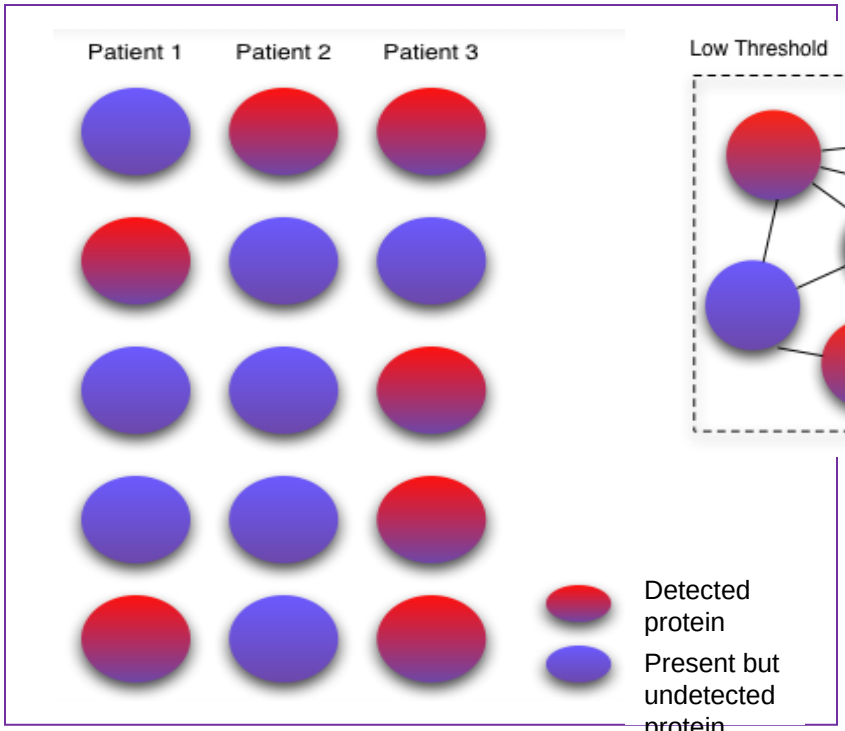
Only 25 out of 800+ proteins are common to all 5 mod-stage HCC patients!

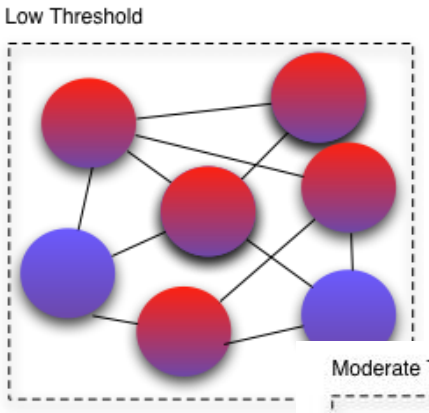
Issues in Proteomic Profiling

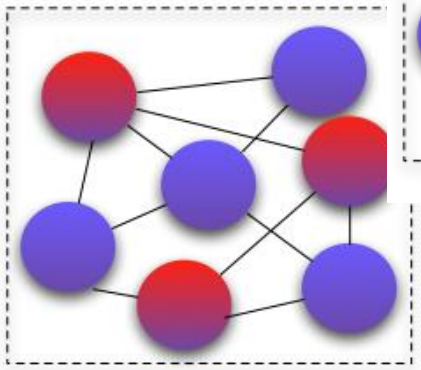
- **Coverage**
- **Consistency**

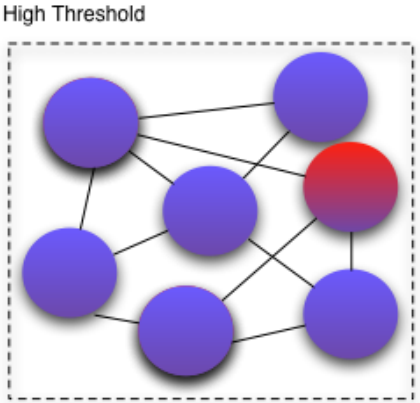
⇒ **Thresholding**

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

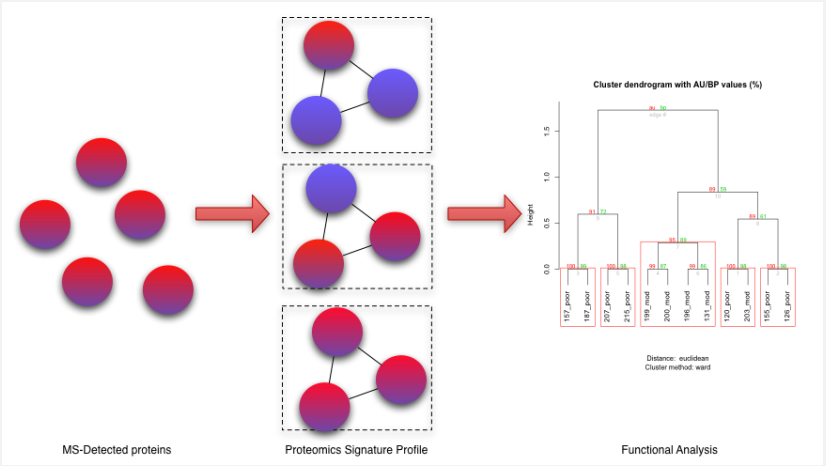








Part 2: Delivering more powerful proteomic profile analysis

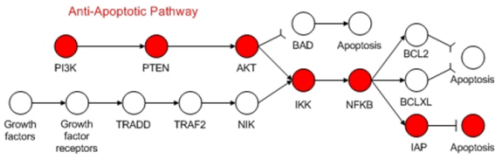


- Common issues in proteomic profile analysis
- **Improving consistency**
  - PSP
  - PDS
- Improving coverage
  - CEA
  - PEP
  - Max Link

An inspiration from gene expression profile analysis

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?

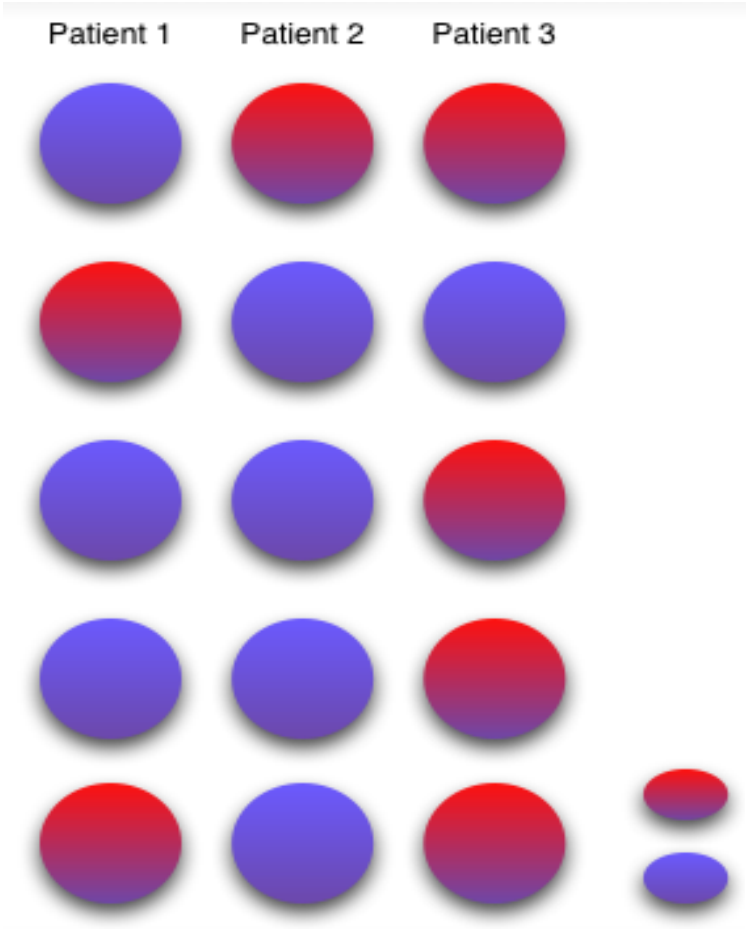
- Prob(group of genes correlated) = $(1/2^6)^5$
  - Good, << $1/2^6$
- ~~# of groups = $^{100000}C_5$~~
- ~~E(# of groups of genes correlated) = $^{100000}C_5 * (1/2^6)^5 = 2.6*10^{12}$~~

