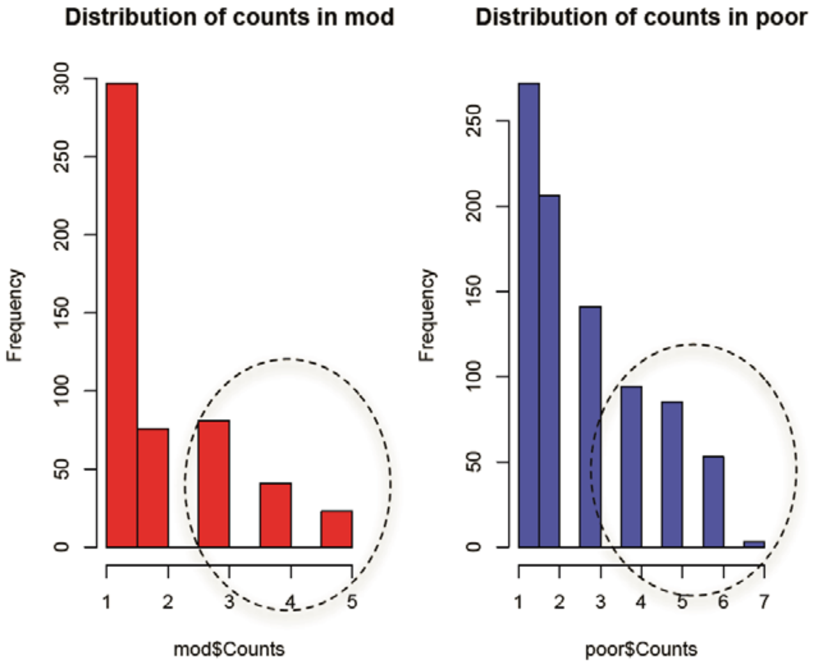
# CS4220: Knowledge Discovery Methods for Bioinformatics
# Unit 3: Proteomic Profiling

**Wong Limsoon**

NUS
National University
of Singapore

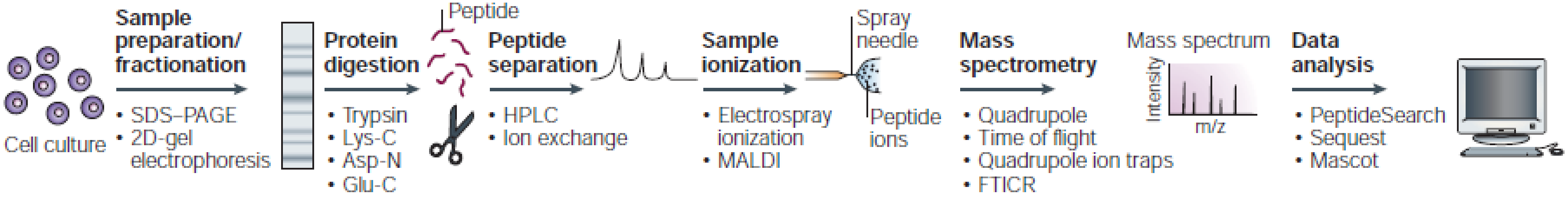
# Delivering more powerful proteomic profile analysis



- **Basic proteomic profile analysis**
- **Common issues**
- **Improving coverage**
- **Improving consistency**
- **Finally, a quantum leap?**

# BASIC PROTEOMIC PROFILE ANALYSIS

# Typical proteomic MS expt



Cell culture

**Sample preparation/ fractionation**
- SDS–PAGE
- 2D-gel electrophoresis

**Protein digestion**
- Trypsin
- Lys-C
- Asp-N
- Glu-C

Peptide

**Peptide separation**
- HPLC
- Ion exchange

**Sample ionization**
- Electrospray ionization
- MALDI

Spray needle

Peptide ions

**Mass spectrometry**
- Quadrupole
- Time of flight
- Quadrupole ion traps
- FTICR

Mass spectrum

Intensity

m/z

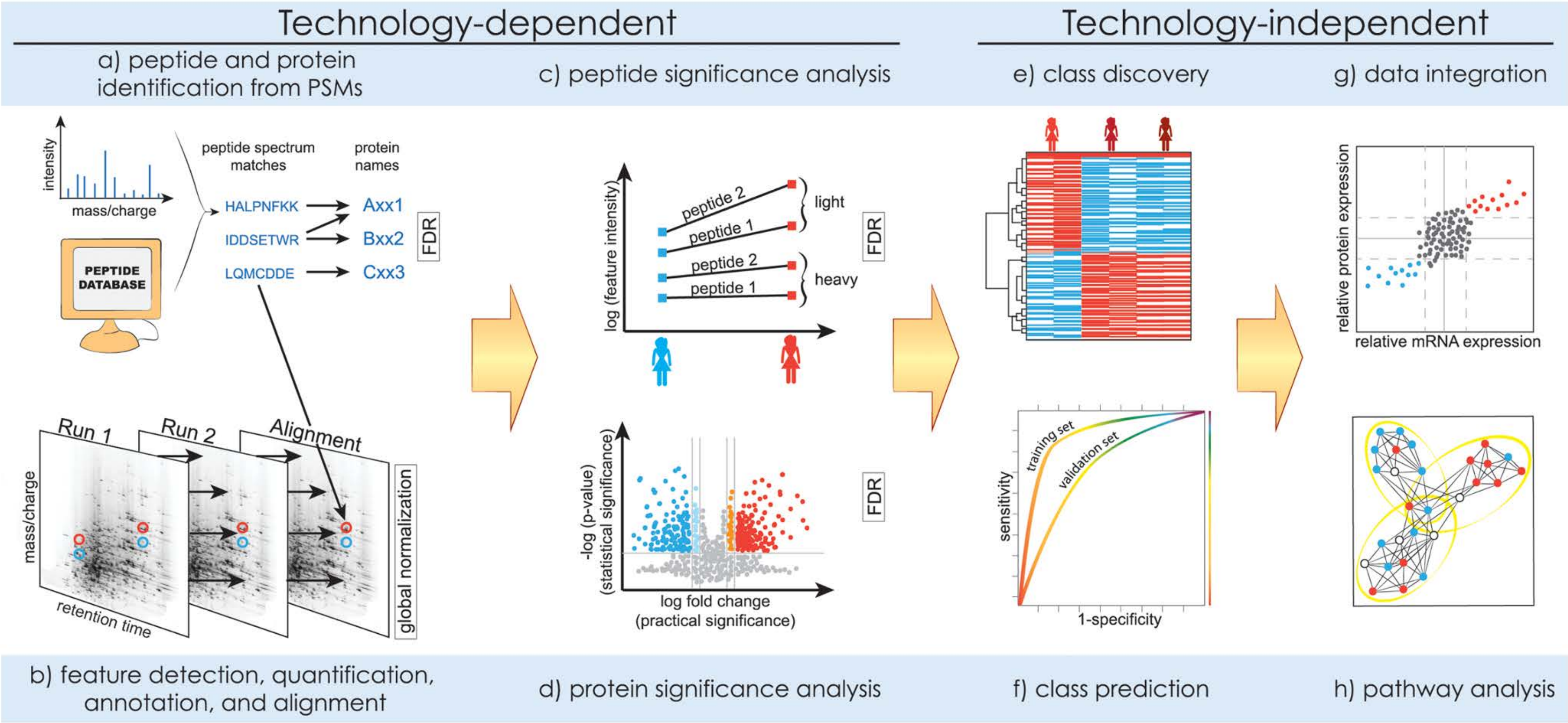
**Data analysis**
- PeptideSearch
- Sequest
- Mascot

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.

See also http://www.slideshare.net/joachimjacob/bits-introduction-to-mass-spec-data-generation

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

A rather nice set of proteomic profiles of leukemia patients

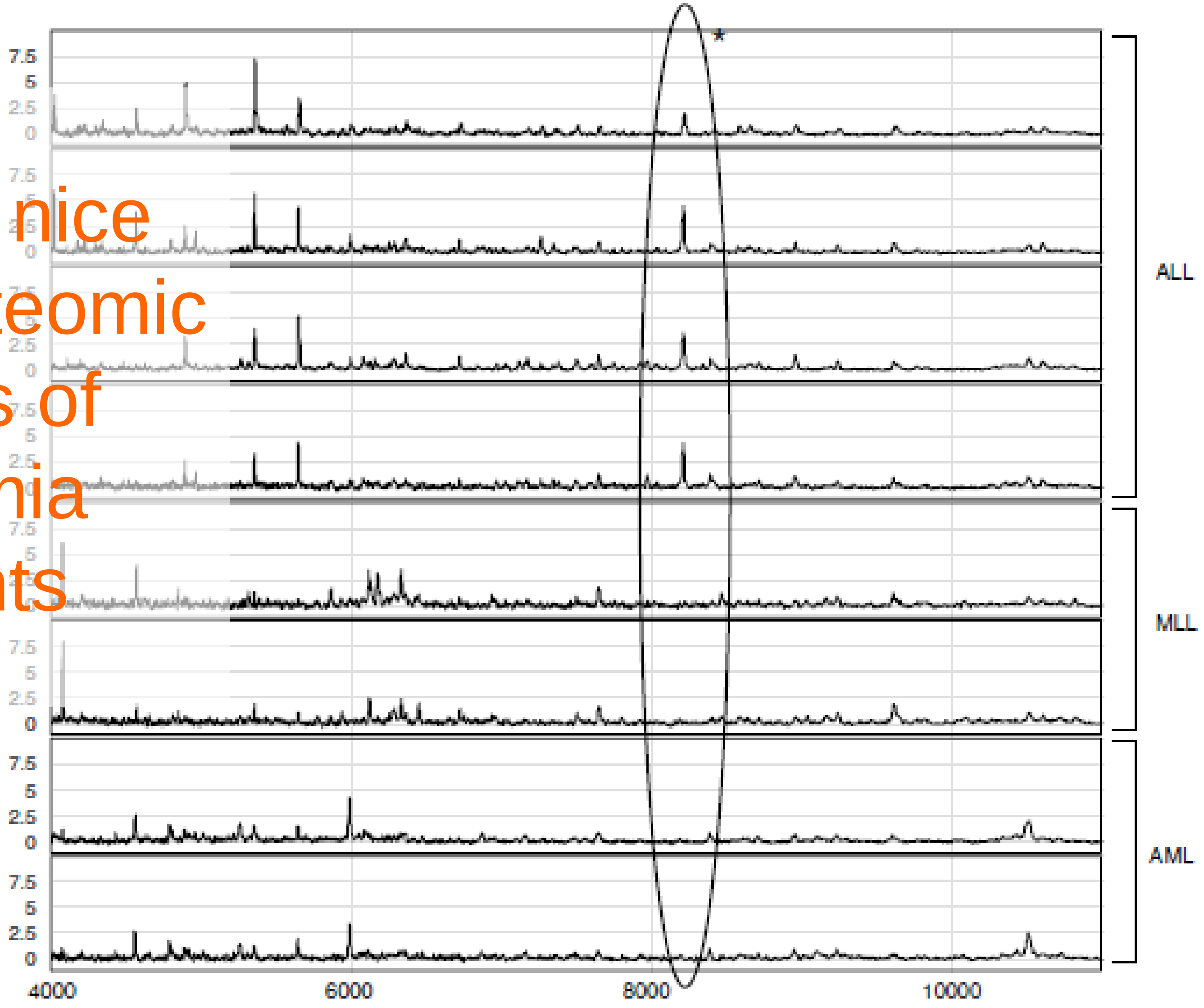
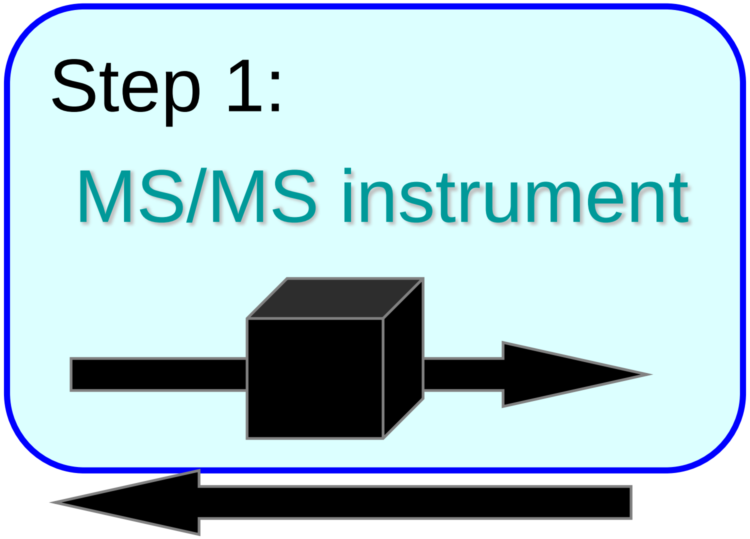


**Figure 1** Spectra from SELDI-TOF MS analysis of REH, 697, MV4;11, and Kasumi cell lines. Protein (4 μg) from each cell type was analyzed on SAX2 ProteinChip® Arrays. ALL cell lines shown are REH and 697, the MLL cell line is MV4;11, and the AML cell line is Kasumi. The asterisk indicates the differentially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

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

# Protein identification by mass spec

Sequence



Step 1:

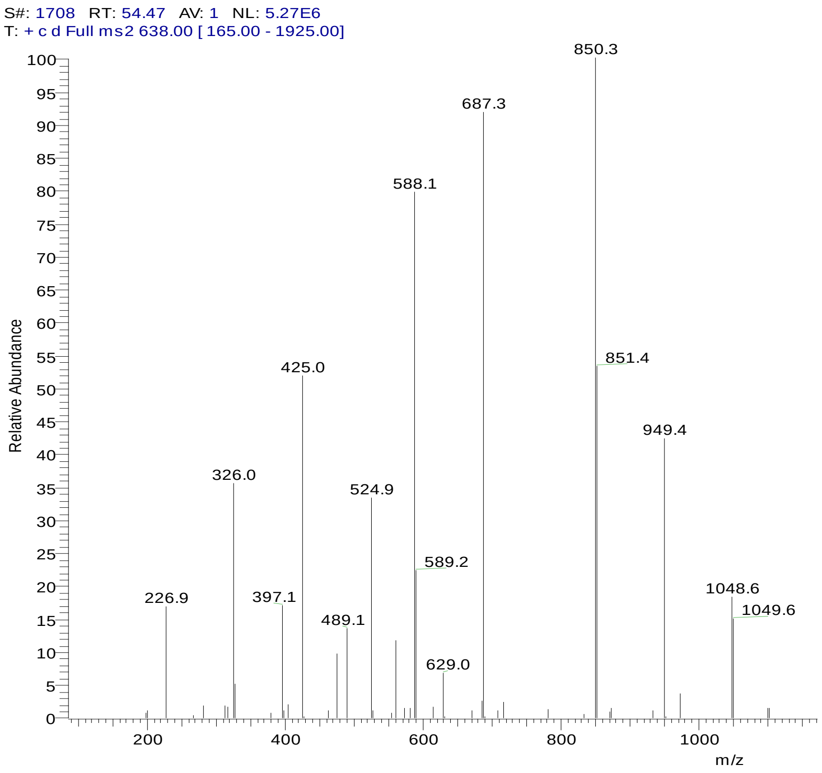
MS/MS instrument

Database search

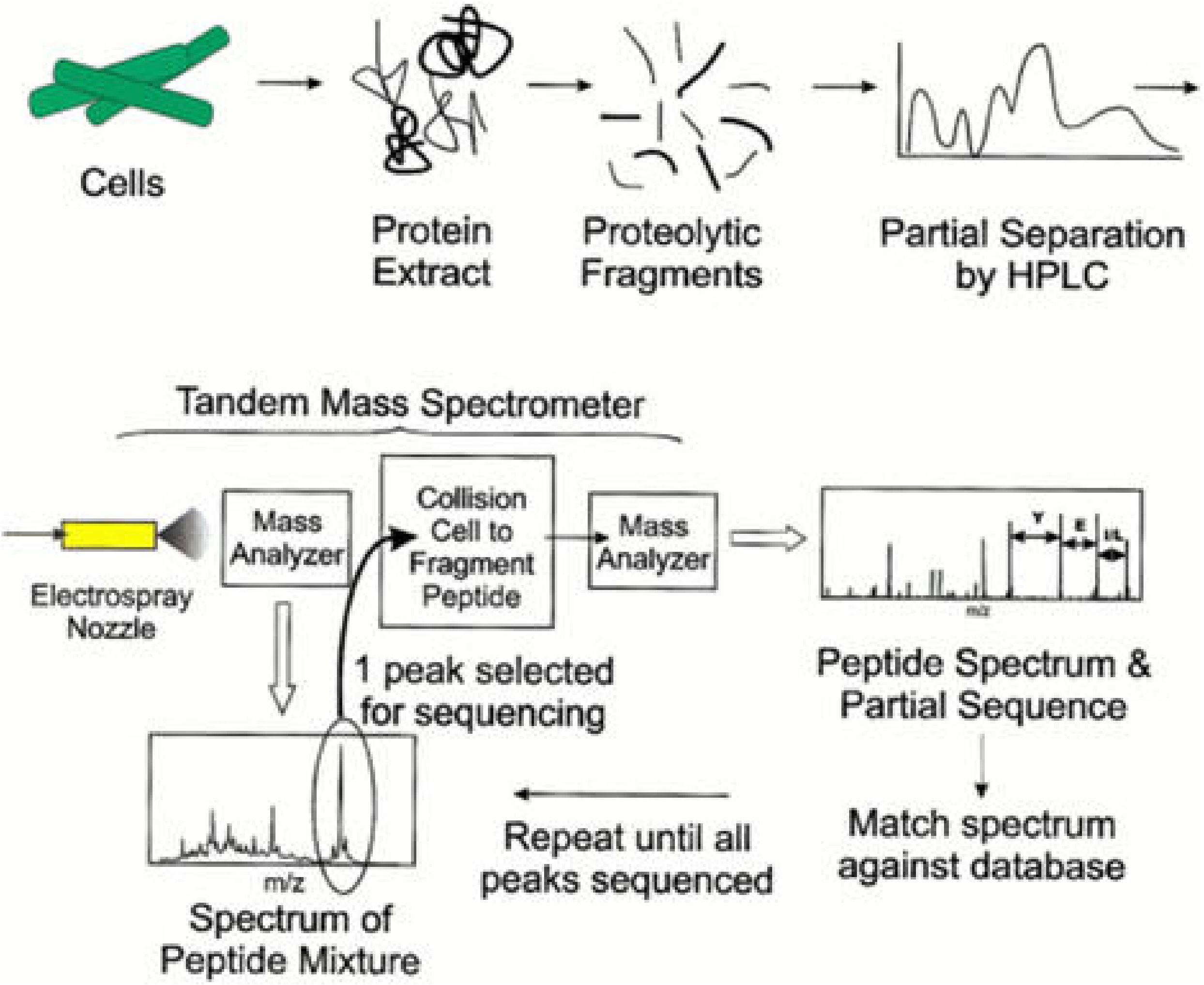
- Sequest, Mascot, InSpect

*de Novo* interpretation

- Lutefisk, Peaks, PepNovo



# Tandem mass spectrometry


Cells
Protein Extract
Proteolytic Fragments
Partial Separation by HPLC
Tandem Mass Spectrometer
Electrospray Nozzle
Mass Analyzer
Collision Cell to Fragment Peptide
Mass Analyzer
1 peak selected for sequencing
m/z
Spectrum of Peptide Mixture
Repeat until all peaks sequenced
m/z
Peptide Spectrum & Partial Sequence
Match spectrum against database

