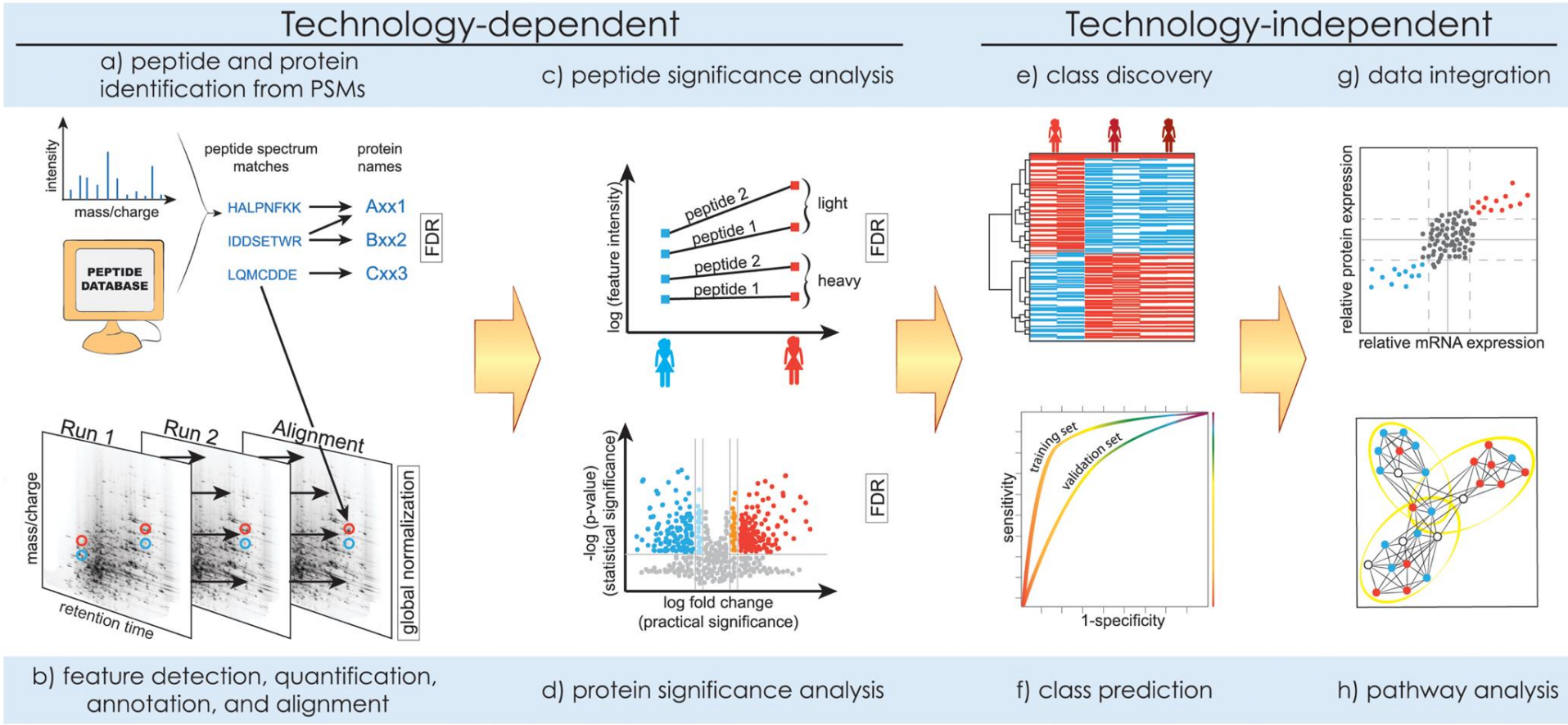
Network-based analysis of proteomic profiles

Limsoon Wong

Proteomics is a system-wide characterization of all proteins



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

Proteomics vs transcriptomics

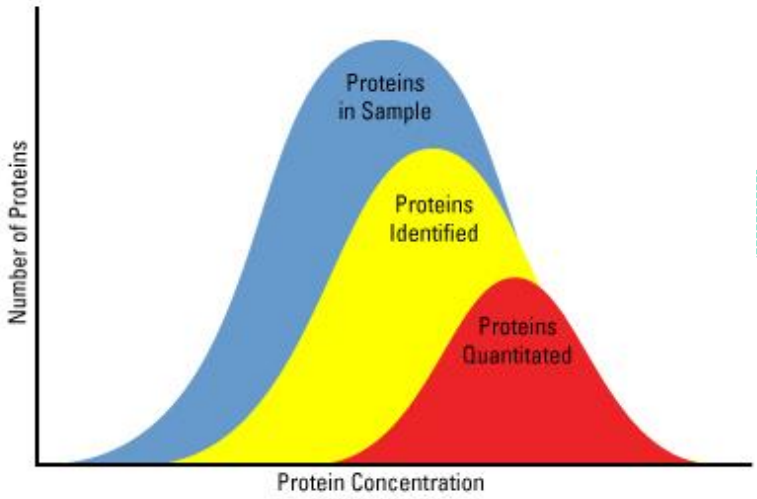
- **Proteomic profile**
  - Which protein is found in the sample
  - How abundant it is
- **Similar to gene expression profile. So typical gene expression profile analysis methods can be (and had been) applied**

- **Key differences**
  - Profiling
    - **Complexity: 20k genes vs 500k proteins**
    - **Dynamic range: > 10 orders of magnitude in plasma. Proteins cannot be amplified**
  - Analysis
    - **Much fewer features**
    - **Difficult to reproduce**
    - **Much fewer samples**
    - **Unstable quantitation**

