Improving Proteomic Profile Analysis by Contextualization

Limsoon Wong

Diagnosis Using Proteomics

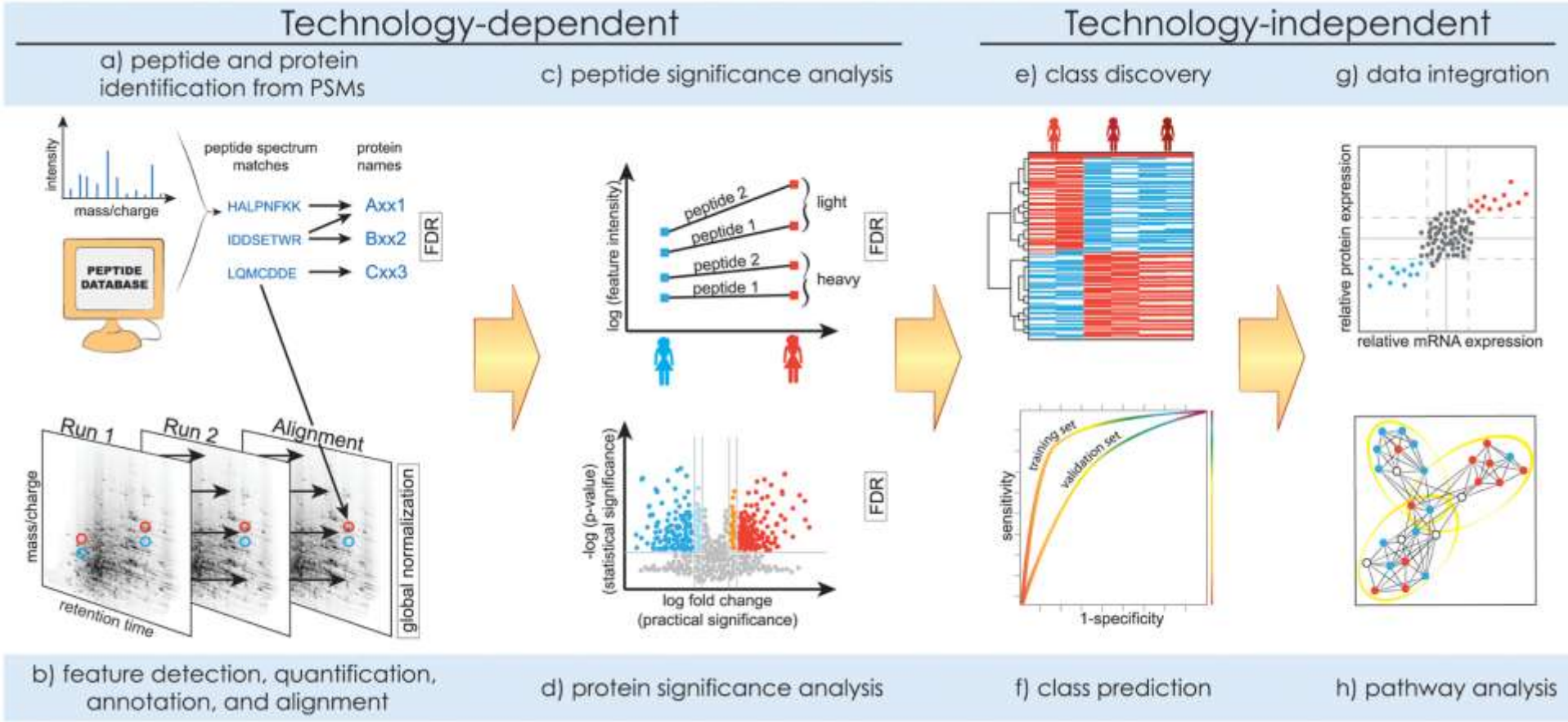


Image credit: Kall and Vitek, *PLoS Comput Biol* , 7(12): e1002277, 2011

Plan

- **Common issues in proteomic profile analysis**
- **Improving consistency**
- **Improving coverage**

Common Issues in Proteomic Profile Analysis

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

NUS National University of Singapore

A concrete example: Bell et al., *Nat Meth*, 6:423-430, 2009

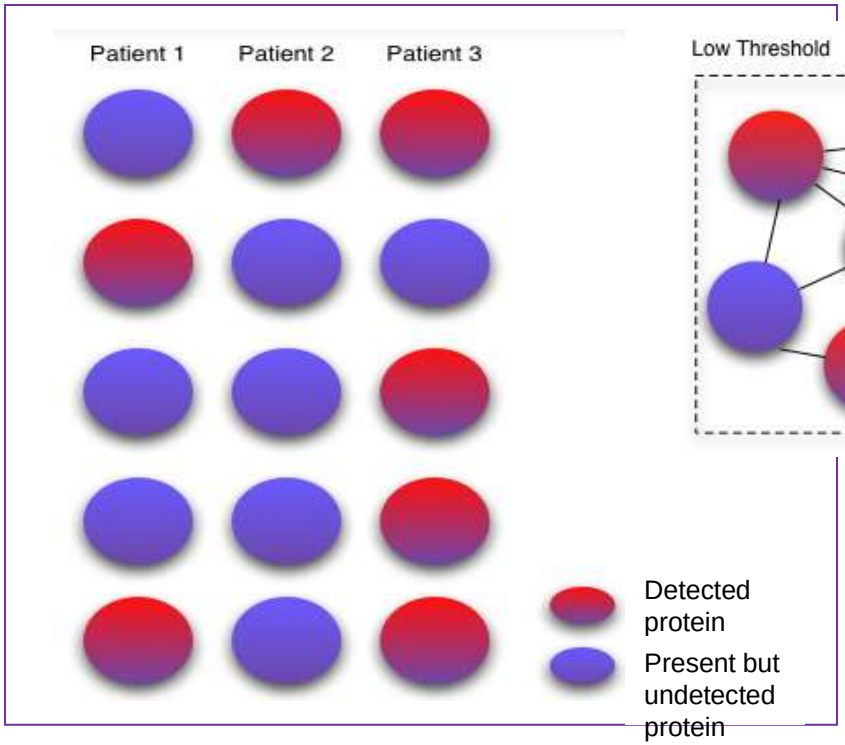
- **"… in a large-scale collaborative study by Bell et al. to assess the extent of reproducibility across different laboratories. The results were striking – only 7 out of 27 laboratories correctly reported all 20 proteins, and only 1 laboratory successfully reported all 22 unique peptides."**