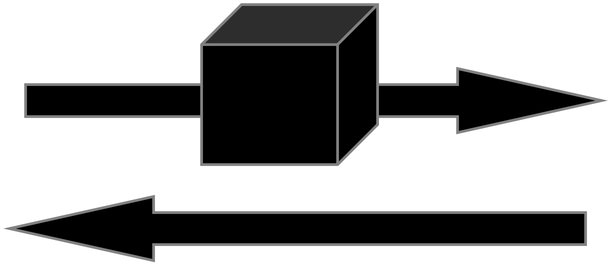
# Breaking protein into peptides, and peptides into fragment ions

- Proteases, e.g. trypsin, break a protein into peptides
- 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

# Peptide identification by mass spec

S
e
q
u
e
n
c
e

## MS/MS instrument

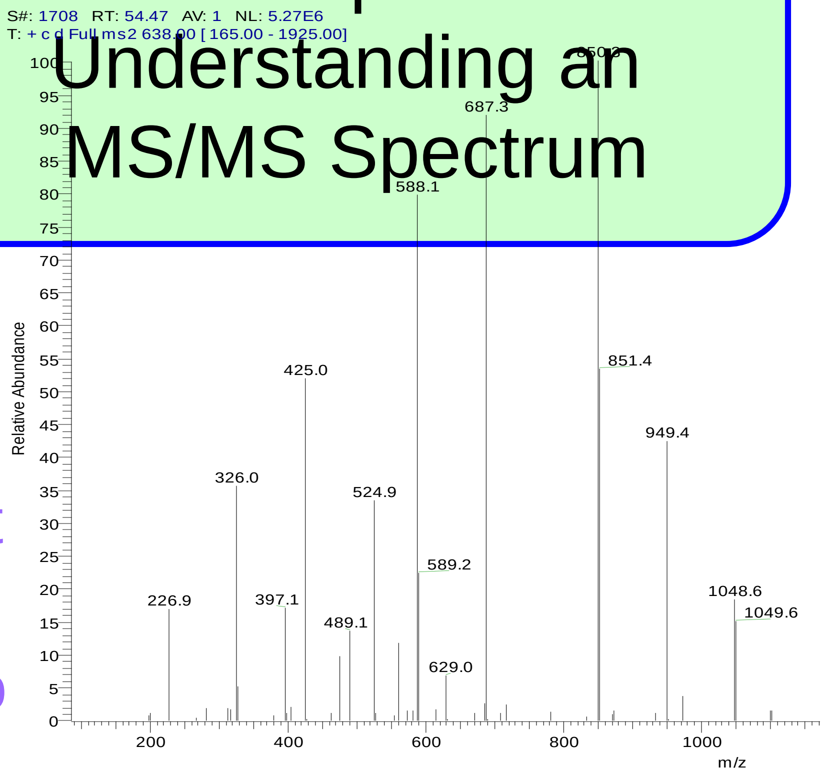


## Database search

- Sequest, Mascot, InSpect

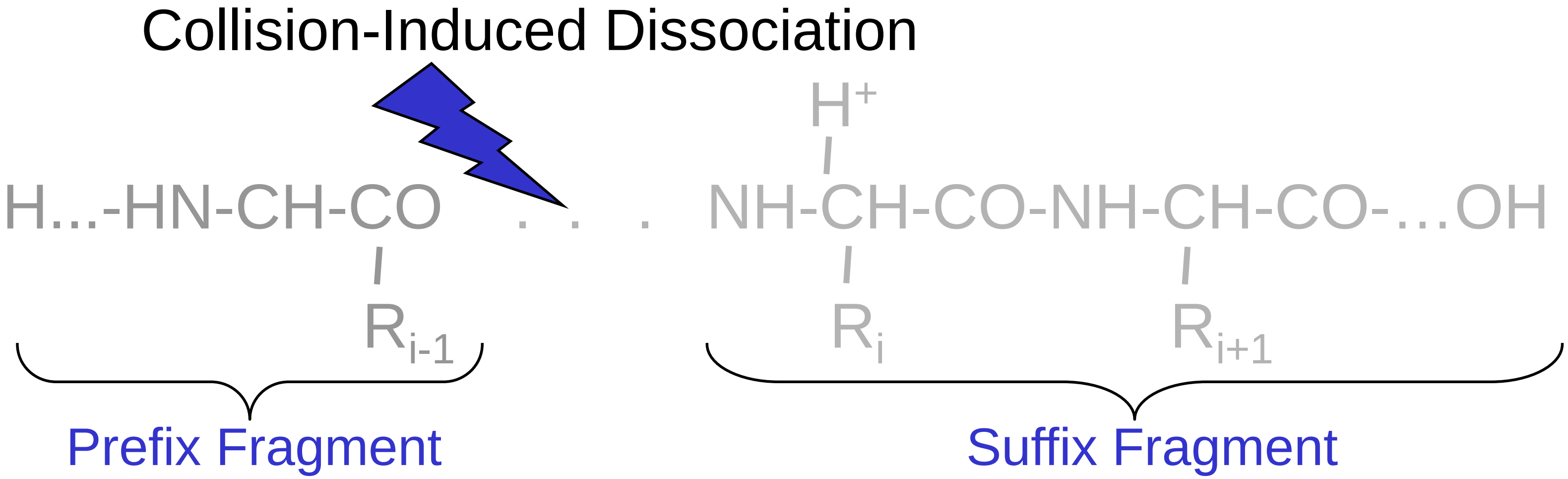
## *de Novo* interpretation

- Lutefisk, Peaks, PepNovo



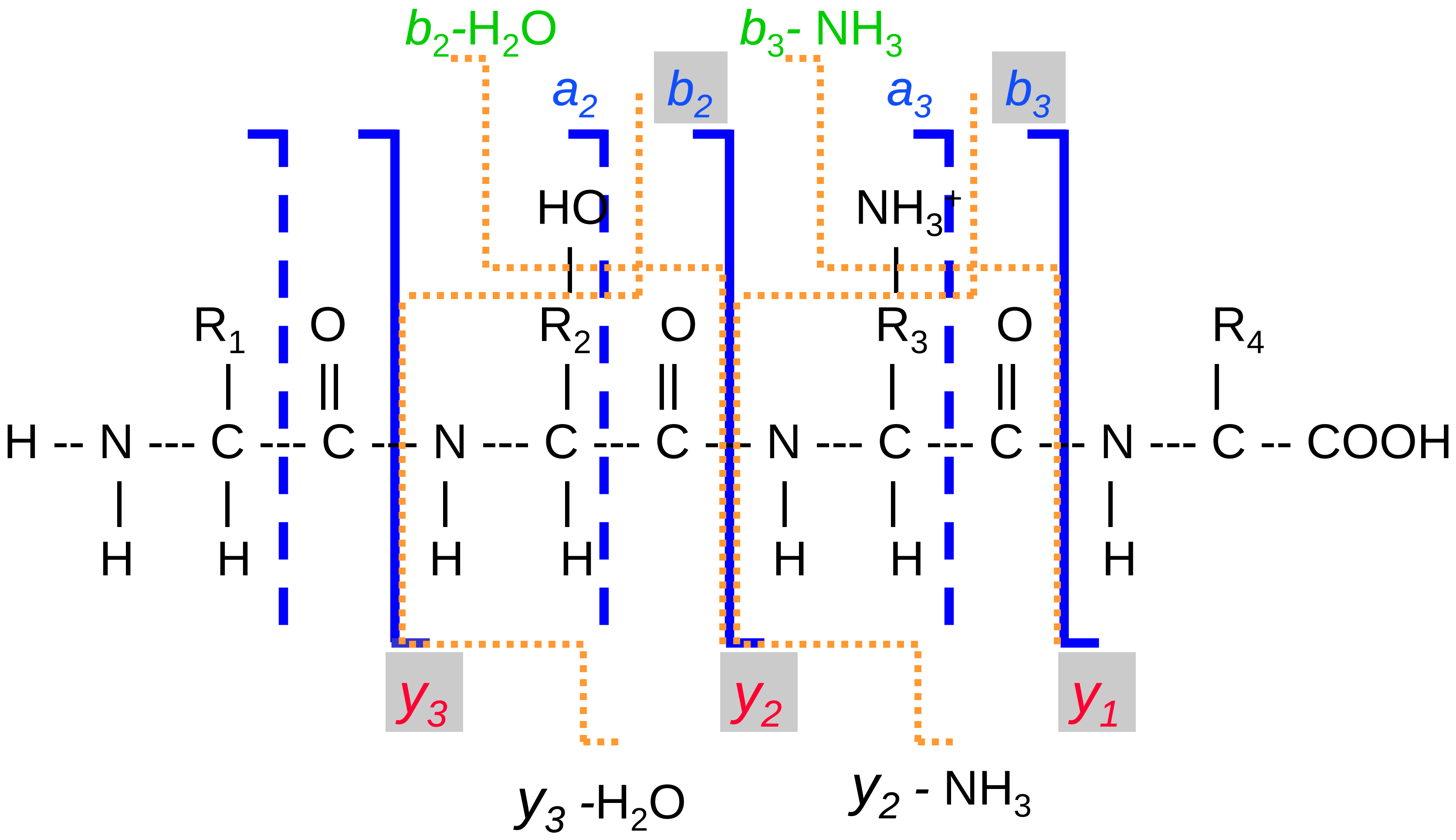
Step 2:
Understanding an MS/MS Spectrum

# Peptide fragmentation

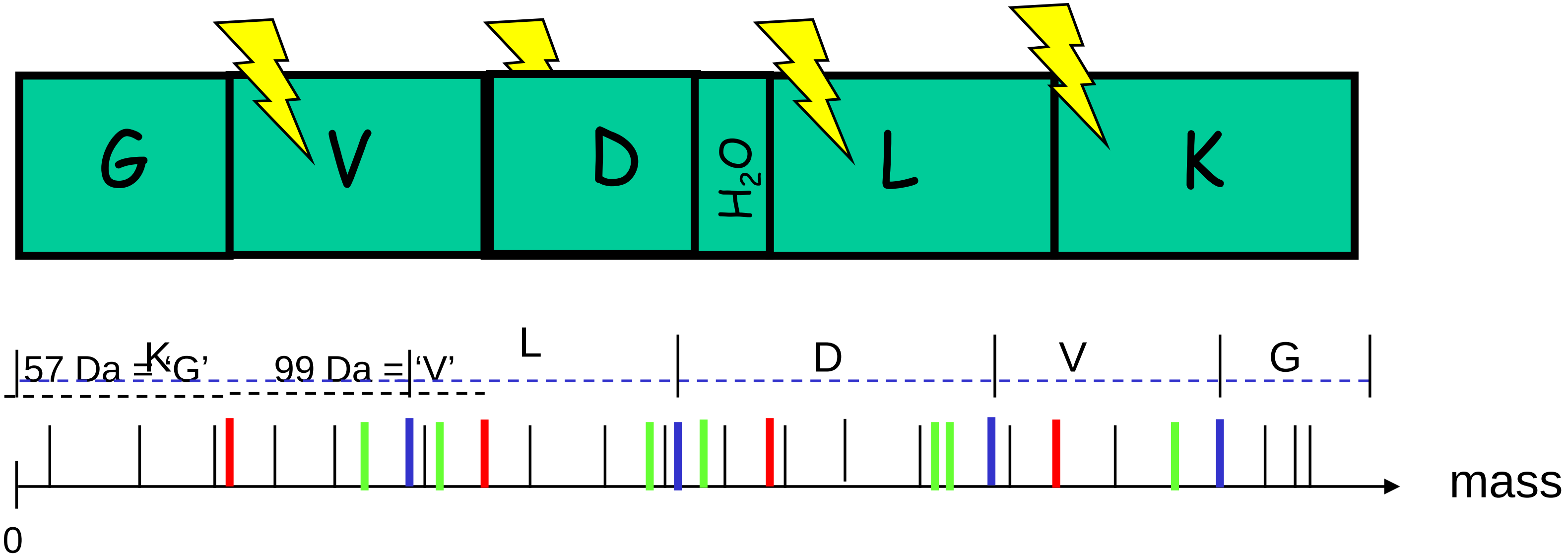

Collision-Induced Dissociation
H+
H...-HN-CH-CO . . . NH-CH-CO-NH-CH-CO-…OH
Ri-1
Ri
Ri+1
Prefix Fragment
Suffix Fragment

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

# … and fragments due to neutral losses



# Mass spectra



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

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 |

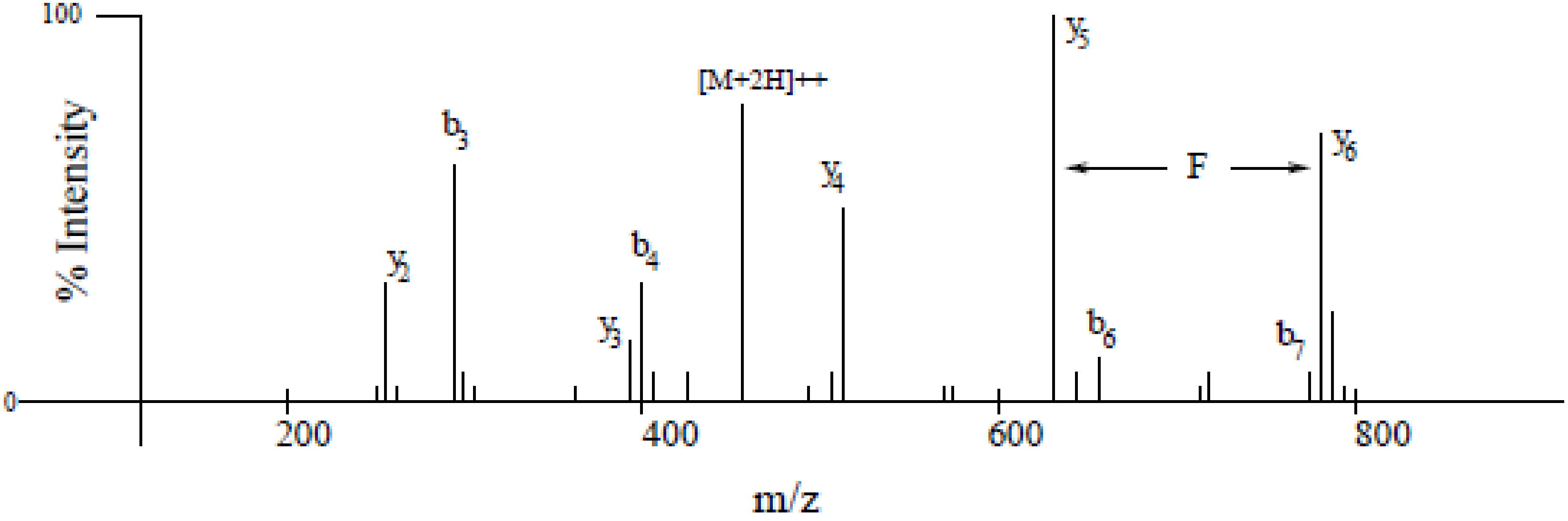
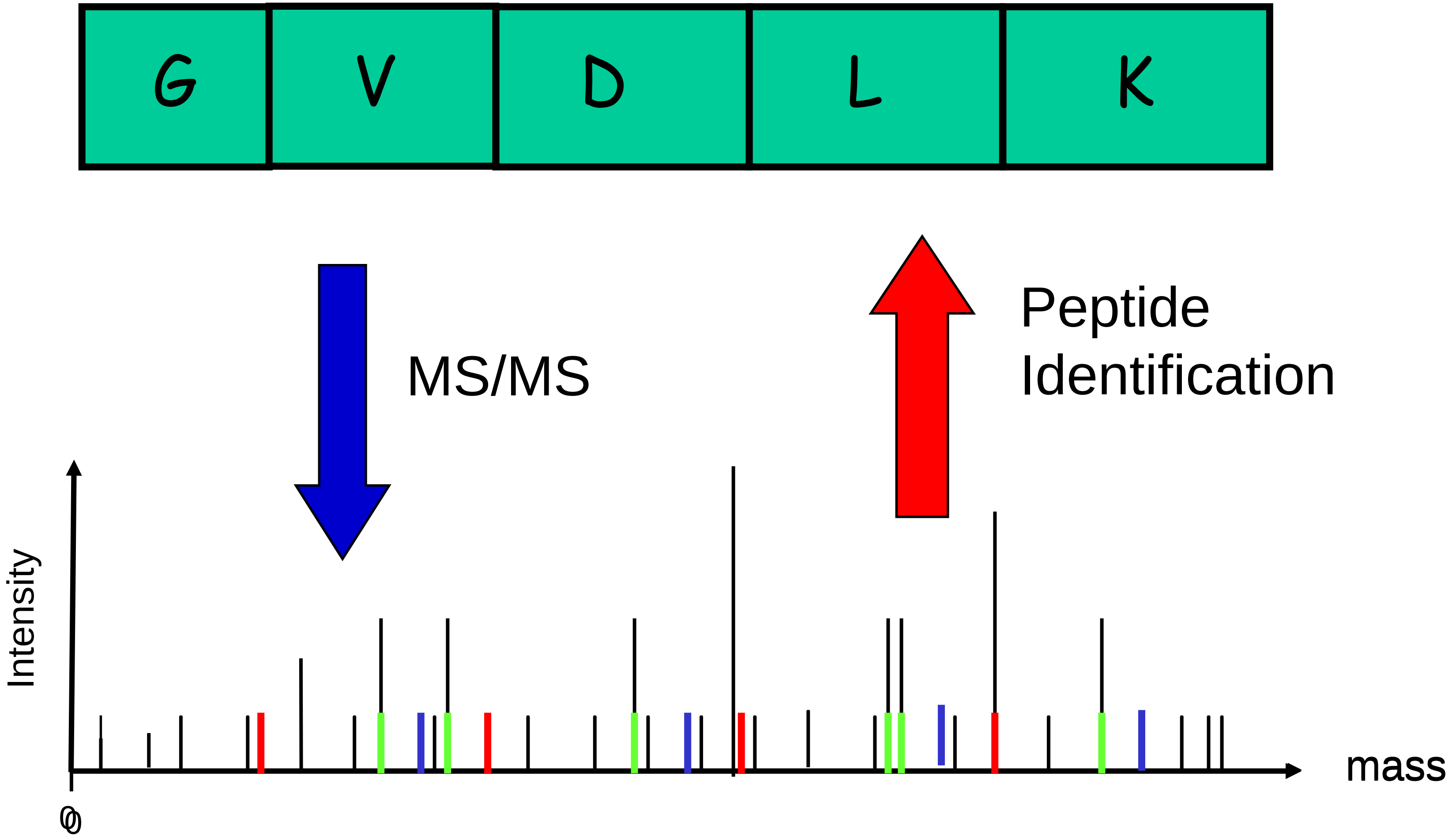


