NANYANG
TECHNOLOGICAL
UNIVERSITY

Statistics for Proteomics

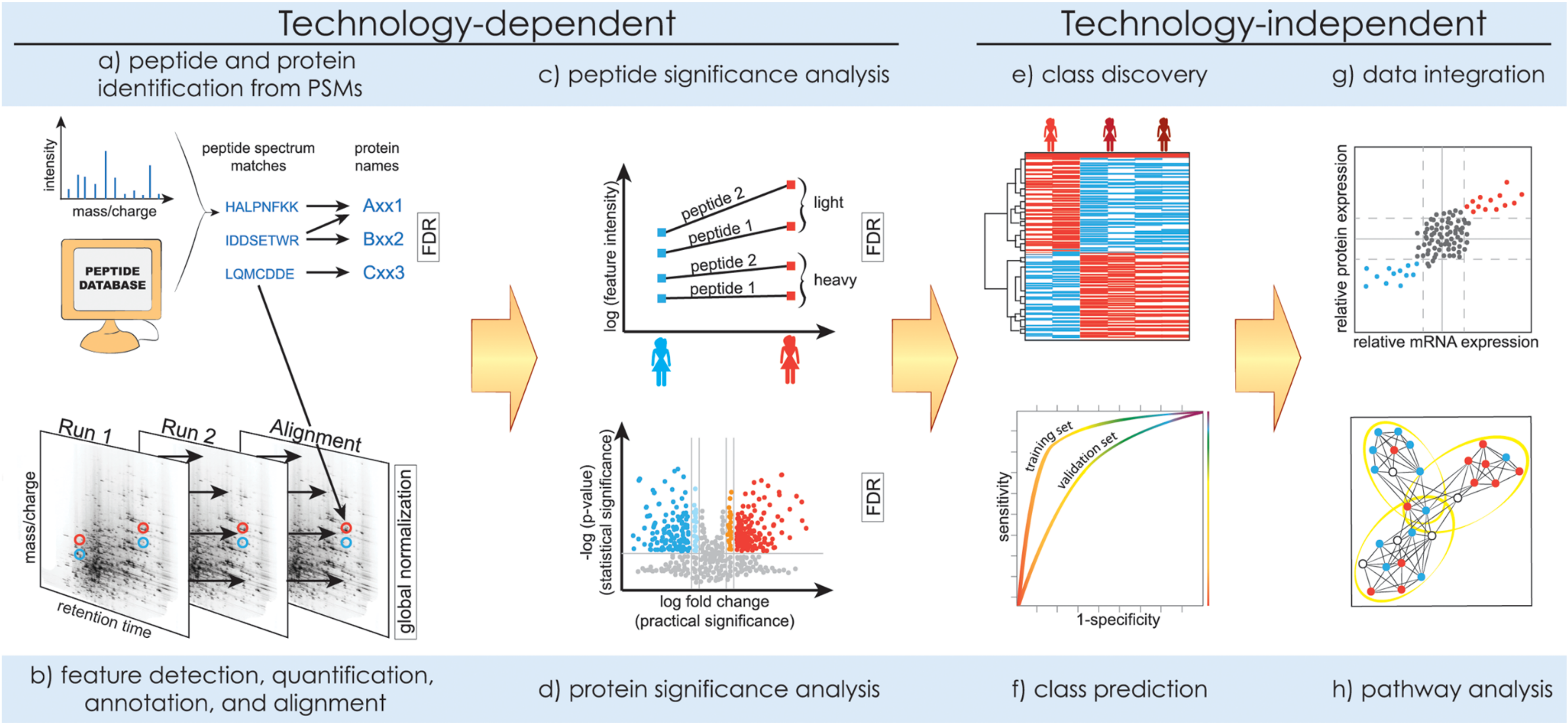
Wilson Wen Bin Goh
School of Biological Sciences, Nanyang Technological University

23 November 2017

Learning objectives

- Describe the various factors affecting PSM quality
- Describe p-values, FDR and PEP
- Describe and evaluate the various decoy library generation strategies (sequence reversal, sequence randomization) for FDR estimation
- Describe coverage and consistency issues in proteomics

Using Proteomics in practical applications



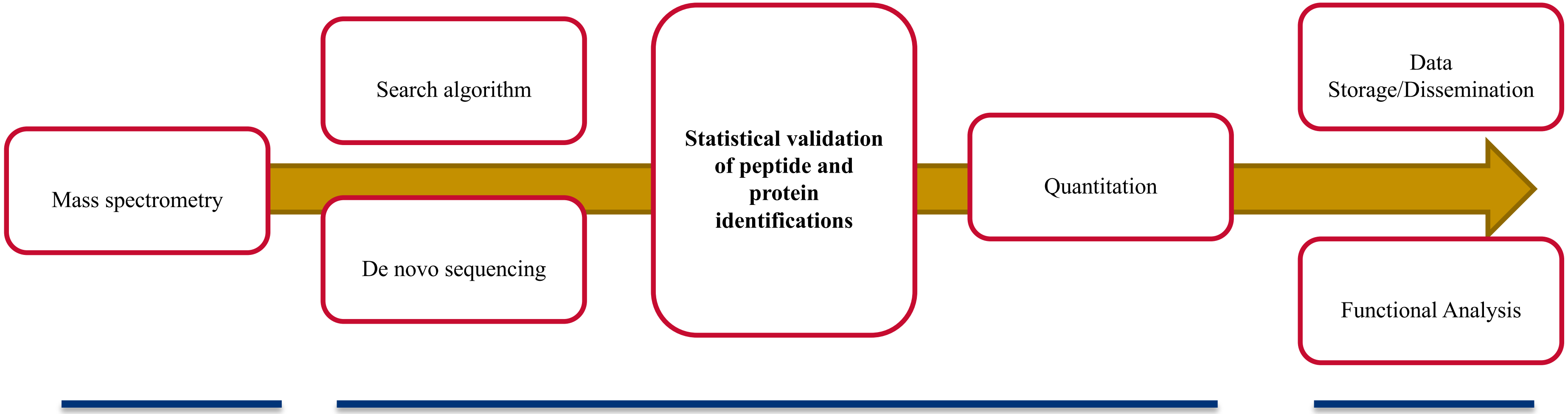
Match Significance

Functional Analysis

Statistics plays key roles in both areas

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

Overview of proteo-informatics



Mass spectrometry

Search algorithm

De novo sequencing

Statistical validation of peptide and protein identifications

Quantitation

Data Storage/Dissemination

Functional Analysis

Data storage/ representation formats

Peptide-spectra matching (PSM)

Functional analysis

What is PSM?

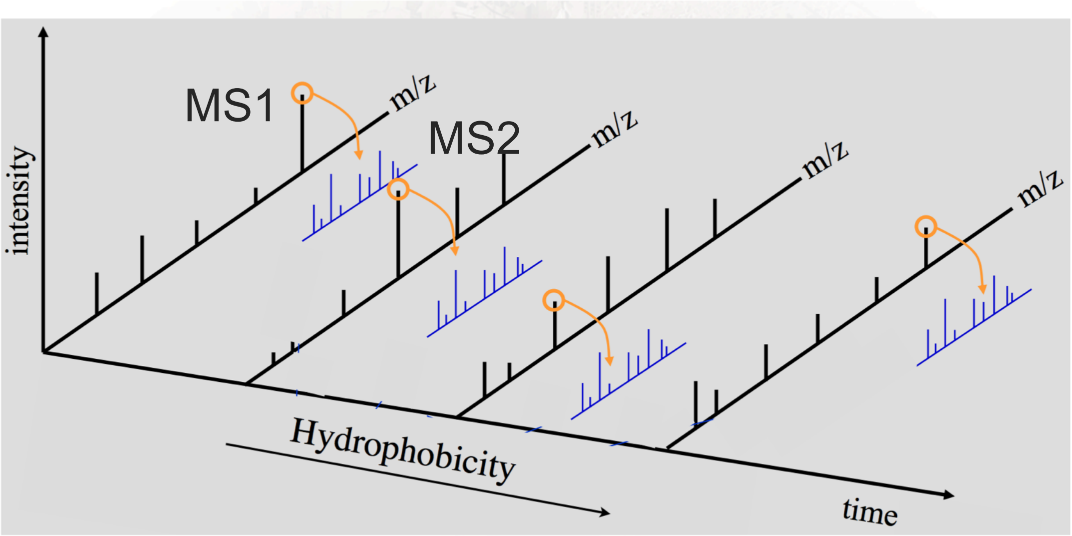
- PSM stands for Peptide-Spectra Match
- It is a pairing of sequence with spectra
- You’ve seen this earlier when we considered library search algorithms
- Earlier we showed a simple scenario where there is only one possible sequence match per spectra…
- But in practice…

What determines whether or not we get a good PSM?

- Search parameters
- Library quality and size
- Spectra quality
- Algorithm scoring method
- Statistical evaluation

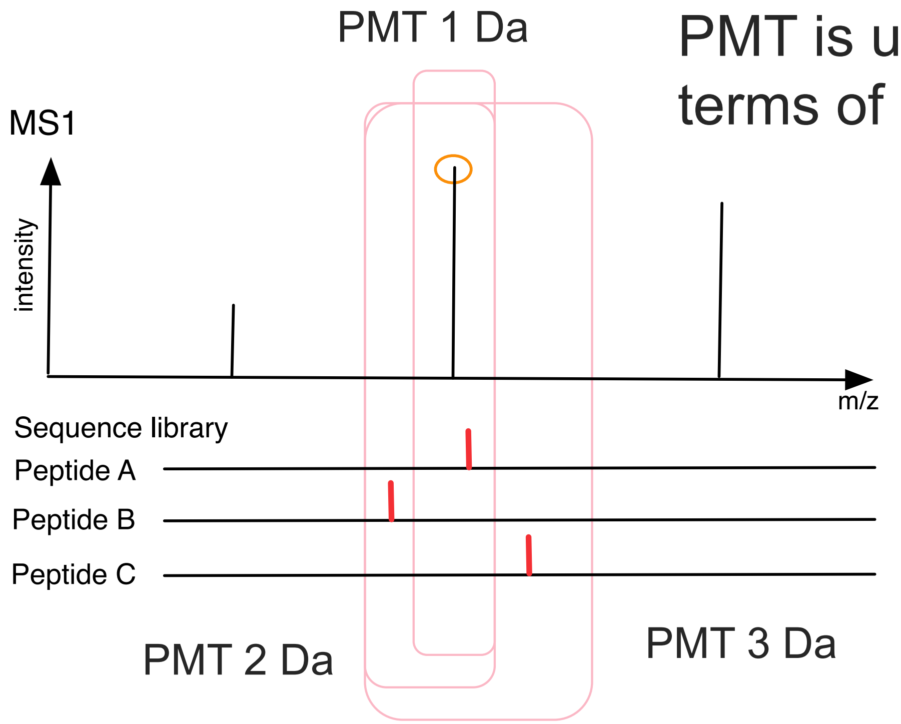
Search parameters

- Precursor mass tolerance (PMT)
- Fragment mass tolerance (FMT)
- Post-translational modification (PTM)



Precursor mass tolerance (PMT)

- PMT deals with MS1 (peptide level)