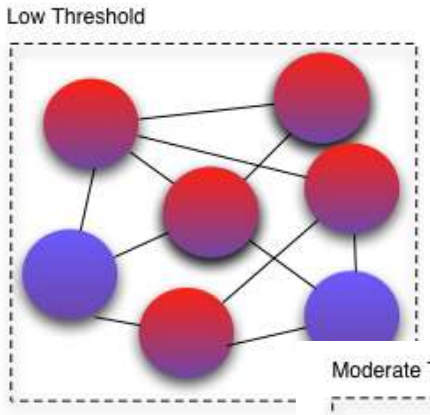
Issues in Proteomic Profiling

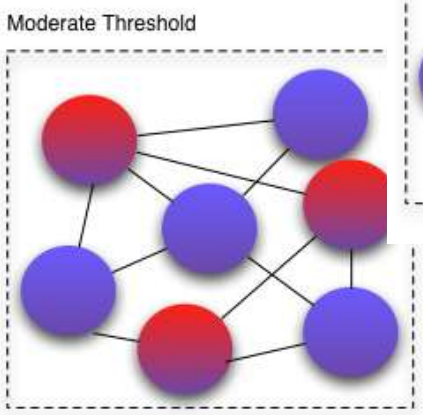
- **Coverage**
- **Consistency**

⇒ **Thresholding**

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







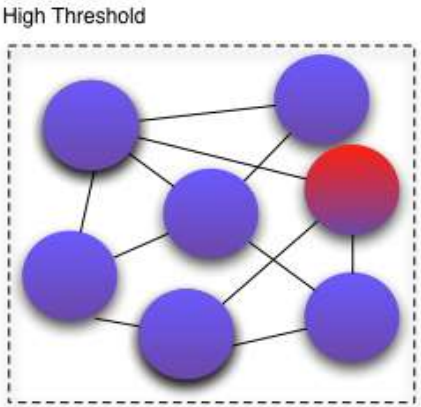
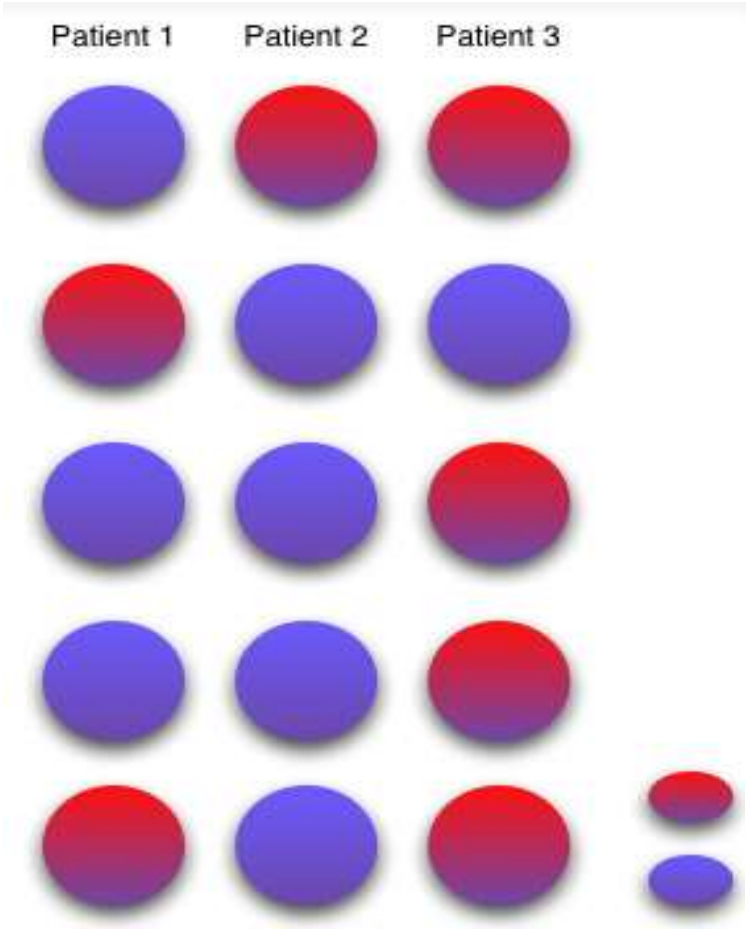