Figure 2: MS/MS spectrum for peptide SGFLEEDK.

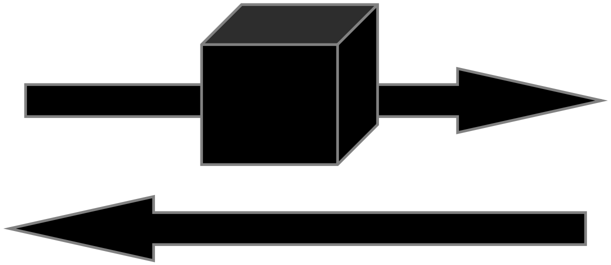
# Protein identification with MS/MS

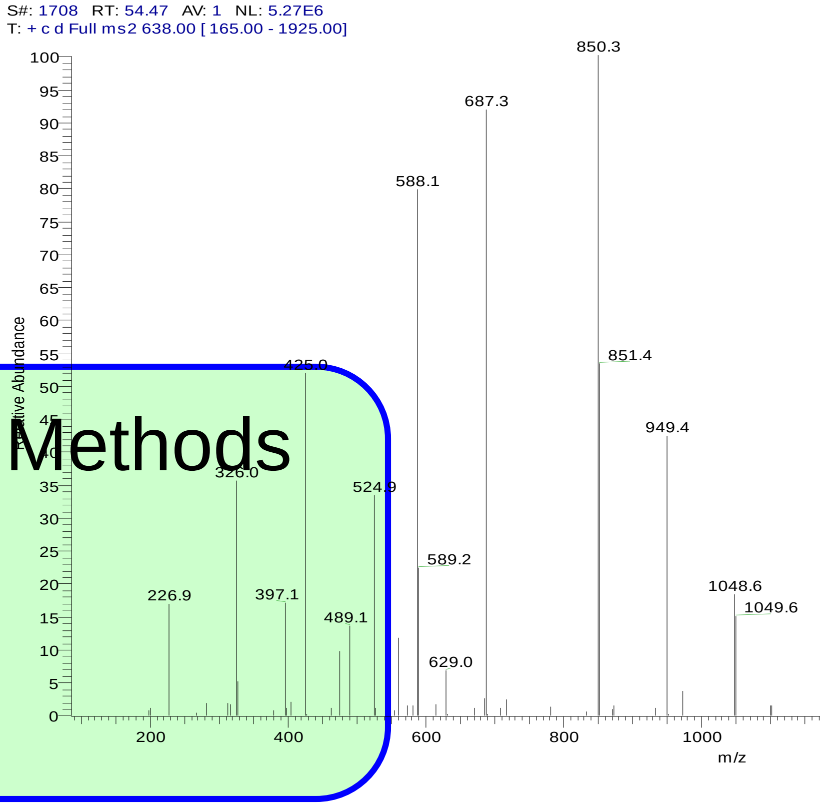


# Peptide identification by mass

Sequence

## MS/MS instrument





### Step 3: Computational Methods

Database search

- Sequest, Mascot

*de Novo* interpretation

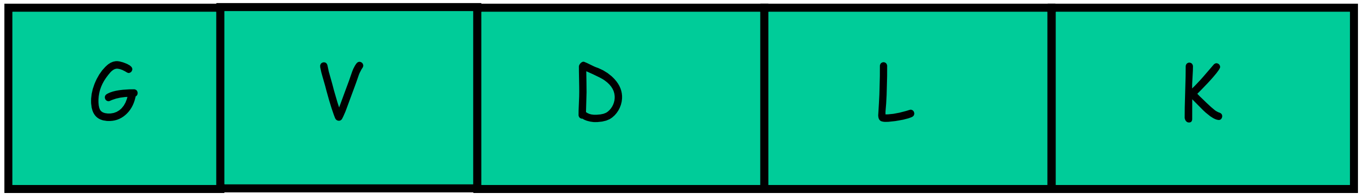
- Lutefisk, Peaks, PepNovo

# Database search algorithms

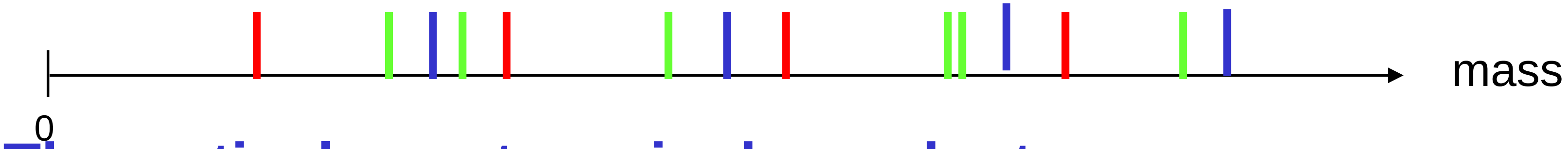
- **Database search**
  - Used for spectrum from known peptides
  - Rely on completeness of database
- **General Approach**
  - Match given spectrum with known peptide
  - Enhanced with advanced statistical analysis and complex scoring functions
- **Methods**
  - SEQUEST, MASCOT, InsPecT, Paragon

# Theoretical spectrum for a peptide

- **Given this peptide**



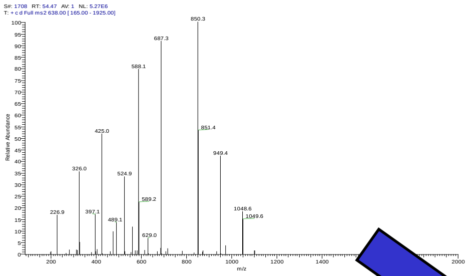
- **Its theoretical spectrum is**



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

# Database search algorithm

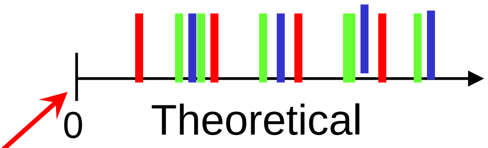
Database Search



Match

Database of known peptides

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



Theoretical spectrum

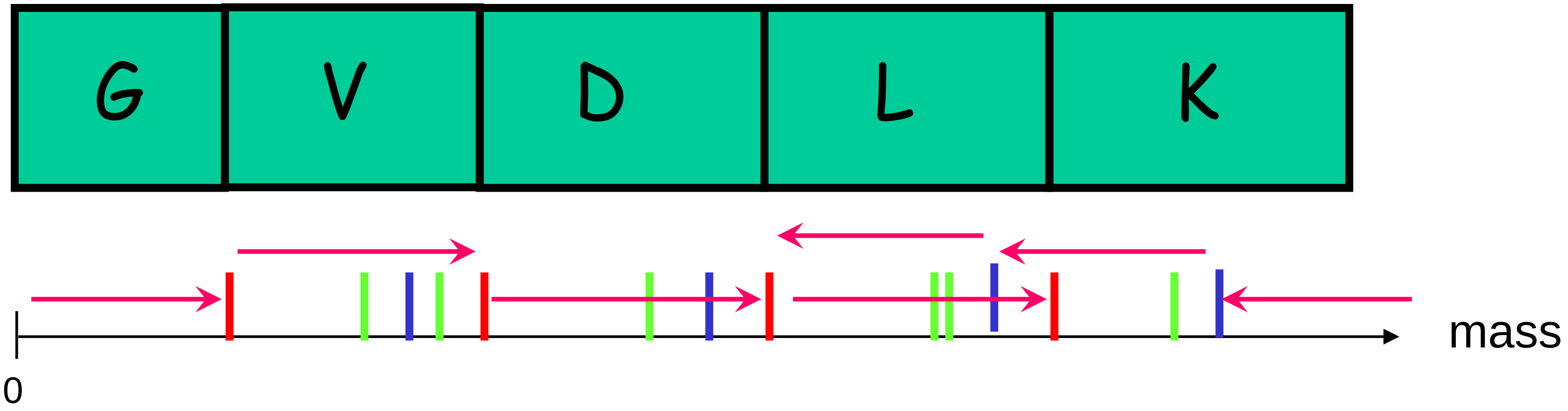
Matching Score for this peptide

Repeat for all the peptides in the Database

# De novo sequencing algorithms

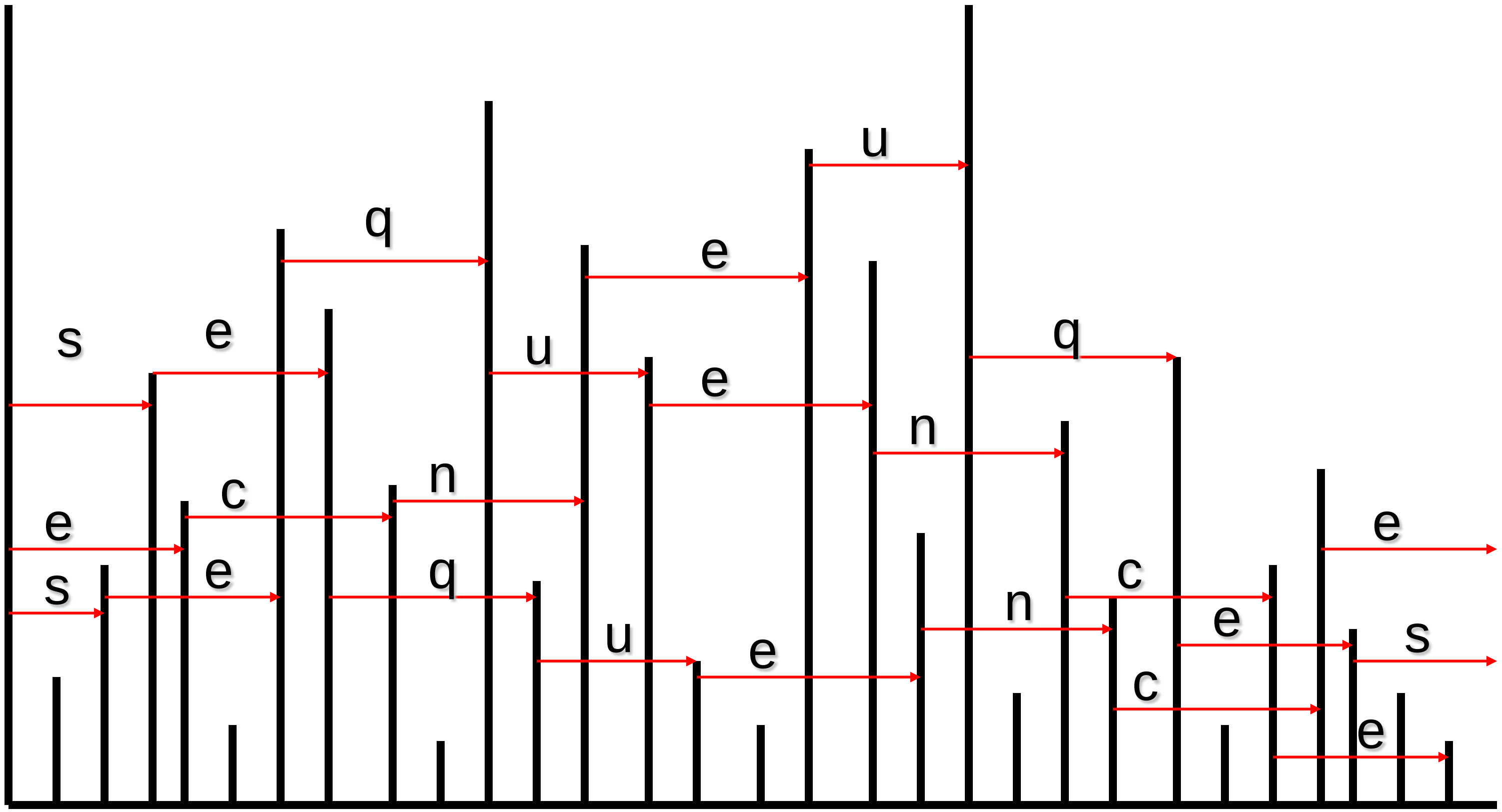
- **Given a spectrum**
  - Build a spectrum graph
  - Peptides are paths in this graph
  - Find the best path

# Spectrum graph for a peptide



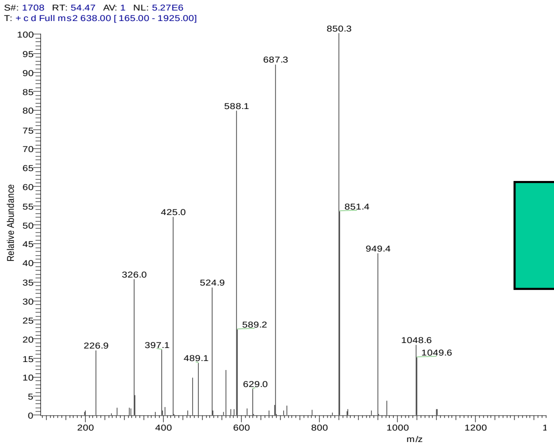
- **Connect peaks together**
  - If their mass difference = mass of an amino acid
- **Theoretical spectrum is dependent on**
  - Set of ion-types considered
  - Larger if multi-charge ions are considered

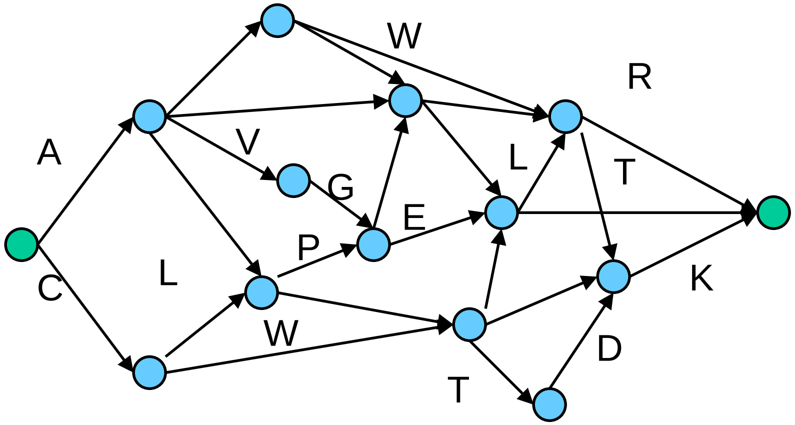
# Building a graph from a spectrum



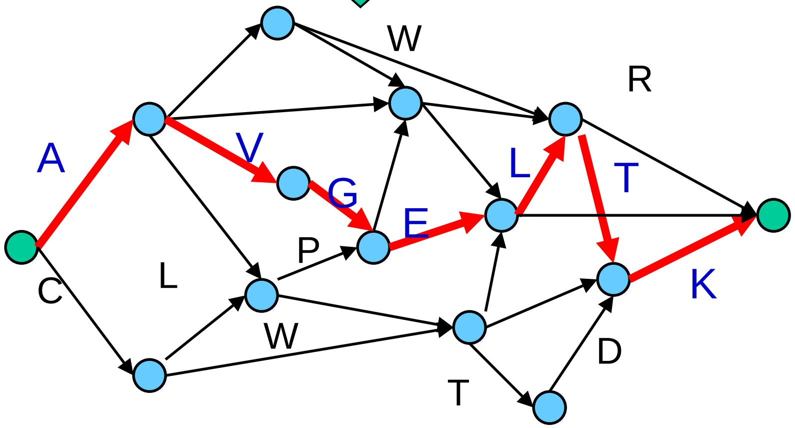
Frank, et al. “De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry”. *J. Proteome Res*. 6:114-123, 2007