Issues in proteomics: Coverage and consistency

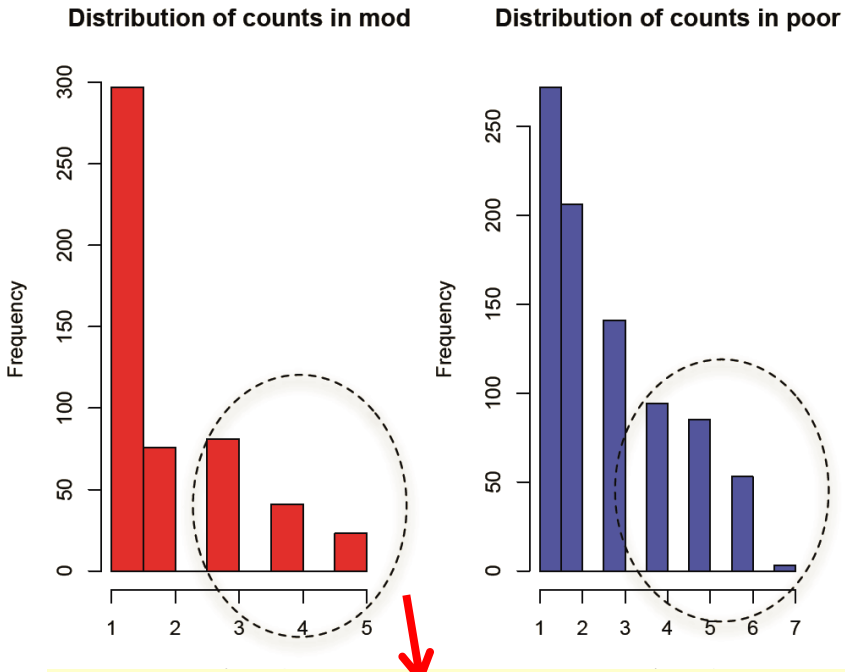
NUS National University of Singapore

Technical incompleteness



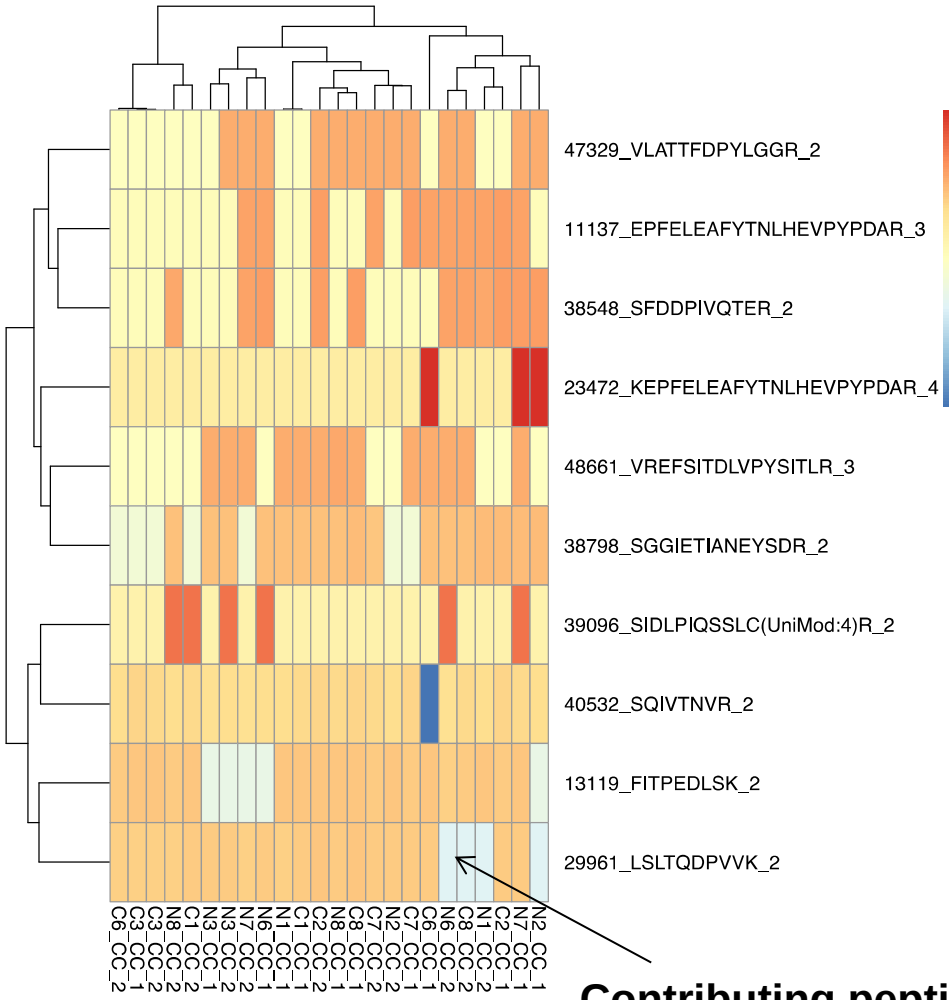
Per sample

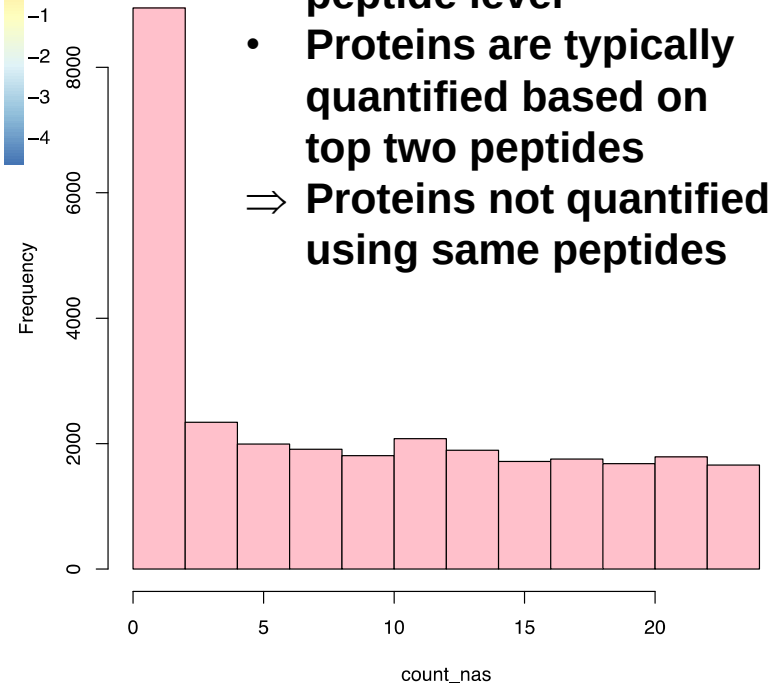
How it affects real data



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

Issues in proteomics: Quantitation noise





- **Many data holes at the peptide level**
- **Proteins are typically quantified based on top two peptides**

⇒ **Proteins not quantified using same peptides**

Contributing peptides do not agree with each other

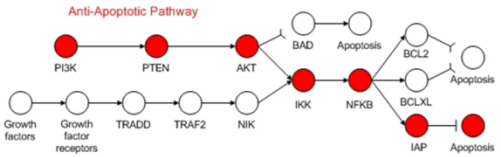
Improving Consistency in Proteomic Profile Analysis

An inspiration from gene expression profile analysis

NUS National University of Singapore

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

Co… t 2011 © Limsoon W

12

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) = $1000 * (1/2^6)^5 = 9.3*10^{-7}$

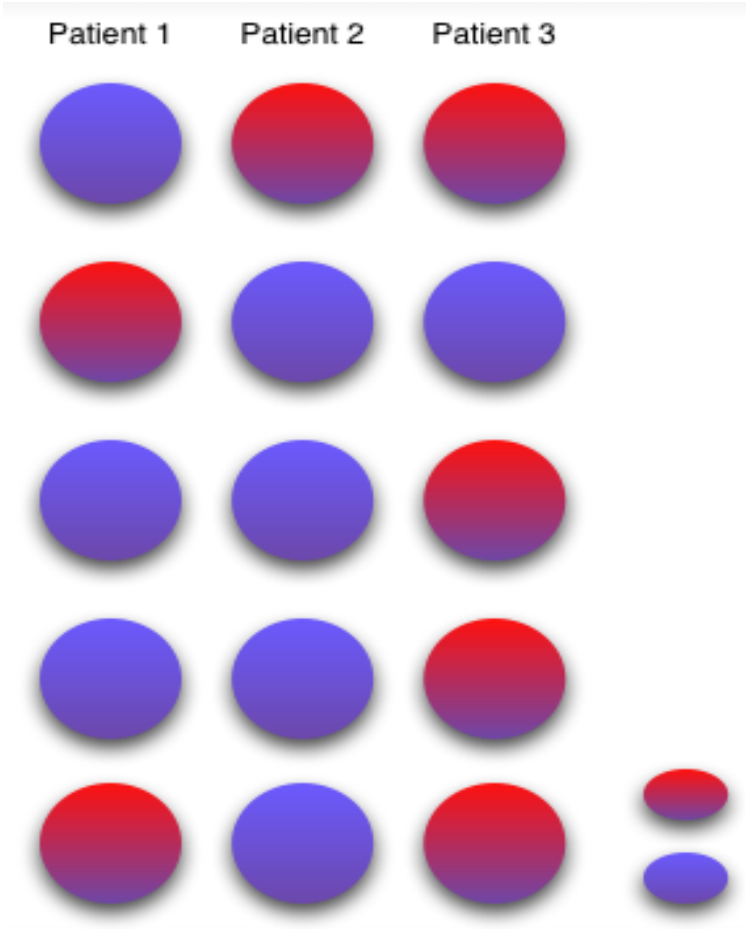
⇒ Even more false positives?
- Perhaps no need to consider every group

Microarray Workshop for Gene Expression Profiling, NUH, 23/9/2011 Copyright 2011 © Limsoon Wong

Copyright 2011 © Limsoon Wong

Contextualization!

Intuition



- **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?

… and the Math

- **Postulate: The chance of a protein complex being present is proportional to the fraction of its constituent proteins being reported in the screen**

- **Suppose proteomics screen has 75% reliability; a complex comprises proteins A, B, C, D, E; and screen reports A, B, C, D only**

⇒ **Complex has 60% (=4/5 * (1-0.25)) chance to be present**

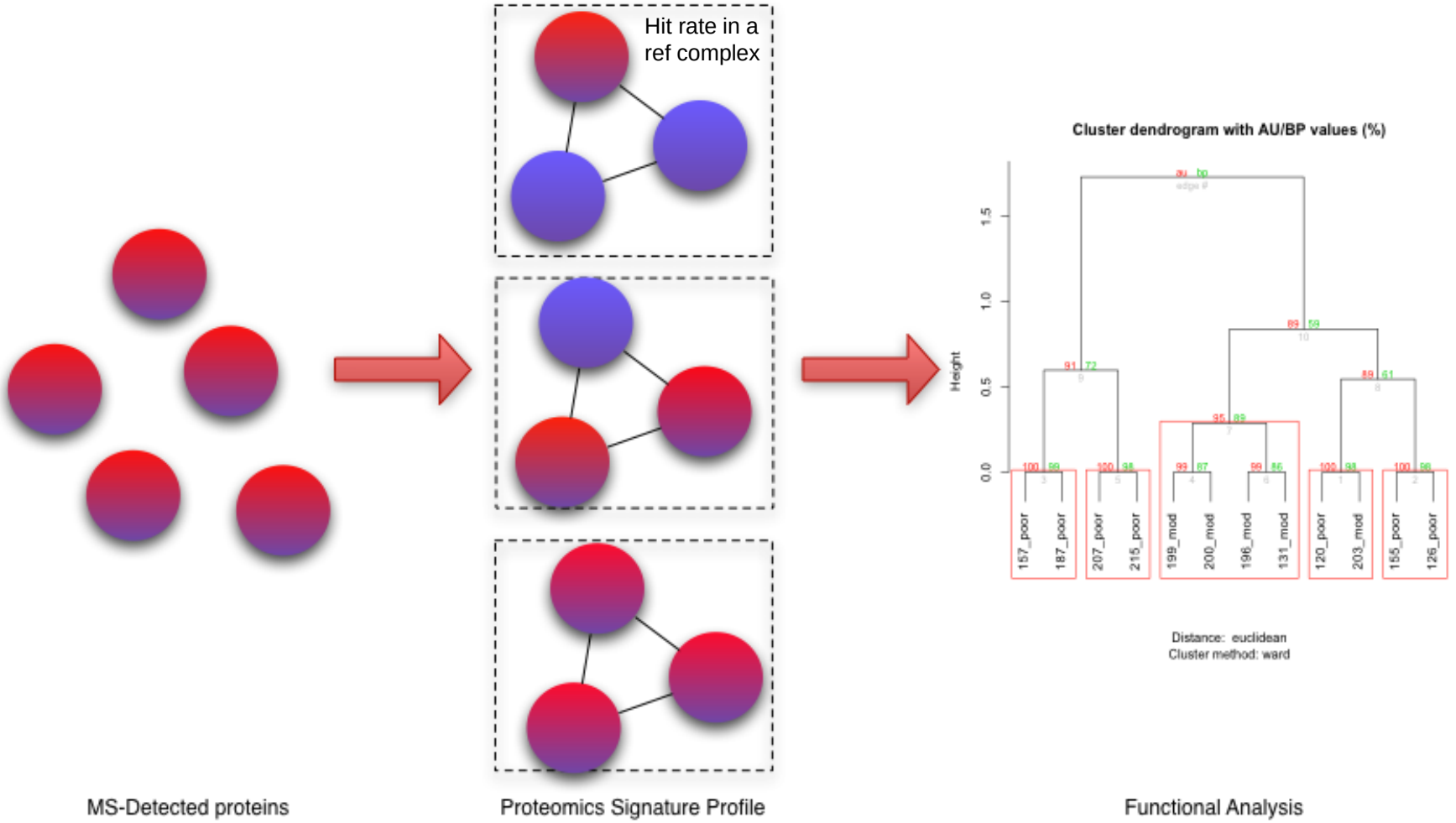
⇒ **The unreported protein E also has 60% chance to be present, as presence of the complex implies presence of all its constituents … improving coverage**

⇒ **Chance of all four reported proteins being true positive is also 60%, rather than 32% (= $(1-0.25)^4$)**