Image credit: Wilson Goh

Improving Consistency in Proteomic Profile Analysis

Intuitive Example



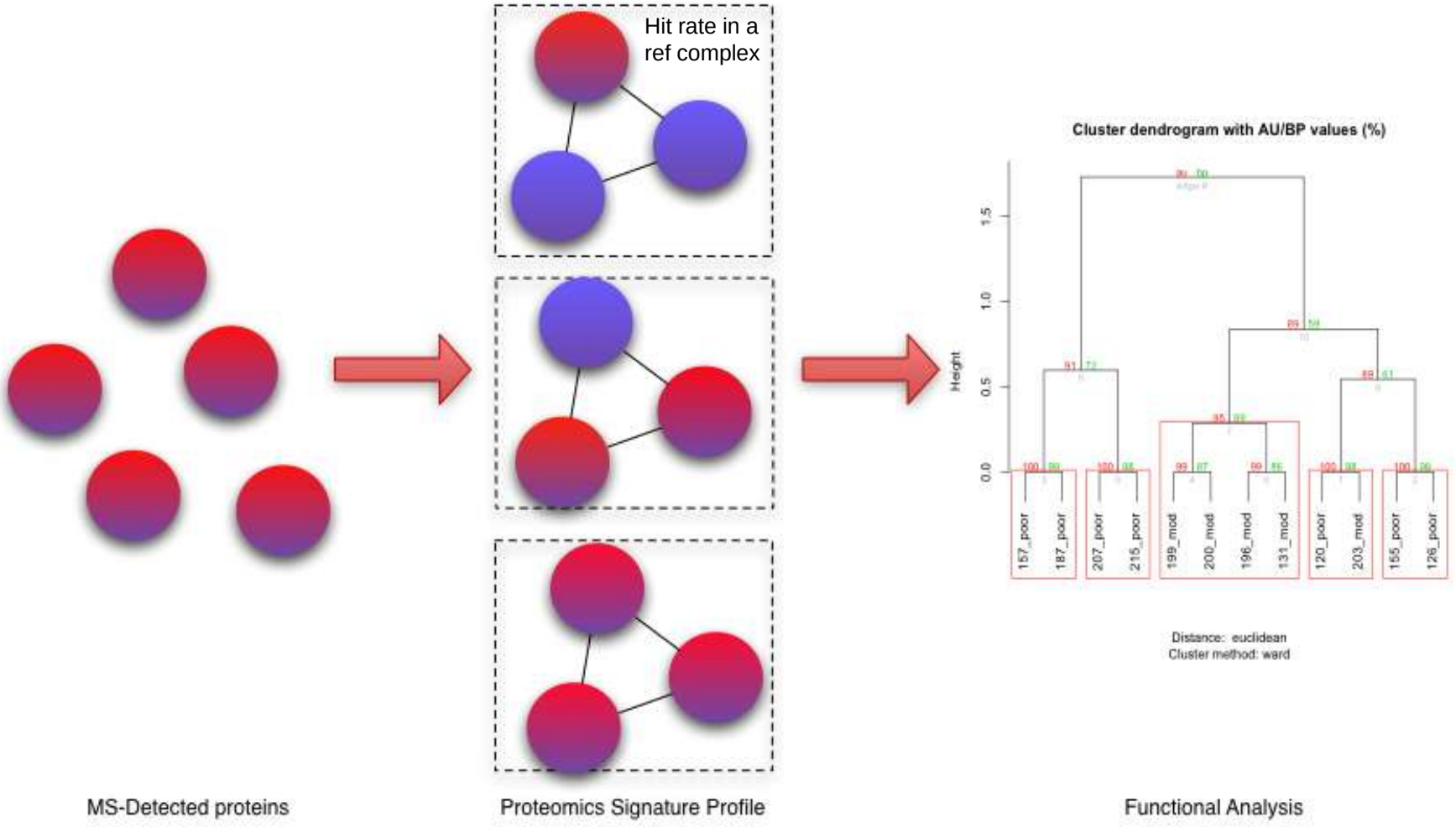
- **Suppose the failure to form a protein complex causes a disease**
  - If any component protein is missing, the complex can't form

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

  - Construct a profile based on complexes?

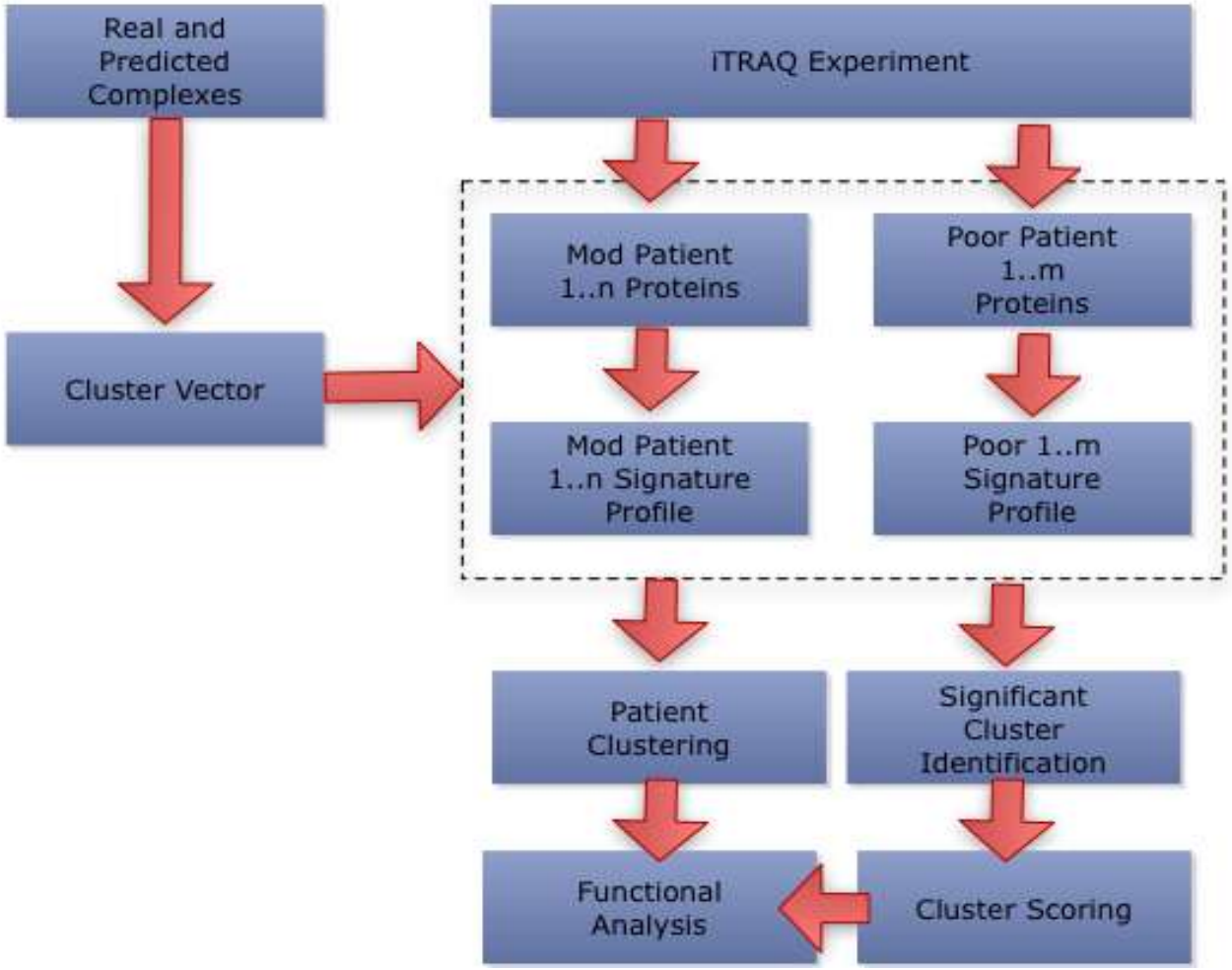
Goh et al. **Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research,*11(3):1571--1581, 2012.