# De novo sequencing algorithms



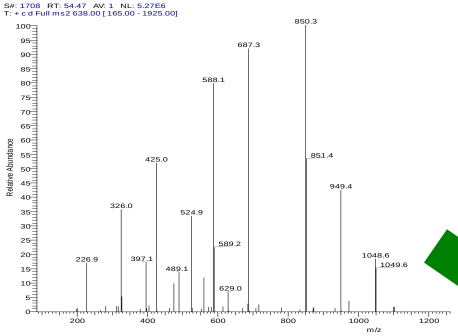


Find longest directed acyclic path



AVGELTK

# De novo vs. database search



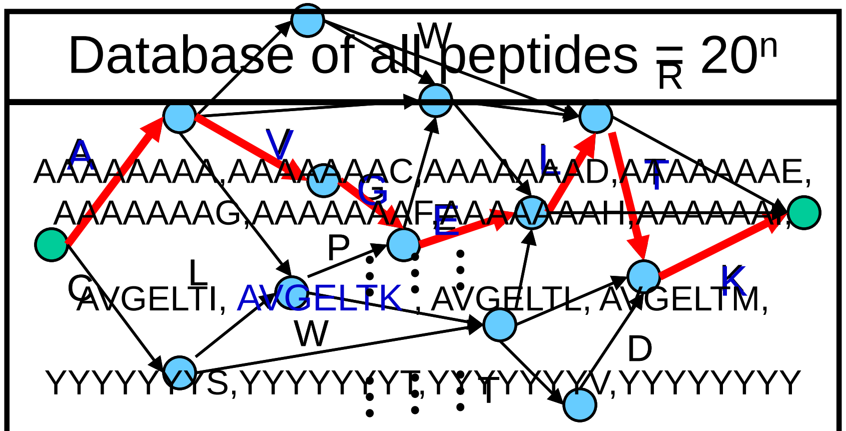
## Database Search

## De Novo

### Database of known peptides

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

### Database of all peptides = $20^n$



AAAAAAAA,AAAAAAAC,AAAAAAAD,AAAAAAAE,
AAAAAAAG,AAAAAAAF,AAAAAAAH,AAAAAAAI,
⋮ ⋮ ⋮
AVGELTI, AVGELTK , AVGELTL, AVGELTM,
⋮ ⋮ ⋮
YYYYYYYS,YYYYYYYT,YYYYYYYV,YYYYYYYY

AVGELTK

# De novo vs. database search: A paradox

- **The database of all peptides is huge ≈ $O(20^n)$**
- **The database of all known peptides is much smaller ≈ $O(10^8)$**
- **However, de novo algorithms can be much faster, even though their search space is much larger!**
  - – A database search scans all peptides in the search space to find best one
  - – De novo eliminates the need to scan all peptides by modeling the problem as a graph search

# 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**

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

# COMMON ISSUES IN PROTEOMIC PROFILE ANALYSIS

# 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 applied in theory…**

- **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**

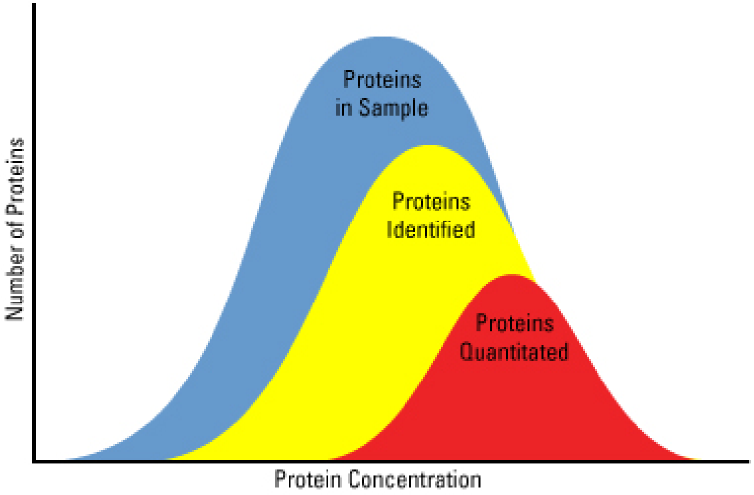
# Peptide & protein identification by MS is still far from perfect

- **"… 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

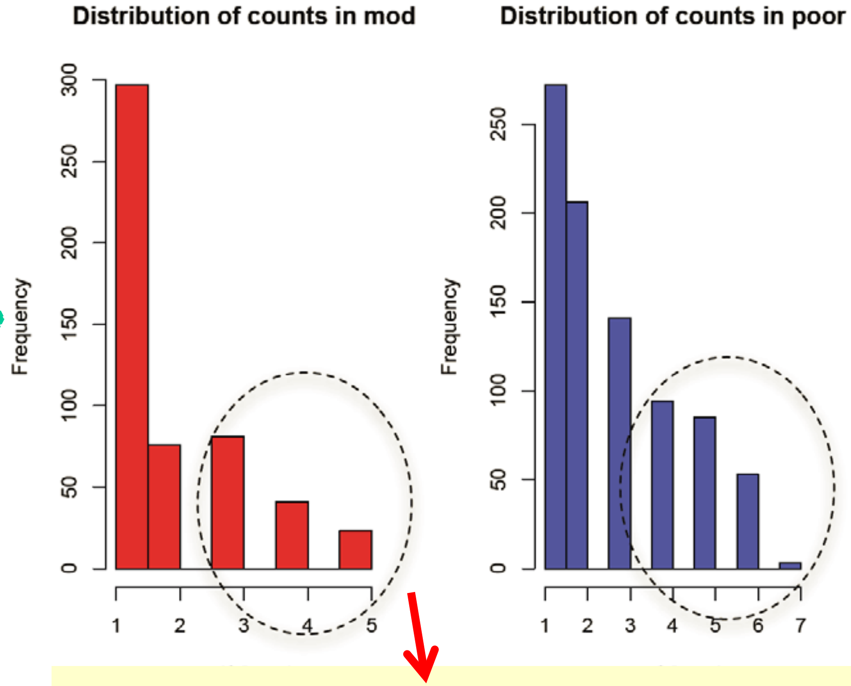
# Issues in proteomics: Coverage and consistency

## Technical incompleteness



Per sample

## How it affects real data



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

# USING PROTEIN COMPLEXES IN PROTEOMICS: BASIC IDEAS

# A postulate and some math

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

- **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% (= 0.75 * 4 / 5) chance to be present**

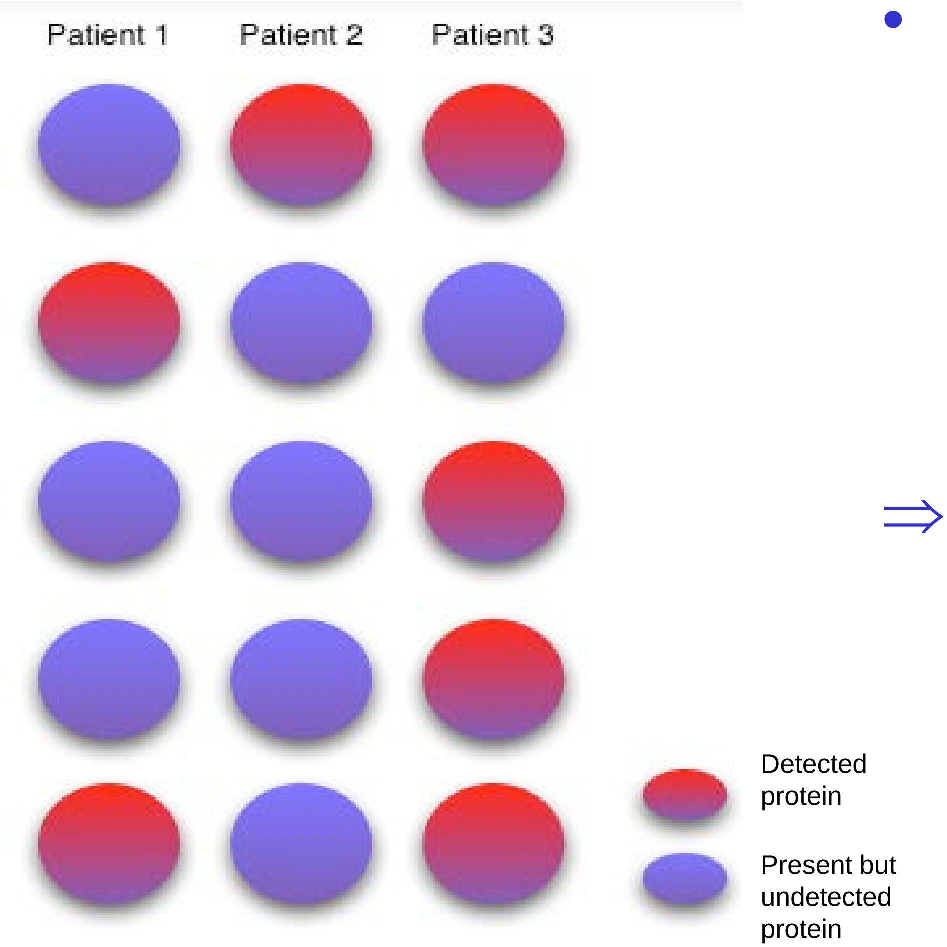
⇒ **The unreported protein E also has ≥ 60% chance to be present, as presence of the complex implies presence of all its constituents**

⇒ improving coverage

⇒ **Each of the reported proteins (A, B, C, and D) individually has 90% (= 100% * 0.6 + 75% * 0.4) chance of being true positive, whereas a reported protein that is isolated has a lower 75% chance of being true positive**

⇒ removing noise

# An 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?

# Reference complexes

- **In this lecture, human complexes (of size at least 5) from CORUM are used as reference complexes**

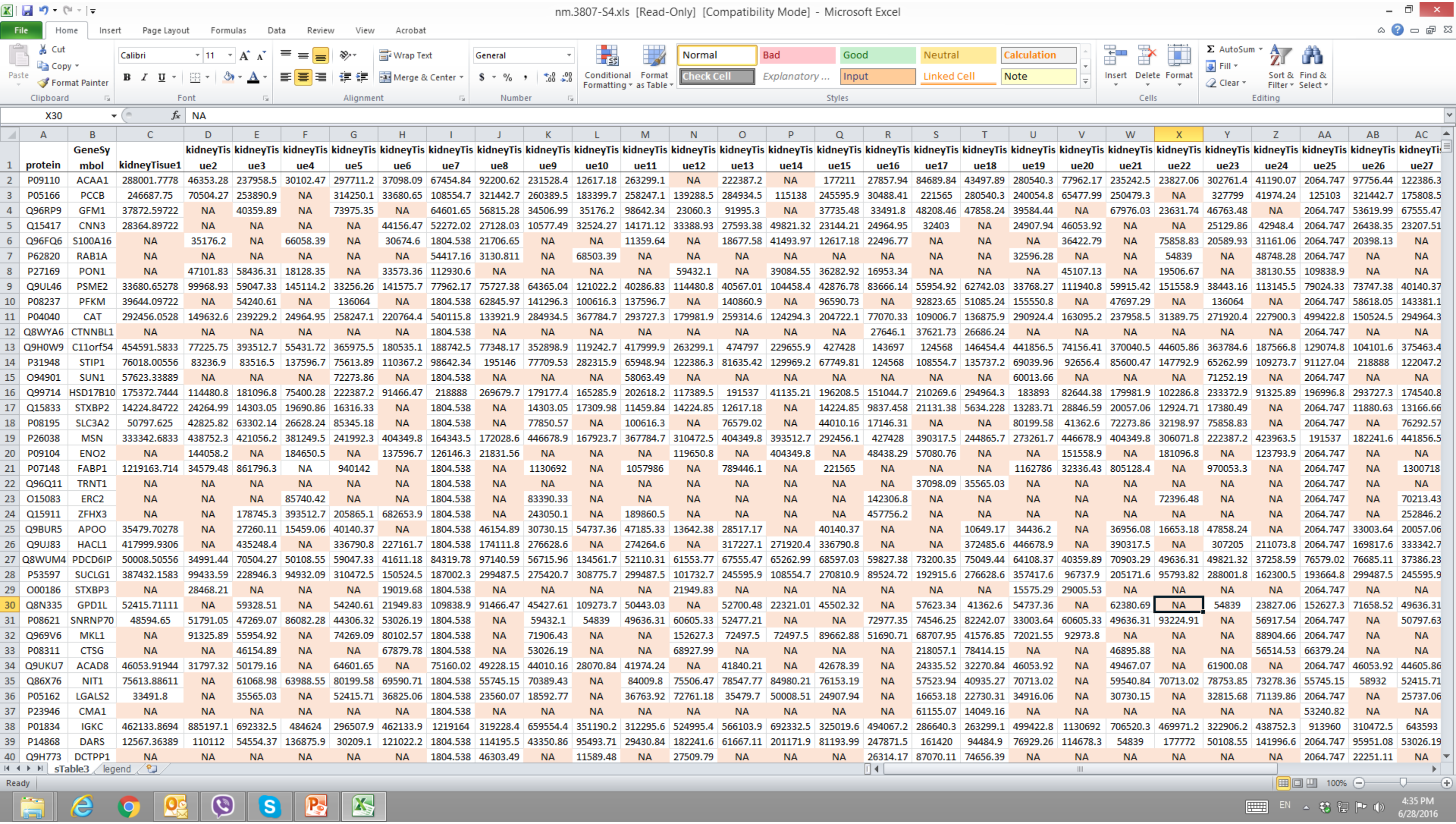
- **It is possible to use subnets generated from pathway and PPI databases. However these such subnets vary significantly depending on network databases and subnet-generation algo used**