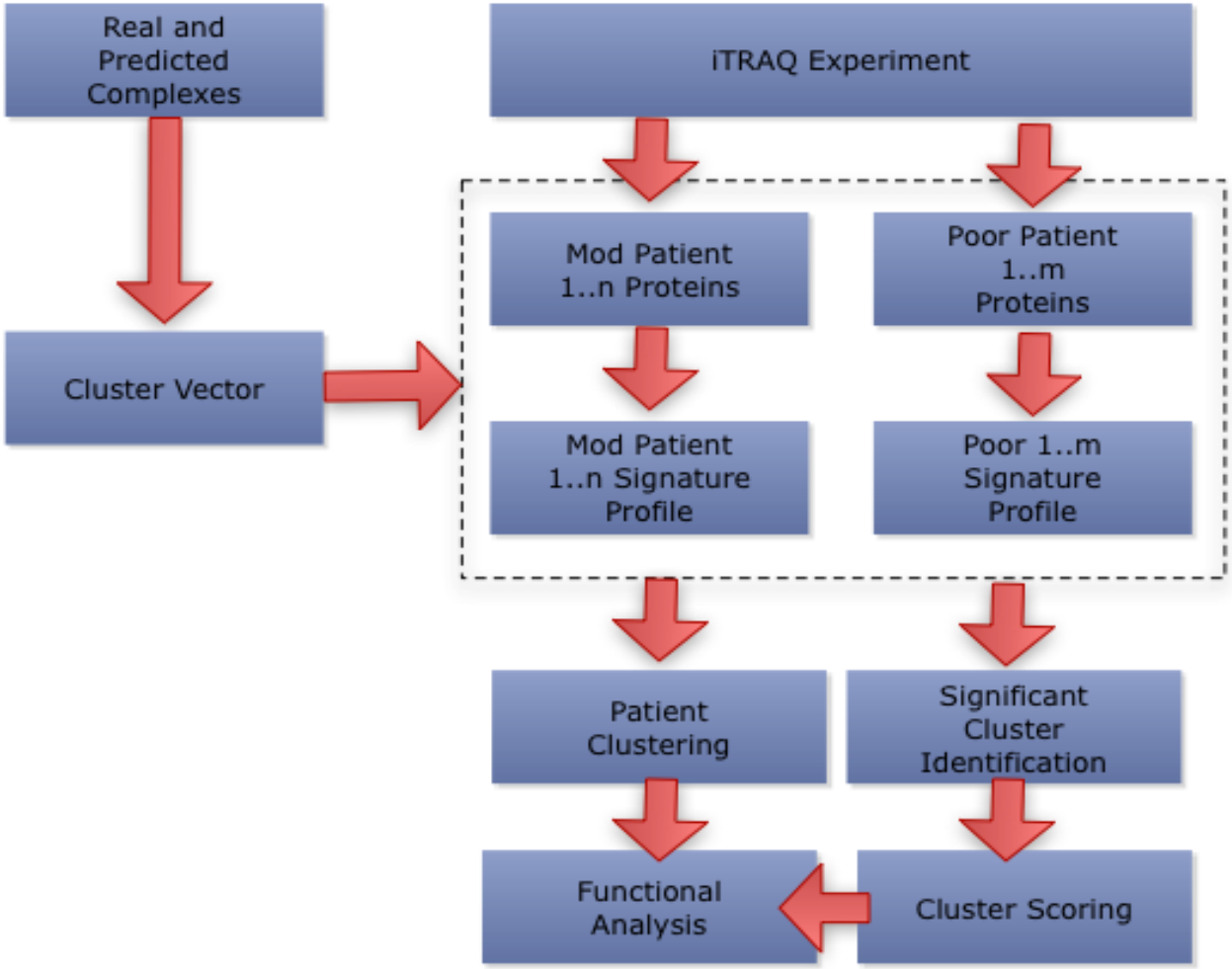
⇒ **Each of A, B, C, and D individually has 88% (= $\sqrt[4]{0.6}$) chance of being true positive, whereas a reported protein that is isolated has a lower 75% (= 1 – 0.25) chance of being true positive … removing noise**