"Threshold-free" Principle of PSP

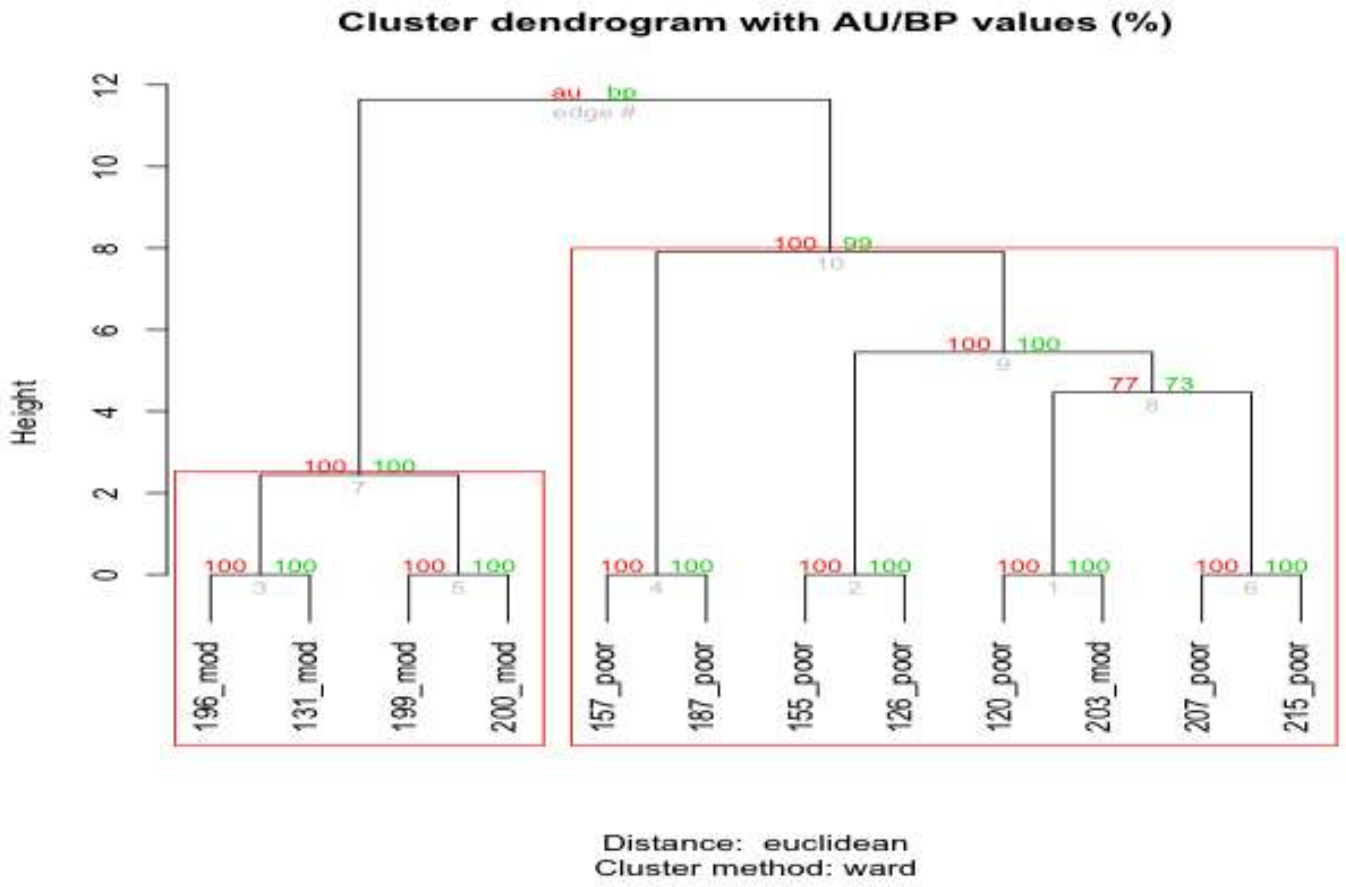


NUS National University of Singapore

Applying PSP to a HCC Dataset



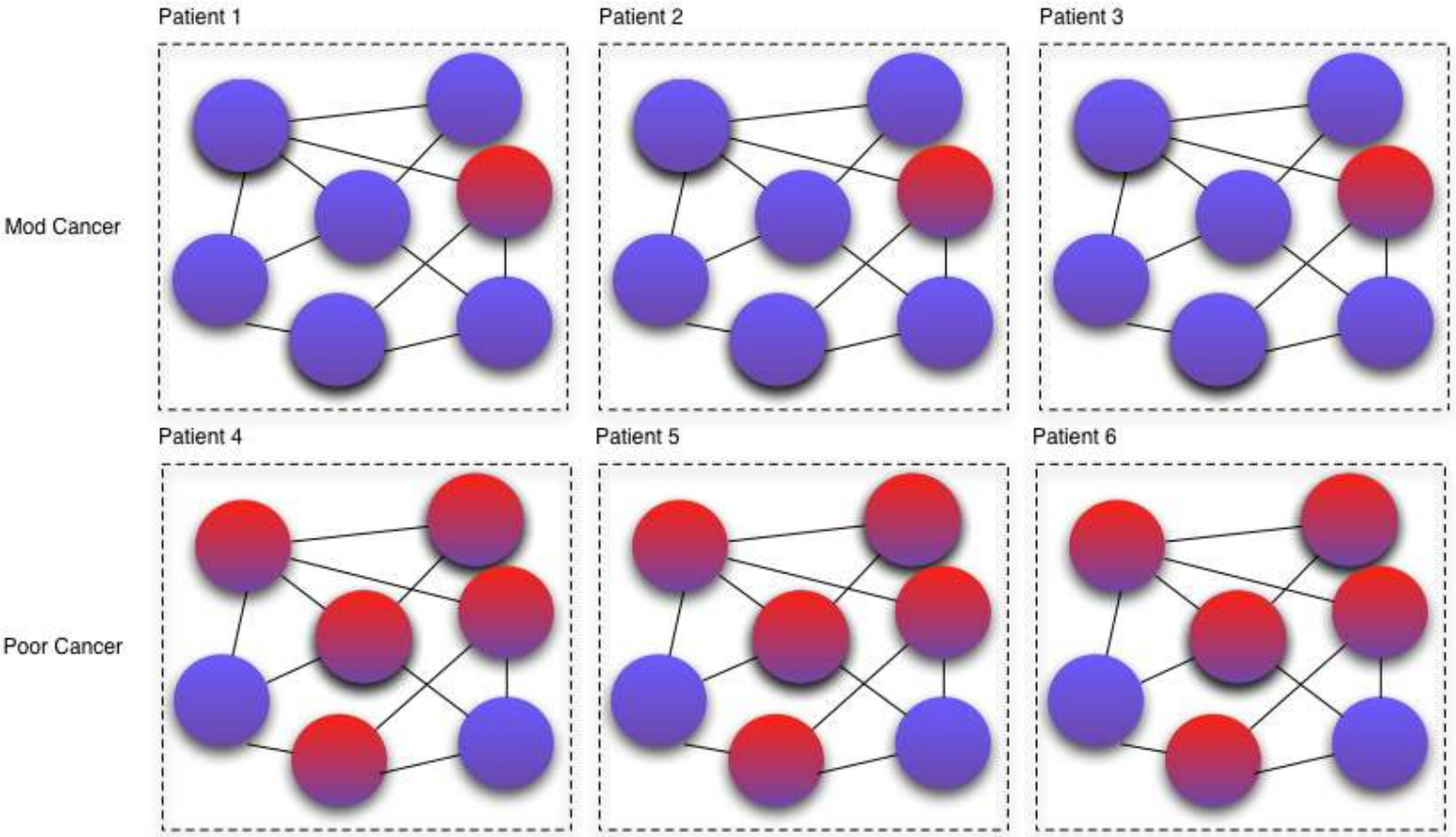
Consistency: Samples segregate by their classes with high confidence



In contrast, at the level of individual proteins, in mod-stage patients, only 25 out of over 800 proteins are common to all 5 patients. Of these 25, all are also reported in poor-stage patients. In poor-stage patients, 