⇒ 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**

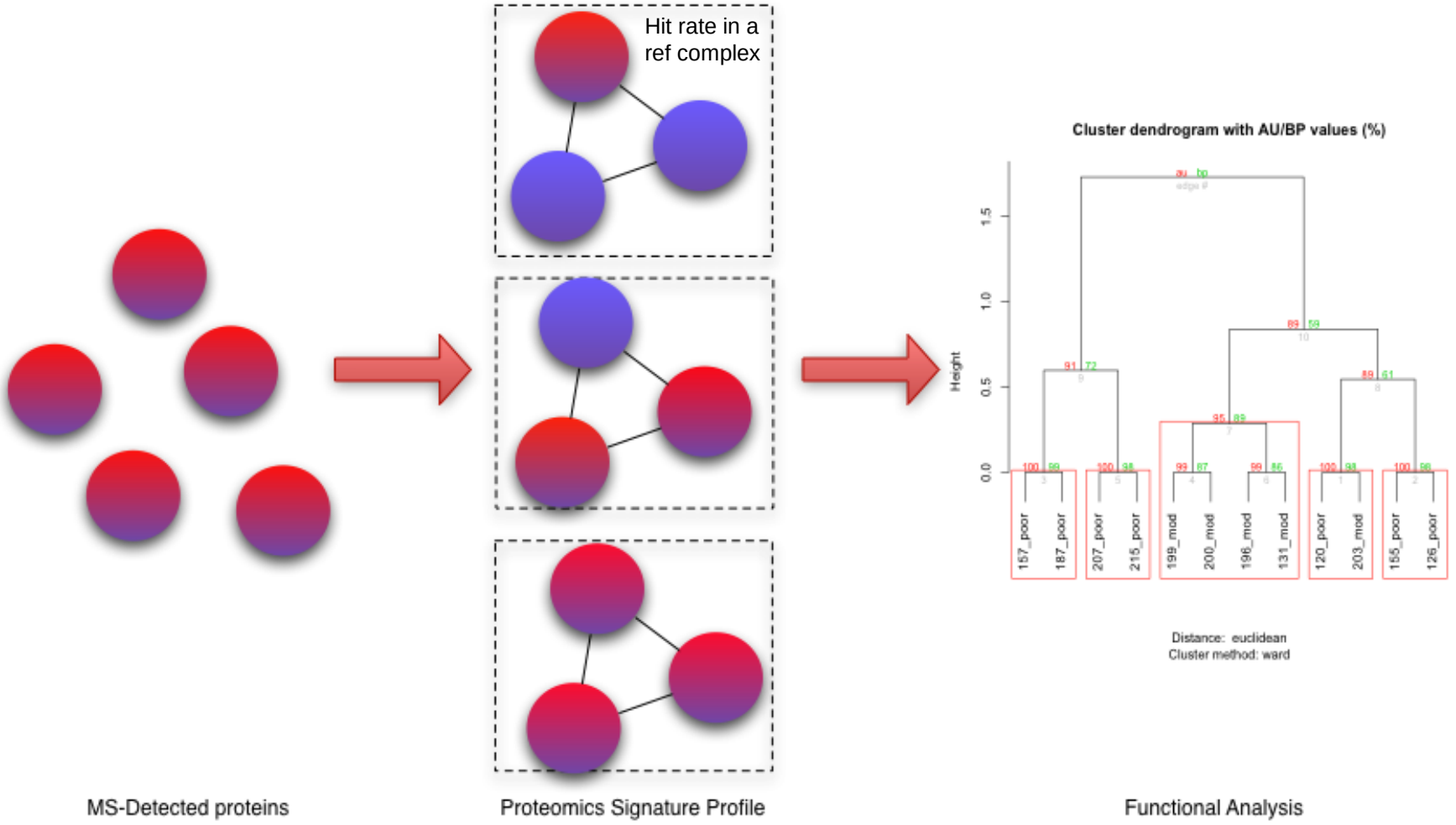
  - Construct a profile based on complexes?

We try an adaptation of SNet on proteomics profiles…

"Proteomic Signature Profiling" (PSP)

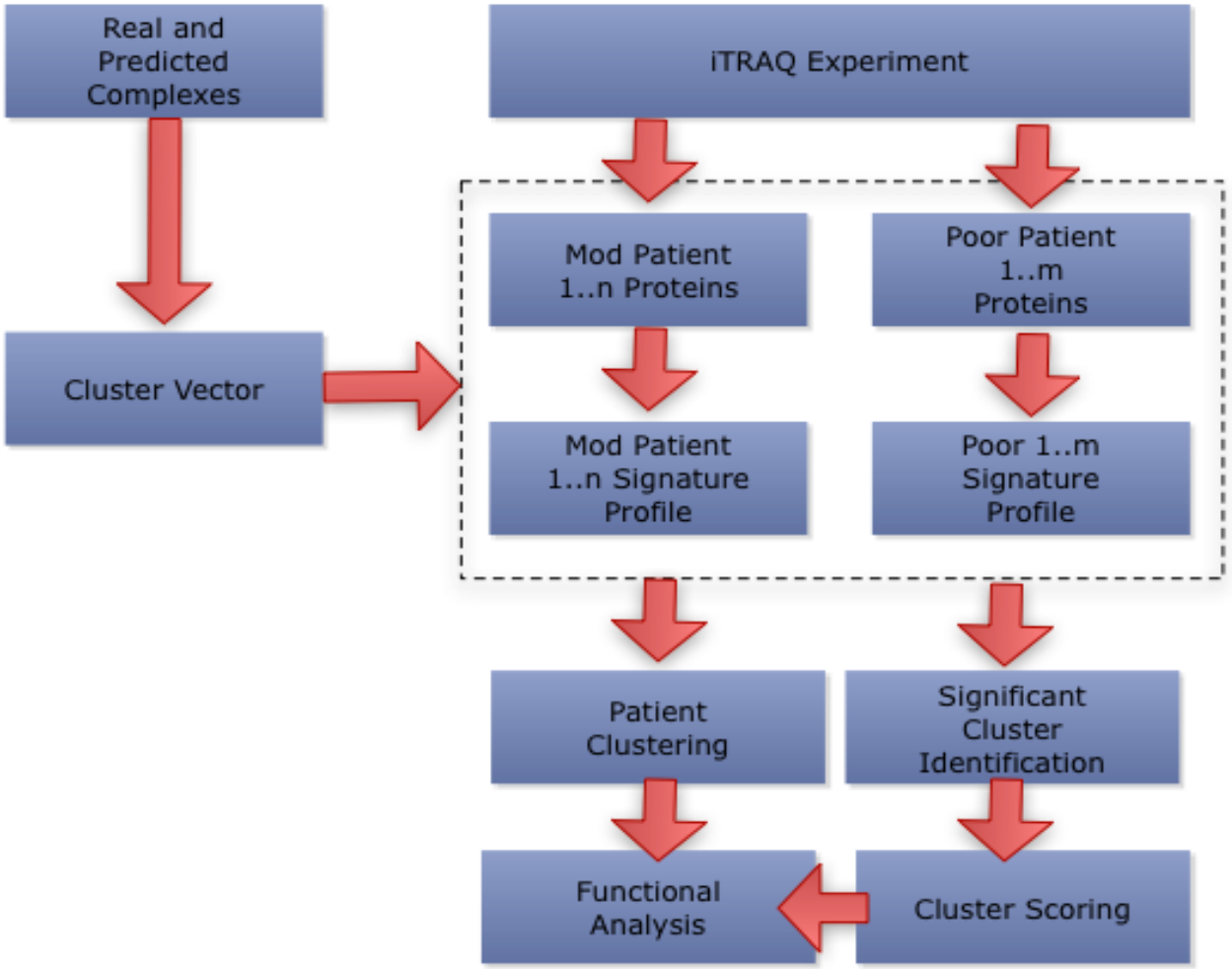
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