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

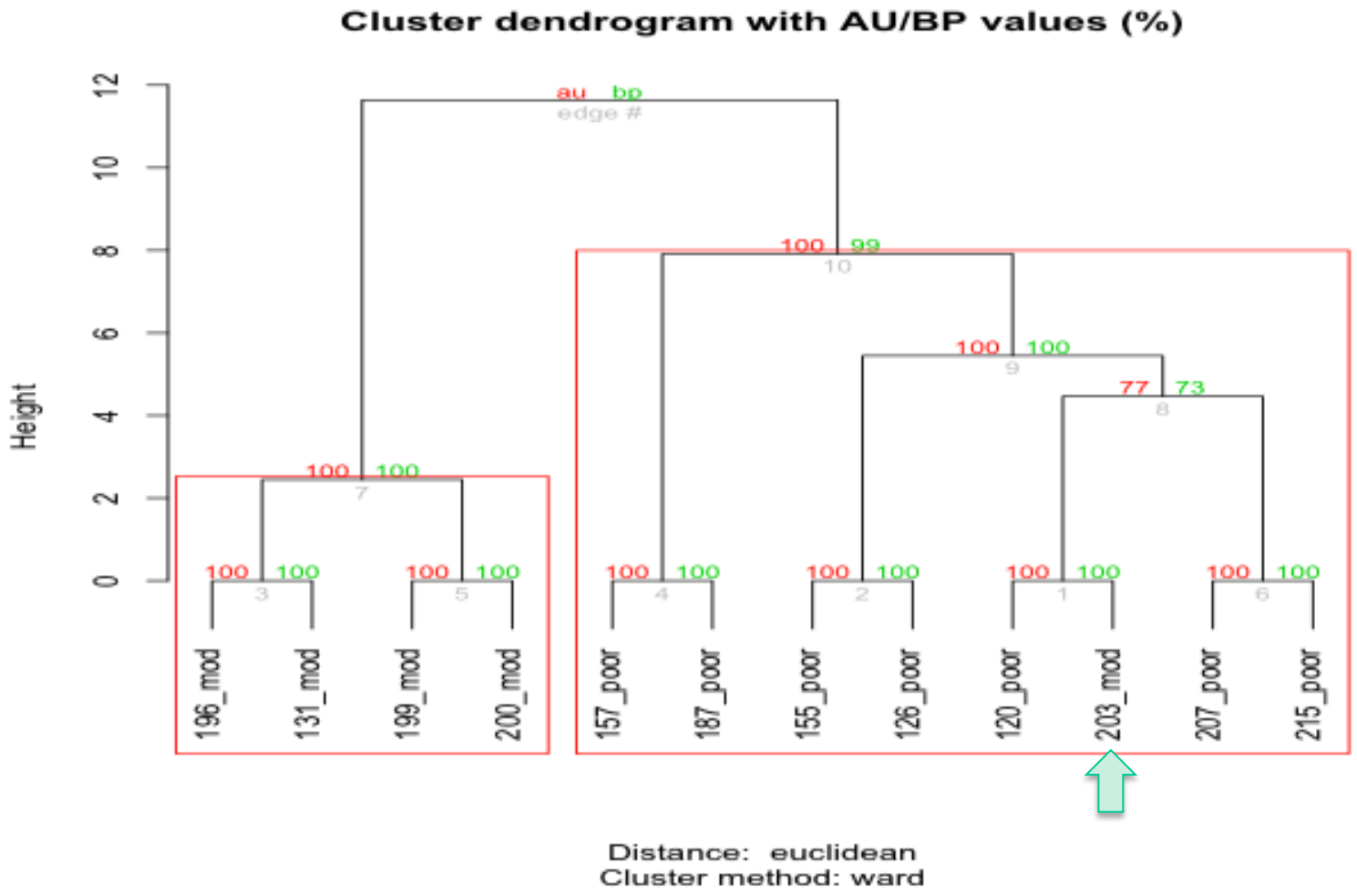
"Threshold-free" principle of PSP



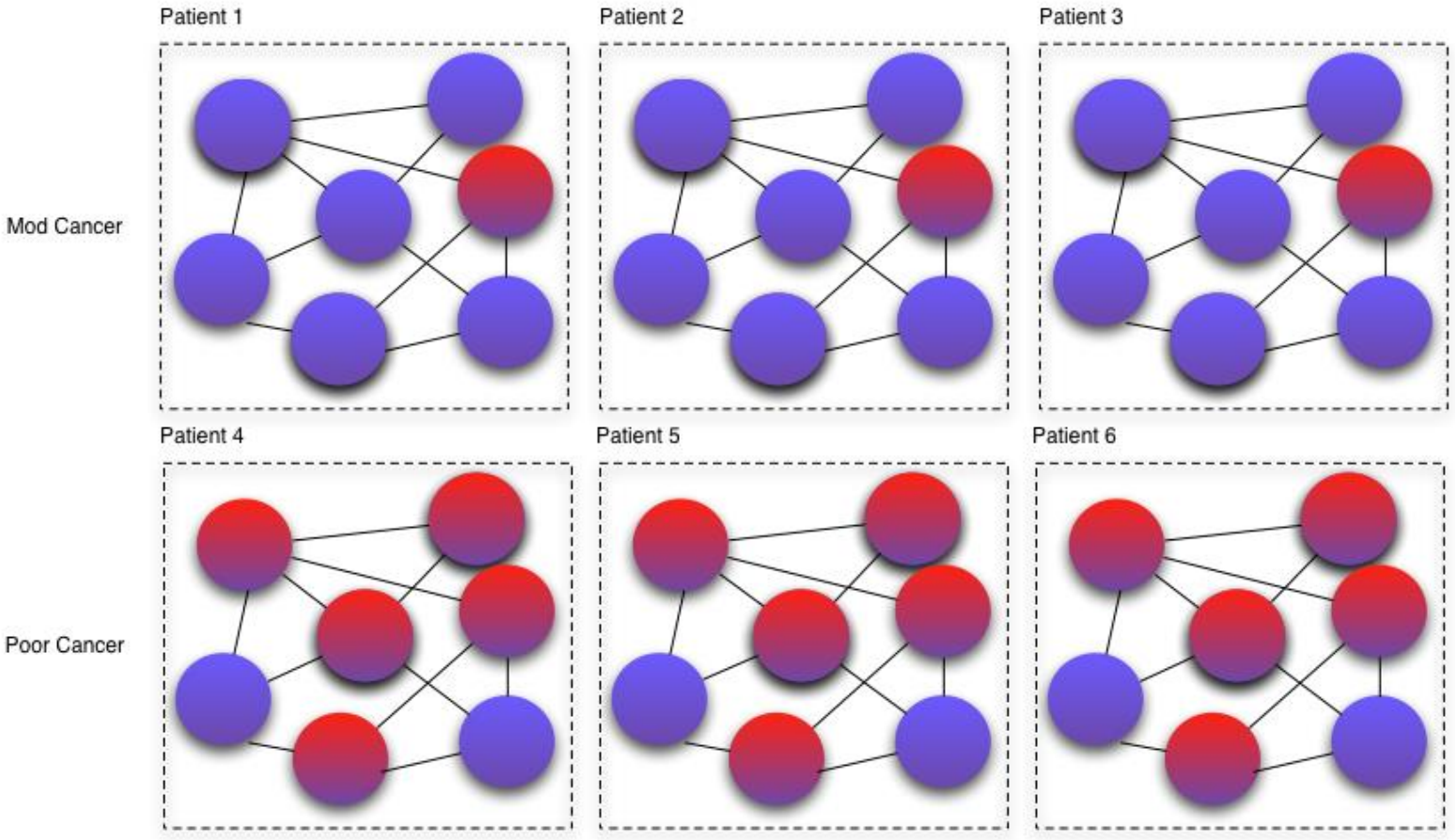
Applying PSP to a HCC dataset



Consistency: Samples segregate by their classes with high confidence



Feature selection



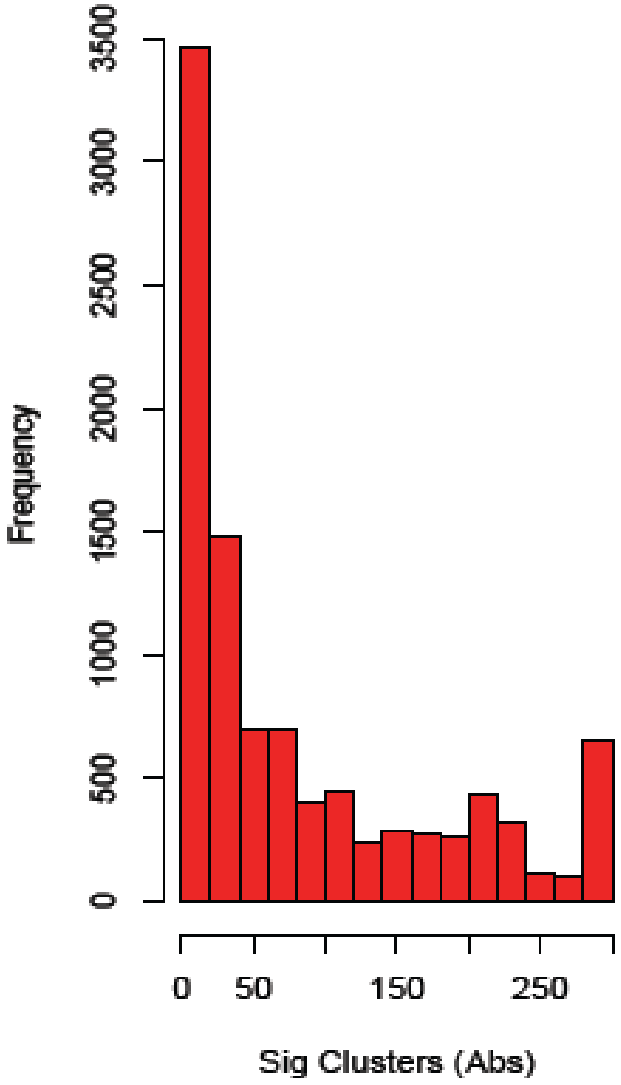
$$t\_score = \frac{\bar{HA} - \bar{HB}}{S_{HA,HB}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

where

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

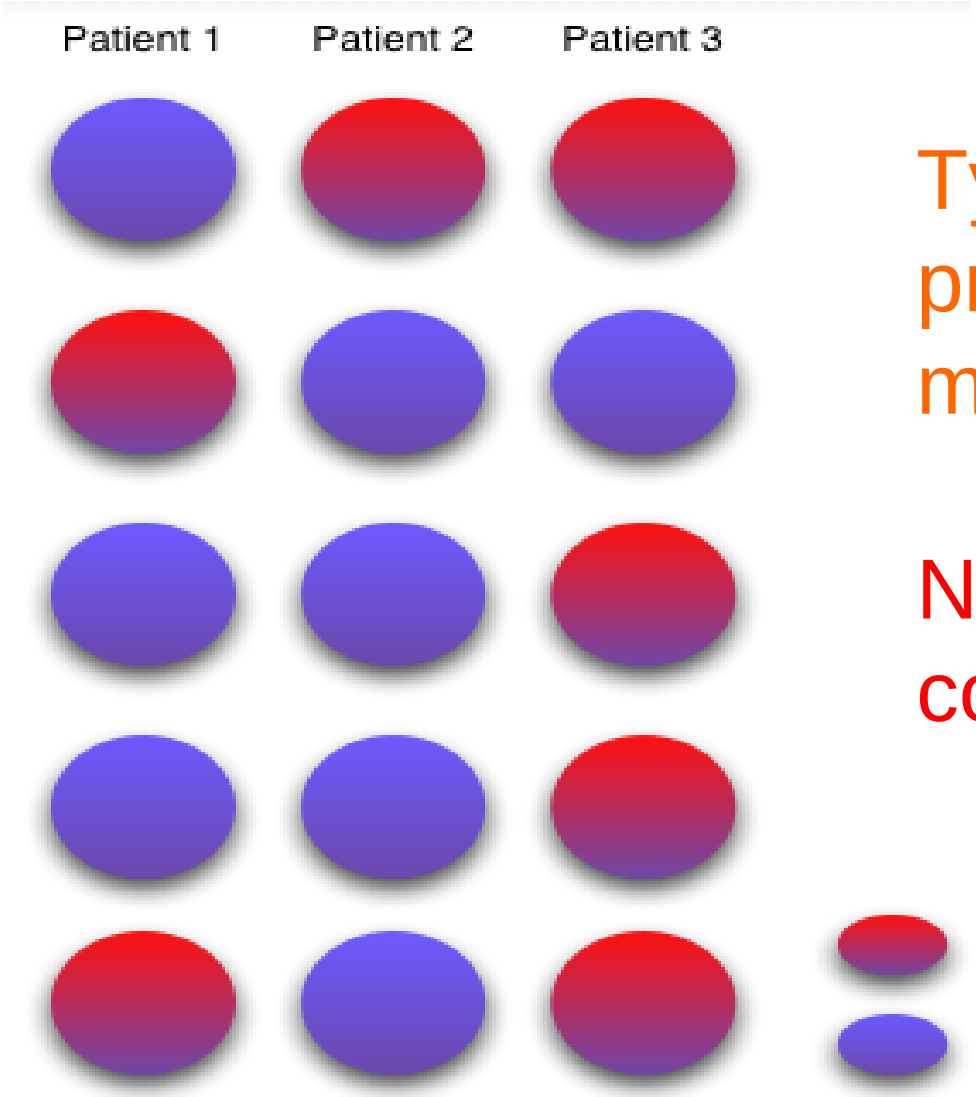
Goh et al. **Enhancing utility of proteomics signature profiling (PSP) with pathway derived subnets (PDSs), performance analysis and specialized ontologies**. *BMC Genomcs,* 14:35, 2013

False-positive rate analysis



- **Divide subjects of one same phenotype into 2 groups**
  - ⇒Significant complexes produced by PSP here are false positives
- **Repeat many times to get "null 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



Typical proteomic profiling misses many proteins

Need to improve coverage!

Goh et al. **Comparative network-based recovery analysis and proteomic profiling of neurological changes in valporic acid-treated mice**. *JPR*, 12(5):2116--2127, 2013

FCS

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

- **Why?**

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

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

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

CEA

- **Generate cliques from PPIN**
- **Rescue undetected proteins from cliques 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 powerful protein 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, 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. *International Journal of Bioinformatics Research and Applications*, 8(3/4):155--170, 2012

MaxLink

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

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

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

Goh et al. **Comparative network-based recovery analysis and proteomic profiling of neurological changes in valporic acid-treated mice**. *JPR*, 12(5):2116--2127, 2013

NUS National University of Singapore

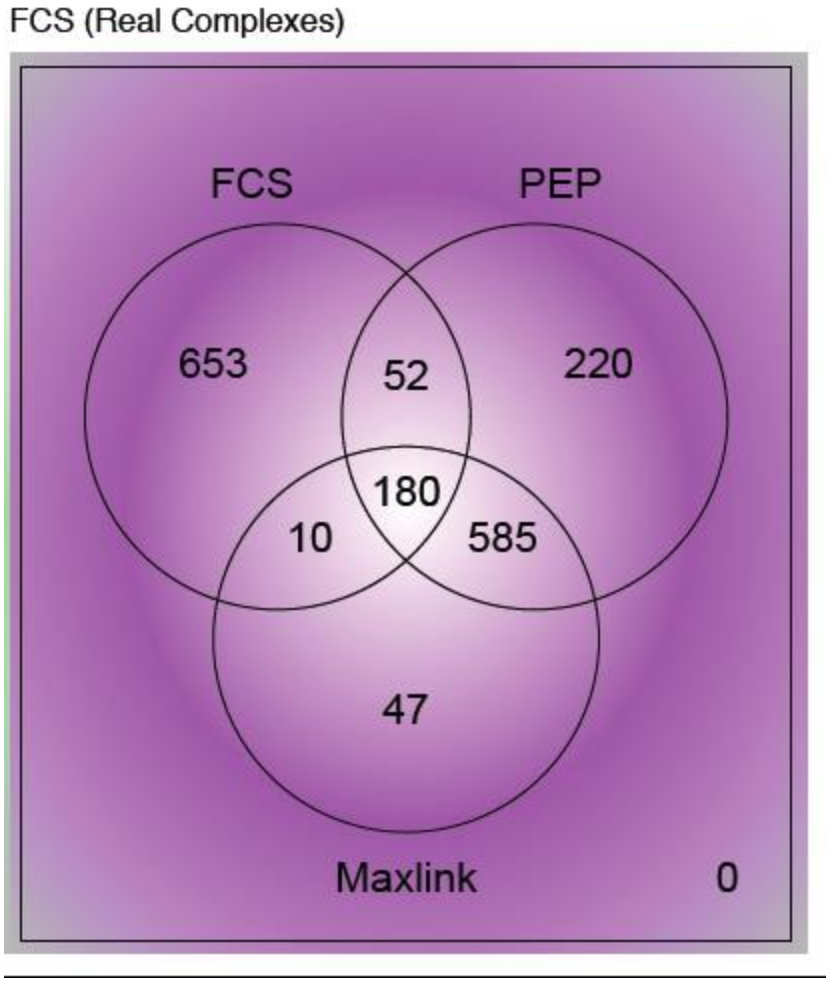
Experiment

- **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



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**

Further Refinement: qPSP

SWATH

- **Traditional iTRAQ-type MS data is sparse**

$\Rightarrow$ **PSP ignores abundance level**

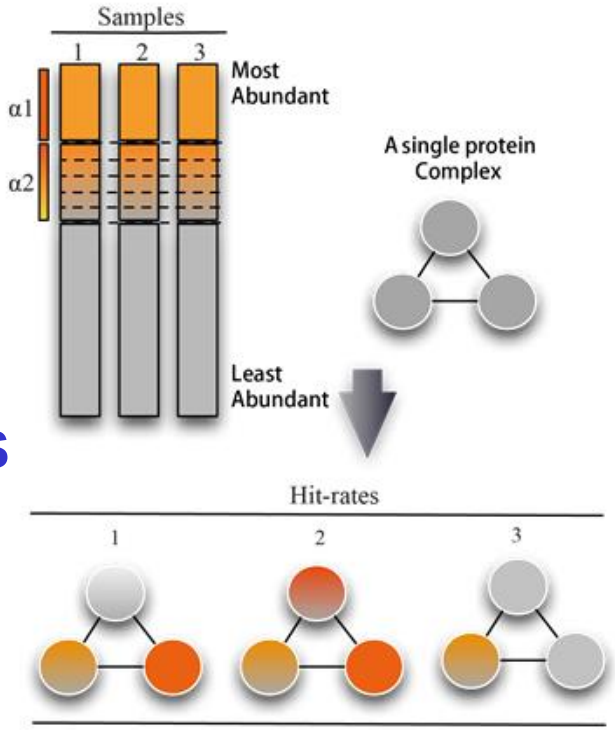
- **Modern SWATH-type MS is much denser**