So I do not consider these…

# IMPROVING COVERAGE IN PROTEOMIC PROFILES

Guo et al. *Nature Medicine*, *21*, 407, 2015

# Lots of missing values in real proteomics datasets



nm.3807-S4.xls [Read-Only] [Compatibility Mode] - Microsoft Excel

X30 | NA

| | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA | AB | AC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | protein | GeneSymbol | kidneyTisue1 | kidneyTisue2 | kidneyTisue3 | kidneyTisue4 | kidneyTisue5 | kidneyTisue6 | kidneyTisue7 | kidneyTisue8 | kidneyTisue9 | kidneyTisue10 | kidneyTisue11 | kidneyTisue12 | kidneyTisue13 | kidneyTisue14 | kidneyTisue15 | kidneyTisue16 | kidneyTisue17 | kidneyTisue18 | kidneyTisue19 | kidneyTisue20 | kidneyTisue21 | kidneyTisue22 | kidneyTisue23 | kidneyTisue24 | kidneyTisue25 | kidneyTisue26 | kidneyTisue27 |
| 2 | P09110 | ACAA1 | 288001.7778 | 46353.28 | 237958.5 | 30102.47 | 297711.2 | 37098.09 | 67454.84 | 92200.62 | 231528.4 | 12617.18 | 263299.1 | NA | 222387.2 | NA | 177211 | 27857.94 | 84689.84 | 43497.89 | 280540.3 | 77962.17 | 235242.5 | 23827.06 | 302761.4 | 41190.07 | 2064.747 | 97756.44 | 122386.3 |
| 3 | P05166 | PCCB | 246687.75 | 70504.27 | 253890.9 | NA | 314250.1 | 33680.65 | 108554.7 | 321442.7 | 260389.5 | 183399.7 | 258247.1 | 139288.5 | 284934.5 | 115138 | 245595.9 | 30488.41 | 221565 | 280540.3 | 240054.8 | 65477.99 | 250479.3 | NA | 327799 | 41974.24 | 125103 | 321442.7 | 175808.5 |
| 4 | Q96RP9 | GFM1 | 37872.59722 | NA | 40359.89 | NA | 73975.35 | NA | 64601.65 | 56815.28 | 34506.99 | 35176.2 | 98642.34 | 23060.3 | 91995.3 | NA | 37735.48 | 33491.8 | 48208.46 | 47858.24 | 39584.44 | NA | 67976.03 | 23631.74 | 46763.48 | NA | 2064.747 | 53619.99 | 67555.47 |
| 5 | Q15417 | CNN3 | 28364.89722 | NA | NA | NA | NA | 44156.47 | 52272.02 | 27128.03 | 10577.49 | 32524.27 | 14171.12 | 33388.93 | 27593.38 | 49821.32 | 23144.21 | 24964.95 | 32403 | NA | 24907.94 | 46053.92 | NA | NA | 25129.86 | 42948.4 | 2064.747 | 26438.35 | 23207.51 |
| 6 | Q96FQ6 | S100A16 | NA | 35176.2 | NA | 66058.39 | NA | 30674.6 | 1804.538 | 21706.65 | NA | NA | 11359.64 | NA | 18677.58 | 41493.97 | 12617.18 | 22496.77 | NA | NA | NA | 36422.79 | NA | 75858.83 | 20589.93 | 31161.06 | 2064.747 | 20398.13 | NA |
| 7 | P62820 | RAB1A | NA | NA | NA | NA | NA | NA | 54417.16 | 3130.811 | NA | 68503.39 | NA | NA | NA | NA | NA | NA | NA | NA | 32596.28 | NA | NA | 54839 | NA | 48748.28 | 2064.747 | NA | NA |
| 8 | P27169 | PON1 | NA | 47101.83 | 58436.31 | 18128.35 | NA | 33573.36 | 112930.6 | NA | NA | NA | NA | 59432.1 | NA | 39084.55 | 36282.92 | 16953.34 | NA | NA | NA | 45107.13 | NA | 19506.67 | NA | 38130.55 | 109838.9 | NA | NA |
| 9 | Q9UL46 | PSME2 | 33680.65278 | 99968.93 | 59047.33 | 145114.2 | 33256.26 | 141575.7 | 77962.17 | 75727.38 | 64365.04 | 121022.2 | 40286.83 | 114480.8 | 40567.01 | 104458.4 | 42876.78 | 83666.14 | 55954.92 | 62742.03 | 33768.27 | 111940.8 | 59915.42 | 151558.9 | 38443.16 | 113145.5 | 79024.33 | 73747.38 | 40140.37 |
| 10 | P08237 | PFKM | 39644.09722 | NA | 54240.61 | NA | 136064 | NA | 1804.538 | 62845.97 | 141296.3 | 100616.3 | 137596.7 | NA | 140860.9 | NA | 96590.73 | NA | 92823.65 | 51085.24 | 155550.8 | NA | 47697.29 | NA | 136064 | NA | 2064.747 | 58618.05 | 143381.1 |
| 11 | P04040 | CAT | 292456.0528 | 149632.6 | 239229.2 | 24964.95 | 258247.1 | 220764.4 | 540115.8 | 133921.9 | 284934.5 | 367784.7 | 293727.3 | 179981.9 | 259314.6 | 124294.3 | 204722.1 | 77070.33 | 109006.7 | 136875.9 | 290924.4 | 163095.2 | 237958.5 | 31389.75 | 271920.4 | 227900.3 | 499422.8 | 150524.5 | 294964.3 |
| 12 | Q8WYA6 | CTNNBL1 | NA | NA | NA | NA | NA | NA | 1804.538 | NA | NA | NA | NA | NA | NA | NA | NA | 27646.1 | 37621.73 | 26686.24 | NA | NA | NA | NA | NA | NA | 2064.747 | NA | NA |
| 13 | Q9H0W9 | C11orf54 | 454591.5833 | 77225.75 | 393512.7 | 55431.72 | 365975.5 | 180535.1 | 188742.5 | 77348.17 | 352898.9 | 119242.7 | 417999.9 | 263299.1 | 474797 | 229655.9 | 427428 | 143697 | 124568 | 146454.4 | 441856.5 | 74156.41 | 370040.5 | 44605.86 | 363784.6 | 187566.8 | 129074.8 | 104101.6 | 375463.4 |
| 14 | P31948 | STIP1 | 76018.00556 | 83236.9 | 83516.5 | 137596.7 | 75613.89 | 110367.2 | 98642.34 | 195146 | 77709.53 | 282315.9 | 65948.94 | 122386.3 | 81635.42 | 129969.2 | 67749.81 | 124568 | 108554.7 | 135737.2 | 69039.96 | 92656.4 | 85600.47 | 147792.9 | 65262.99 | 109273.7 | 91127.04 | 218888 | 122047.2 |
| 15 | O94901 | SUN1 | 57623.33889 | NA | NA | NA | 72273.86 | NA | 1804.538 | NA | NA | NA | 58063.49 | NA | NA | NA | NA | NA | NA | NA | 60013.66 | NA | NA | NA | 71252.19 | NA | 2064.747 | NA | NA |
| 16 | Q99714 | HSD17B10 | 175372.7444 | 114480.8 | 181096.8 | 75400.28 | 222387.2 | 91466.47 | 218888 | 269679.7 | 179177.4 | 165285.9 | 202618.2 | 117389.5 | 191537 | 41135.21 | 196208.5 | 151044.7 | 210269.6 | 294964.3 | 183893 | 82644.38 | 179981.9 | 102286.8 | 233372.9 | 91325.89 | 196996.8 | 293727.3 | 174540.8 |
| 17 | Q15833 | STXBP2 | 14224.84722 | 24264.99 | 14303.05 | 19690.86 | 16316.33 | NA | 1804.538 | NA | 14303.05 | 17309.98 | 11459.84 | 14224.85 | 12617.18 | NA | 14224.85 | 9837.458 | 21131.38 | 5634.228 | 13283.71 | 28846.59 | 20057.06 | 12924.71 | 17380.49 | NA | 2064.747 | 11880.63 | 13166.66 |
| 18 | P08195 | SLC3A2 | 50797.625 | 42825.82 | 63302.14 | 26628.24 | 85345.18 | NA | 1804.538 | NA | 77850.57 | NA | 100616.3 | NA | 76579.02 | NA | 44010.16 | 17146.31 | NA | NA | 80199.58 | 41362.6 | 72273.86 | 32198.97 | 75858.83 | NA | 2064.747 | NA | 76292.57 |
| 19 | P26038 | MSN | 333342.6833 | 438752.3 | 421056.2 | 381249.5 | 241992.3 | 404349.8 | 164343.5 | 172028.6 | 446678.9 | 167923.7 | 367784.7 | 310472.5 | 404349.8 | 393512.7 | 292456.1 | 427428 | 390317.5 | 244865.7 | 273261.7 | 446678.9 | 404349.8 | 306071.8 | 222387.2 | 423963.5 | 191537 | 182241.6 | 441856.5 |
| 20 | P09104 | ENO2 | NA | 144058.2 | NA | 184650.5 | NA | 137596.7 | 126146.3 | 21831.56 | NA | NA | NA | 119650.8 | NA | 404349.8 | NA | 48438.29 | 57080.76 | NA | NA | 151558.9 | NA | 181096.8 | NA | 123793.9 | 2064.747 | NA | NA |
| 21 | P07148 | FABP1 | 1219163.714 | 34579.48 | 861796.3 | NA | 940142 | NA | 1804.538 | NA | 1130692 | NA | 1057986 | NA | 789446.1 | NA | 221565 | NA | NA | NA | 1162786 | 32336.43 | 805128.4 | NA | 970053.3 | NA | 2064.747 | NA | 1300718 |
| 22 | Q96Q11 | TRNT1 | NA | NA | NA | NA | NA | NA | 1804.538 | NA | NA | NA | NA | NA | NA | NA | NA | NA | 37098.09 | 35565.03 | NA | NA | NA | NA | NA | NA | 2064.747 | NA | NA |
| 23 | O15083 | ERC2 | NA | NA | NA | 85740.42 | NA | NA | 1804.538 | NA | 83390.33 | NA | NA | NA | NA | NA | NA | 142306.8 | NA | NA | NA | NA | NA | 72396.48 | NA | NA | 2064.747 | NA | 70213.43 |
| 24 | Q15911 | ZFHX3 | NA | NA | 178745.3 | 393512.7 | 205865.1 | 682653.9 | 1804.538 | NA | 243050.1 | NA | 189860.5 | NA | NA | NA | NA | 457756.2 | NA | NA | NA | NA | NA | NA | NA | NA | 2064.747 | NA | 252846.2 |
| 25 | Q9BUR5 | APOO | 35479.70278 | NA | 27260.11 | 15459.06 | 40140.37 | NA | 1804.538 | 46154.89 | 30730.15 | 54737.36 | 47185.33 | 13642.38 | 28517.17 | NA | 40140.37 | NA | NA | 10649.17 | 34436.2 | NA | 36956.08 | 16653.18 | 47858.24 | NA | 2064.747 | 33003.64 | 20057.06 |
| 26 | Q9UJ83 | HACL1 | 417999.9306 | NA | 435248.4 | NA | 336790.8 | 227161.7 | 1804.538 | 174111.8 | 276628.6 | NA | 274264.6 | NA | 317227.1 | 271920.4 | 336790.8 | NA | NA | 372485.6 | 446678.9 | NA | 390317.5 | NA | 307205 | 211073.8 | 2064.747 | 169817.6 | 333342.7 |
| 27 | Q8WUM4 | PDCD6IP | 50008.50556 | 34991.44 | 70504.27 | 50108.55 | 59047.33 | 41611.18 | 84319.78 | 97140.59 | 56715.96 | 134561.7 | 52110.31 | 61553.77 | 67555.47 | 65262.99 | 68597.03 | 59827.38 | 73200.35 | 75049.44 | 64108.37 | 40359.89 | 70903.29 | 49636.31 | 49821.32 | 37258.59 | 76579.02 | 76685.11 | 37386.23 |
| 28 | P53597 | SUCLG1 | 387432.1583 | 99433.59 | 228946.3 | 94932.09 | 310472.5 | 150524.5 | 187002.3 | 299487.5 | 275420.7 | 308775.7 | 299487.5 | 101732.7 | 245595.9 | 108554.7 | 270810.9 | 89524.72 | 192915.6 | 276628.6 | 357417.6 | 96737.9 | 205171.6 | 95793.82 | 288001.8 | 162300.5 | 193664.8 | 299487.5 | 245595.9 |
| 29 | O00186 | STXBP3 | NA | 28468.21 | NA | NA | NA | 19019.68 | 1804.538 | NA | NA | NA | NA | 21949.83 | NA | NA | NA | NA | NA | NA | 15575.29 | 29005.53 | NA | NA | NA | NA | 2064.747 | NA | NA |
| 30 | Q8N335 | GPD1L | 52415.71111 | NA | 59328.51 | NA | 54240.61 | 21949.83 | 109838.9 | 91466.47 | 45427.61 | 109273.7 | 50443.03 | NA | 52700.48 | 22321.01 | 45502.32 | NA | 57623.34 | 41362.6 | 54737.36 | NA | 62380.69 | NA | 54839 | 23827.06 | 152627.3 | 71658.52 | 49636.31 |
| 31 | P08621 | SNRNP70 | 48594.65 | 51791.05 | 47269.07 | 86082.28 | 44306.32 | 53026.19 | 1804.538 | NA | 59432.1 | 54839 | 49636.31 | 60605.33 | 52477.21 | NA | NA | 72977.35 | 74546.25 | 82242.07 | 33003.64 | 60605.33 | 49636.31 | 93224.91 | NA | 56917.54 | 2064.747 | NA | 50797.63 |
| 32 | Q969V6 | MKL1 | NA | 91325.89 | 55954.92 | NA | 74269.09 | 80102.57 | 1804.538 | NA | 71906.43 | NA | NA | 152627.3 | 72497.5 | 72497.5 | 89662.88 | 51690.71 | 68707.95 | 41576.85 | 72021.55 | 92973.8 | NA | NA | NA | 88904.66 | 2064.747 | NA | NA |
| 33 | P08311 | CTSG | NA | NA | 46154.89 | NA | NA | 67879.78 | 1804.538 | NA | 53026.19 | NA | NA | 68927.99 | NA | NA | NA | NA | 218057.1 | 78414.15 | NA | NA | 46895.88 | NA | NA | 56514.53 | 66379.24 | NA | NA |
| 34 | Q9UKU7 | ACAD8 | 46053.91944 | 31797.32 | 50179.16 | NA | 64601.65 | NA | 75160.02 | 49228.15 | 44010.16 | 28070.84 | 41974.24 | NA | 41840.21 | NA | 42678.39 | NA | 24335.52 | 32270.84 | 46053.92 | NA | 49467.07 | NA | 61900.08 | NA | 2064.747 | 46053.92 | 44605.86 |
| 35 | Q86X76 | NIT1 | 75613.88611 | NA | 61068.98 | 63988.55 | 80199.58 | 69590.71 | 1804.538 | 55745.15 | 70389.43 | NA | 84009.8 | 75506.47 | 78547.77 | 84980.21 | 76153.19 | NA | 57523.94 | 40935.27 | 70713.02 | NA | 59540.84 | 70713.02 | 78753.85 | 73278.36 | 55745.15 | 58932 | 52415.71 |
| 36 | P05162 | LGALS2 | 33491.8 | NA | 35565.03 | NA | 52415.71 | 36825.06 | 1804.538 | 23560.07 | 18592.77 | NA | 36763.92 | 72761.18 | 35479.7 | 50008.51 | 24907.94 | NA | 16653.18 | 22730.31 | 34916.06 | NA | 30730.15 | NA | 32815.68 | 71139.86 | 2064.747 | NA | 25737.06 |
| 37 | P23946 | CMA1 | NA | NA | NA | NA | NA | NA | 1804.538 | NA | NA | NA | NA | NA | NA | NA | NA | NA | 61155.07 | 14049.16 | NA | NA | NA | NA | NA | NA | 53240.82 | NA | NA |
| 38 | P01834 | IGKC | 462133.8694 | 885197.1 | 692332.5 | 484624 | 296507.9 | 462133.9 | 1219164 | 319228.4 | 659554.4 | 351190.2 | 312295.6 | 524995.4 | 566103.9 | 692332.5 | 325019.6 | 494067.2 | 286640.3 | 263299.1 | 499422.8 | 1130692 | 706520.3 | 469971.2 | 322906.2 | 438752.3 | 913960 | 310472.5 | 643593 |
| 39 | P14868 | DARS | 12567.36389 | 110112 | 54554.37 | 136875.9 | 30209.1 | 121022.2 | 1804.538 | 114195.5 | 43350.86 | 95493.71 | 29430.84 | 182241.6 | 61667.11 | 201171.9 | 81193.99 | 247871.5 | 161420 | 94484.9 | 76929.26 | 114678.3 | 54839 | 177772 | 50108.55 | 141996.6 | 2064.747 | 95951.08 | 53026.19 |
| 40 | Q9H773 | DCTPP1 | NA | NA | NA | NA | NA | NA | 1804.538 | 46303.49 | NA | 11589.48 | NA | 27509.79 | NA | NA | NA | 26314.17 | 87070.11 | 74656.39 | NA | NA | NA | NA | NA | NA | 2064.747 | 22251.11 | NA |

Webb-Robertson, *JPR*, 14(5):1993-2001, 2015

# Missing values are not due mostly to low-abundance proteins

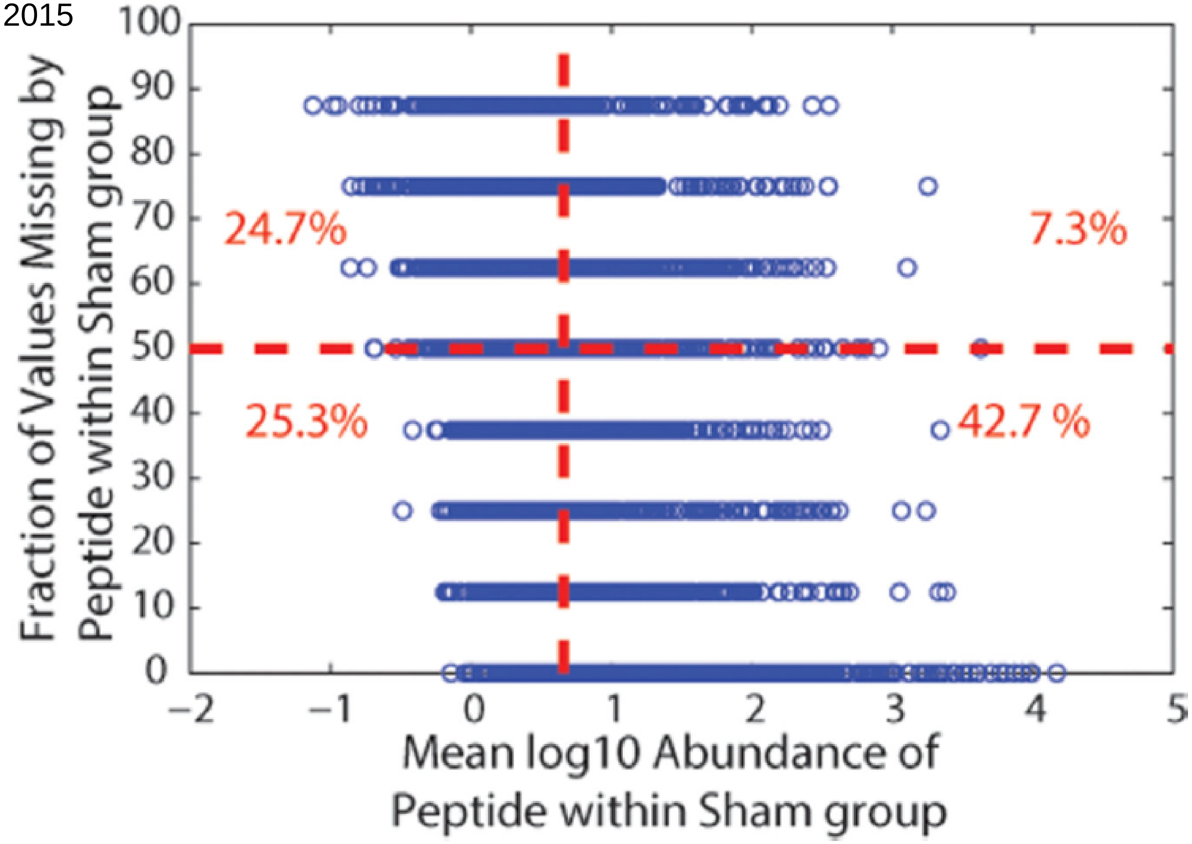
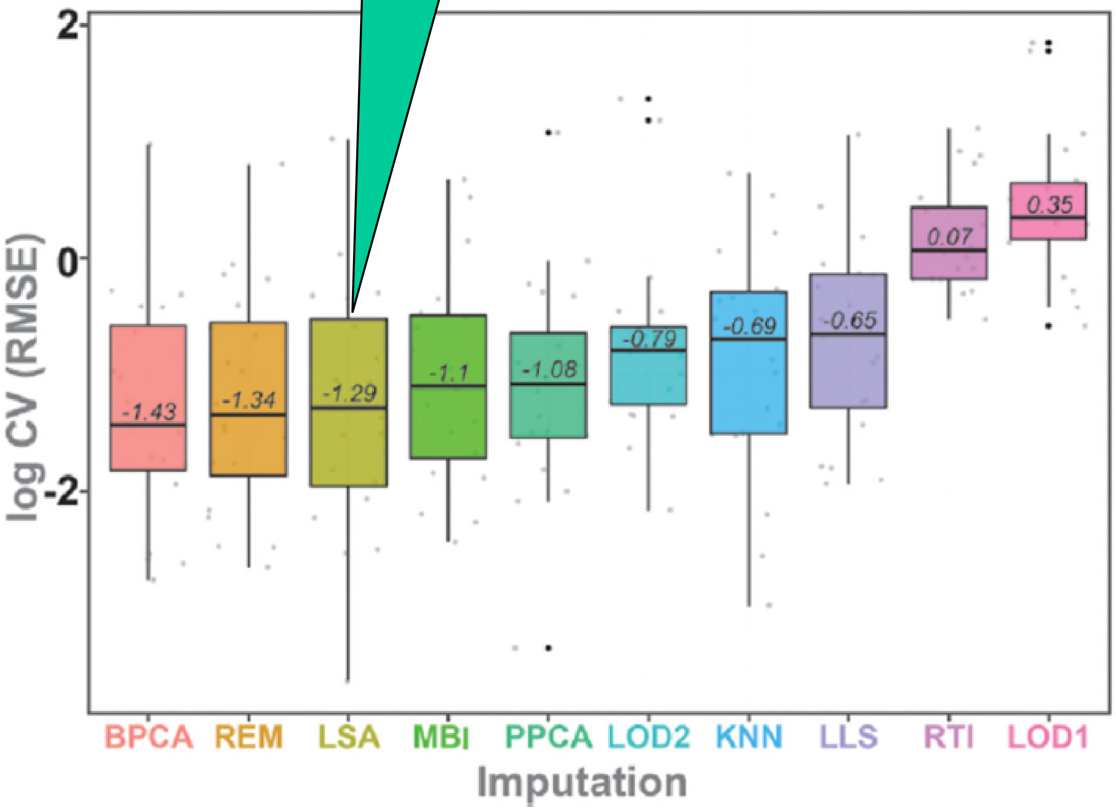