"Threshold-free" Principle of PSP



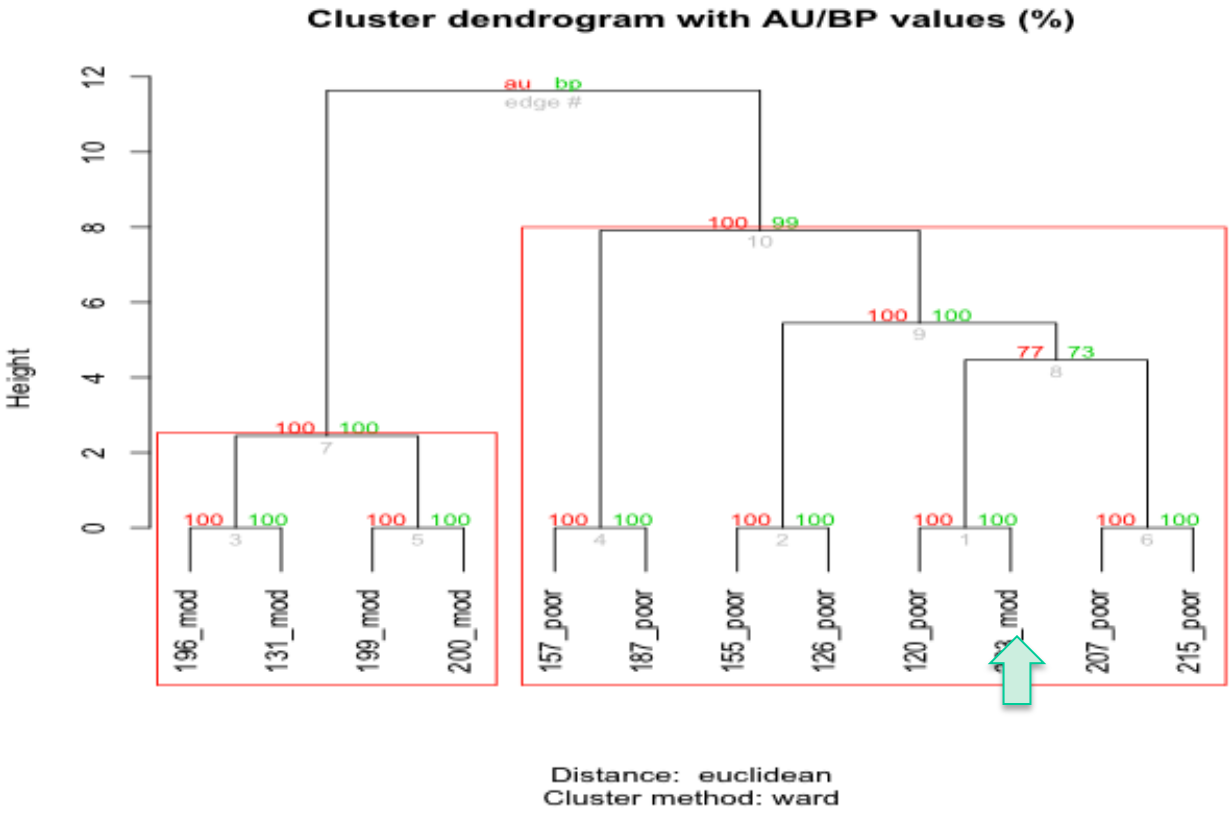
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

Applying PSP to a HCC Dataset



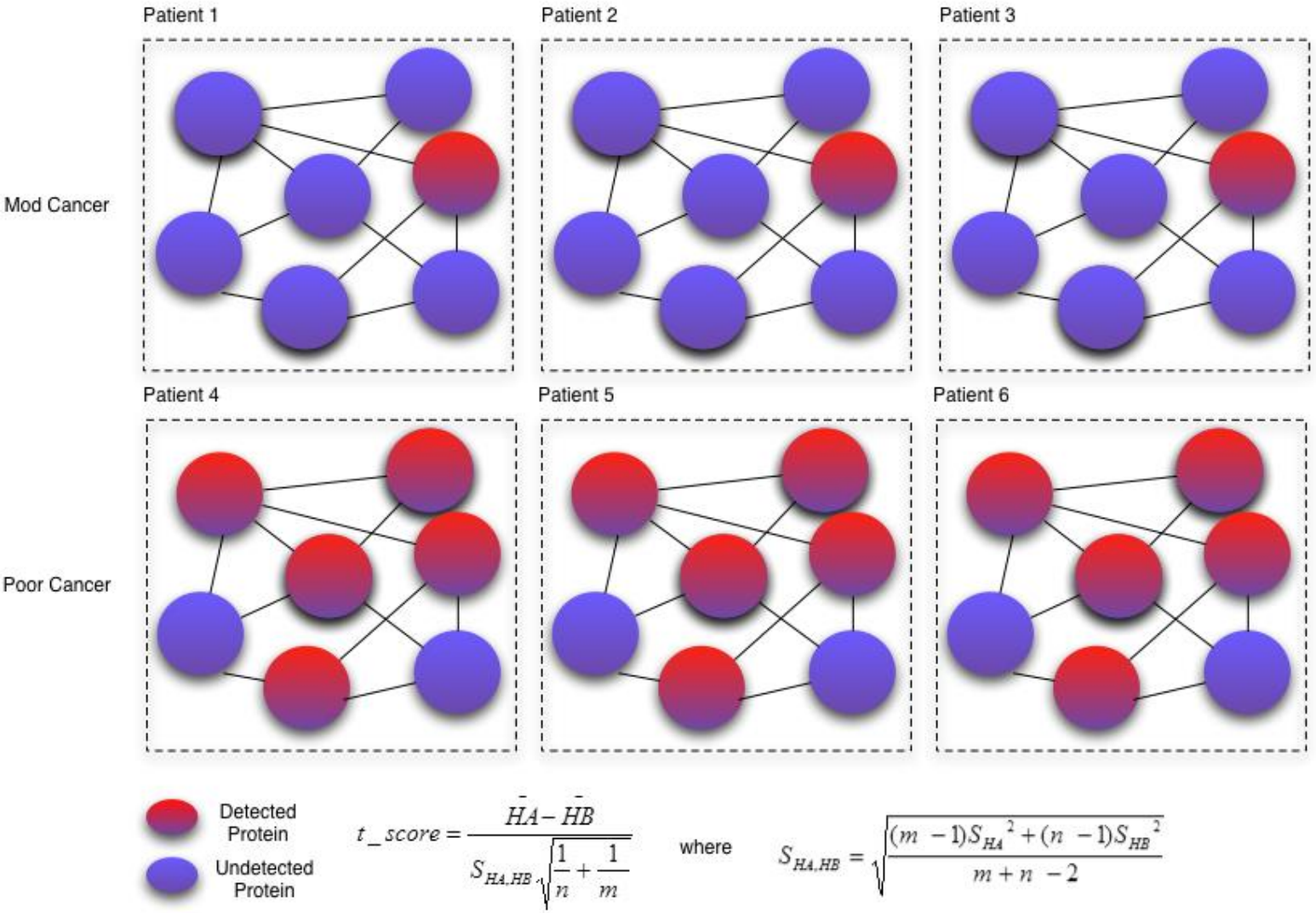
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research*. accepted.