$\Rightarrow$ **Can we refine PSP to take abundance level into consideration?**

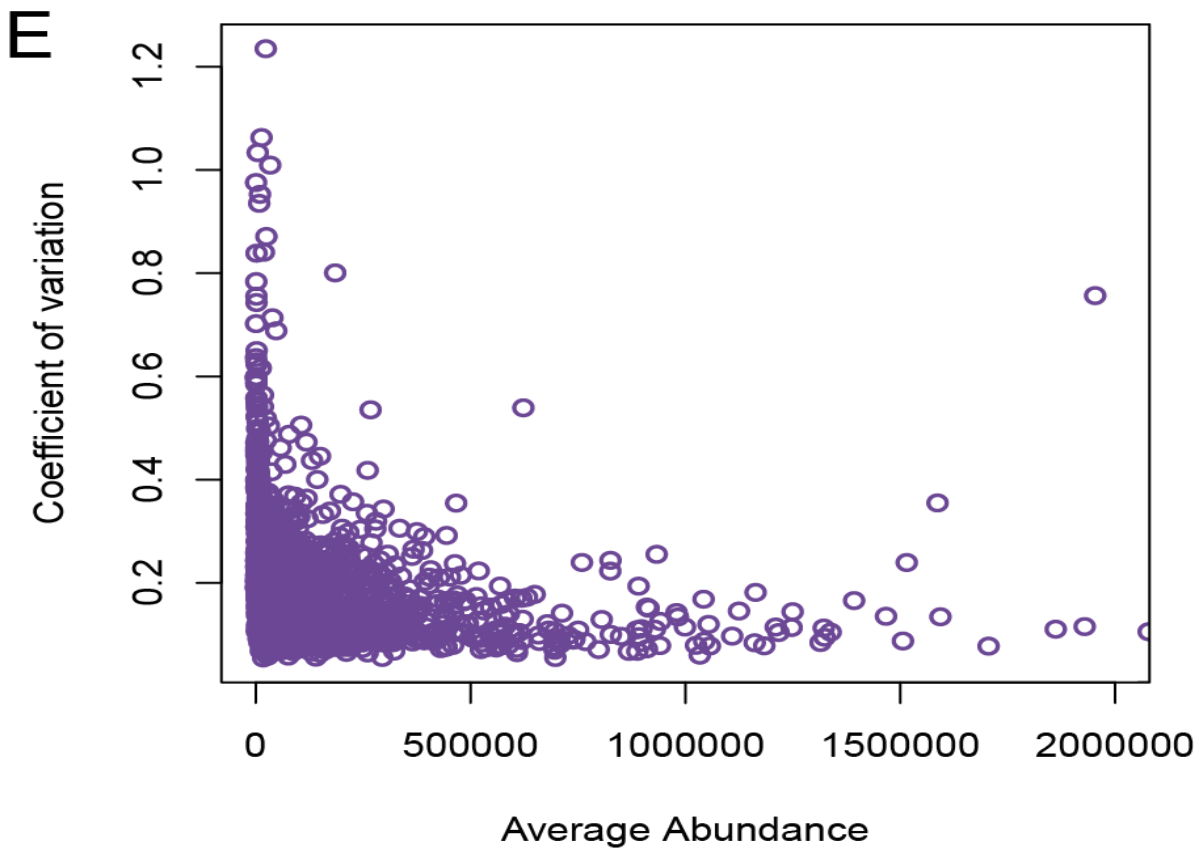
Goh et al. **Quantitative proteomics signature profiling based on network contextualization**. *Biology Direct*, accepted

qPSP

- **In a sample, assign weight to proteins as follow**
  - Most-abundant 10% of proteins, wt = 1
  - Proteins at 10-12.5%, wt = 0.8
  - Proteins at 12.5-15%, wt = 0.6
  - Proteins at 15-17.5%, wt = 0.4
  - Proteins at 17.5-20%, wt = 0.2
  - All other proteins, wt = 0
- **Hit rate of a complex C wrt a sample S is sum of the wt of proteins in C in S**
- **All other steps, same as PSP**

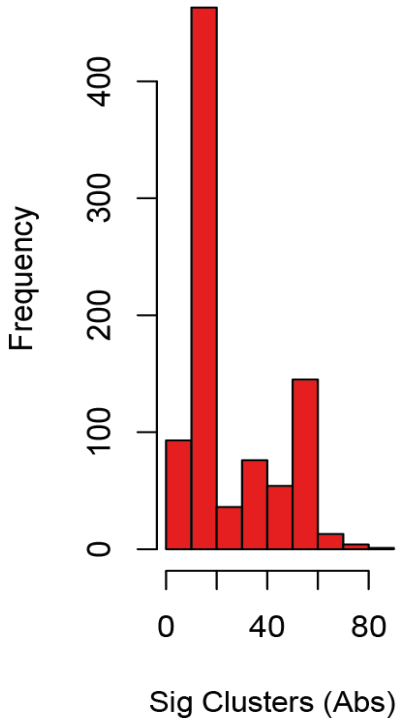


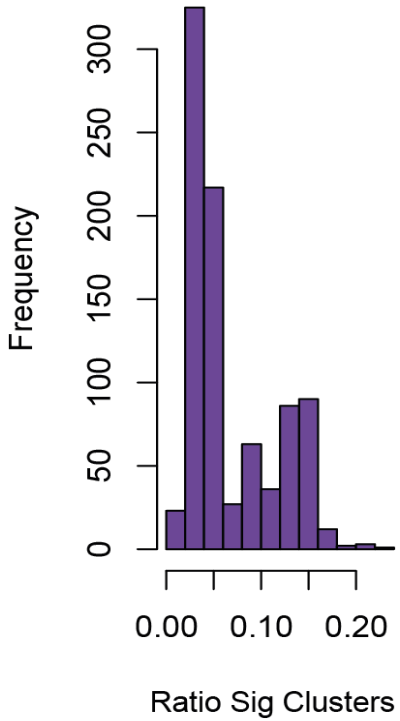
Why qPSP is based on the most abundant (top 10-20%) proteins



False-positive rate analysis

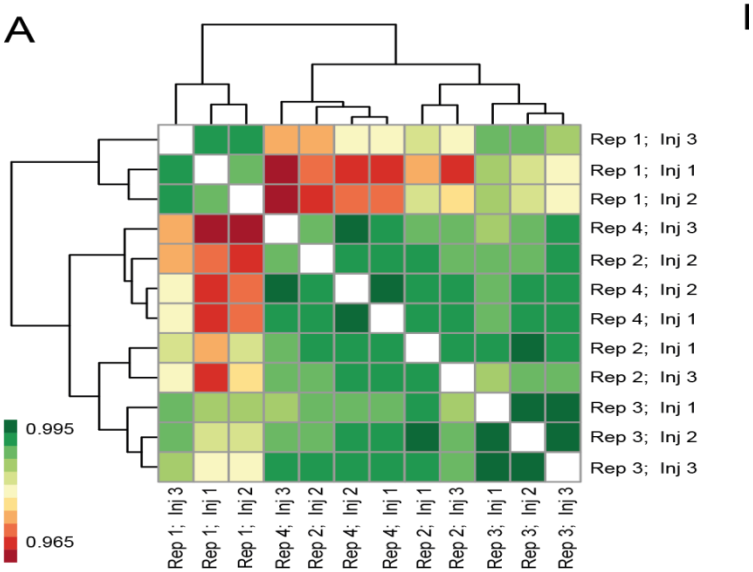
- **12 kidney controls randomly assigned into two groups of equal size, and qPSP analysis performed many rounds**
- **# of significant clusters (5% FDR) determined each round**
- **False-positive rate well within the expectation levels**
  - Sig Clusters (Abs)
    - **Expect: 19, Observed: 16**
  - Sig Clusters (Ratio)
    - **Expect: 0.05, Observed: 0.04**