3 out of over 1000 proteins are common to all 7. Of these, 2 are reported in mod-stage patients.

NUS National University of Singapore

Feature Selection



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

where

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

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

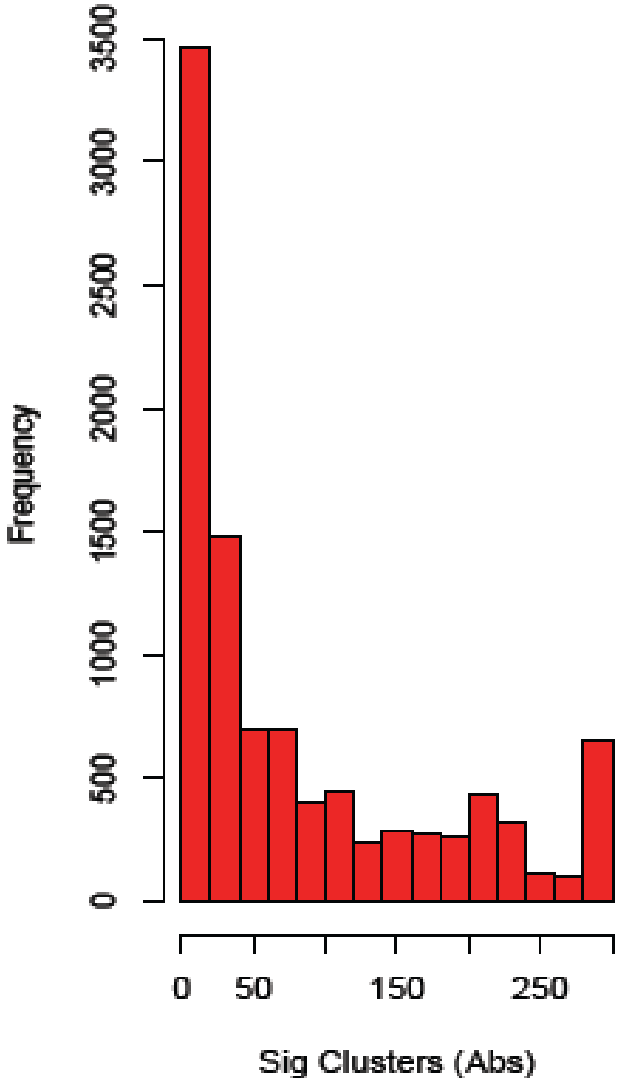
NUS National University of Singapore

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,* 14:35, 2013.