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



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

NUS National University of Singapore

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.

NUS National University of Singapore

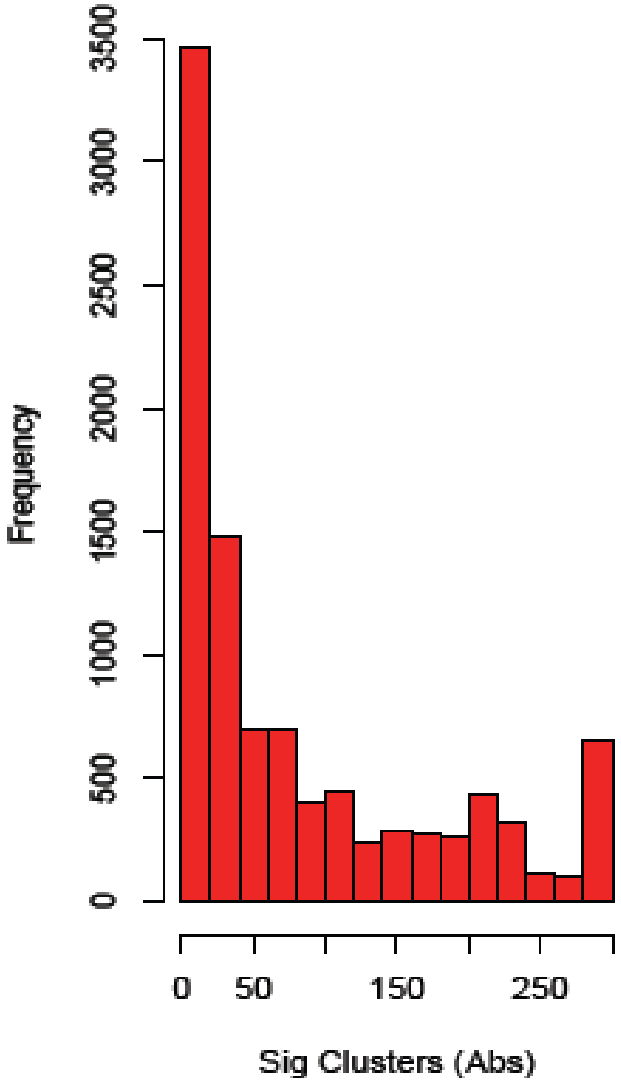
Top-Ranked GO Terms

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.*

NUS National University of Singapore

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?**

NUS National University of Singapore

Another adaptation of SNet on proteomics profiles…

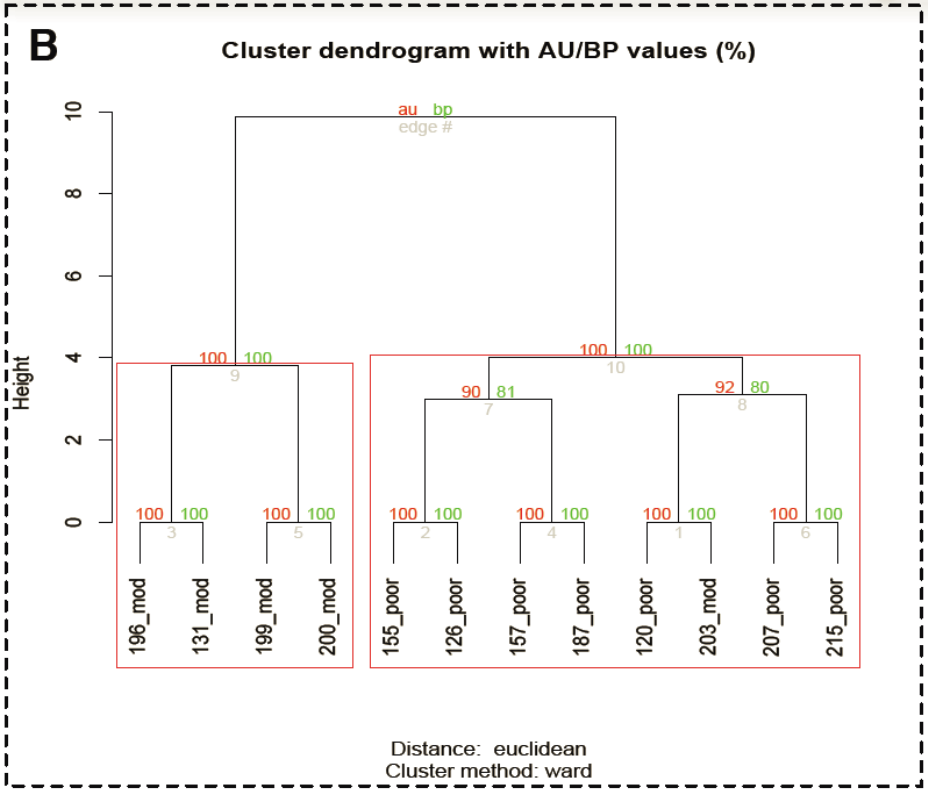
"Pathway-Derived Subnets" (PDS)