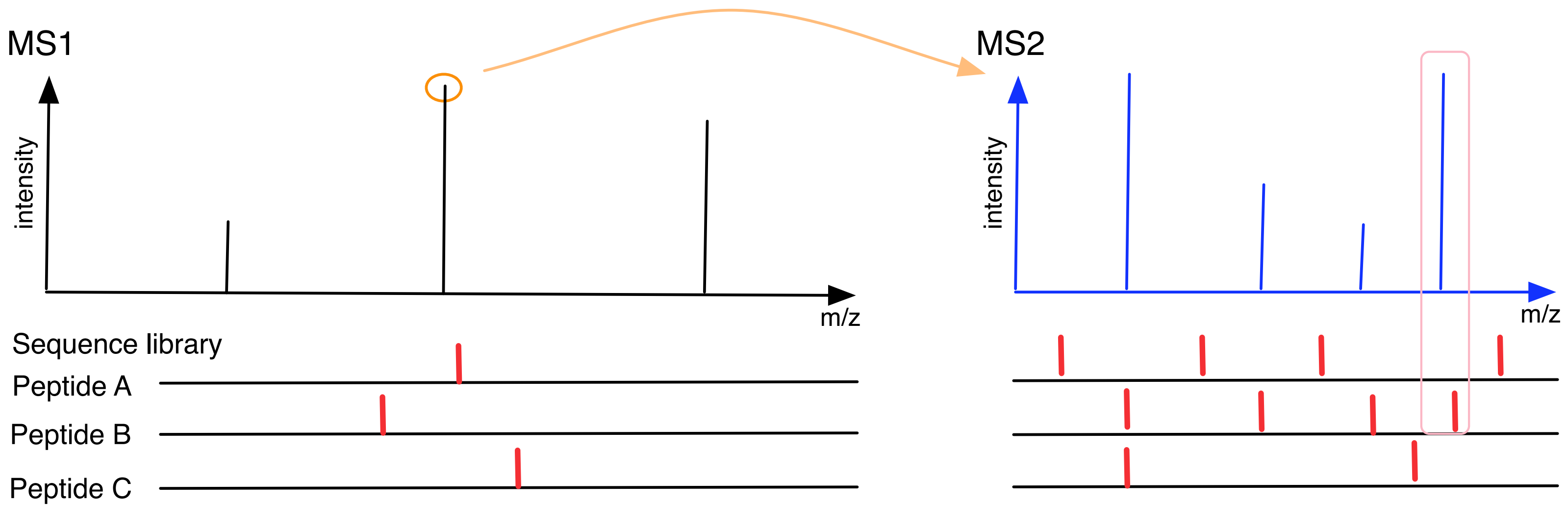


PMT is usually specified in terms of daltons

What happens when the PMT window size is increased?

Fragment mass tolerance (FMT)

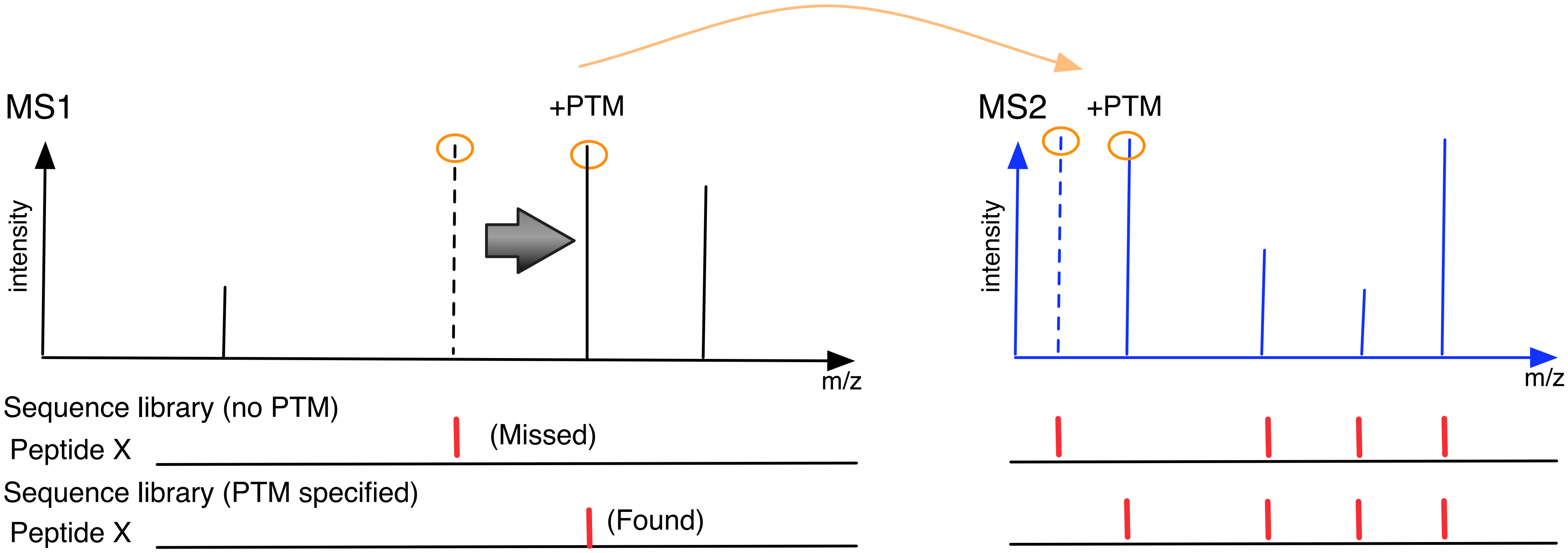
- FMT deals with MS2 (identification level)



Which looks like the correct answer?

Post-translational modification (PTM)

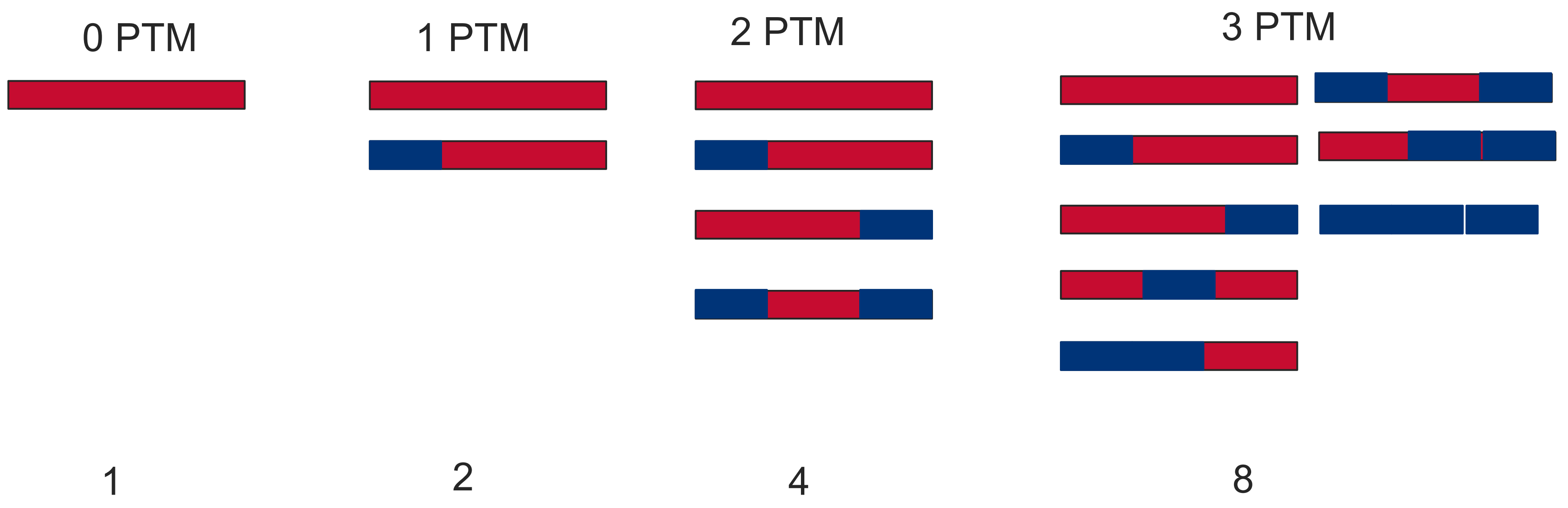
- Some peptide sequences undergo chemical modifications resulting in mass shifts
- If the PTMs are not specified in the search space, then the corresponding PSM may not be detectable



Some fragments will contain the PTM in MS2

Search space

- The set of candidate peptides to be considered for potential match to spectra
- Without PTMs, the search space is simply the set of peptides
- With PTMs, the search space effectively doubles for every PTM to be considered.



Examples of typical PTMs

- Phosphorylation
- Ubiquitination
- O-GlcNAcylation
- Methylation
- Acetylation
- Succinylation
- SUMOylation
- Citrullination

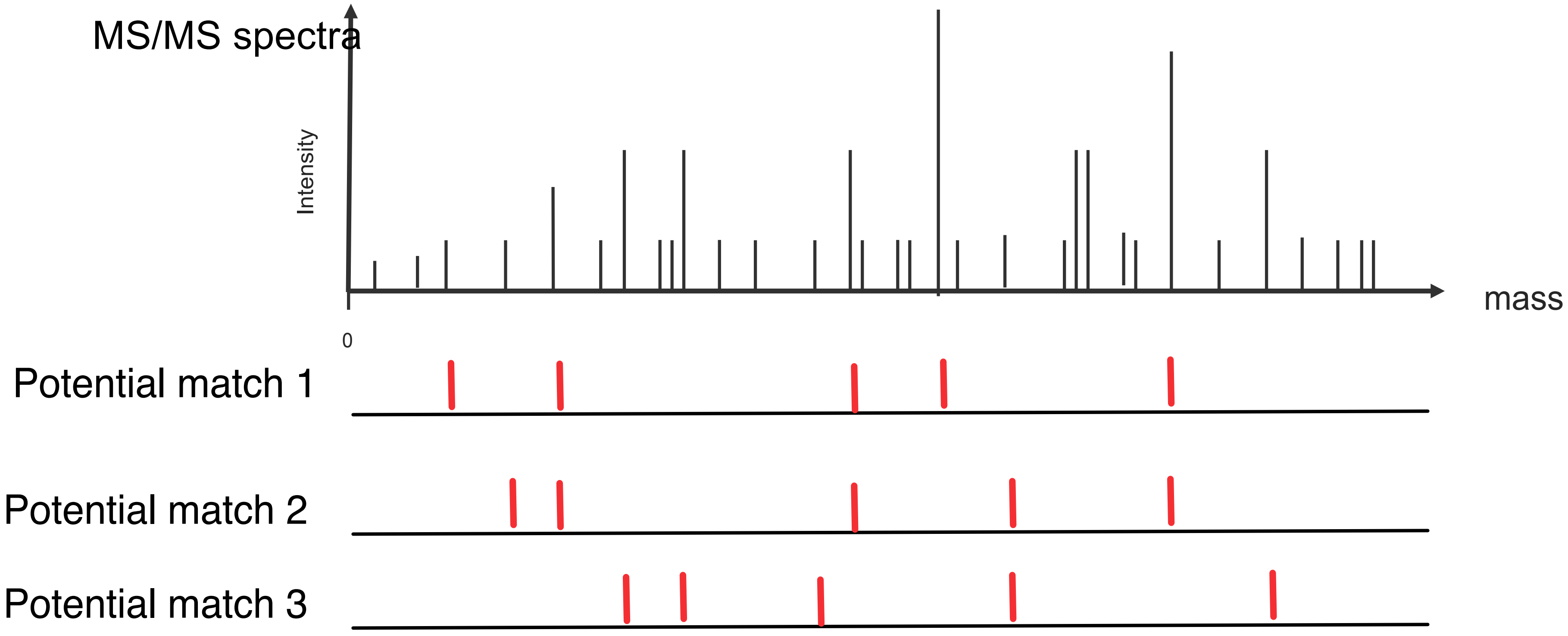
Some 260 000 PTM sites that have been identified in the human proteome thus far, but only a few have been assigned to key regulatory and/or other biological roles!

It is difficult to pin-point exact locations of PTMs as well. And incorporating all possibilities (where there is only 1 or few right matches)… can lead to high false positive rates (we will see how later).

Library quality and size

- UniProt sequence library has 2 databases
  - SwissProt (manually curated and reviewed) - > 500K sequences
  - TrEMBL (Automatic annotation, no review) - > 90M sequences

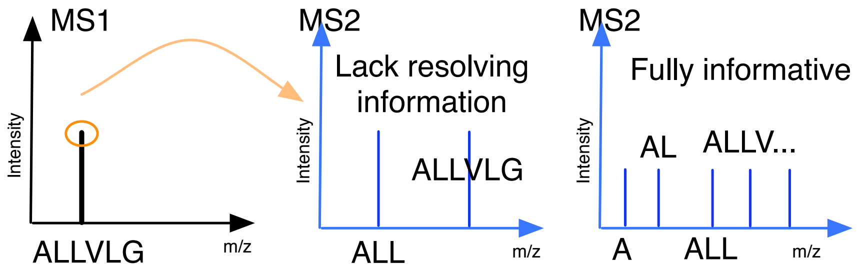
Consider this scenario…



What are potential explanations for 1, 2 and 3?