**Figure 1.**
Average $\log_{10}$ intensity as measured by peptide peak area in the control group versus fraction of missing values and peptide counts associated with bins corresponding to the fraction of missing data comparing phenotypes and exposures for datasets from (A) human plasma and (B) mouse lung. The control group for the human plasma is the normal glucose tolerant (NGT) samples, and the sham group for the mouse lung is the regular weight mice with no lipopolysaccharide (LPS) exposure. The vertical red line represents median average intensity, and the horizontal red line represents the point that 50% of the values are missing.

Webb-Robertson, *JPR*, 14(5):1993-2001, 2015

# Current imputation methods don't work very well



CV(RMSE) ~ 20% at 75th percentile

**Figure 2.**

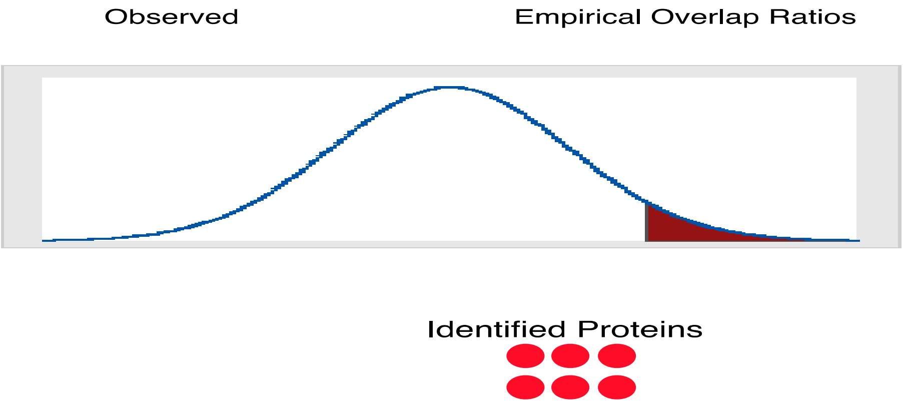
Boxplot of the average $\log_{10}$ CV(RMSE) for the imputed dilution series datasets (Table 1) at the (A) peptide and (B) protein levels. The lower line represents the 25th percentile, the upper line of the box represents the 75th percentile, and the inner line corresponds to the median $\log_{10}$ CV(RMSE).

# FCS

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

- **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**

Basic idea: 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.



# Other methods for rescuing missing proteins

- **CEA**
  - Generate cliques from PPIN
  - Rescue missing proteins from cliques containing lots of high-confidence proteins
  - Li et al. Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. Biol.*, 5:303, 2009

- **MaxLink**
  - Map high-confidence proteins (“seeds”) to PPIN
  - Rescue proteins that interact many seeds but few non-seeds
  - Goh et al. *Int J Bioinformatics Research and Applications*, 8(3/4):155-170, 2012

- **PEP**
  - Map high-confidence proteins to PPIN
  - Extract neighbourhood & predict protein complexes using CFinder
  - Rescue undetected proteins from high-ranking predicted complexes
  - Goh et al. A Network-based pipeline for analyzing MS data---An application towards liver cancer. *J. Proteome Research*, 10(5):2261-2272, 2011

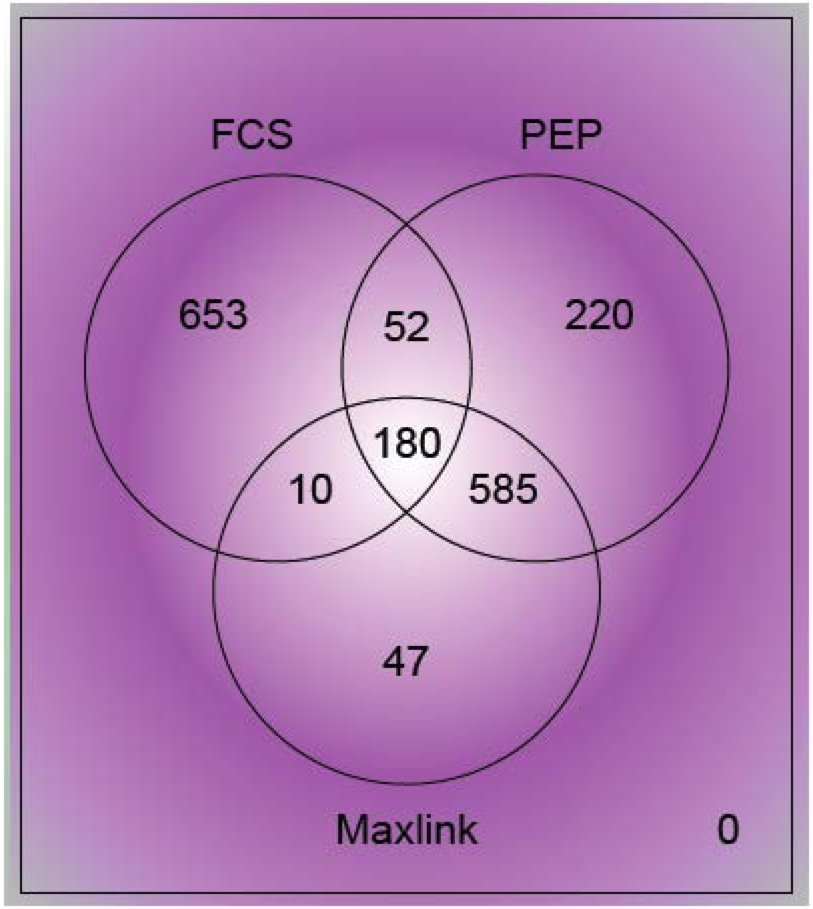
Goh et al. **Comparative network-based recovery analysis and proteomic profiling of neurological changes in valporic acid-treated mice**. *JPR*, 12(5):2116-2127, 2013

# iTRAQ experiment

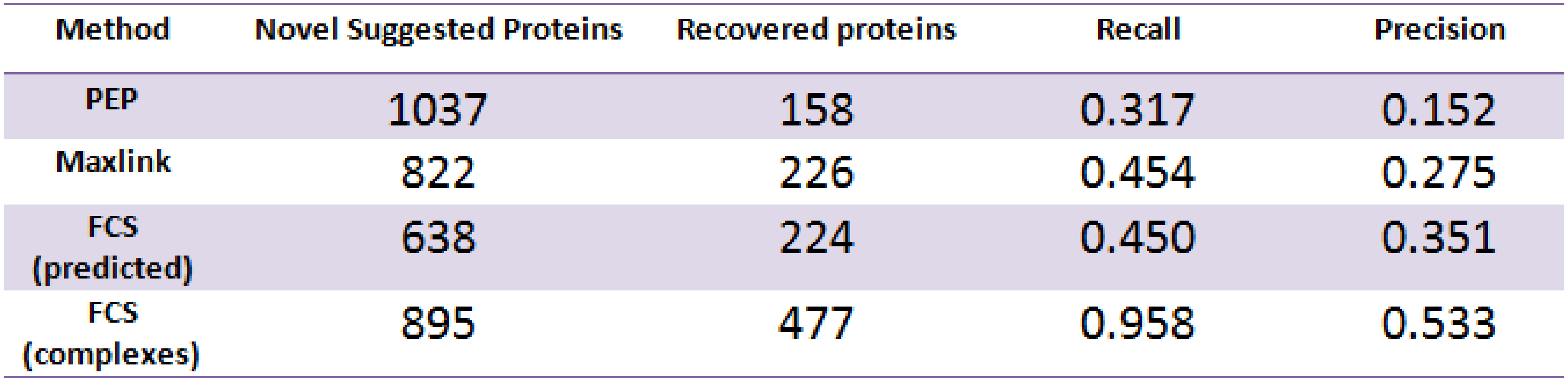
- **Valporic acid (VPA)-treated mice vs control**
  - VPA or vehicle injected every 12 hours into postnatal day-56 adult mice for 2 days
  - Role of VPA in epigenetic remodeling
- **MS was scanned against IPI rat db in round #1**
  - 291 proteins identified
- **MS was scanned against UniProtkb in round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

Moderate level of agreement of reported proteins between various recovery methods

FCS (Real Complexes)



# Performance comparison



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

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

# SWATH experiment

- **If there are technical replicates, they should have reported the same proteins. So we can run FCS on one replica, and see whether the predicted missing proteins show up in other replicas**
- **If there are multiple biological replicates (i.e. patients of the same phenotype), we can run FCS on one of them, and check on the others**
- **Proteomics data used: Renal cancer**
  - Guo et al. *Nature Medicine*, 21(4):407-413, 2015
  - 6 pairs of normal vs cancer ccRCC tissues
  - SWATH in duplicates

# ~20% of predicted missing proteins are supported by ≥1 reported peptide in the screen

# ~20% of predicted missing proteins are supported by ≥1 reported peptide in the replicate

# But ~25% of predicted missing proteins are supported by peptides in the screen or replicate

**C** st union of self and other replicate)

# Time for Exercise #1

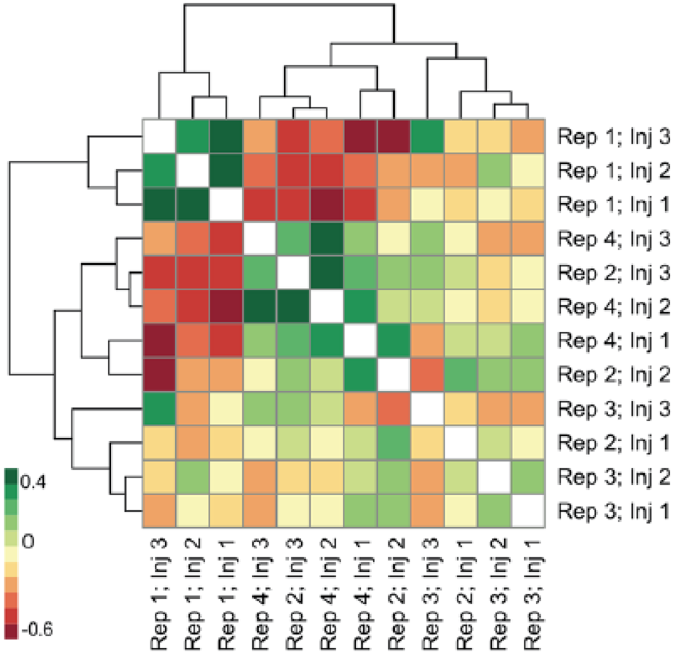
- **About 25% FCS-predicted missing proteins are supported by peptides in screen/replicate. Can we do better? In particular, suggest some ways to rank missing proteins, so that those which are more likely to be really present are ranked higher**

# IMPROVING CONSISTENCY IN PROTEOMIC PROFILE ANALYSIS

# Proteomic profiles generally not consistent, even for technical replicates

- **A human kidney tissue**
  - Guo et al. *Nature Medicine*, 21(4):407-413, 2015
  - Digested in quadruplicates
  - Analyzed in triplicates

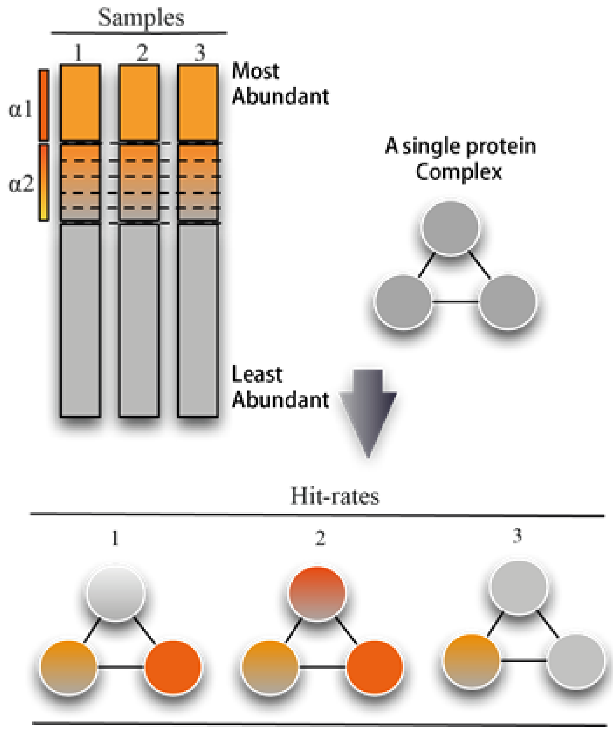
- **Clustering by proteins**
  - Correlation betw replicates is not good (~0.4)
  - Technical replicates of the same biological replicate are not tightly clustered



Goh et al. **Quantitative proteomics signature profiling based on network contextualization**. *Biology Direct*, 10:71. 2015

# qPSP

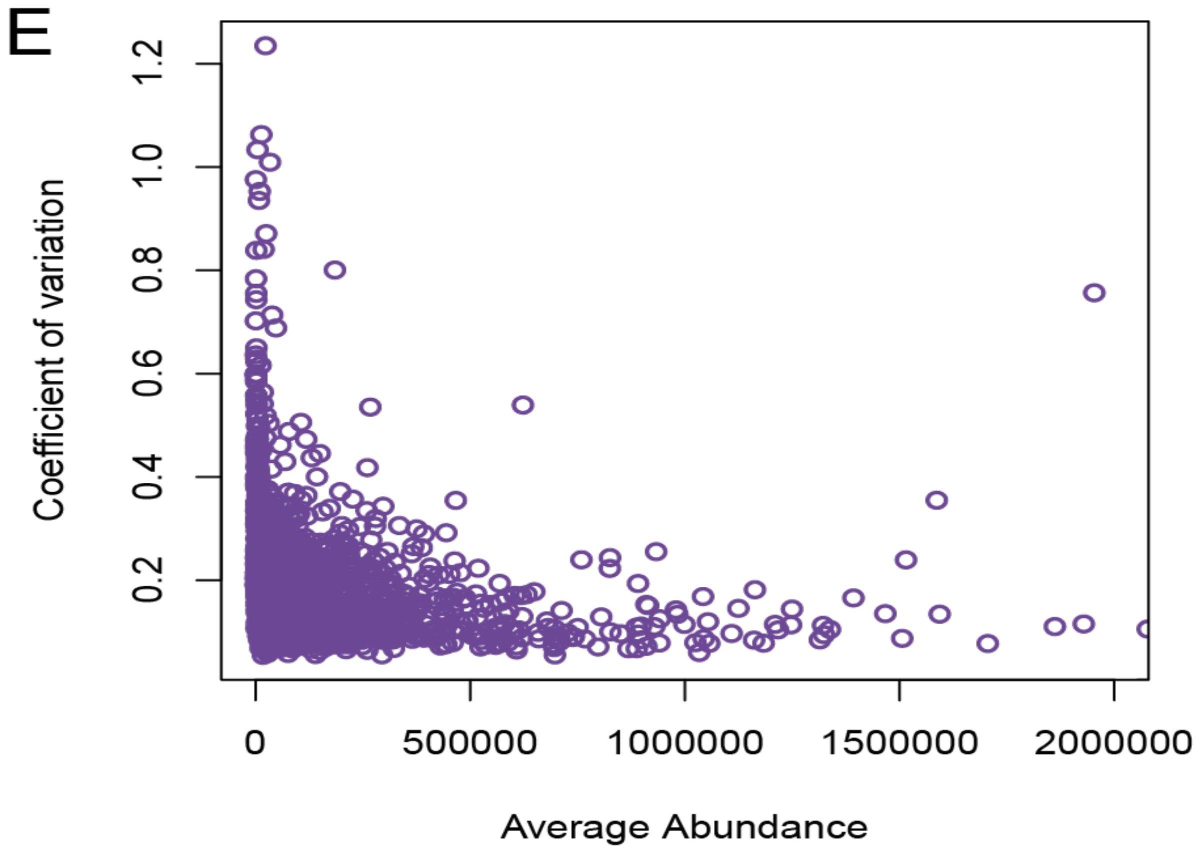
- **In a sample, assign weight to proteins as follow**
  - Most-abundant 10% of proteins, wt = 1
  - Proteins at 10-12.5%, wt = 0.8
  - Proteins at 12.5-15%, wt = 0.6
  - Proteins at 15-17.5%, wt = 0.4
  - Proteins at 17.5-20%, wt = 0.2
  - All other proteins, wt = 0