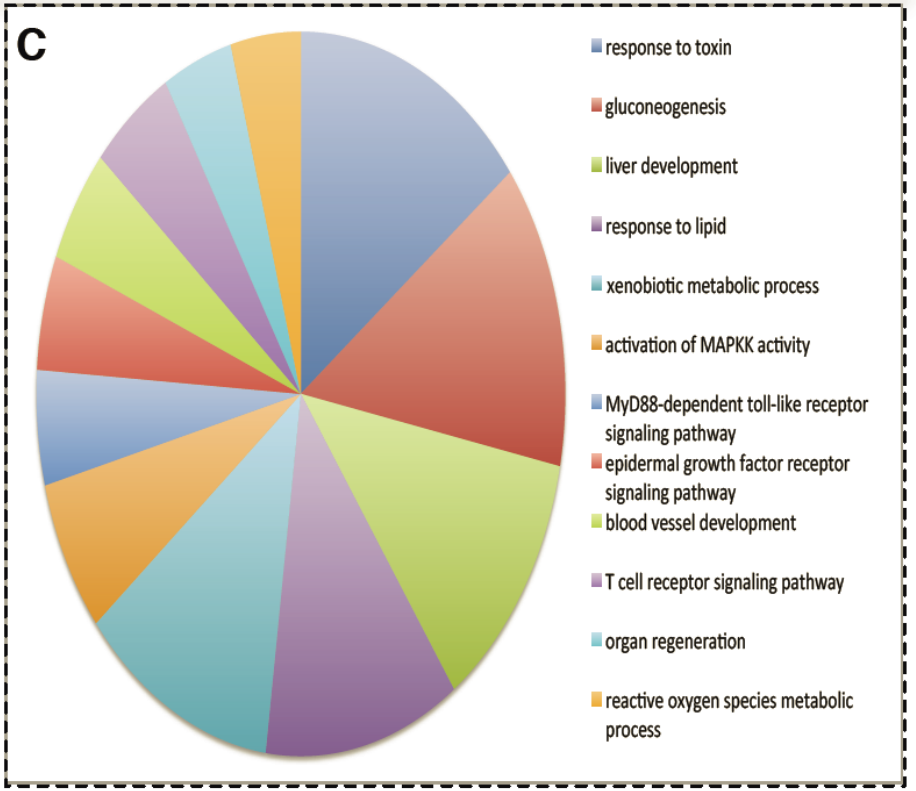
Pathway-Derived Subnets (PDS)

- **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, …, 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

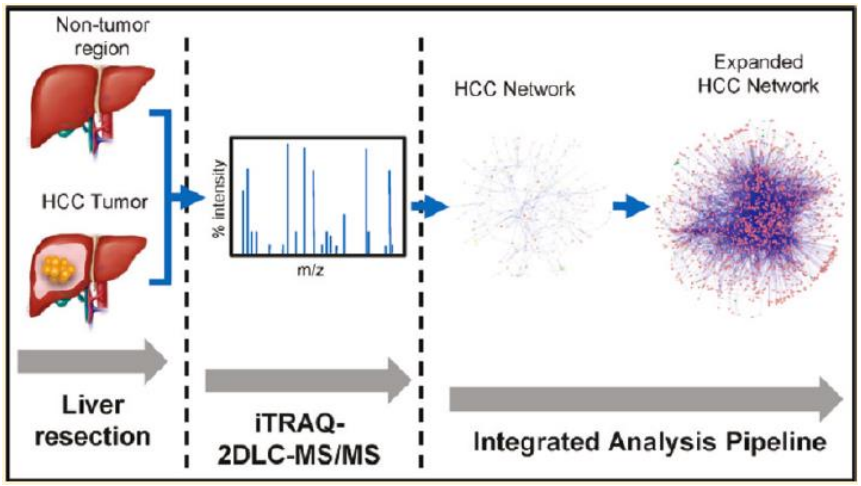




What have we learned?

- **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



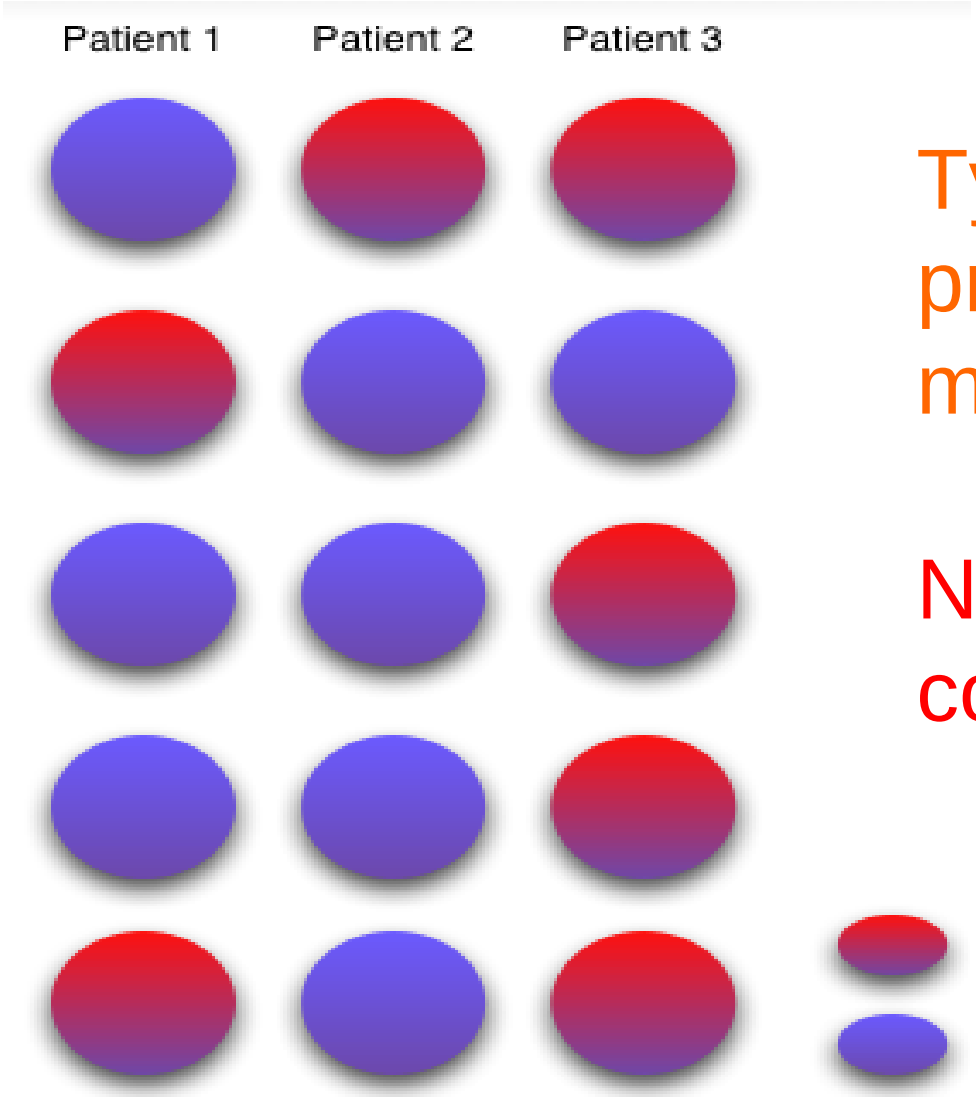
- **Common issues in proteomic profile analysis**
- **Improving consistency**
  - PSP, PDS
- **Improving coverage**
  - FCS,
  - CEA, PEP
  - Max Link

NUS National University of Singapore

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!

NUS National University of Singapore

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.