Spectra Quality

Incomplete fragmentation

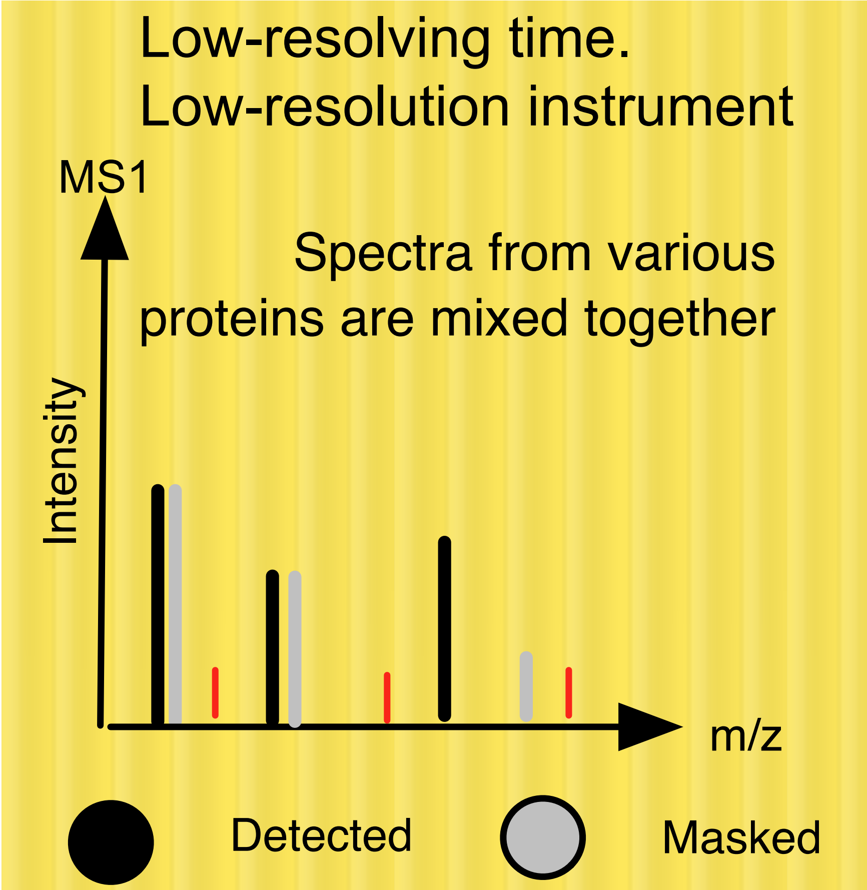


Lack resolving information

Fully informative

Complete MS2 profile allows confident identification of spectra

Mixed signals

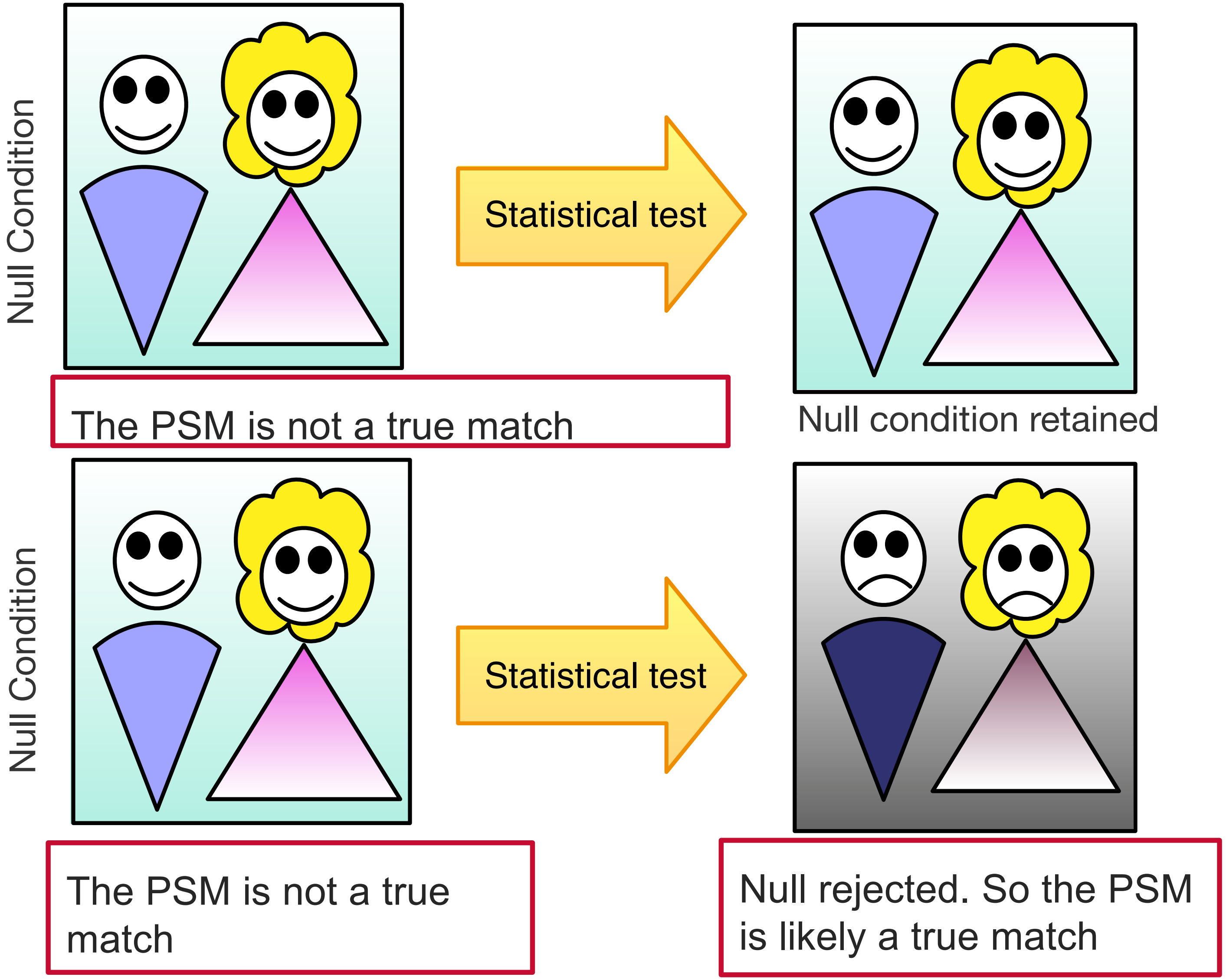


Low-resolving time.
Low-resolution instrument

Spectra from various proteins are mixed together

Statistical testing

The elements of null hypothesis statistical testing



Null Condition

Statistical test

The PSM is not a true match

Null condition retained

Null Condition

Statistical test

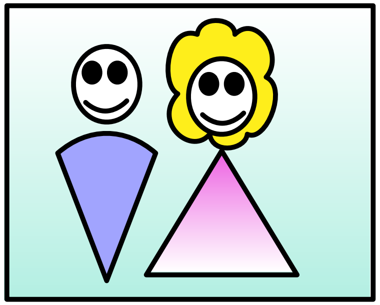
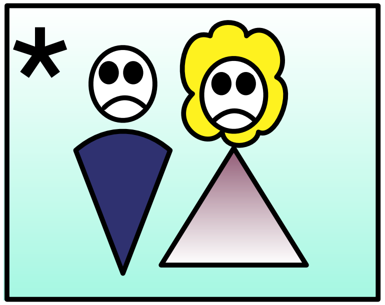
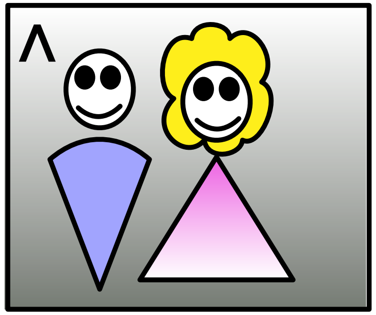
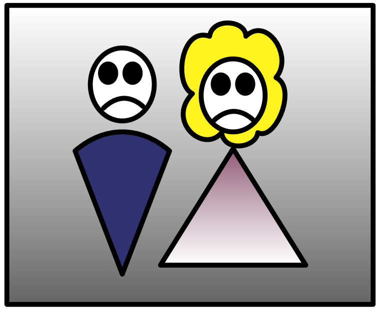
The PSM is not a true match

Null rejected. So the PSM is likely a true match

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology 2018

Possible outcomes from a statistical test

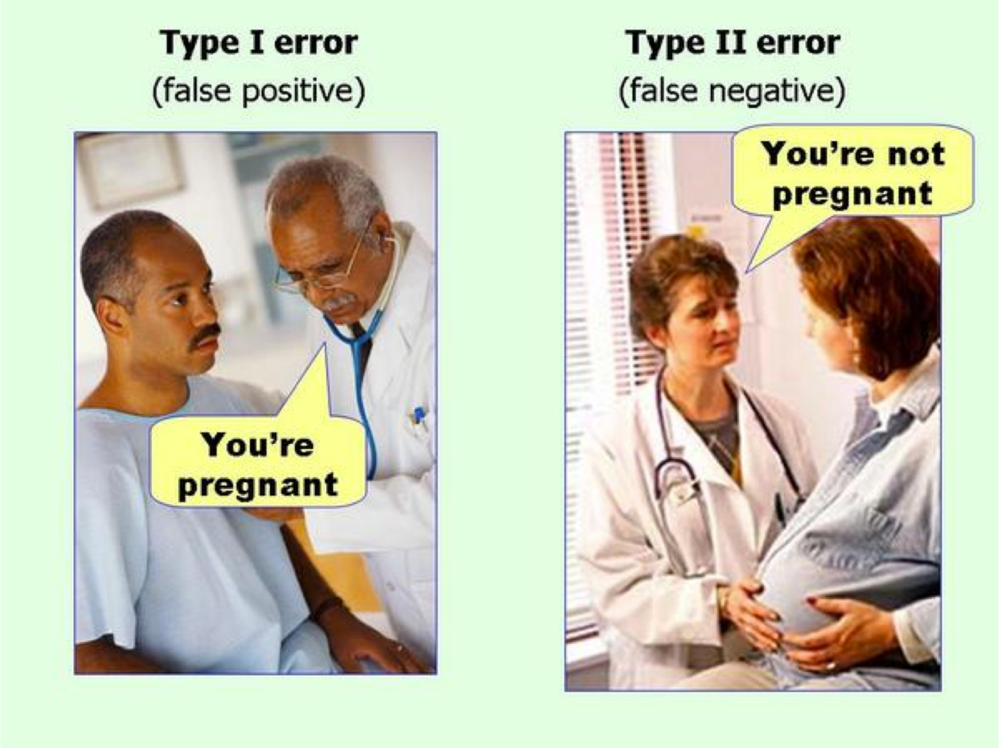
4 Possible outcomes

		Predicted as	
		Happy (Irrelevant)	Unhappy (Relevant)
Reality	Happy (Irrelevant)	 True Negative	*  False Positive
	Unhappy (Relevant)	^  False Negative	 True Positive

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 2018

NANYANG TECHNOLOGICAL UNIVERSITY

How to remember?



Do you recall type I and II statistical errors?