- **Hit rate of a complex C wrt a sample S is sum of the wt of proteins in C in S**
  - score(C,Si) = Σp∊C fs(p,Si) / |C|

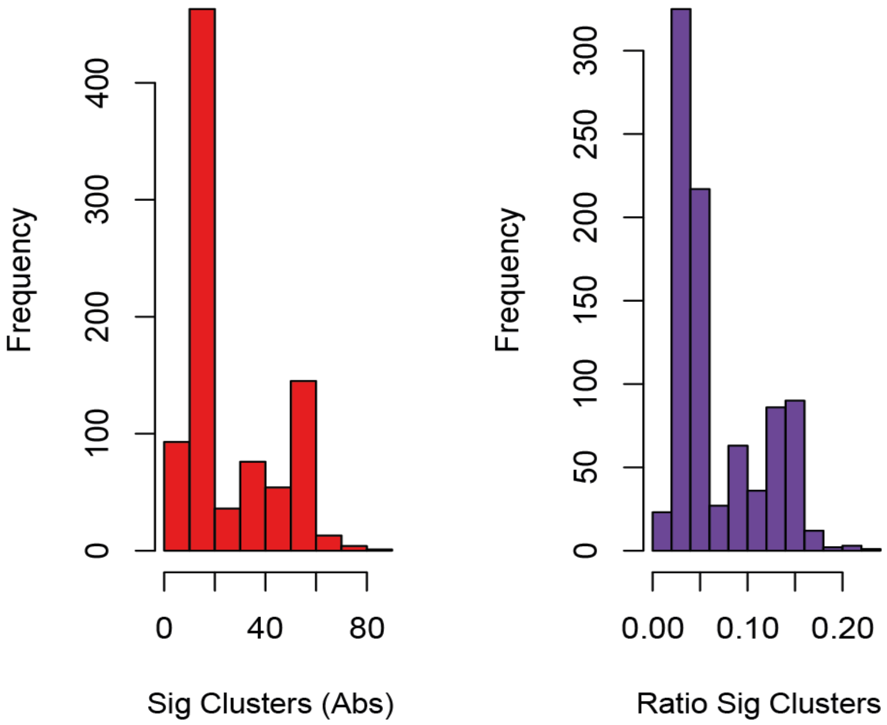
- **Complex C is significant if {score(C,$S_i$) | $S_i$ ∊ A} is very different by t-test from {score(C,$S_i$) | $S_i$ ∊ B}**

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



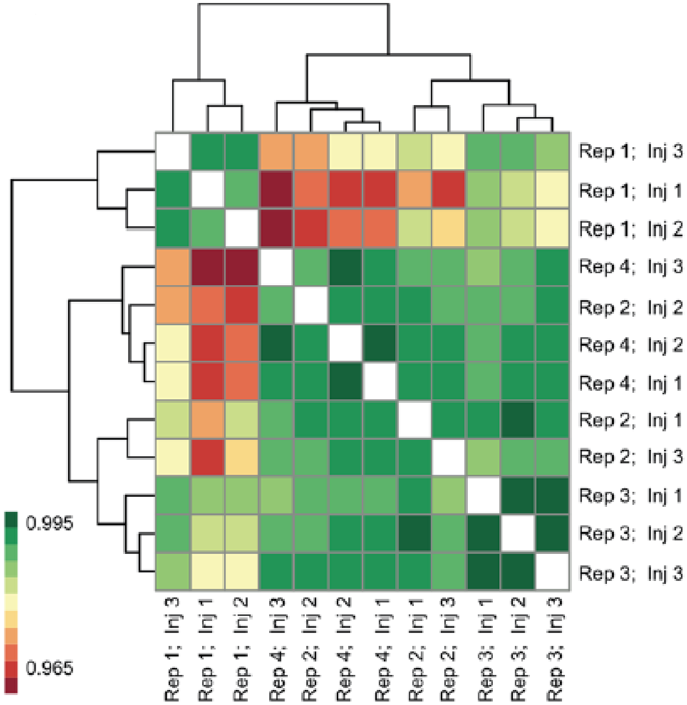
# False-positive analysis

- **12 kidney controls were randomly assigned into two groups of equal size, and qPSP analysis was performed 10000 rounds**
- **For each round, # of significant clusters (5% FDR) was determined.**
- **Histograms showed that the false-positive rates were well within the expectation levels**
  - Sig Clusters Abs E-level: 19, Observed level: 16
  - Ratio Sig Clusters E-level: 0.05, Observed level: 0.04



# Consistency of qPSP

- **Clustering of benchmarking control data based on protein complexes (i.e. qPSP)**
  - Correlation betw replicates is >0.95
    - **Cf. 0.4 based on proteins**
  - Technical replicates are better clustered



# Application to renal & colorectal cancers

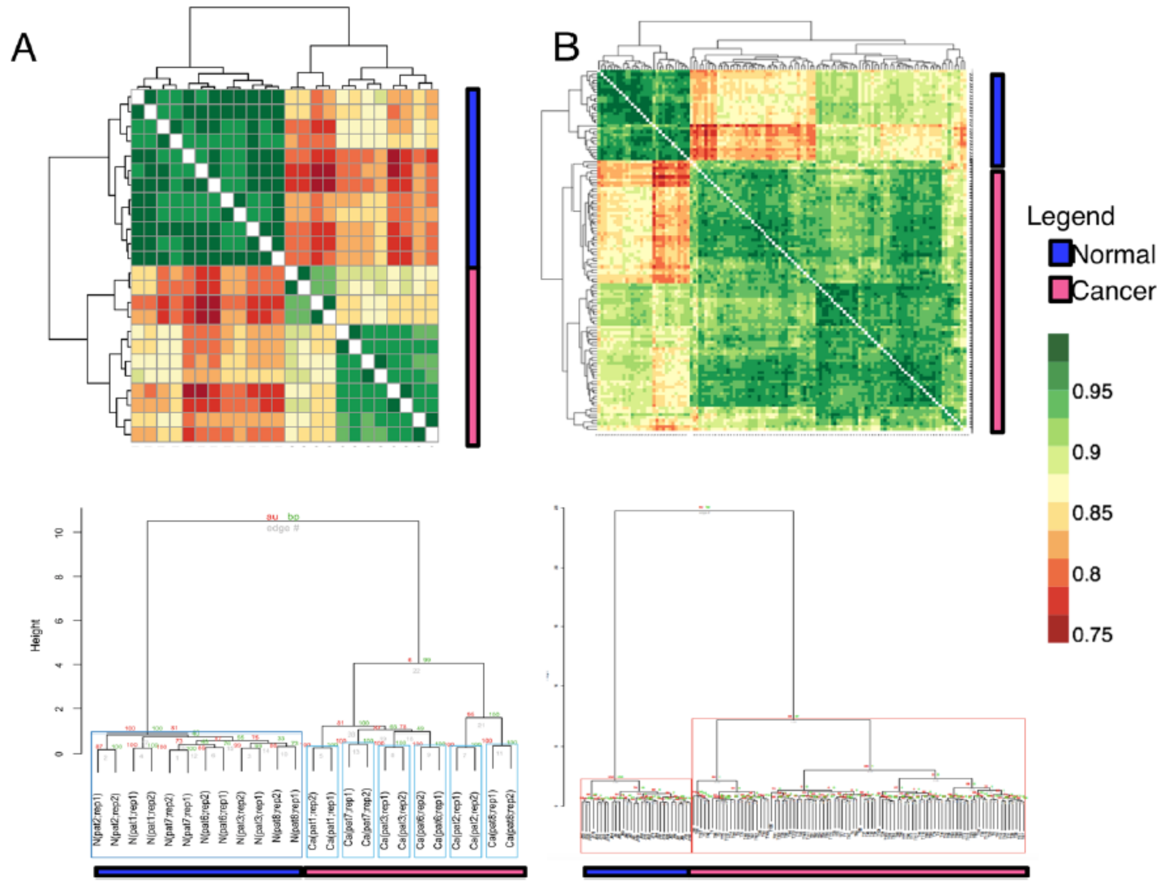


**Fig. 3** qPSP strongly discriminates sample classes for renal cancer (**a**) and colorectal cancer (**b**). Clustered similarity maps at the top row showed specific and consistent segregation of non-cancer and cancer samples. The trees below the heatmaps are from bootstrap analysis (PVCLUST), which demonstrates that the discrimination between sample classes based on qPSP hit-rates is highly stable

# ESSNET: A QUANTUM LEAP?

# ESSNet, adapted for proteomics

- Let $g_i$ be a protein in a given protein complex
- Let $p_j$ be a patient
- Let $q_k$ be a normal

- Let $\Delta_{i,j,k} = \text{Expr}(g_i,p_j) - \text{Expr}(g_i,q_k)$

- 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

⇒ Potential to reliably detect low-abundance but differential proteins

Goh & Wong. **Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms**. *Journal of Proteome Research*, 15(9):3167--3179, 2016

# Five methods to compare with

- **Network-based methods**
  - Hypergeometric enrichment (HE)
  - Direct group analysis (DG), similar to GSEA
  - qPSP, Goh et al., *Biology Direct*, 10:71, 2015
  - PFSNET, Goh & Wong, *JBCB,* 14(5):16500293, 2016
  - ESSNET, Lm et al., *JBCB*, 13(4):1550018, 2015 & Goh & Wong, *JPR*, 15(9):3167-3179 2016

- **Standard t-test on individual proteins (SP)**

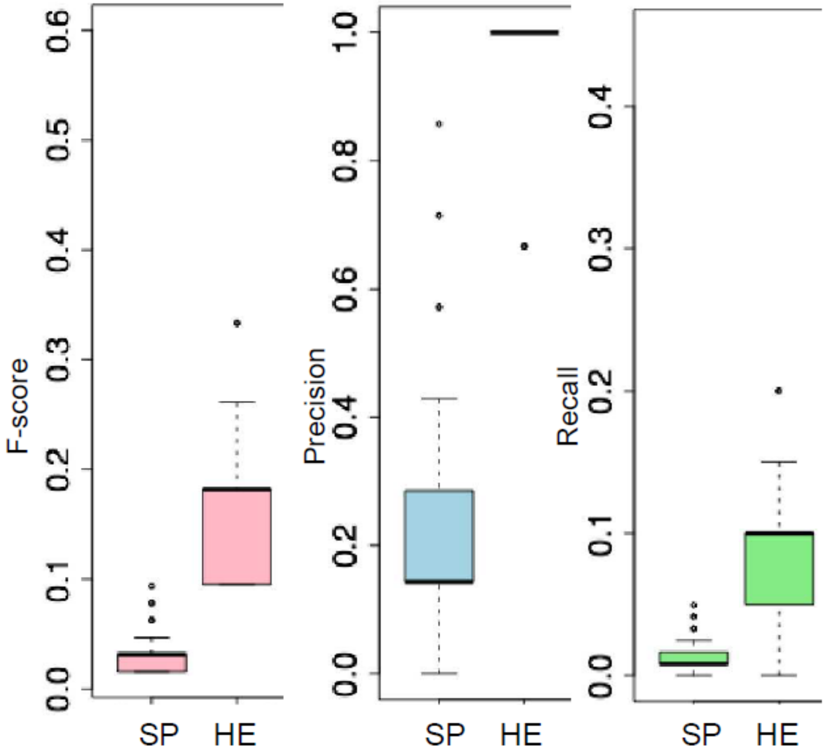
Langley & Mayr, *J. Proteomics*, 129:83-92, 2015

# Simulated data

- **Simulated datasets from Langley and Mayr**
  - D.1.2 is from study of proteomic changes resulting from addition of exogenous matrix metallopeptidase (3 control, 3 test)
  - D2.2 is from a study of hibernating arctic squirrels (4 control, 4 test)
- **Both D1.2 and D2.2 have 100 simulated datasets, each with 20% significant features**
  - Effect sizes of these differential features are sampled from one out of five possibilities (20%, 50%, 80%, 100% and 200%), increased in one class and not in the other
- **Significant artificial complexes are constructed with various level of purity (i.e. proportion of significant proteins in the complex)**
  - Equal # of non-significant complexes are constructed as well

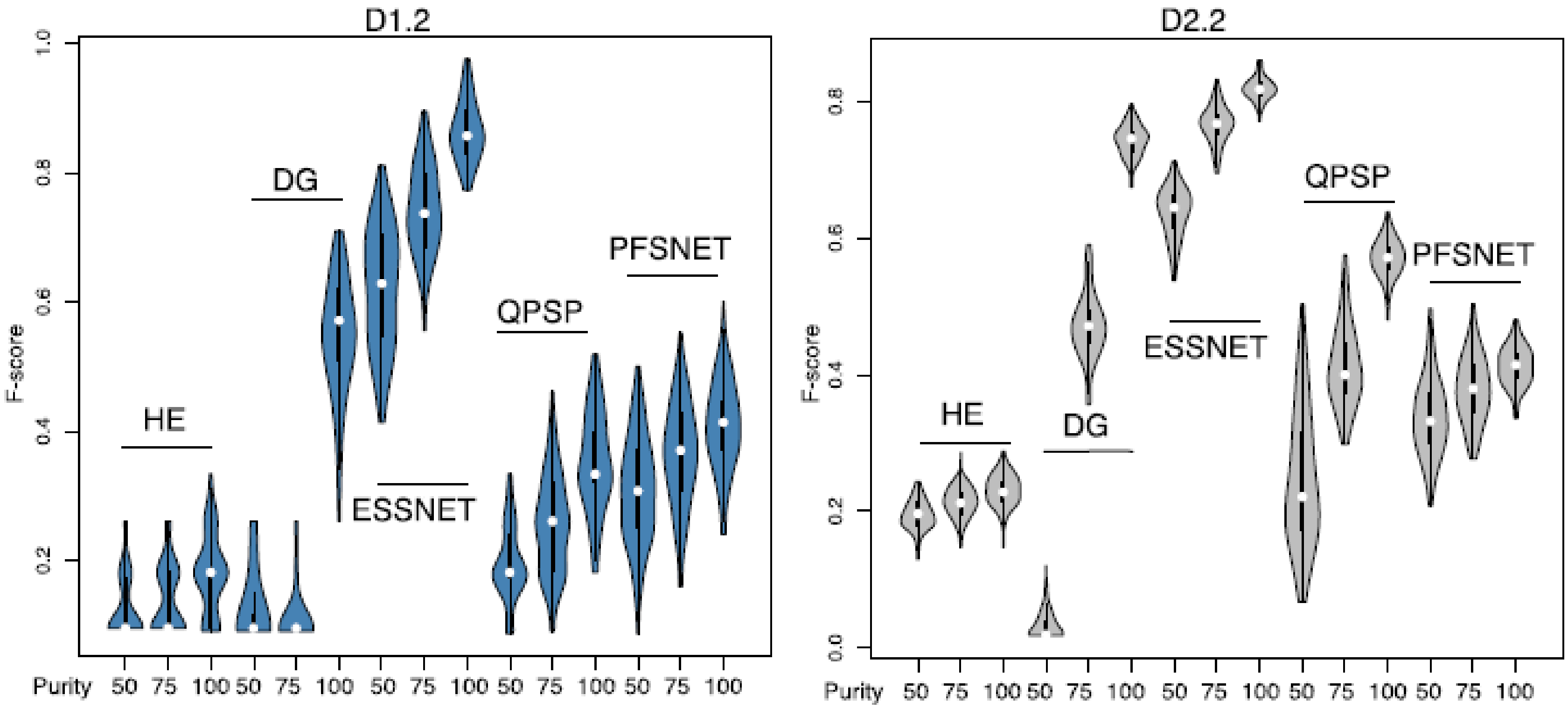
# SP shows poor performance on simulated data.

# Can network-based methods do better?



**Supplementary Figure 1 Single protein (SP) precision-recall performance on D1.2.** The f-score (pink), precision (blue) and recall (green) shows that SP performs abysmally on simulated data. HE is shown next to SP as a reference.