NUS National University of Singapore

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

NUS National University of Singapore

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

NUS National University of Singapore

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 …

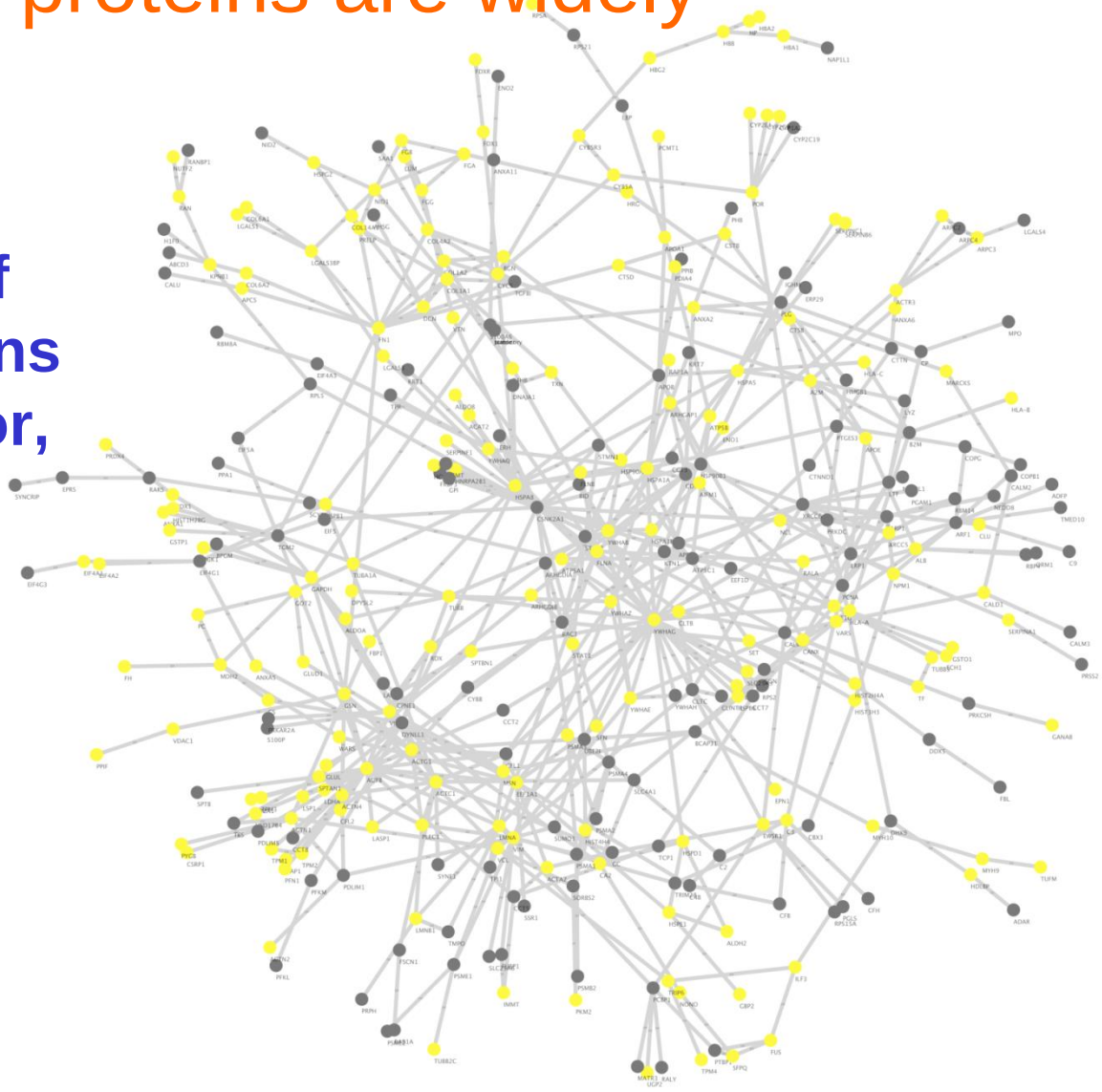
Background

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

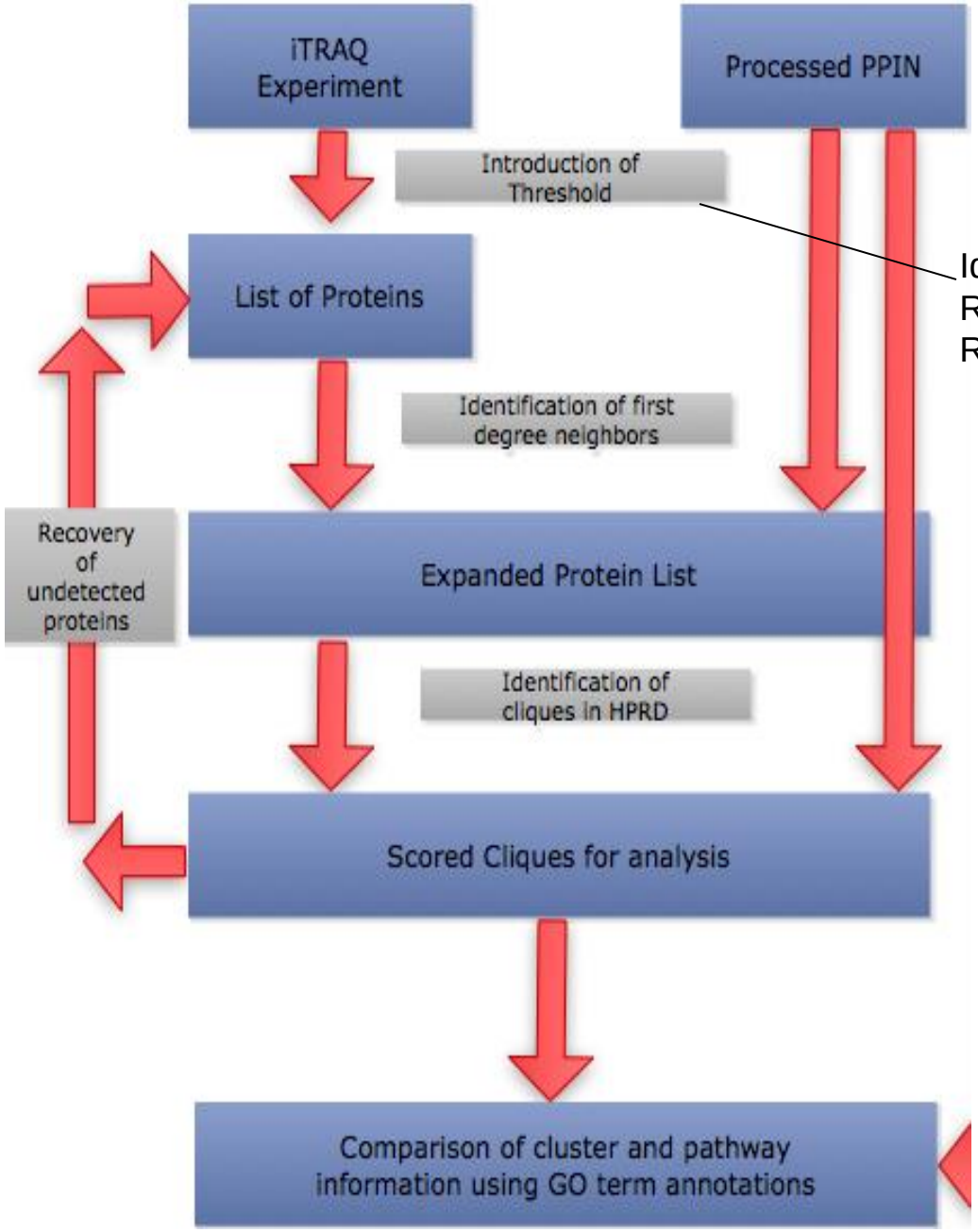
- **Mass Spectrometry**
  - iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)
  - Coupled with 2D LC MS/MS
  - Popular because of ability to run 8 concurrent samples in one go