Type I: Reject the null when the null is true
Type II: Fail to reject the null when the null is not true

True positive

True negative

Possible outcomes given the PSMs

The PSM is…

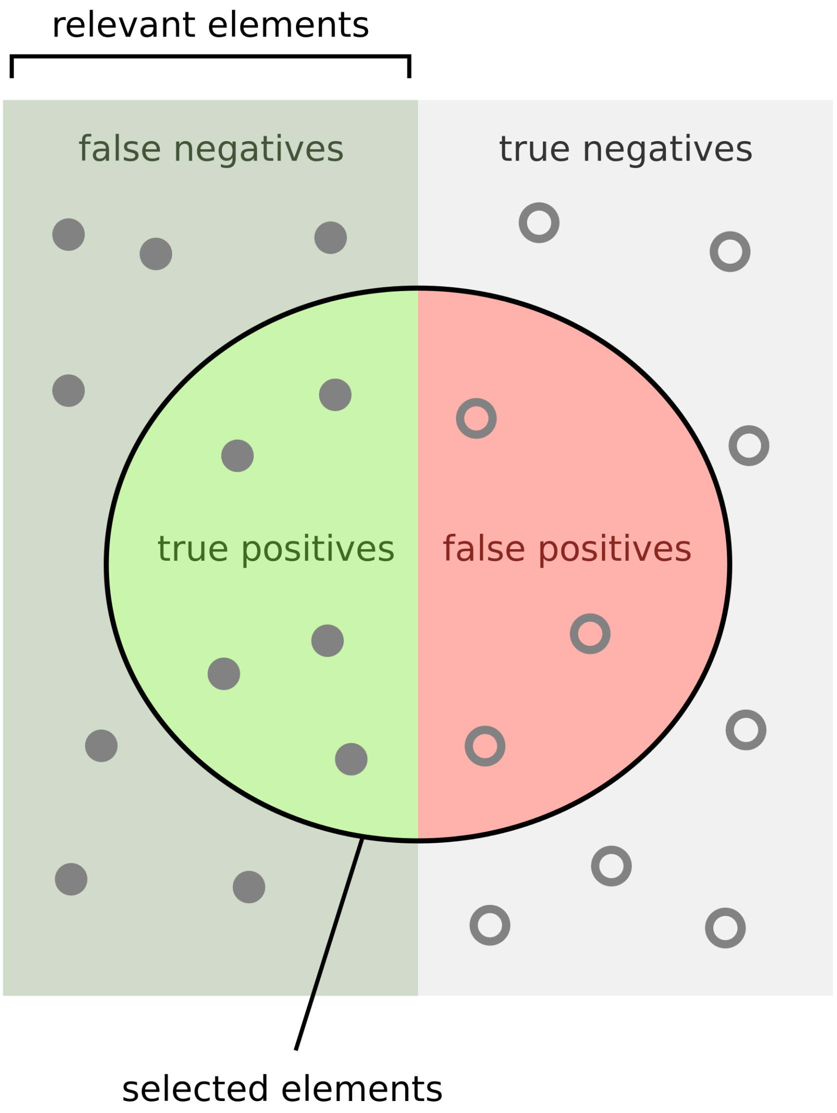
Predicted as

BUT IN Reality..

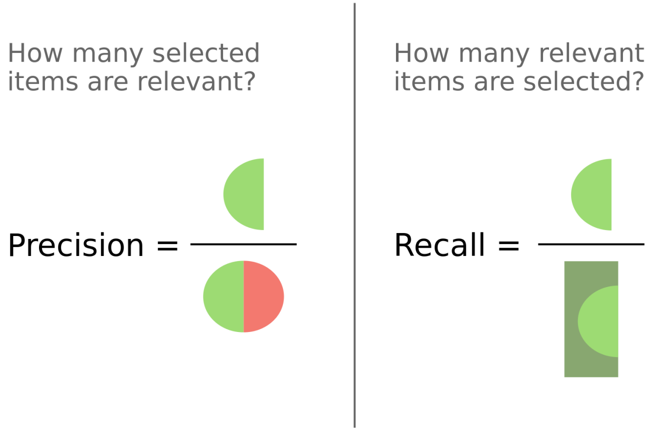
	CORRECT	WRONG
CORRECT		
WRONG		

Imagine we do this for every spectra…

Recall, Precision and the F-score



e.g. let’s say we set a p-value cutoff of 0.05


How many selected items are relevant?
Precision =
How many relevant items are selected?
Recall =

Precision: Of the selected feature,
How many are correct?

Recall: Of the selected feature,
What is the proportion of all the correct ones we got?

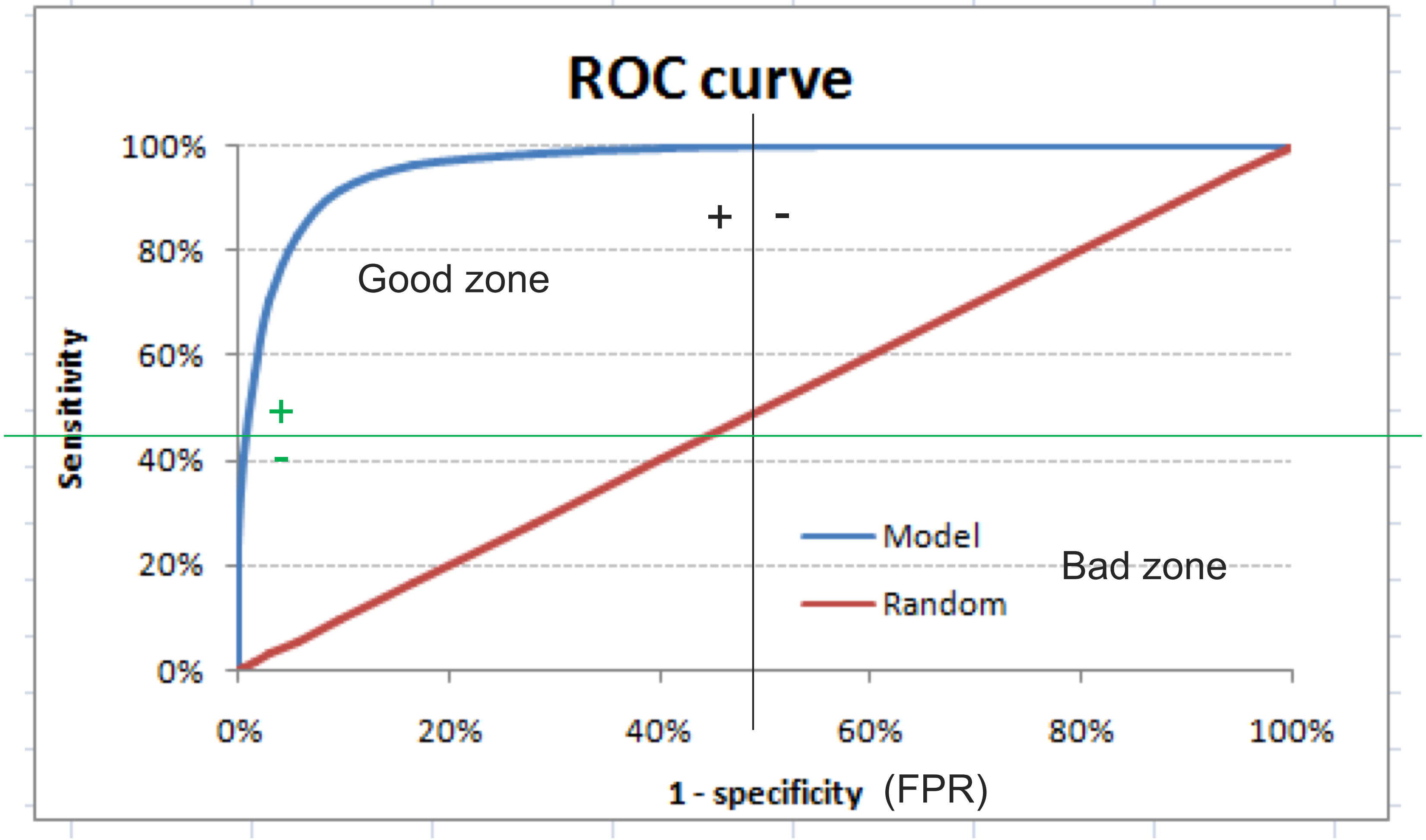
Precision and recall can be combined as:

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Recall is also often called sensitivity or True positive rate

Precision and recall works against each other

The Receiver Operator Characteristic (ROC) curve

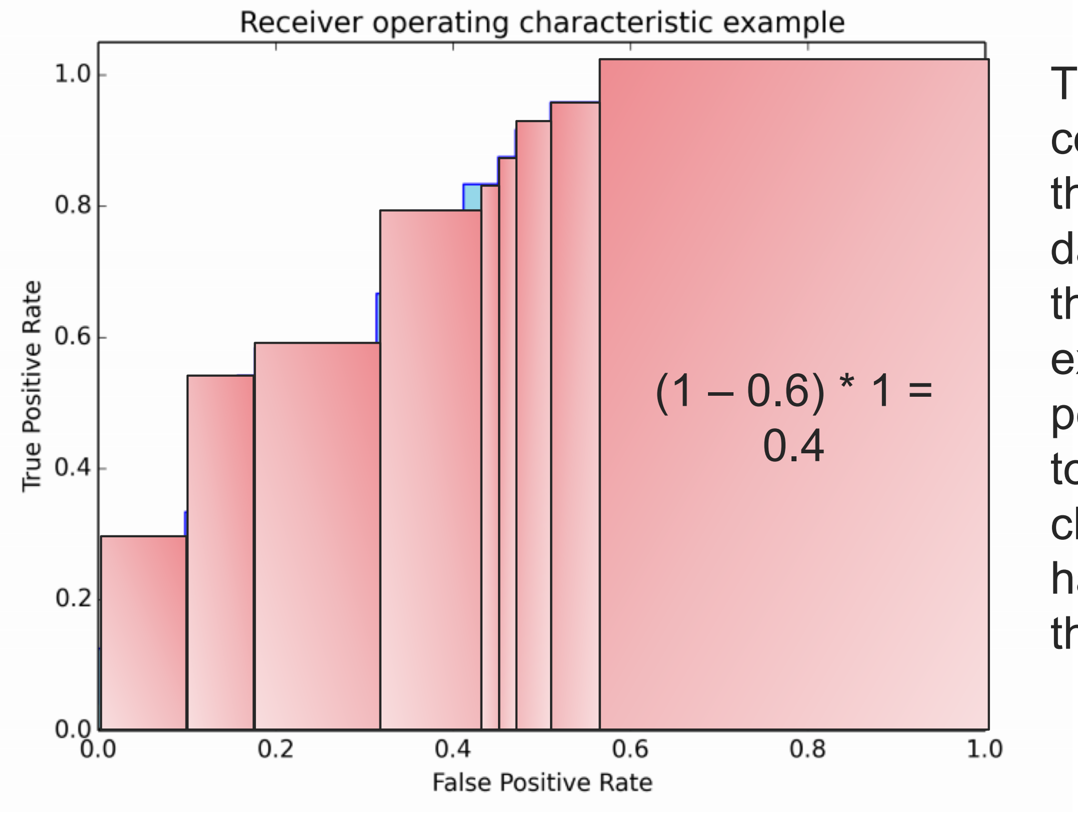


$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

Can you write down the formula for FPR yourself?