# ESSNET shows excellent recall/precision on simulated data



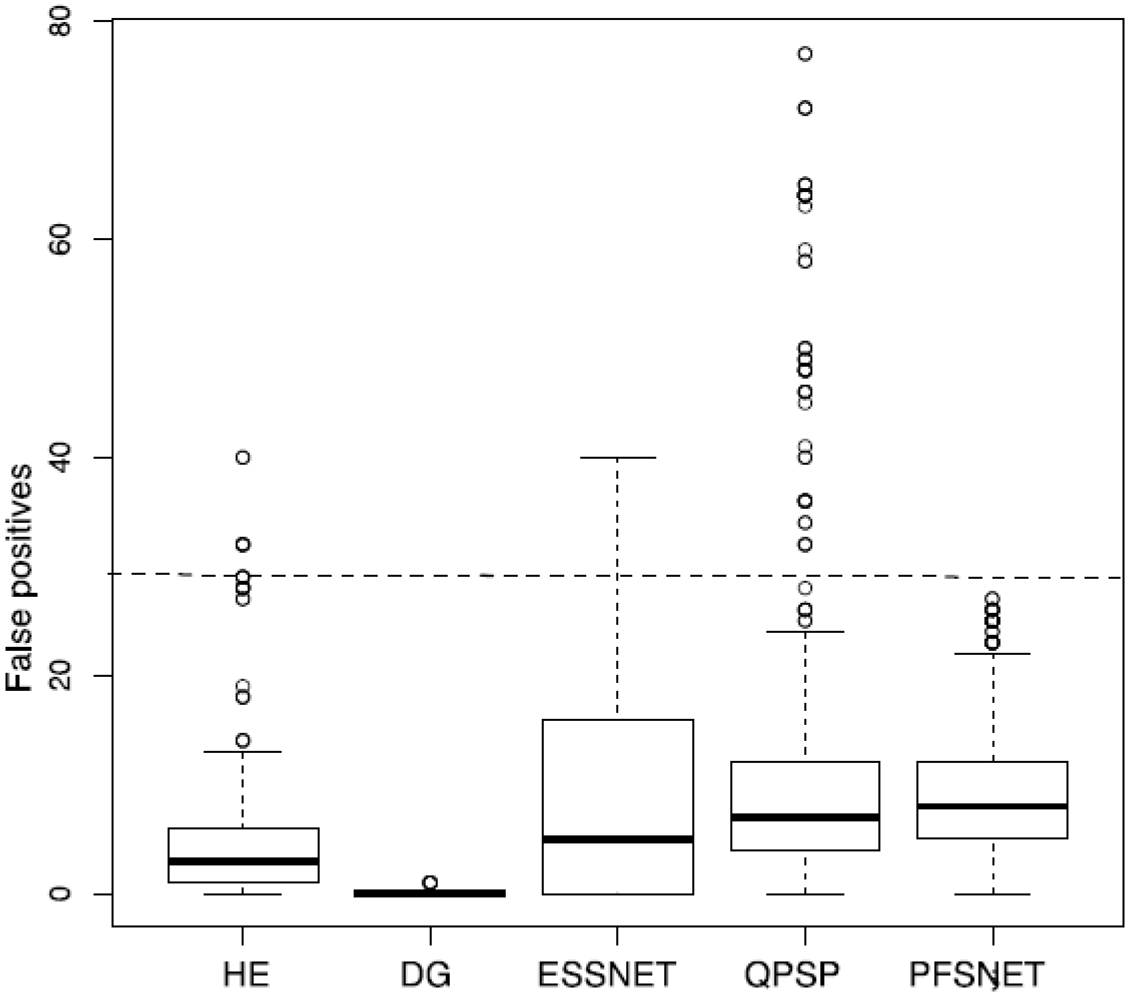
Guo et al. *Nature Medicine*, 21(4):407-413, 2015

# Renal cancer control data (RCC)

- **12 runs originating from a human kidney tissue digested in quadruplicates and analyzed in triplicates**

- **Excellent for evaluating false-positive rates of feature-selection methods**
  - Randomly split the 12 runs into two groups. Report of any significant features between the groups must be false positives

All methods control false positives well



Dash line corresponds to expected # of false positives at alpha 0.05 (~30 complexes)

Guo et al. *Nature Medicine*, 21(4):407-413, 2015

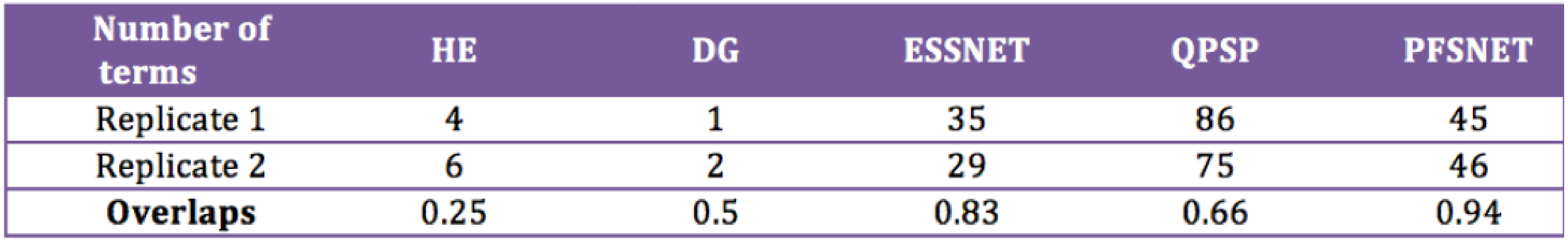
# Renal cancer data (RC)

- **12 samples are run twice so that we have technical replicates over 6 normal and 6 cancer tissues**

- **Excellent opportunity for testing reproducibility of feature-selection methods**
  - A good method should report similar feature sets between replicates

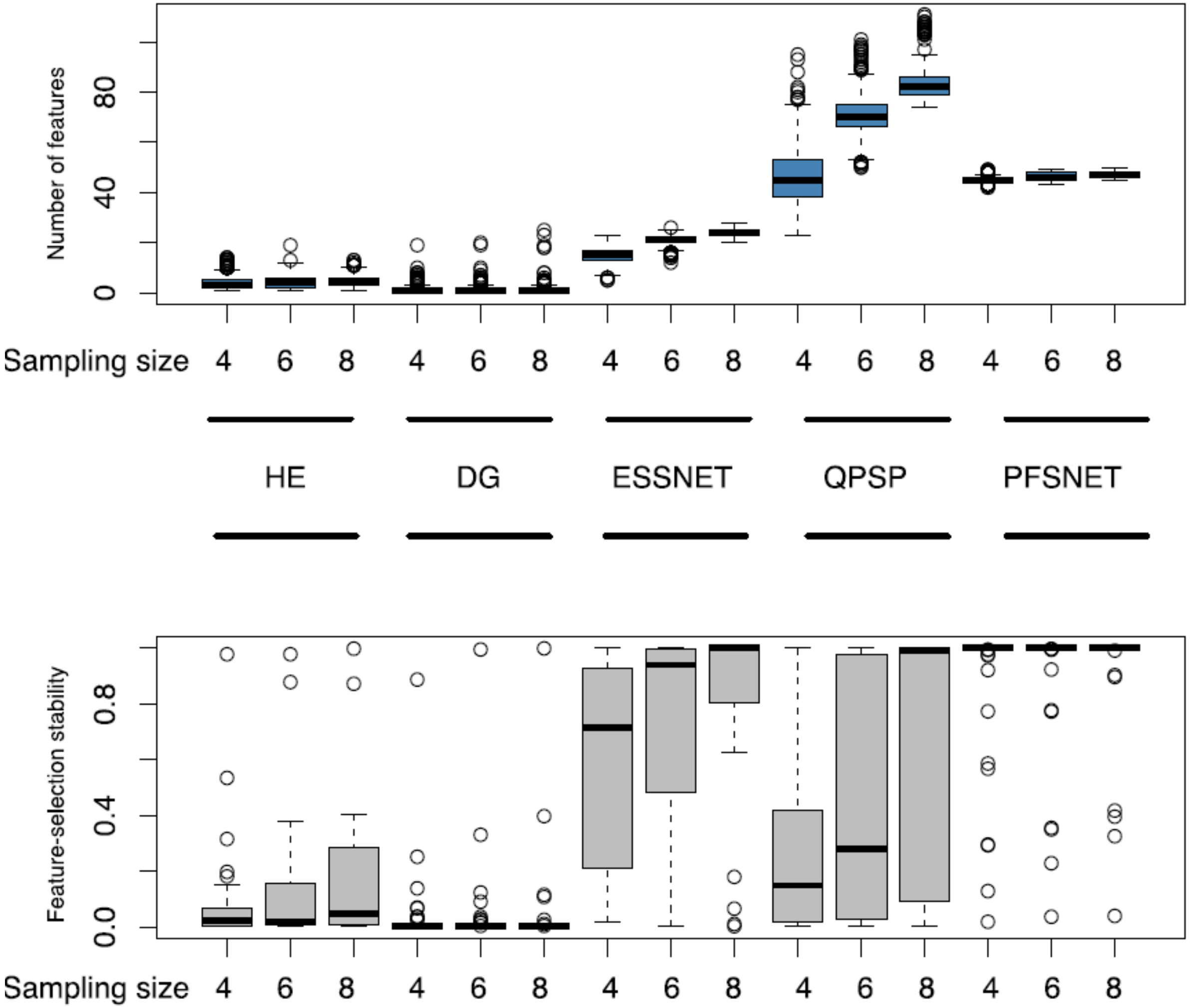
- **Can also test feature-selection stability**
  - Apply feature-selection method on subsamples and see whether the same features get selected

# ESSNET & PFSNET show excellent reproducibility

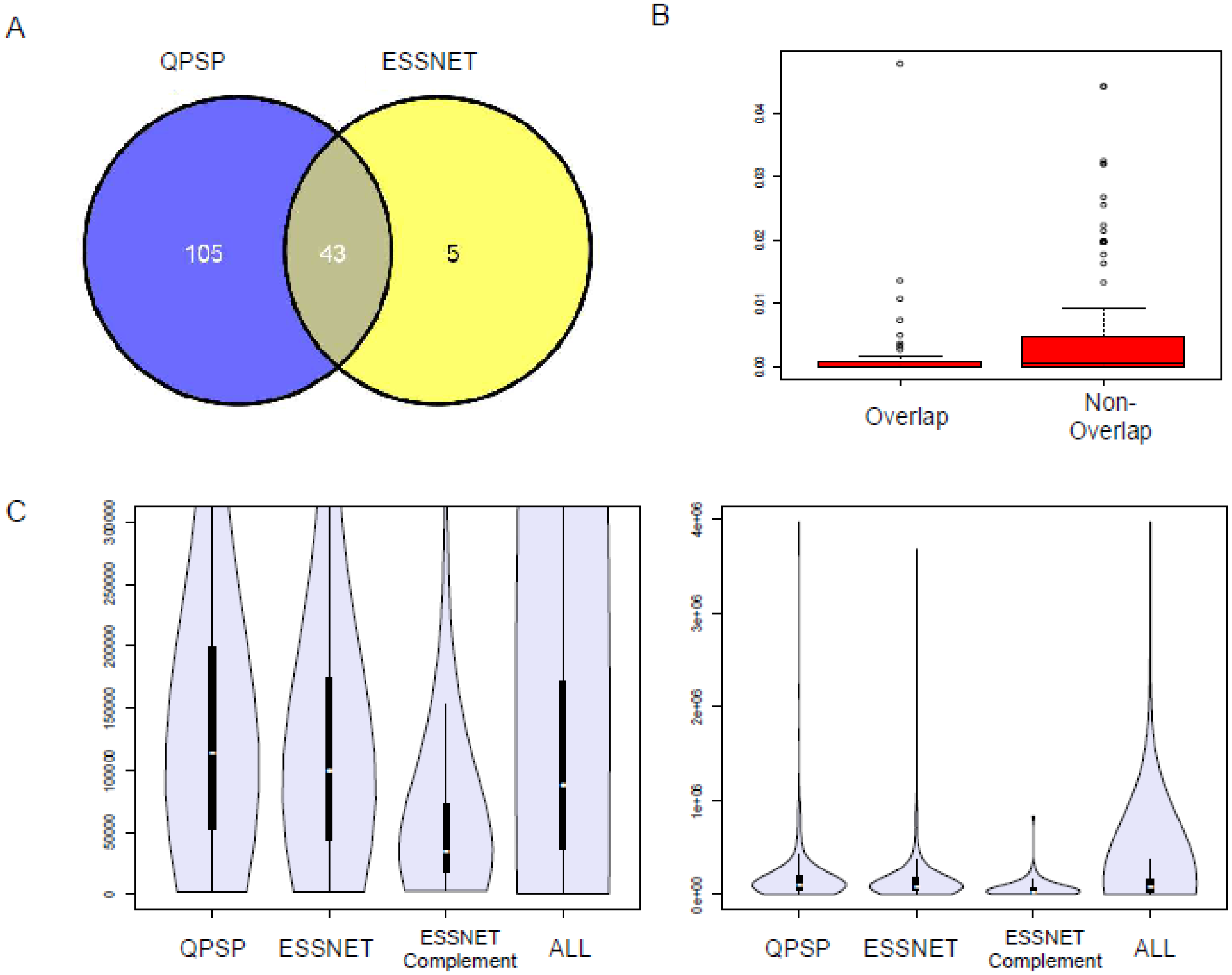


| Number of terms | HE | DG | ESSNET | QPSP | PFSNET |
|---|---|---|---|---|---|
| Replicate 1 | 4 | 1 | 35 | 86 | 45 |
| Replicate 2 | 6 | 2 | 29 | 75 | 46 |
| **Overlaps** | 0.25 | 0.5 | 0.83 | 0.66 | 0.94 |

# ESSNET & PFSNET show excellent stability



# ESSNET can assay low-abundance complexes that qPSP cannot

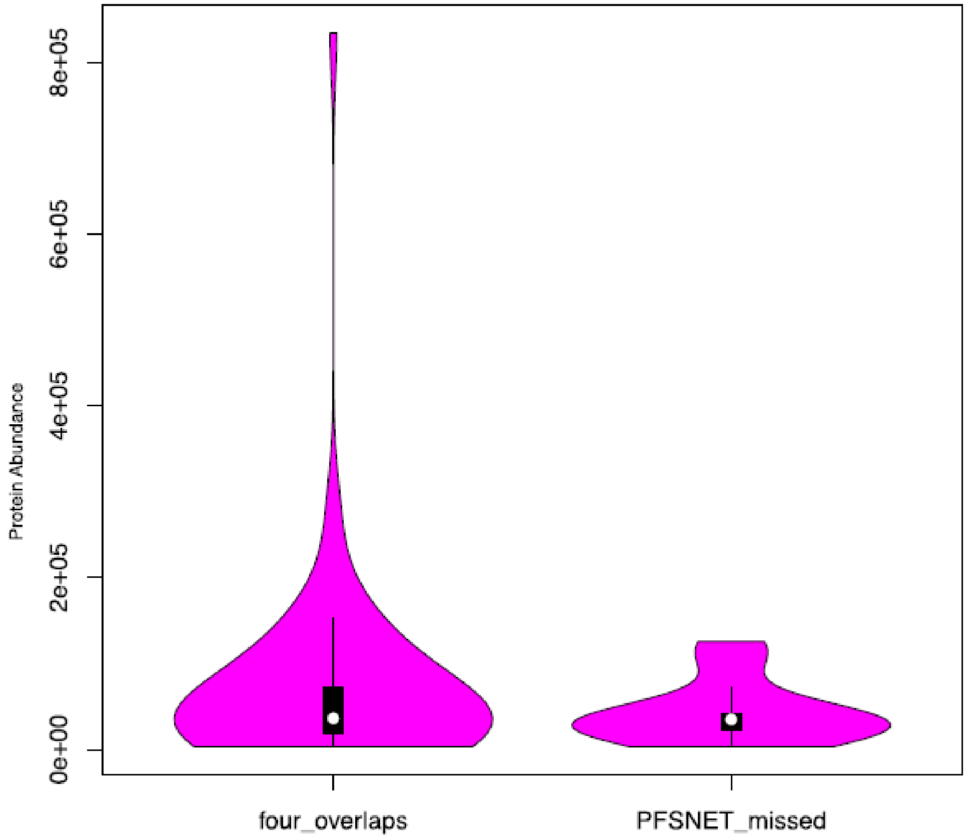


A: QPSP-ESSNET significant-complex overlaps

B: P-value distribution for overlapping and non-overlapping QPSP complexes.

C: Sampling abundance distribution. The left panel is a zoom-in of the right. The y-axis is the protein abundance while the four categories are the distribution of abundances of complexes found in QPSP, ESSNET, ESSNET unique (complement), and all proteins in RC.

# ESSNET can assay low-abundance complexes that PFSNET cannot



Of the 5 ESSNET-unique complexes, PFSNET can detect 4; the missed complex consists entirely of low-abundance proteins.

If p-value threshold is adjusted by Benjamini-Hochberg 5% FDR, PFSNET can detect only 3 of the 5 ESSNET-unique complexes while ESSNET continues to detect them all.

# CONCLUDING REMARKS

# In conclusion…

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

# Acknowledgements

- **The slides on peptide identification were adapted from those given to me by A/P Leong Hon Wai**
- **A lot of the slides on PSP, qPSP, and ESSNet came from the work of Wilson Goh**

**Leong Hon Wai**

**Wilson Goh**

# Must read

- Steen & Mann. **The ABC's and XYZ's of peptide sequencing.** *Nature Reviews Molecular Cell Biology*, 5:699-711, 2004
- Cottrell. **Protein identification using MS/MS data**. *Journal of Proteomics*, 74:1842-1851, 2011
- [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*, 10:71, 2015
- [ESSNET] Goh & Wong. **Advancing clinical proteomics via analysis based on biological complexes: A tale of five paradigms**. *Journal of Proteome Research*, 15(9):3167-3179, 2016

# Good to Read

- Frank, et al. **De Novo Peptide Sequencing and Identification with Precision Mass Spectrometry.** *J. Proteome Res*. 6:114-123, 2007
- Sung. **Chap. 12: Peptide sequencing**. *Algorithms in Bioinformatics: A Practical Introduction*. CRC Press, 2010
- Käll & Vitek. **Computational mass spectrometry–based proteomics**. *PLoS Comput Biol* , 7(12): e1002277, 2011