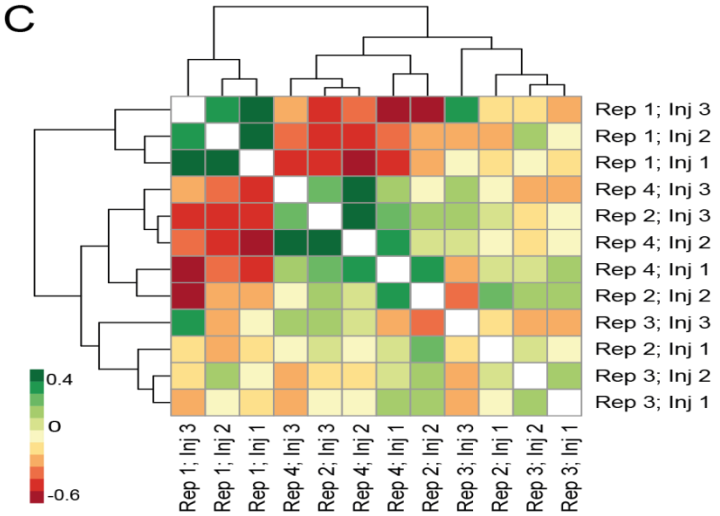


Stability of qPSP

- **Similarity of qPSP's of control samples is >90%**

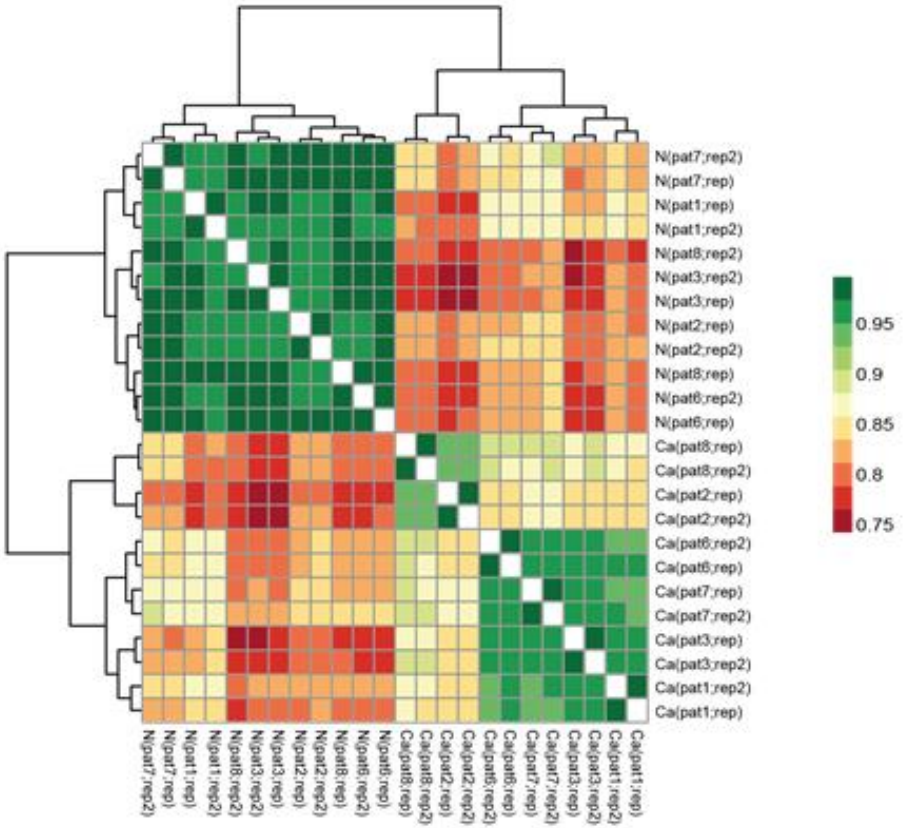


- **Similarity of proteomic profiles of control samples is <40%**



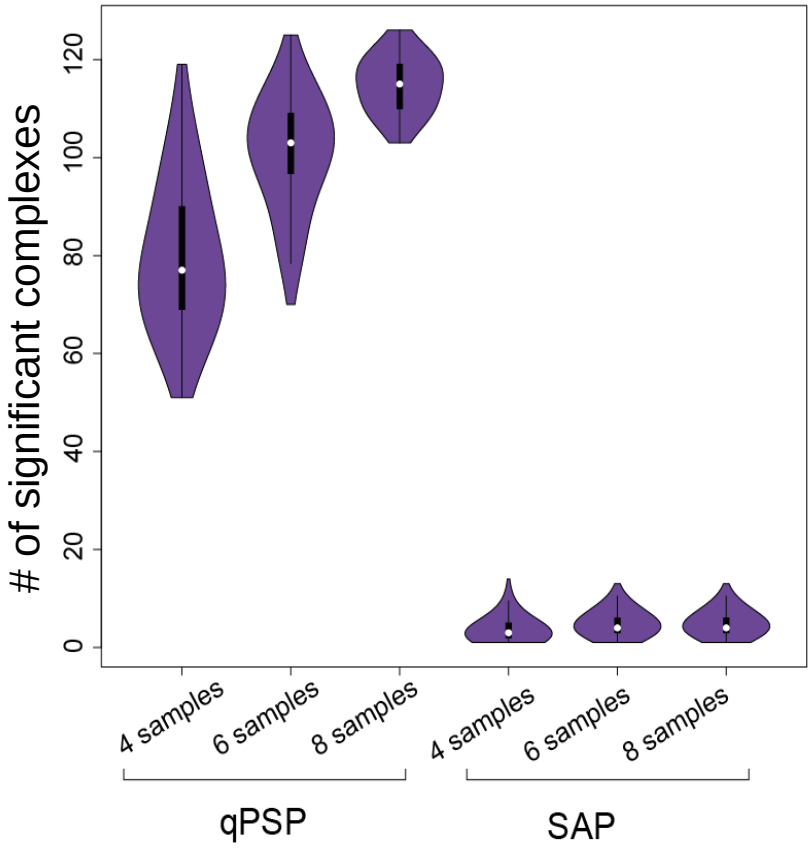
Test on 12 paired tumour / control tissues of a cancer

- **Clustering shows specific & consistent segregation of non-cancer and cancer samples**
- **qPSP ws able to detect sub-clusters within the cancer**
  - The smaller sub-cluster verified to belong to the two patients who died

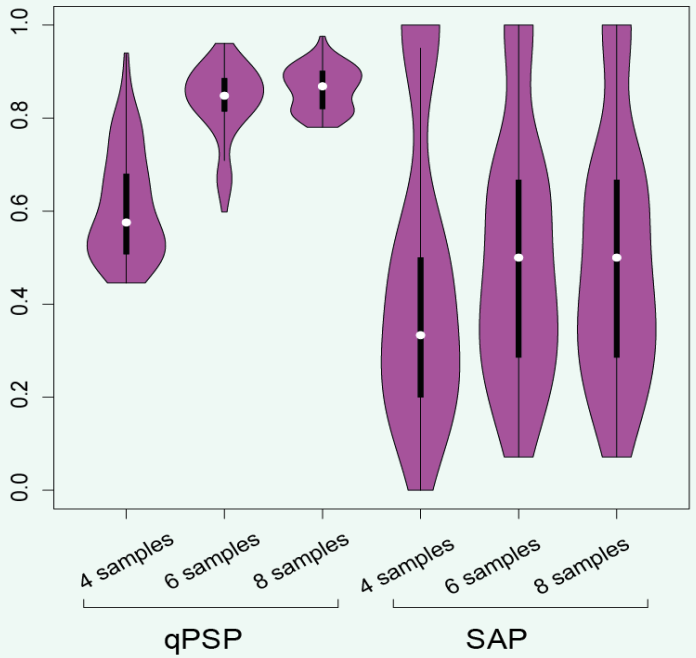


qPSP predicts more biological complexes than SAP

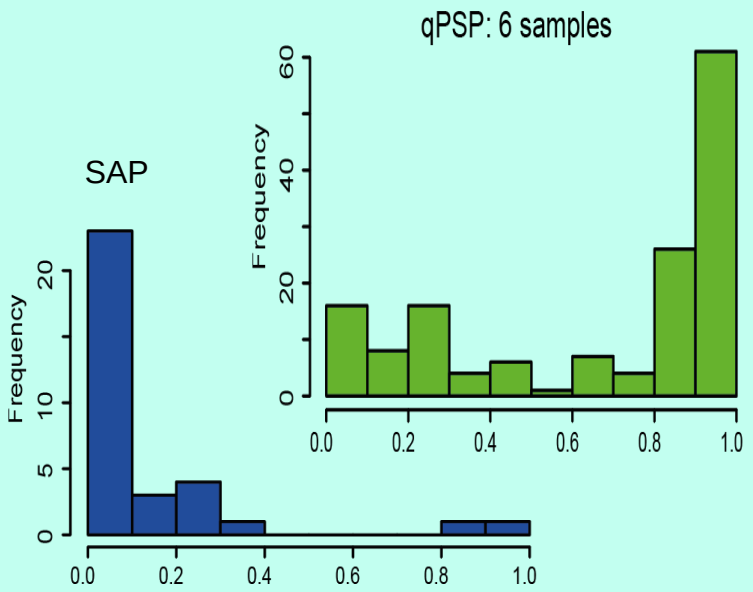
- **A standard analytical procedure (SAP)**
  - t-test for differential-protein selection
  - Complex-enrichment analysis by hypergeometric test



qPSP is more stable than SAP

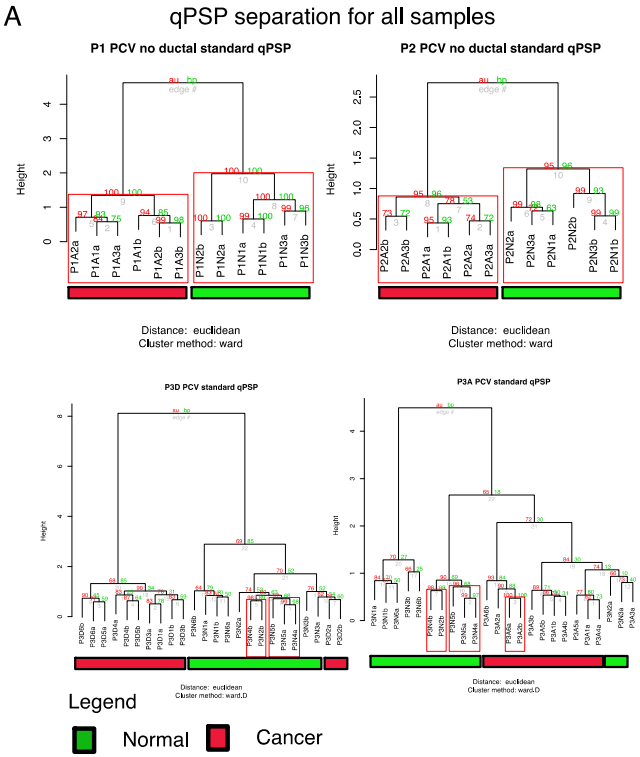


Pair-wise analysis of simulations to calculate agreement levels (using Jaccard Score, 0 for complete disagreement, 1 for complete agreement) across complexes showed that qPSP is more consistent than SAP

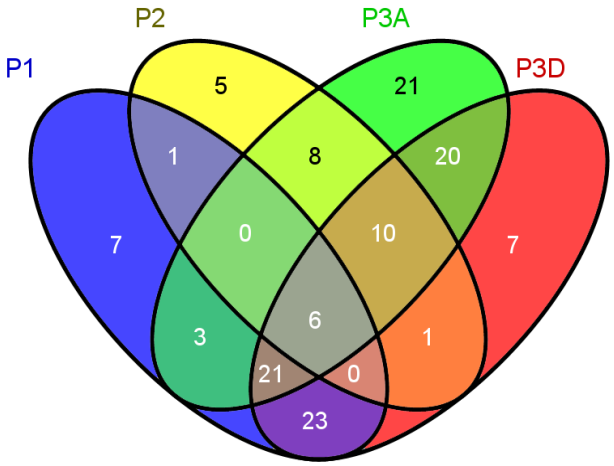


Distribution of significant complex agreements (On the x-axis, a score of 1 means complete agreement across all simulations, the y-axis is a frequency measurement, and its sum adds up to all complexes observed to be significant at least once)