The Area Under Curve (AUC)

More rightfully called AUROC (Area under the ROC curve)



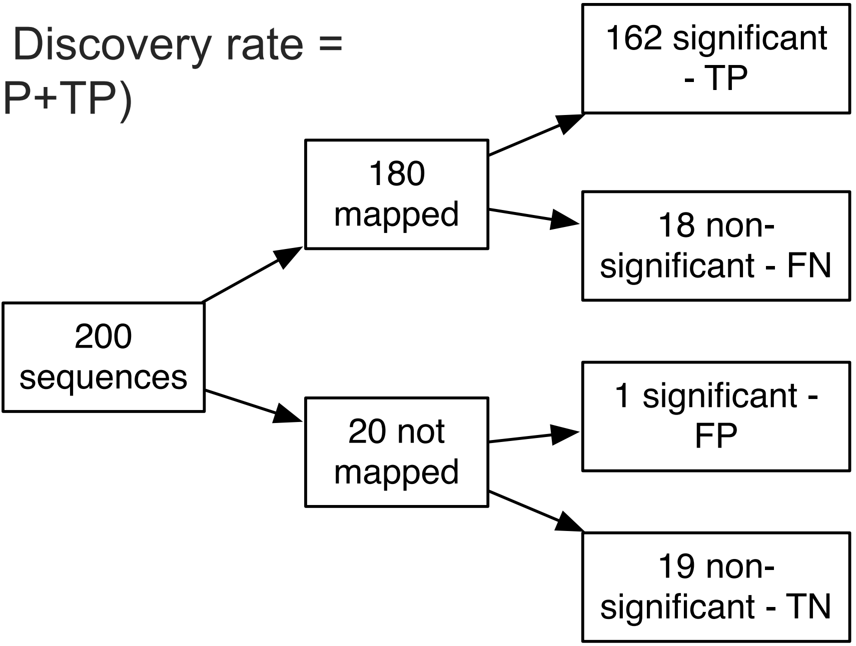
The blue area corresponds to the AUROC. The dashed line in the diagonal is expected performance due to random chance (so we have to be better than chance)

Total area = 1 x 1 = 1
Half area under the diagonal = ½ = 0.5
One simple way to get the AUROC is to simply calculate the area using simple length x breadth. But of course one may use calculus.

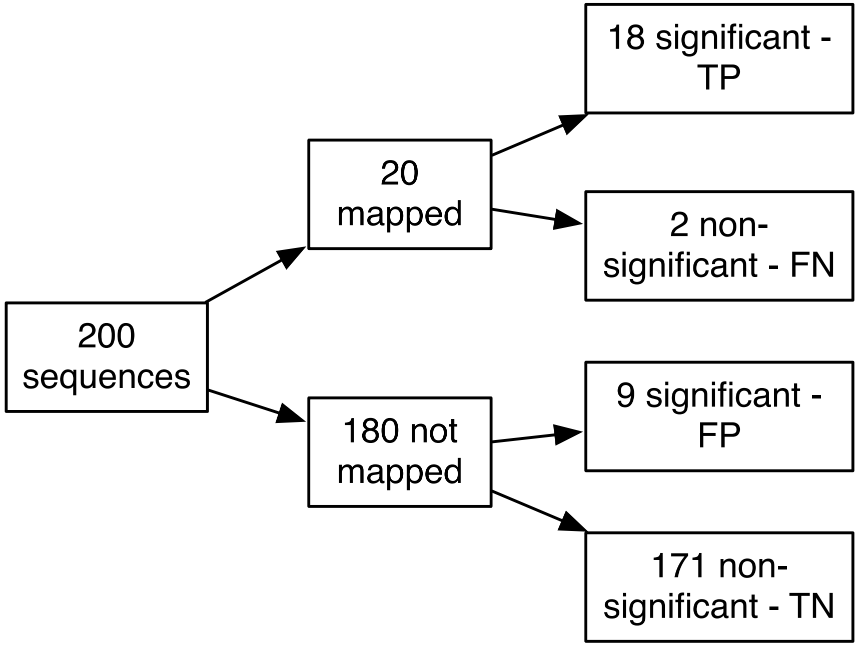
The False Discovery Rate

The FDR relates to the proportion of errors amongst predictions. It is equals to 1 - precision

False Discovery rate = FP/(FP+TP)


200 sequences
180 mapped
20 not mapped
162 significant - TP
18 non-significant - FN
1 significant - FP
19 non-significant - TN

1/163


200 sequences
20 mapped
180 not mapped
18 significant - TP
2 non-significant - FN
9 significant - FP
171 non-significant - TN

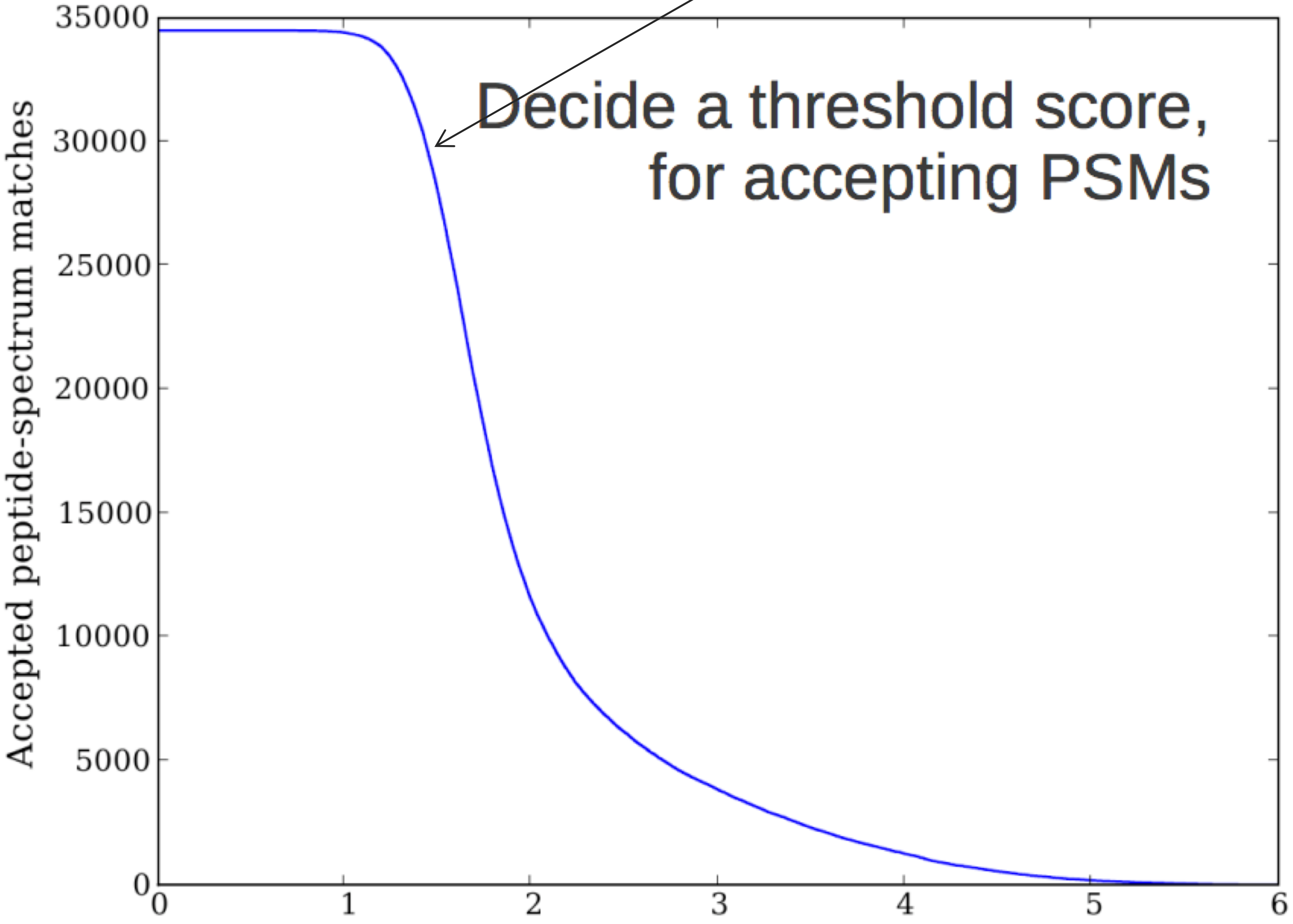
9/27=1/3

The FDR is sensitive to the proportion of true features in the data.

How to get all the good quality matches?
What if we just use our eye power?

PSM quality score

Can do it by eye?



Decide a threshold score,
for accepting PSMs

PSM Score

Statistics provides a more objective manner of evaluation

Statistics help us determine the best match

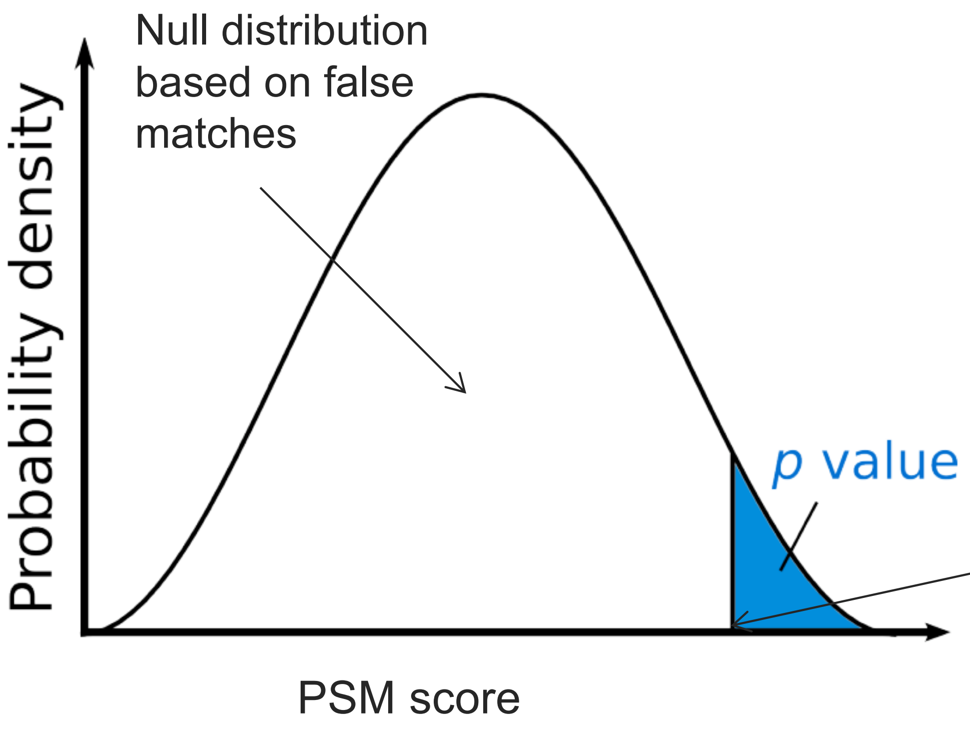
- p-values
- False Discovery Rate (FDR)
- Posterior Error Probability (PEP)

The p-value

We can set up our PSMs as a statistical test (based on the following hypothesis statements)

H0: A PSM is incorrect H1: A PSM is correct

We may reject the null hypothesis with a certain degree of error:
Type I: Falsely reject the null hypothesis (False positive)
Type II: Falsely accept the null hypothesis (False negative)



Null distribution based on false matches

E.g. p value = 0.02
This can be taken to mean that 2% of null have a value equal or more extreme than the observed
In proteomics, this can be taken to mean that 2% of all incorrect PSMs have an equal or better quality score

The score corresponding to our PSM of interest

The probability of obtaining an equal or more extreme result, assuming the null hypothesis is true (Type I error)

The p-value

- We are comparing the observed score against a distribution of “null scores”
- The null distribution are comprised of the natural distribution of values when there is no signal i.e., when a PSM is incorrect (or the null statement is true)
- Why does this make sense?
- Because under this setup, a small p-value would imply that the observed PSM score is very significant (unlikely to arise due to chance).