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**

Improving Coverage in Proteomic Profile Analysis

FCS

- **Rescue undetected proteins from high-scoring protein complexes produced by e.g. FCS**

- **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**

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

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**

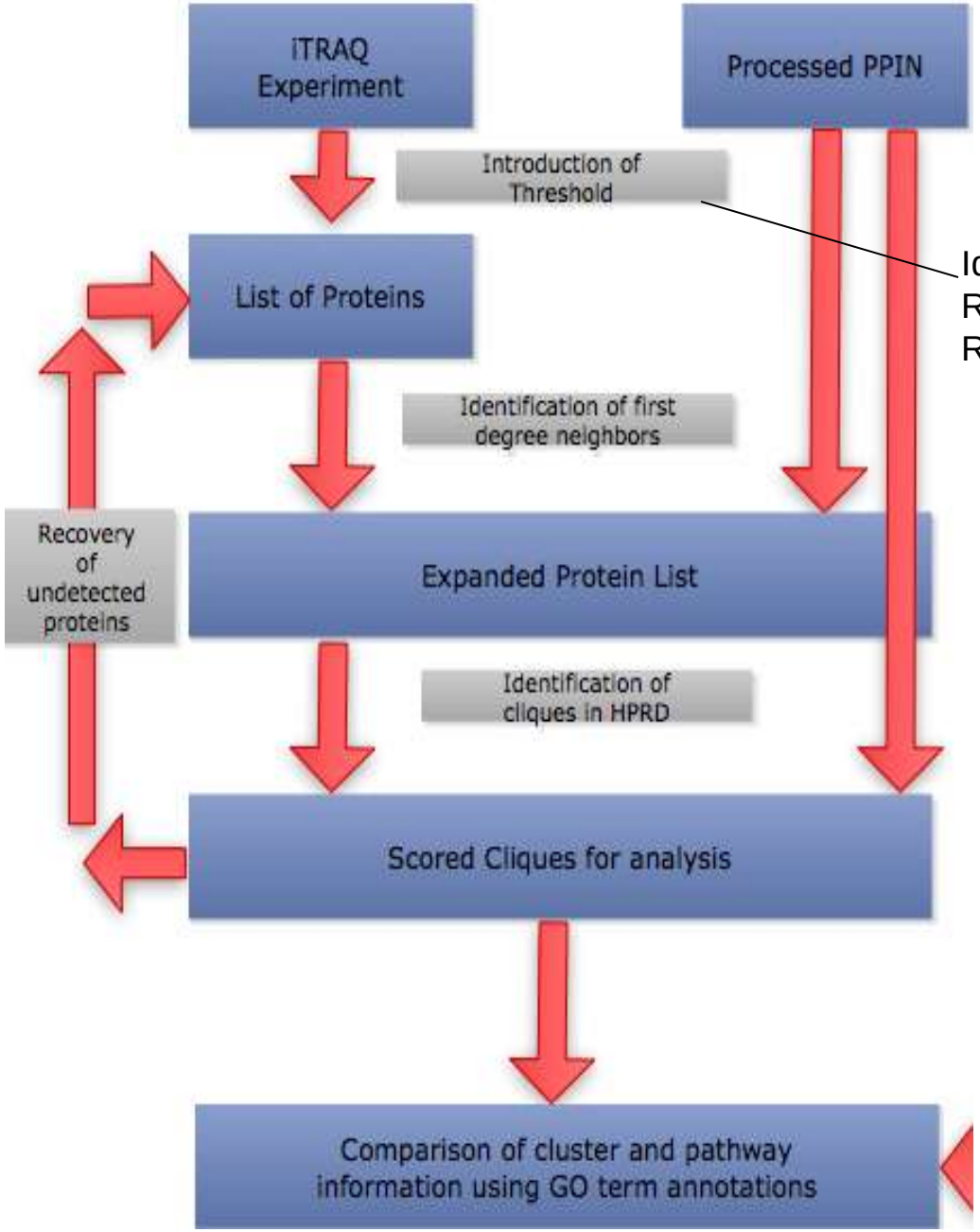
- **Reason: Proteins interacting with many seeds are likely to be part of the same complex as these seeds**