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



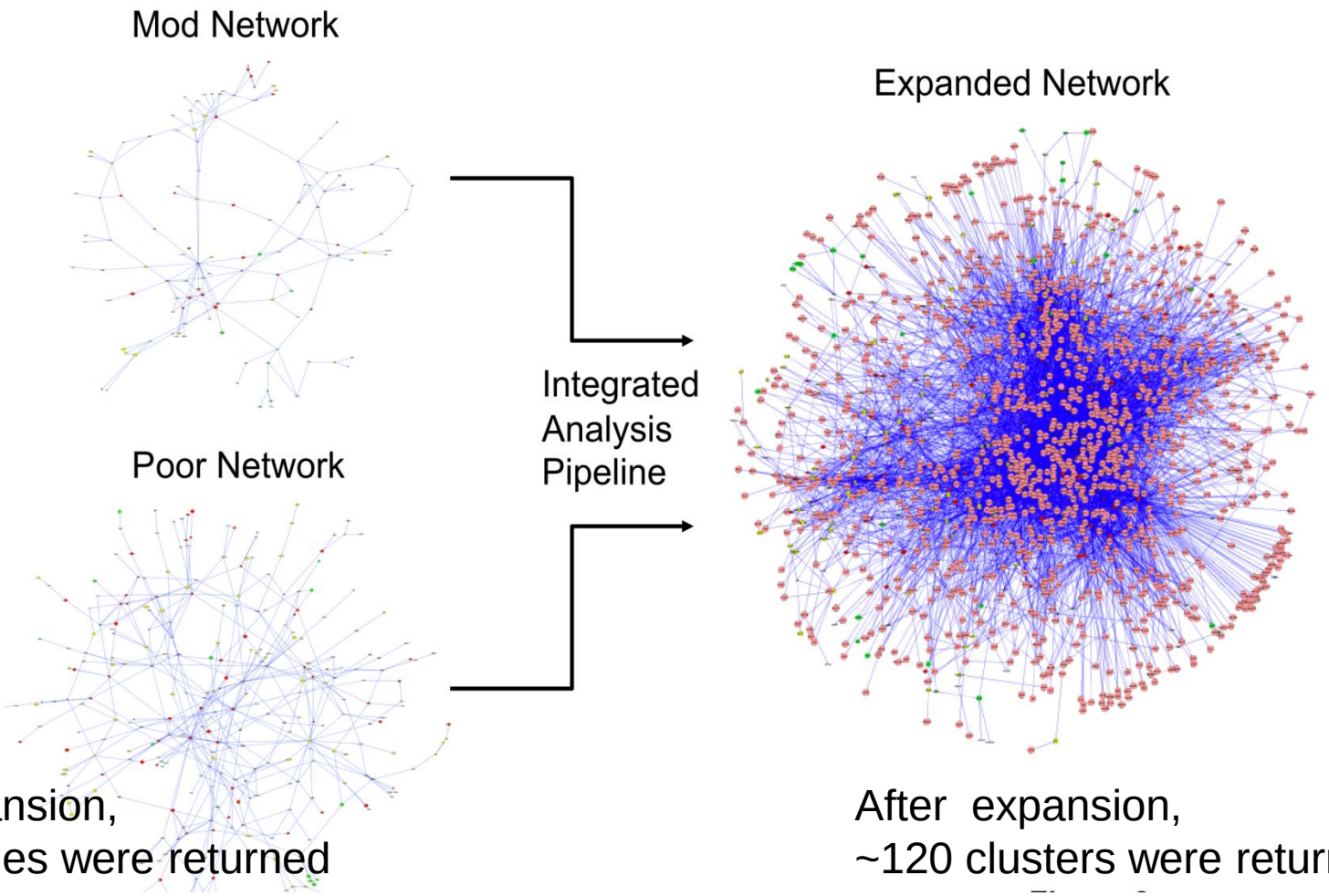
Identify the "seeds"
Ratio < 0.8 and > 1.25 for Mod (min 3 patients)
Ratio < 0.8 and > 1.25 for Poor (min 4 patients)

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

NUS National University of Singapore

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

NUS National University of Singapore

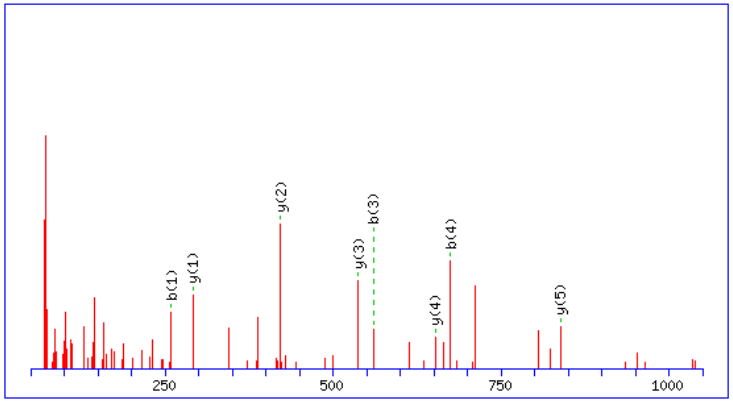
Successful Verification

ACTR2

1068. IPI00005159 Mass: 48707 Score: 39 Queries matched: 3
Tax_Id=9606 Gene_Symbol=ACTR2 Actin-like protein 2
Check to include this hit in error tolerant search or archive report

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Expect	Rank	Peptide
739	1096.54	1095.53	1095.44	0.10	0	39	0.018	1	R.NWDDMK.M
2711	1410.79	1409.78	1409.65	0.13	1	10	11	3	K.LNIDTRNCK.I
5797	1912.02	1911.01	1911.00	0.01	1	7	20	8	K.ILLTEPPMNPTKNR.E

Proteins matching the same set of peptides:
IPI00470573 Mass: 49610 Score: 39 Queries matched: 3
Tax_Id=9606 Gene_Symbol=ACTR2 actin-related protein 2 isoform a
IPI00749250 Mass: 49499 Score: 39 Queries matched: 3
Tax_Id=9606 Gene_Symbol=ACTR2 45 kDa protein



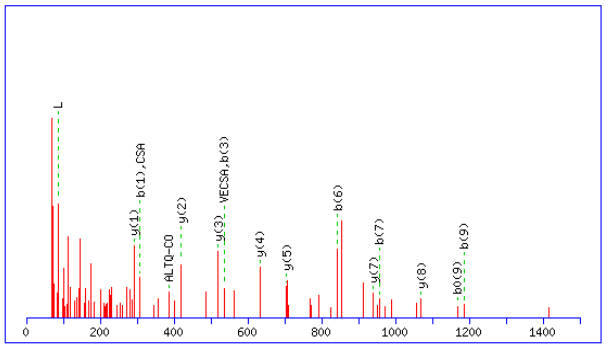
MONOISOTOPIC mass of neutral peptide Mr(calc): 1095.44
Fixed modifications: MMTS (C),(N-TERM)_iTRAQ,Lysine(K)_iTRAQ
Ions Score: 39 **Expect:** 0.018
Matches (Bold Red): 8/57 fragment ions using 15 most intense peaks