The p-value

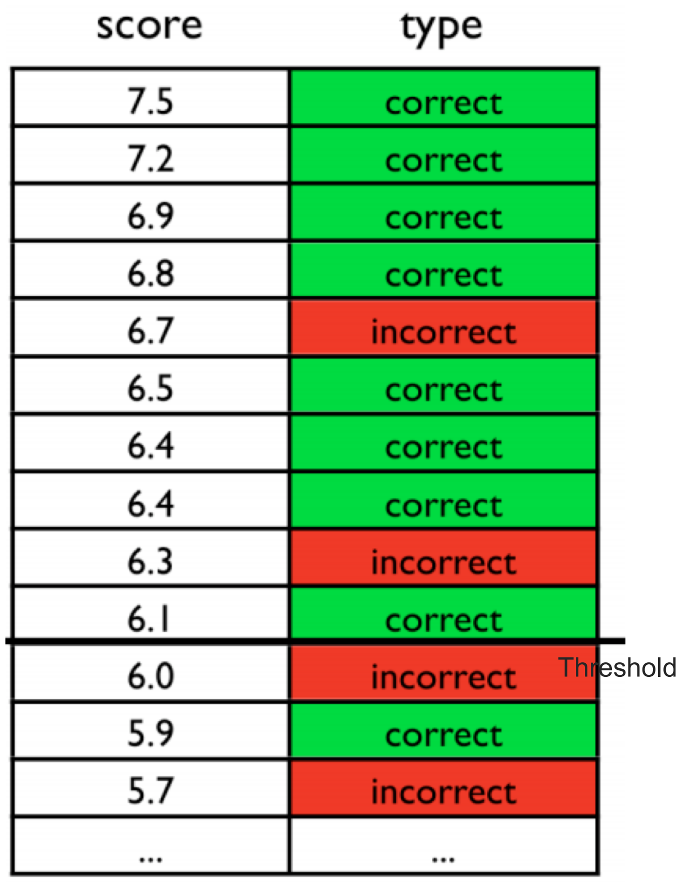
- For each hypothesis tested. Suppose we use a statistical cutoff at 0.01, we should therefore expect 1 in 100 times the result is a false positive
- Suppose 100 tests are performed, then we should expect 100 * 0.01 = 1 false positive
- To control for this, a multiple test correction can be used. For example, to maintain 0.01 FPs given 100 tests, the cutoff can be reduced from 0.01 to 0.01/100 = 0.0001

The p-value (in proteomics)

- Is a local measure, meaning that it is confined specifically to the particular PSM under consideration (it is therefore self-contained)
- Global measures on the other hand, considers all PSMs scores concurrently and relative to each other (they are therefore not self-contained).
- Lets say we observe a PSM with a score of 1, we can build an empirical reference distribution of similar false/random sequences and find out what are their respective PSM scores. If the observed PSM does better than at least some alpha threshold, then we can say that this PSM is statistically significant, and so we reject the null hypothesis for the alternative.
- This is computationally very intensive. ALSO… what is a reasonable null?

False Discovery Rate (FDR)

False Discovery Rate (FDR): The expected fraction of false positives among the significant test statistics. (FP/FP+TP)
Compare this against the false positive rate which is FP/(FP+TN)



score	type
7.5	correct
7.2	correct
6.9	correct
6.8	correct
6.7	incorrect
6.5	correct
6.4	correct
6.4	correct
6.3	incorrect
6.1	correct
6.0	incorrect
5.9	correct
5.7	incorrect
...	...

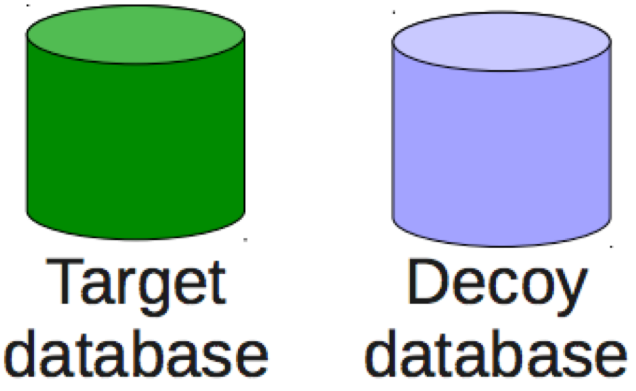
So how do we look at this?
Let's say we have a set of PSM scores and decide to draw the line at 6, i.e., we accept all PSMs with scores > 6.
Let's also assume we have perfect knowledge of correct and wrong matches.
We note that 10 PSMs are retained.
Of these, 2 are wrong. So the FDR is therefore

$$FDR = 2/10 = 20\% = 0.2$$

This seems great. But in reality, we don't know which ones are wrong. This is similar to the null problem in p-value generation. So how do we create something which we know to be wrong or sure?

Benjamini & Hochberg, JSTOR, 1995; Storey & Tibshirani, PNAS, 2003

The target-decoy analysis



Estimating FDR:
How to purposely create your incorrect PSMs

Target database

Protein sequences of the studied organism.

Decoy database

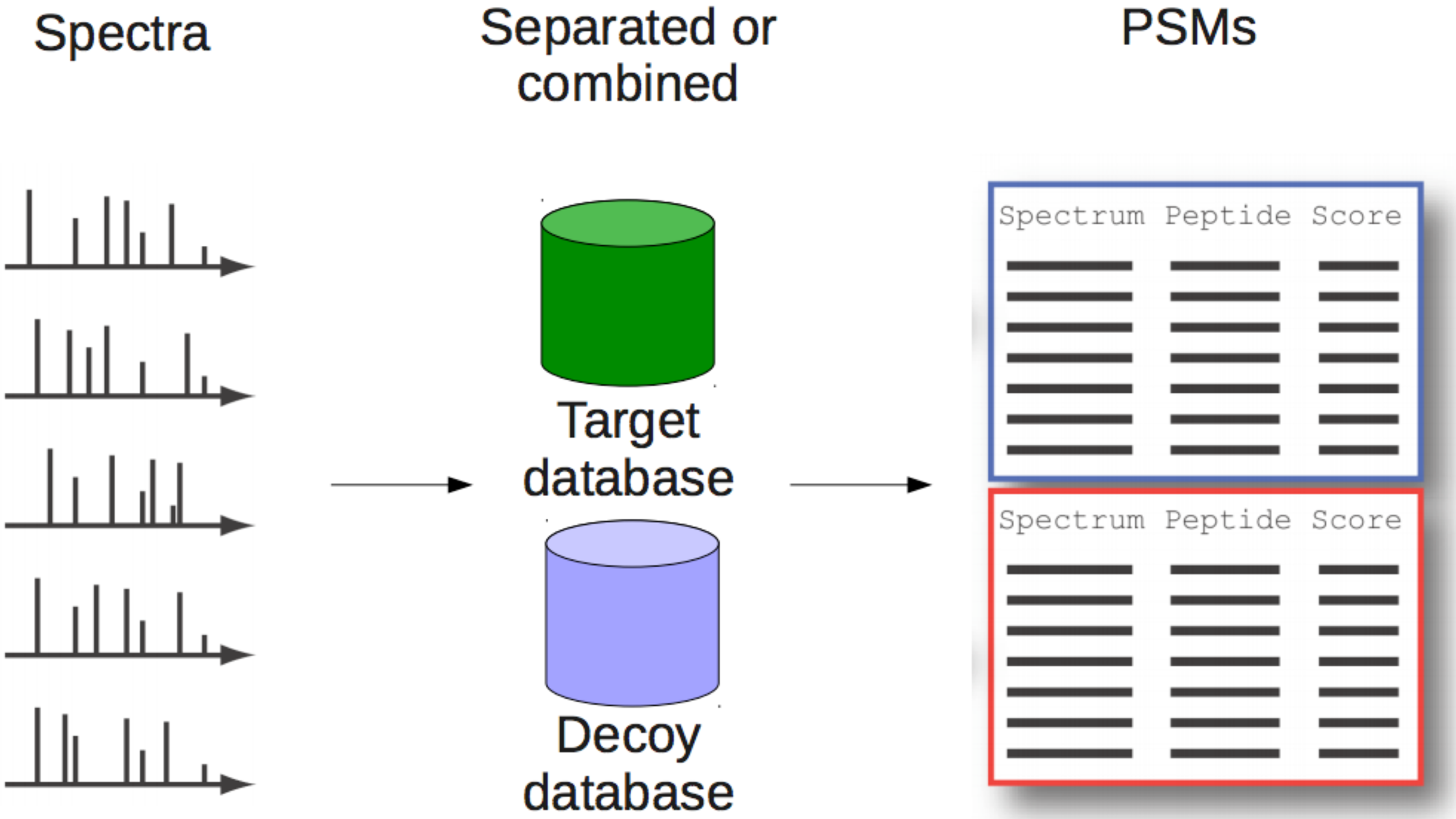
Reversed or shuffled sequences.

Assumption

Spectra matched to the decoy database are good models of **incorrect** matches to the target database.

In other words, all matches to decoy are false positives

The target-decoy analysis



Elias and Gygi. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. Methods Mol Biol. 2010.

Target-decoy searching steps

- Construct a concatenated target-decoy sequence list, marking decoy sequences with a text flag in their annotation.
- Use a MS/MS search engine to interpret input MS/MS spectra using target-decoy sequence list.
- Evaluate the relative proportion of target and decoy sequences in the search space to derive the multiplicative factor required to estimate false positives, if necessary.
- Estimate false positive-related statistics.
- Use decoy hits to guide the establishment of filtering criteria.
- Report statistics for filtered data set.

Decoy construction rules

- **Similar** amino acid distributions as target protein sequences.
- **Similar** protein length distribution as target protein sequence list.
- **Similar** numbers of proteins as target protein list.
- **Similar** numbers of predicted peptides as target protein list.
- **No** predicted peptides **in common** between target and decoy sequence lists.

False Discovery Rate (FDR)

Reversal

Advantages

- Simple
- Preserve general features of the target sequence list e.g. same inter-protein redundancy
- Defined transformation therefore repeatable

Disadvantages

- Non-random transformation is less statistically rigorous
- Cannot be used for peptides with low sequence complexity

Shuffling

Advantages

- Simple
- Has desired stochastic properties

Disadvantages

- Redundancies and homologies between protein entries will not be preserved, so many more decoy peptides than originally present in the target sequence list

Random Proteins

Advantages

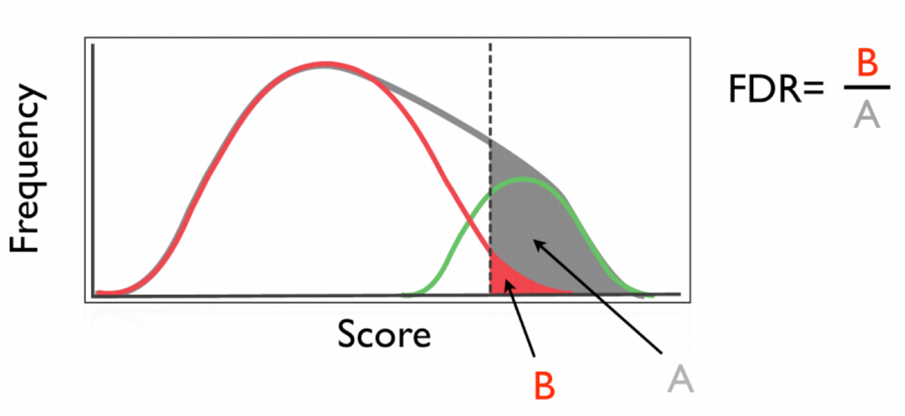
- Has desired stochastic properties
- Can preserve amino acid bias and protein length distribution

Disadvantages

- Redundancies and homologies between protein entries will not be preserved, so many more decoy peptides than originally present in the target sequence list

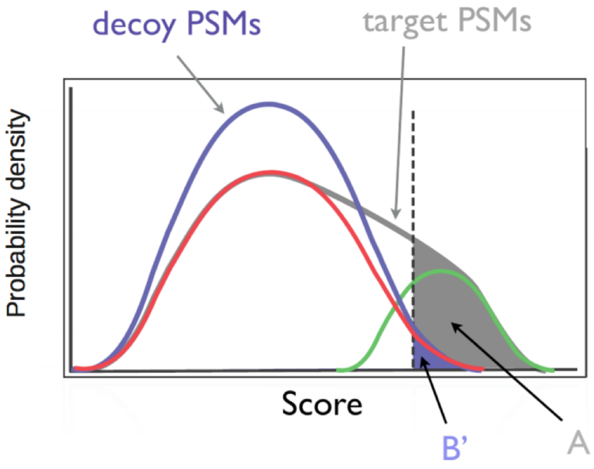
FDR estimation based on decoy

No decoy



$$\text{FDR} = \frac{B}{A}$$

With decoy



$$\text{FDR} = \frac{B}{A} = \frac{\pi_0 \; B'}{A}$$

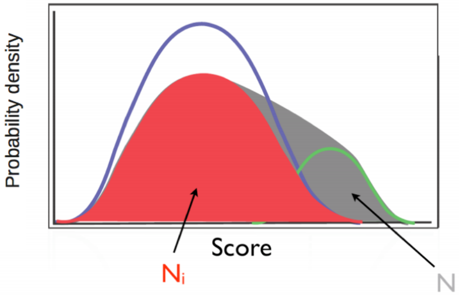
π_0 is the fraction of incorrect target PSMs among target PSMs