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

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

An Experiment: PEP

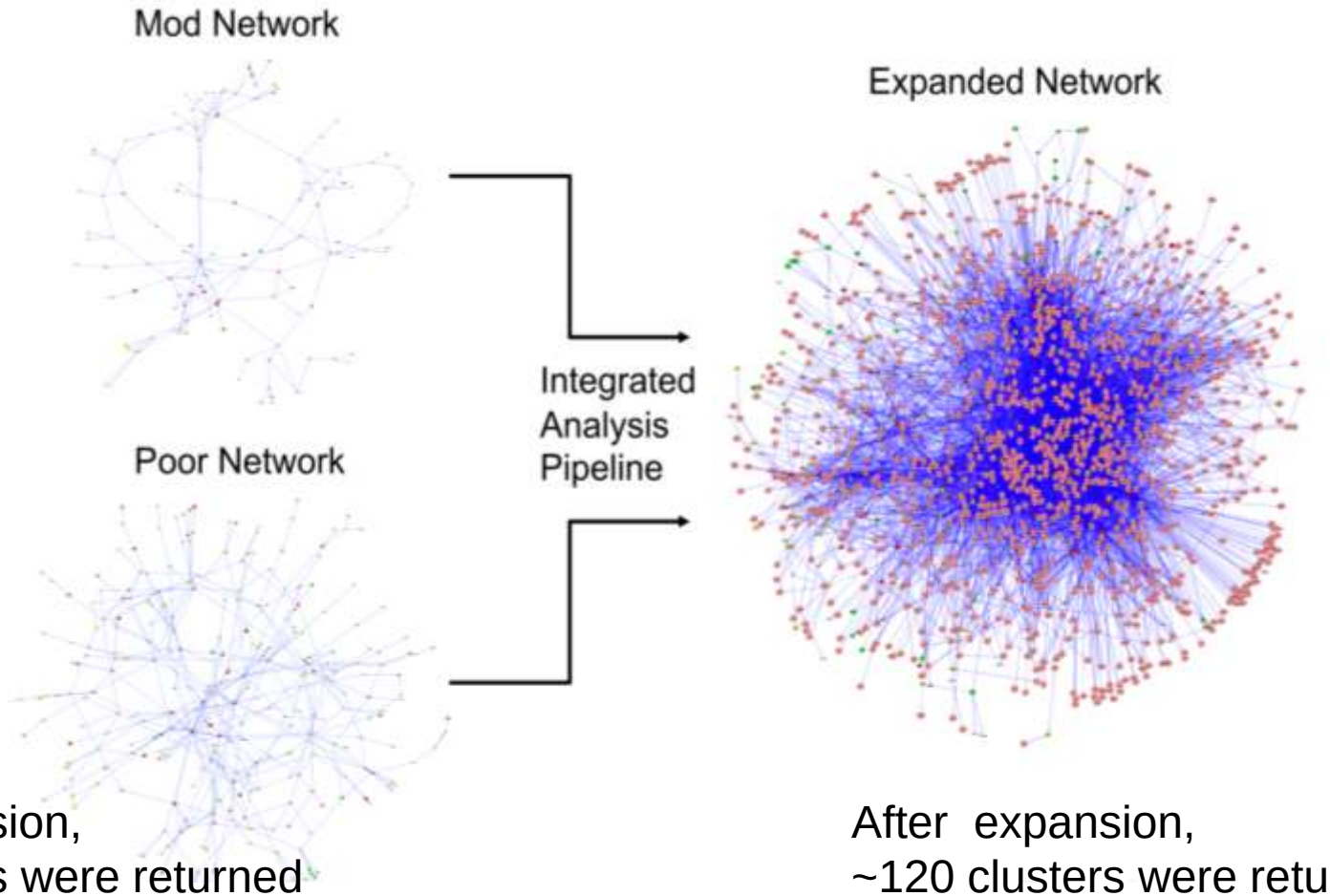
- **HCC (Hepatocellular carcinoma)**
  - Classified into 3 phases: differentiated, moderately differentiated and poorly differentiated
- **Mass Spectrometry**
  - iTRAQ (Isobaric Tag for Relative and Absolute Quantitation)
  - Coupled with 2D LC MS/MS
  - Popular because of ability to run 8 concurrent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



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

Expansion to include neighbors greatly improves coverage



W/o expansion,
4 k3 cliques were returned

After expansion,
~120 clusters were returned

"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

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**

Another Experiment: Comparison

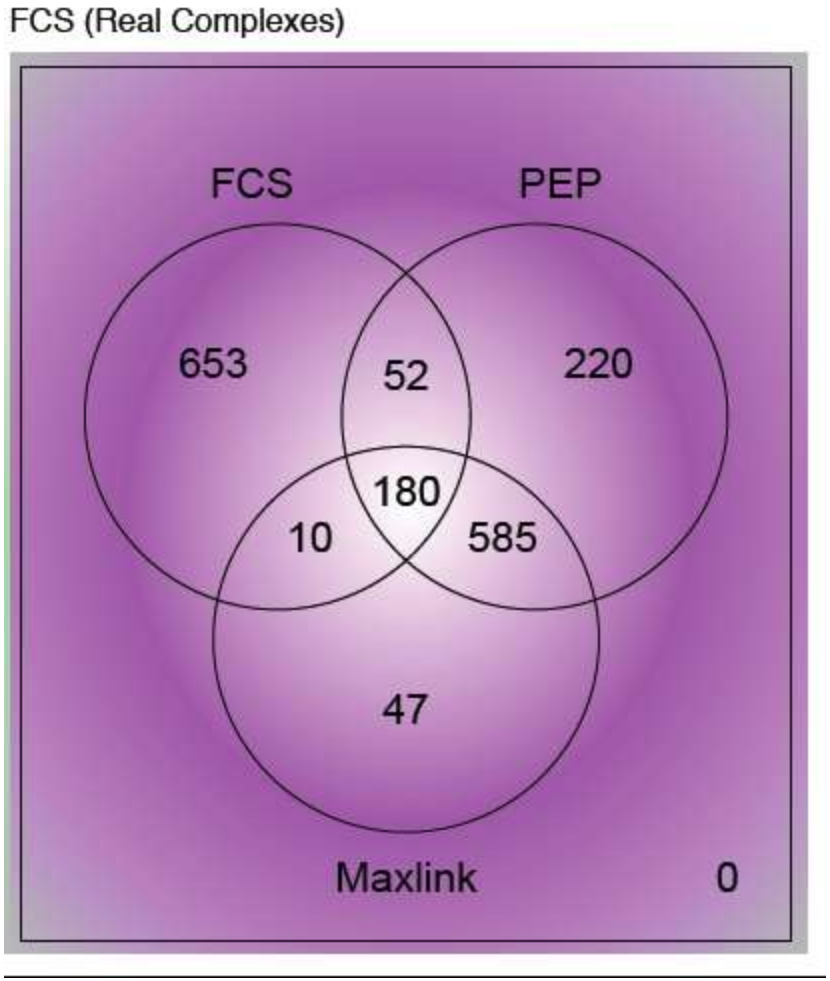
- **Valporic acid (VPA)-treated mice vs control**
  - VPA or vehicle injected every 12 hours into postnatal day-56 adult mice for 2 days
  - Role of VPA in epigenetic remodeling