A



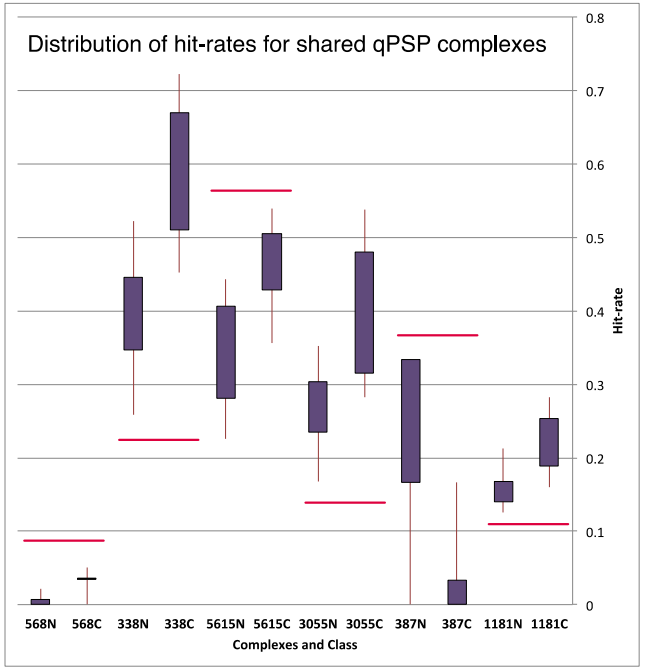
B



C

Complex ID	Complex Name	Function	Complex size
338	40S ribosomal subunit, cytoplasmic	Protein biosynthesis	31
387	MCM complex	DNA replication	6
568	Nuclear pore complex	Protein transport	28
1181	C complex spliceosome	RNA splicing	80
3055	Nop56p-associated pre-rRNA complex	ribosome biogenesis	104
5615	Emerin complex 52	intracellular signaling cascade	23

D



qPSP overcomes incongruity issues

Test	T-test (p<0.05)		T-test (p<0.01)		Significant intersection	
Training'>Validation	Proteins	qPSP	Proteins	qPSP	Proteins	qPSP
P1|P2'>P3	0.56	0.83	0.61	0.88	0.83	0.94
P1|P3'>P2	0.66	0.5	0.33	0.5	0.5	0.83
P2|P3'>P1	0.58	1	0.83	0.92	0.75	1

qPSP overcomes incongruity issues

ESSNet: A Quantum Leap?

ESSNet, adapted for SWATH

- Let g_i be a protein in a given protein complex
- Let p_j be a patient
- Let q_k be a normal
- Let $\Delta_{i,j,k} = Expr(g_i,p_j) – Expr(g_i,q_k)$
- Test whether $\Delta_{i,j,k}$ is a distribution with mean 0

- Null hypothesis is "Complex C is irrelevant to the difference between patients and normals, and the proteins in C behave similarly in patients and normals"
- No need to restrict to most abundant proteins

$\Rightarrow$ Potential to reliably detect low-abundance but differential proteins

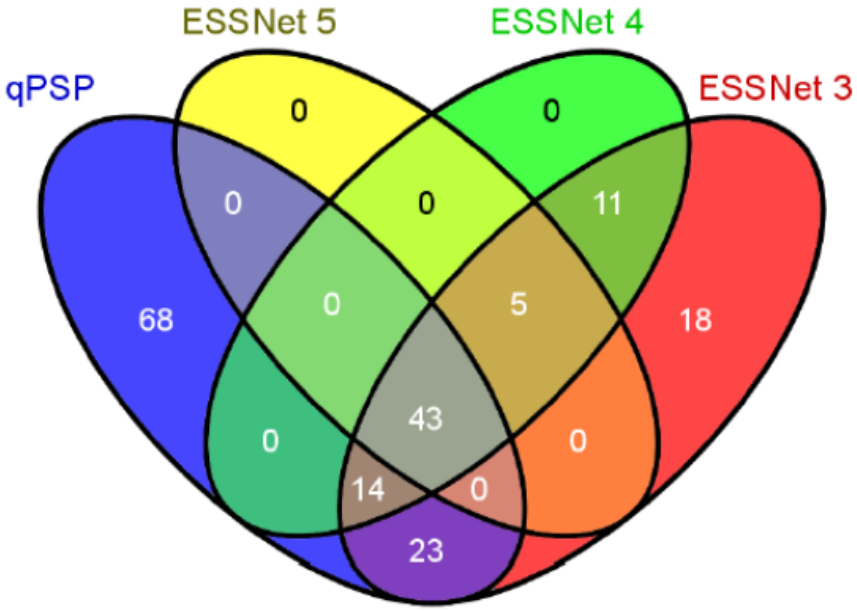
Lim et al. **A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small**. *JBCB*, 13(4):1550018, 2015

Significant complexes reported by qPSP and ESSNet for a cancer dataset

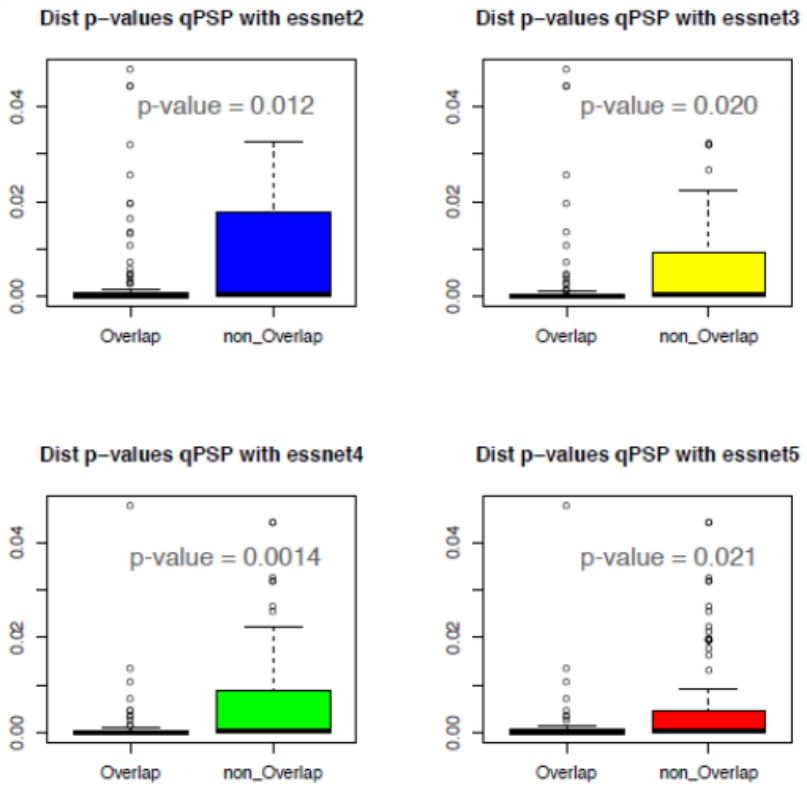
NUS National University of Singapore

Those qPSP complexes that are also ESSNet complexes have lower p-value than those that aren't