Target-decoy analysis

Combined searches

Target and decoy database are searched togethe

$\widehat{\text{FDR}}$ = {#decoys over threshold} /
{#targets over threshold}



$$\text{FDR} = \frac{B}{A} = \frac{\pi_0 \; B'}{A}$$

$$\pi_0 = \frac{N_i}{N}$$

π_0 is the fraction of incorrect target PSMs among target PSMs

I.e., π_0 is 1

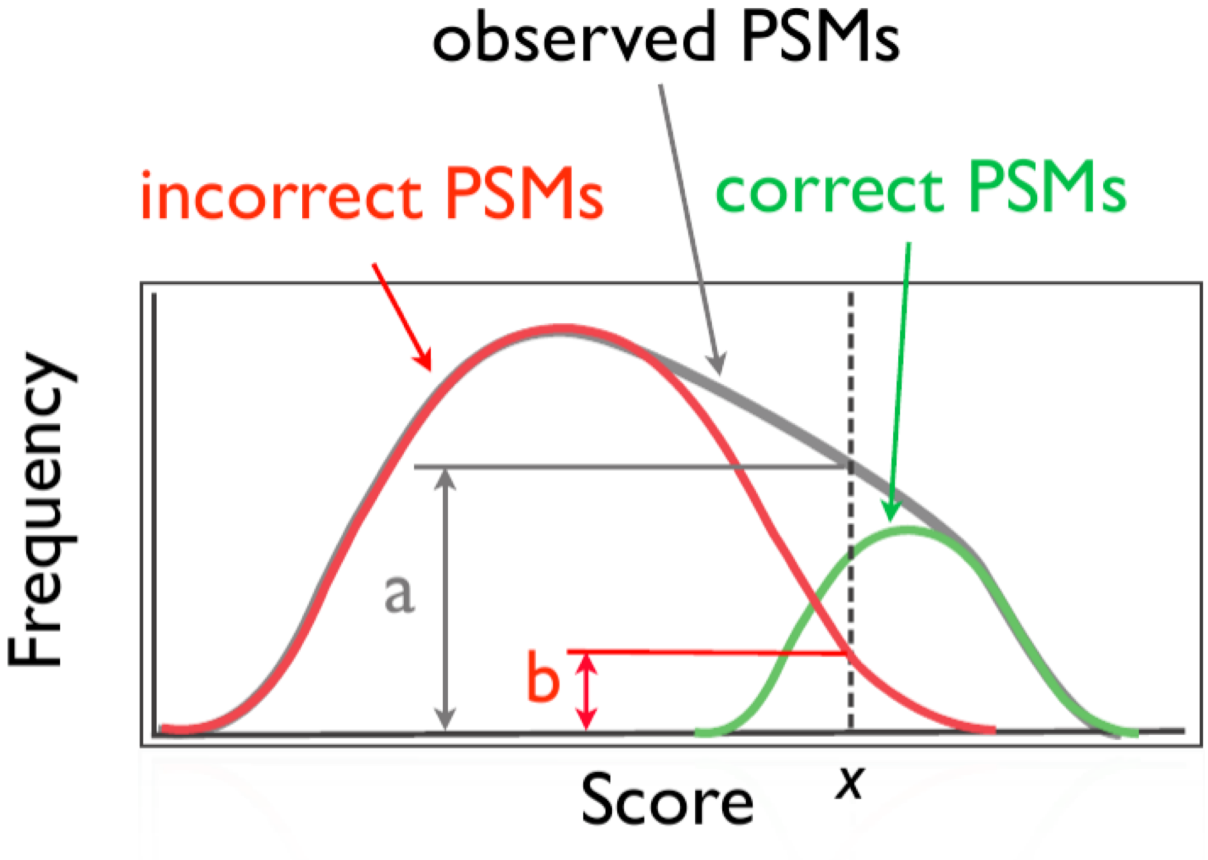
Simpler. Since estimating π_0 can be tricky.

Posterior Error Probability (PEP)

Posterior Error Probability (PEP): The probability that the null hypothesis is true for a particular test statistic

In proteomics, it can be taken to mean the probability that a given PSM is wrong.

PEP is sometimes called “local FDR”



$$\text{PEP}=\frac{b}{a}$$

A PEP is the probability that a PSM scoring x is incorrect

Statistical evaluation: when to use what?

P-value

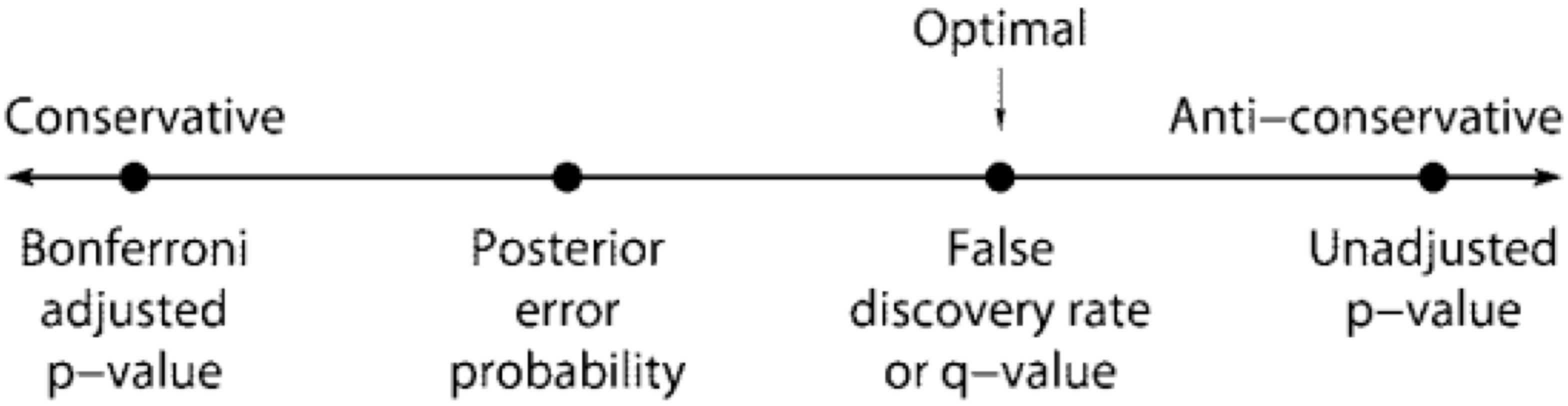
In experiments with one, or very few, spectra

FDR/q value

For the confidence in sets, of PSMs, peptides or proteins

PEP

For the confidence in a particular PSM, peptide or proteins

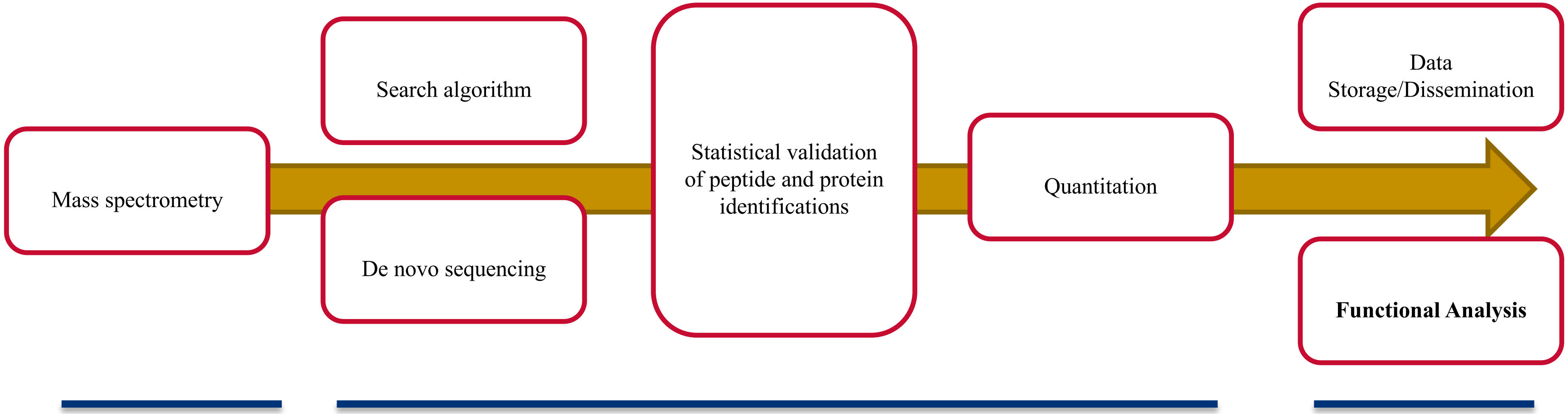


Optimal
Conservative
Anti−conservative
Bonferroni adjusted p−value
Posterior error probability
False discovery rate or q−value
Unadjusted p−value

Figure 3. Methods for assigning statistical significance

Käll et al. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. JPR, 2008

Overview of proteo-informatics


Mass spectrometry
Search algorithm
De novo sequencing
Statistical validation of peptide and protein identifications
Quantitation
Data Storage/Dissemination
Functional Analysis

Data storage/ representation formats

Peptide-spectra matching (PSM)

Functional analysis

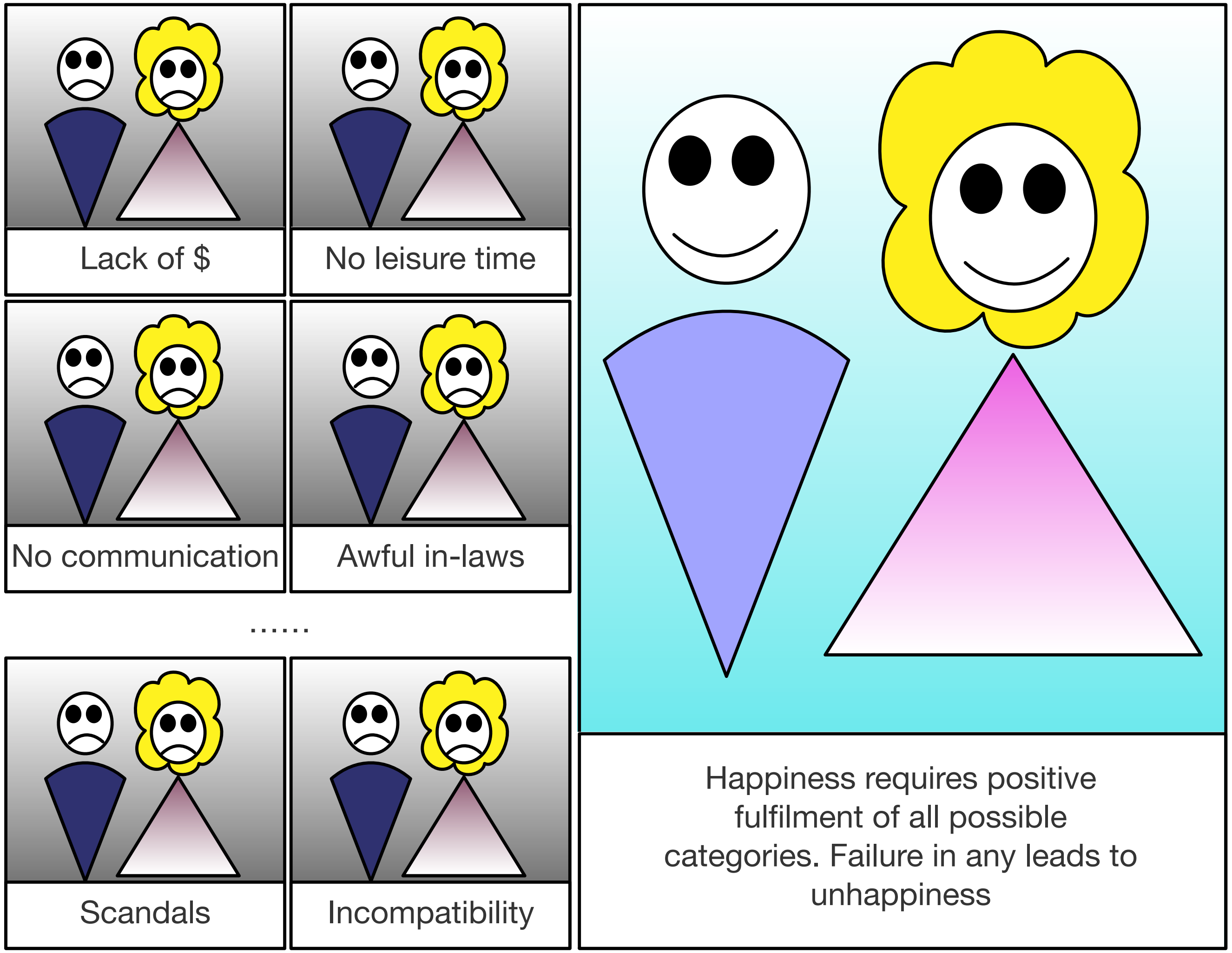
Functional analysis 1 (Comparative analysis)

- The process of creating knowledge and insight from biological data
- Comparative analysis (group A vs B)
  - Assumption: the differences between two groups are phenotypically relevant and can be used to construct mechanistic explanation
  - This is a fallacious assumption.