- **MS was scanned against IPI rat db in round #1**
  - 396 proteins identified
- **MS was scanned against UniProtkb in round #2**
  - 393 additional proteins identified

- **All recovery methods ran on round #1 data and the recovered proteins checked against round #2**

NUS National University of Singapore

Moderate level of agreement of reported proteins between various recovery methods



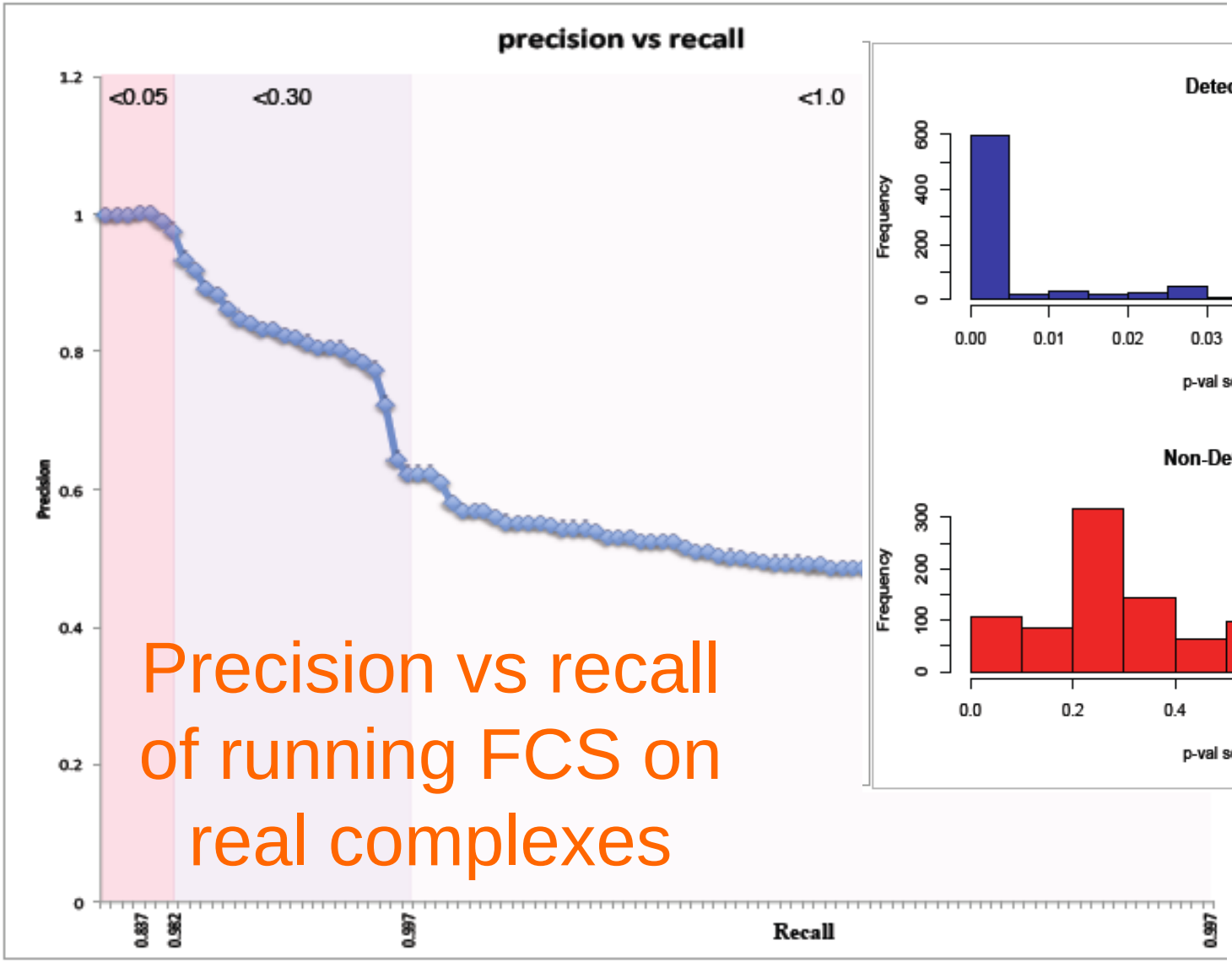
NUS National University of Singapore

Performance Comparison

Method	Novel Suggested Proteins	Recovered proteins
PEP	375	158
Maxlink	910	226
FCS (predicted)	678	224
FCS (complexes)	789	775

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

NUS National University of Singapore



Precision vs recall of running FCS on real complexes

Remarks

NUS National University of Singapore

What have we learned?

- **Contextualization (into complexes and pathways) can deal with consistency issues in proteomics**
- **GO term analysis also indicates that context-based methods select clusters that play integral roles in cancer**
- **Context-based methods reveal many potential clusters and are not constrained by any prior arbitrary filtering which is a common first step in conventional analytical approaches**

Acknowledgements & References

- This talk is based on joint work with

Wilson Goh

[PSP] Goh et al**. Proteomics signature profiling (PSP): A novel contextualization approach for cancer proteomics**. *Journal of Proteome Research,*11(3):1571--1581, 2012.

[PEP] 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

[MaxLink] 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.

Goh et al**. Enhancing utility of proteomics signature profiling (PSP) with pathway derived subnets (PDSs), performance analysis and specialized ontologies**. *BMC Genomcs,* 14:35, 2013