#	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 2 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
4313	1590.84	1589.83	1589.75	0.08	0	8	18	3	K.TCLLISYTTNK.F
4880	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

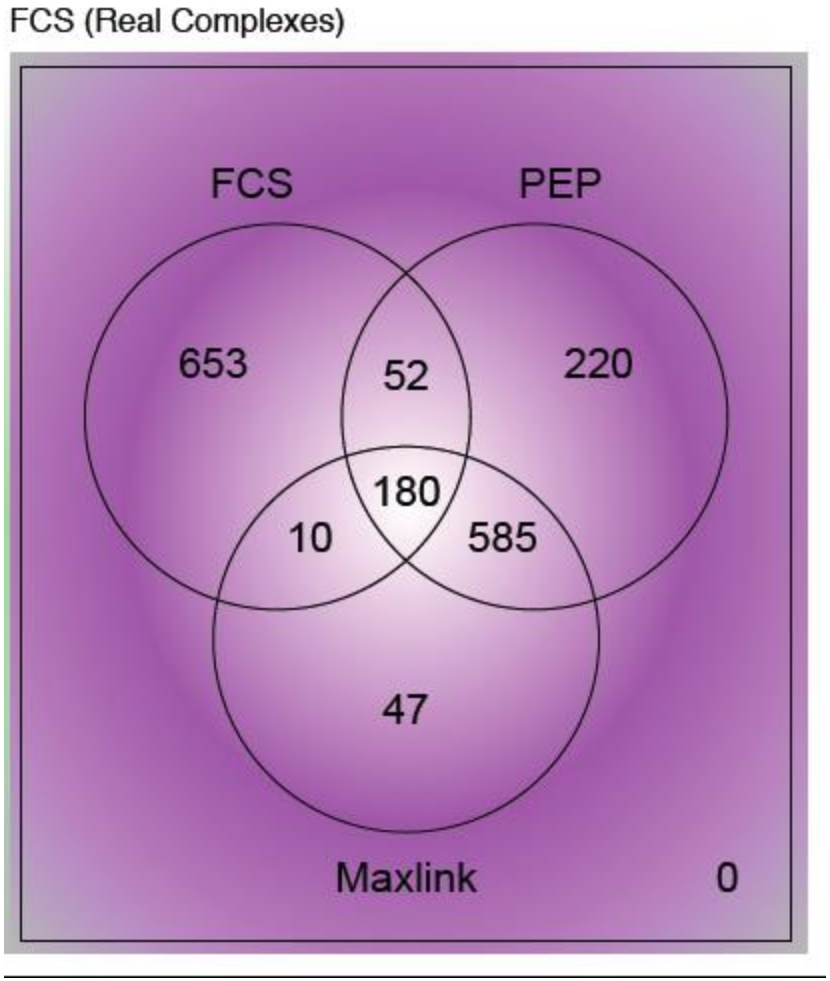
Another Experiment

- **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

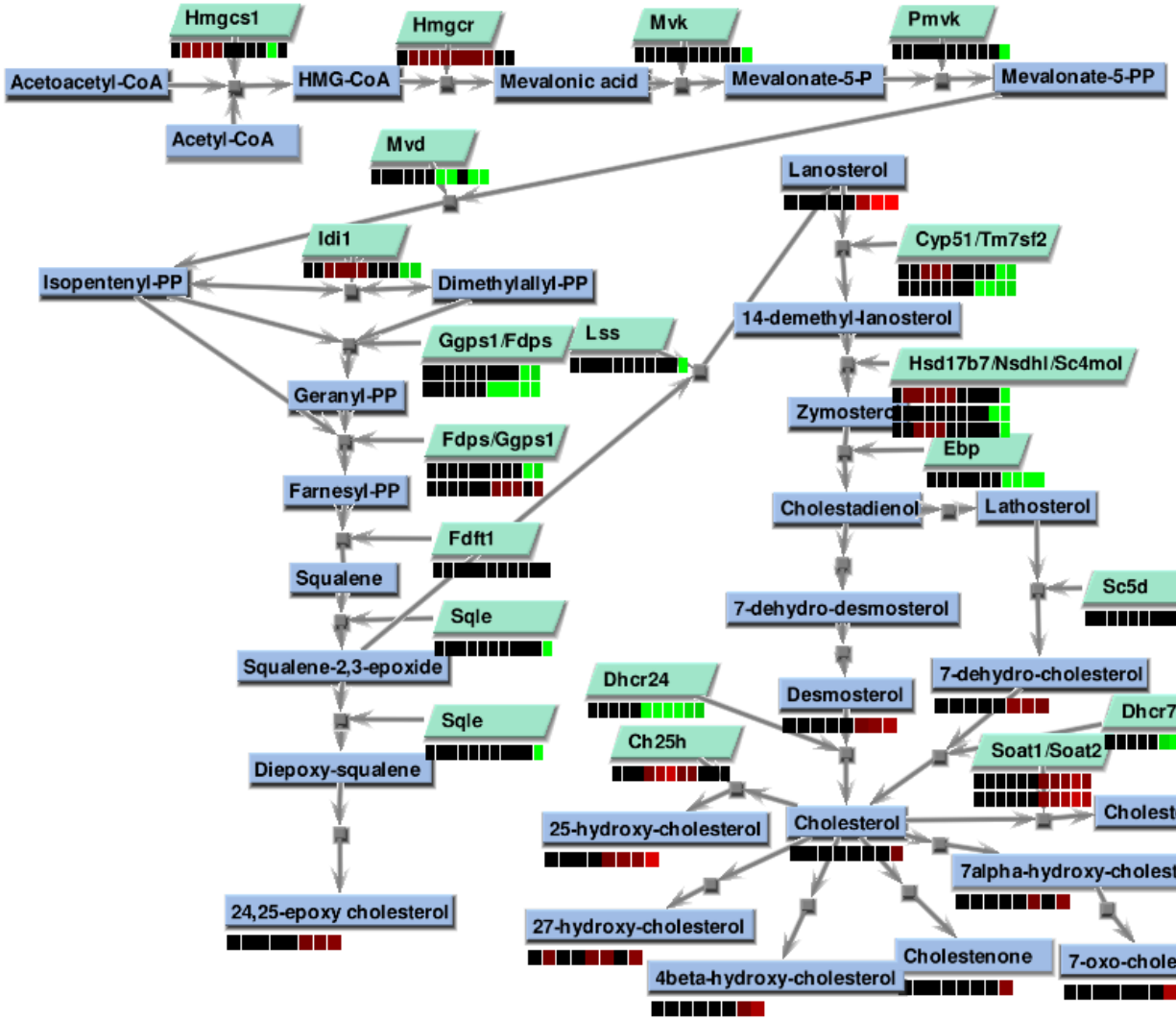


NUS National University of Singapore

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**



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

Acknowledgements

- **Chuan Hock Koh**
- **Kevin Lim**
- **Donny Soh**
- **Wilson Goh**

- **Singapore funding agencies**
  - A*STAR
  - Ministry of Education
  - National Research Foundation
- **UK funding agencies**
  - Wellcome Trust scholarship