A

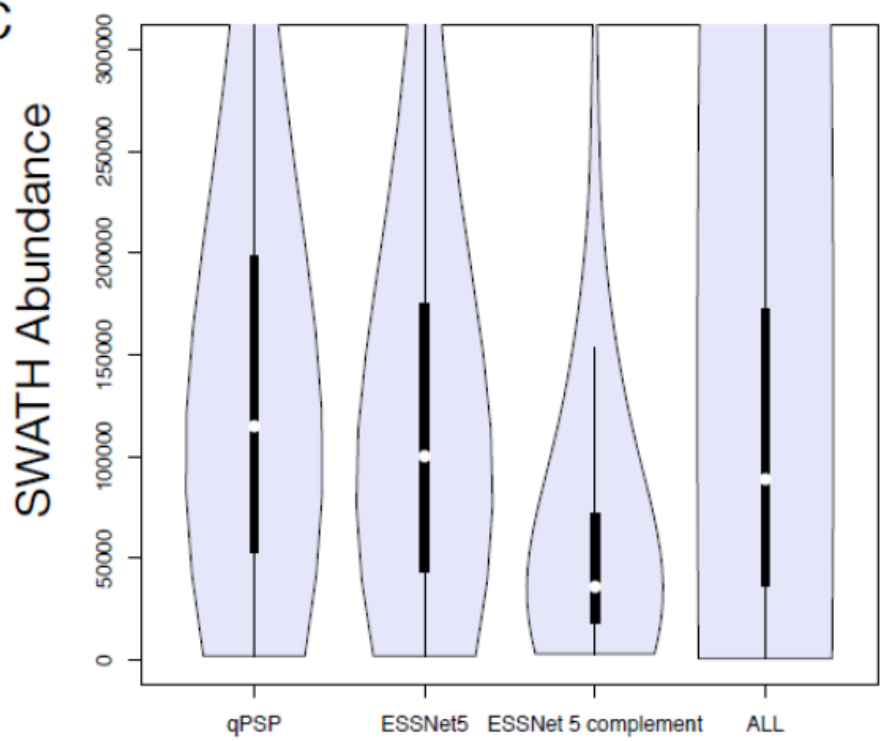


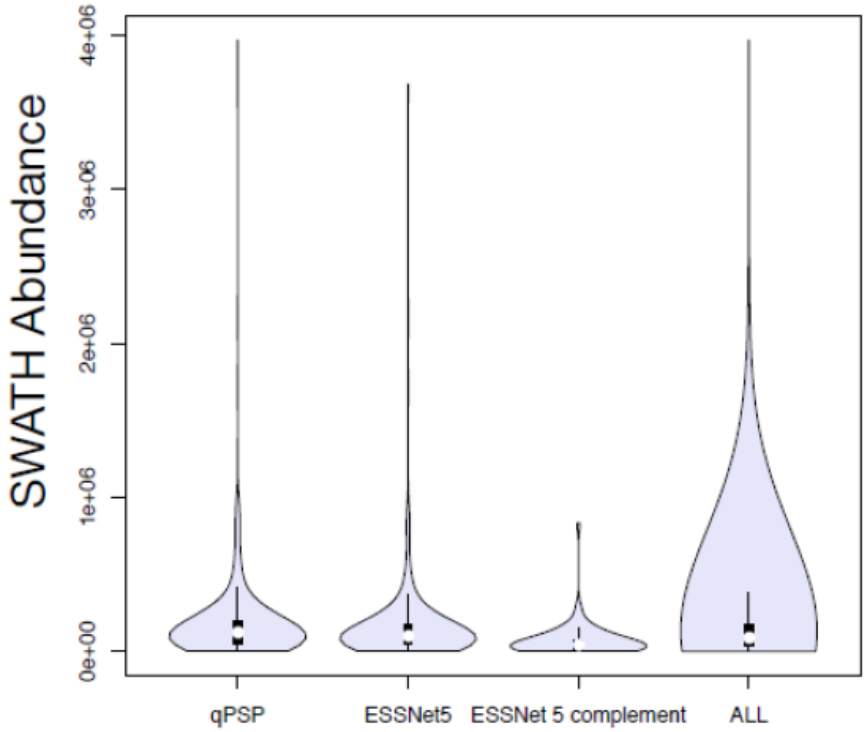
B



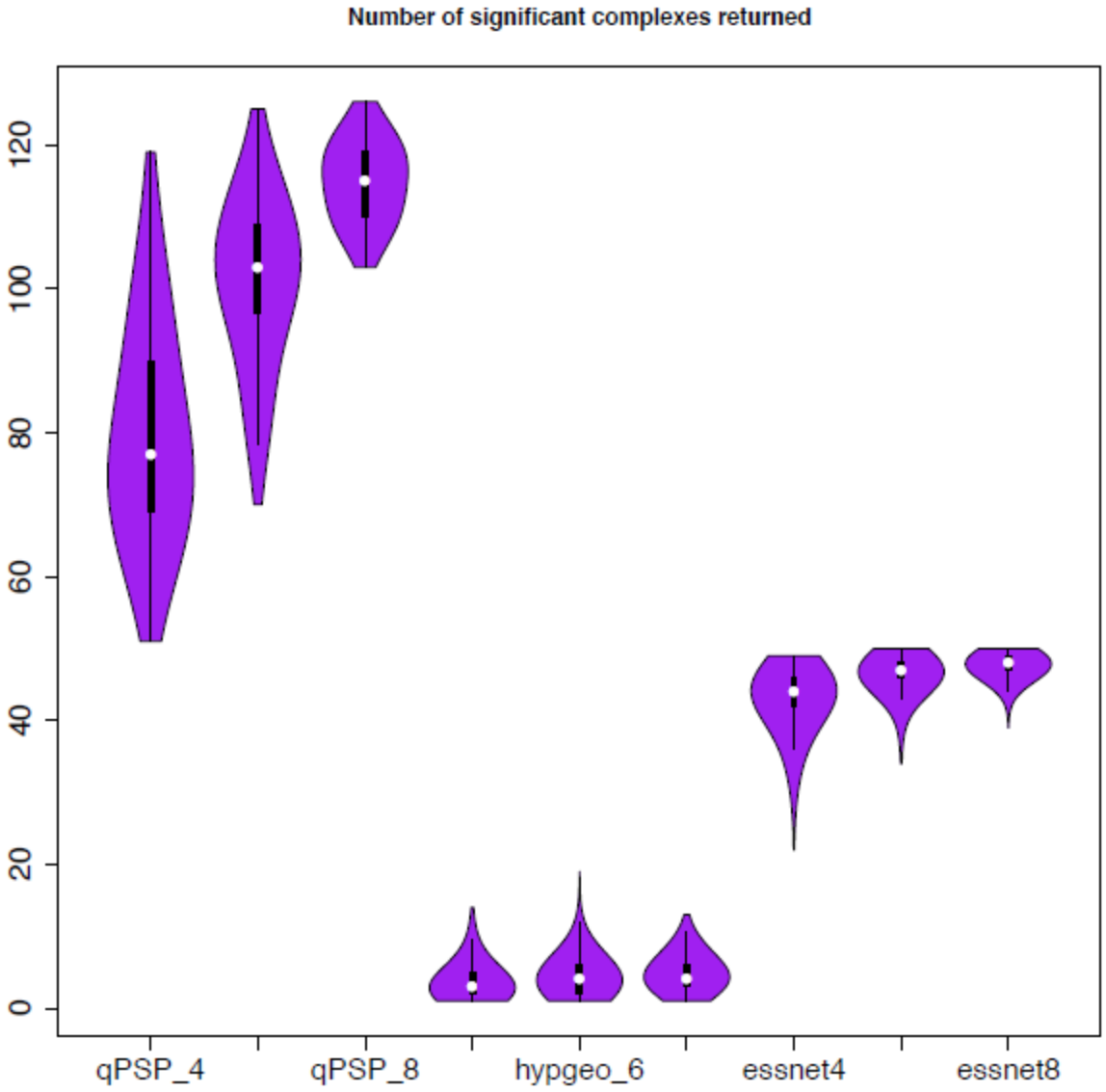
The 5 ESSNet-only complexes are low abundance ones (below 25th percentile of all SWATH proteins)!

C

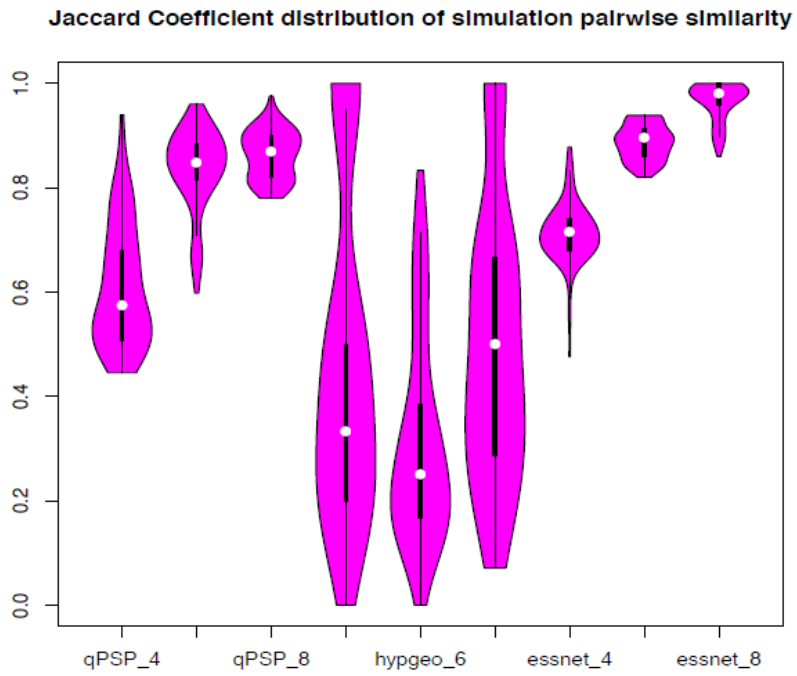


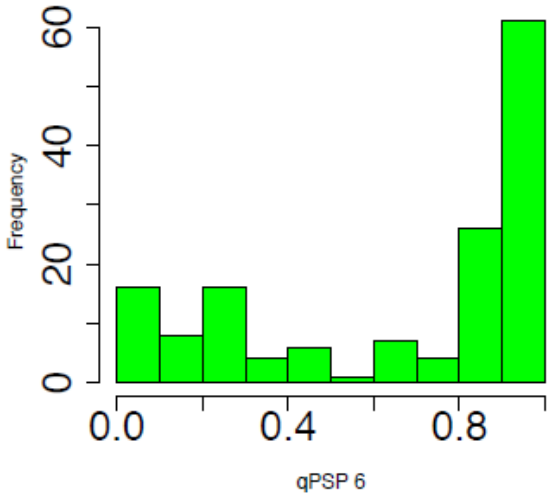


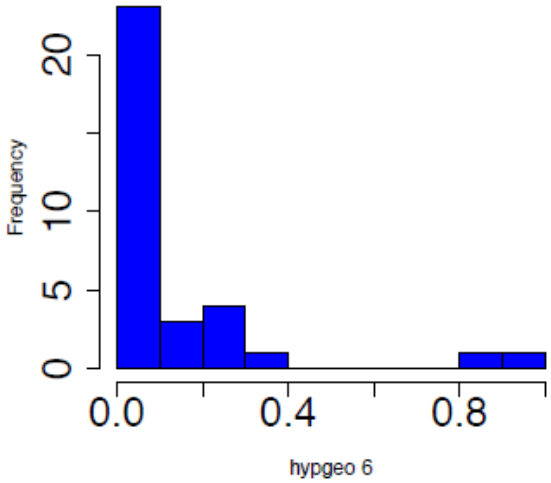
ESSNet reports more complexes than SAP but less than qPSP

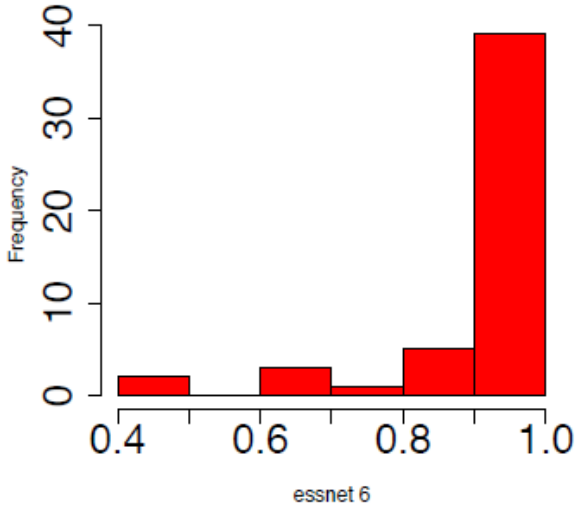


Complexes reported by ESSNet are more stable than those by SAP & qPSP









In conclusion…

Contextualization (into complexes) can deal with coverage, consistency, and incongruity issues in proteomics

References

- 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**. *Journal of Proteome Research*, 11(3):1571-1581, 2012
- [FCS] Goh et al. **Comparative network-based recovery analysis and proteomic profiling of neurological changes in valporic acid-treated mice.** *Journal of Proteome Research***,** 12(5):2116-2127, 2013
- [qPSP] Goh et al. **Quantitative proteomics signature profiling based on network contextualization**. *Biology Direct*, accepted.
- [ESSNET] Lim et al. **A quantum leap in the reproducibility, precision, and sensitivity of gene expression profile analysis even when sample size is extremely small**. *Journal of Bioinformatics and Computational Biology*, 13(4):1550018, 2015

Acknowledgements

Wilson Goh

Kevin Lim

- **Singapore Ministry of Education**