The Anna Karenina Principle

- Happy families are all alike, every unhappy family is unhappy in its own way --- Leo Tolstoy
- Interpreted as: There are many ways to violate the null hypothesis, but only one way that is pertinent to the outcome of interest

Happiness does not come easy



Lack of $

No leisure time

No communication

Awful in-laws

……

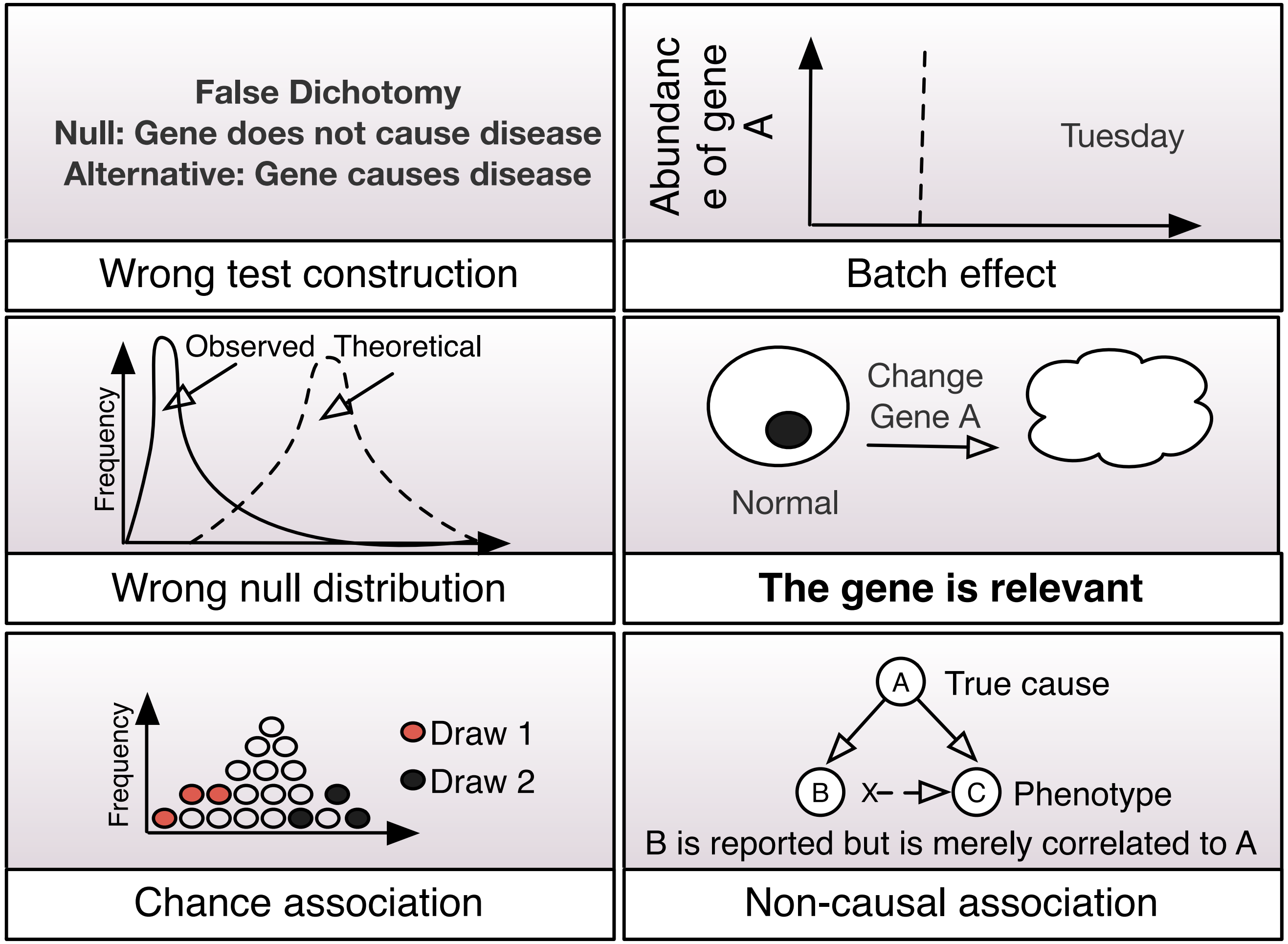
Scandals

Incompatibility

Happiness requires positive fulfilment of all possible categories. Failure in any leads to unhappiness

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 2018

Anna Karenina in comparative proteomics



False Dichotomy
Null: Gene does not cause disease
Alternative: Gene causes disease

Wrong test construction

Batch effect

Wrong null distribution

The gene is relevant

Chance association

B is reported but is merely correlated to A

Non-causal association

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 2018

Dealing with the Anna Karenina

Causes

- A careless null/alternative hypothesis due to forgotten assumptions:
  - Distributions of the feature of interest in the two samples are identical to the two corresponding populations
  - Features not of interest are equalized/controlled for in the two samples
  - No other explanation for the significance of the test
  - Null distribution models the real world
- These make it easy to reject the carelessly stated null hypothesis and accept an incorrect alternative hypothesis.

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 2018

Dealing with the Anna Karenina

Good Practices to Avoid Wrong Conclusions and Get Deeper Insight

- Check for sampling bias:
  - Are the distributions of the feature of interest in the two samples same as that in the two populations?
- Check for exceptions:
  - Are there large subpopulations for which the test outcome is opposite?
  - Are there large subpopulations for which the test outcome becomes much more significant?
- Check for validity of the null distribution:
  - Is there evidence that suggests the null distribution is inappropriate?
- Check the hypothesis statement construction
  - Are the hypothesis statements being framed correctly (as opposed to a statement that is prone to being rejected for the wrong reasons)?

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 2018

Dealing with the Anna Karenina

Good Practices to Avoid Wrong Conclusions and Get Deeper Insight

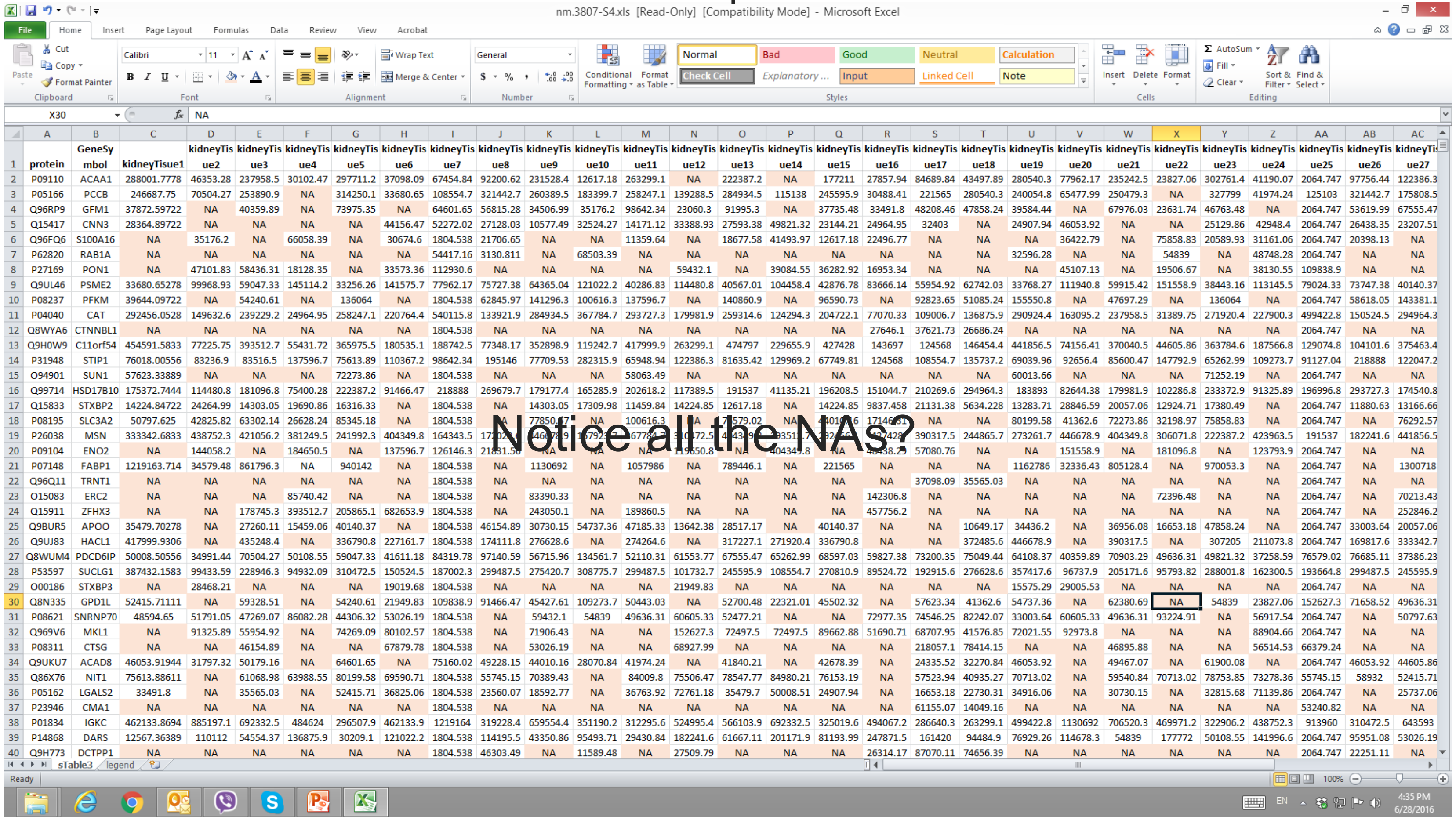
- Check your assumptions
  - Are the right assumptions being made (e.g. the independence of measured variables)?
- Check if appropriate summary statistics are used
  - If an event is extremely rare, then using mean/median-based statistics will miss it; ditto if many similar events are present, but only one is relevant

- **Note**: Even if all (or most) of the above points are addressed, it still does not ensure phenotypic relevance, only correlation.

Goh and Wong. Dealing with confounders in -omics analysis. Trends in Biotechnology, 2018

Functional analysis 2 (Missing proteins)

Samples



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	77850.47	NA	100616.3	NA	71579.02	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	[illegible]	446678.9	[illegible]	367784.7	310472.5	[illegible]	[illegible]	[illegible]	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

sTable3 legend

Ready

Notice all the NAs?

Proteins

Cannot do analysis properly with such data

Missing proteins

- Any gene sequence whose respective protein has never been observed is an MP.
- Alongside various initiatives---e.g. GPMDB, PeptideAtlas and neXtProt ---the goal is to establish a genome-proteome bridge.

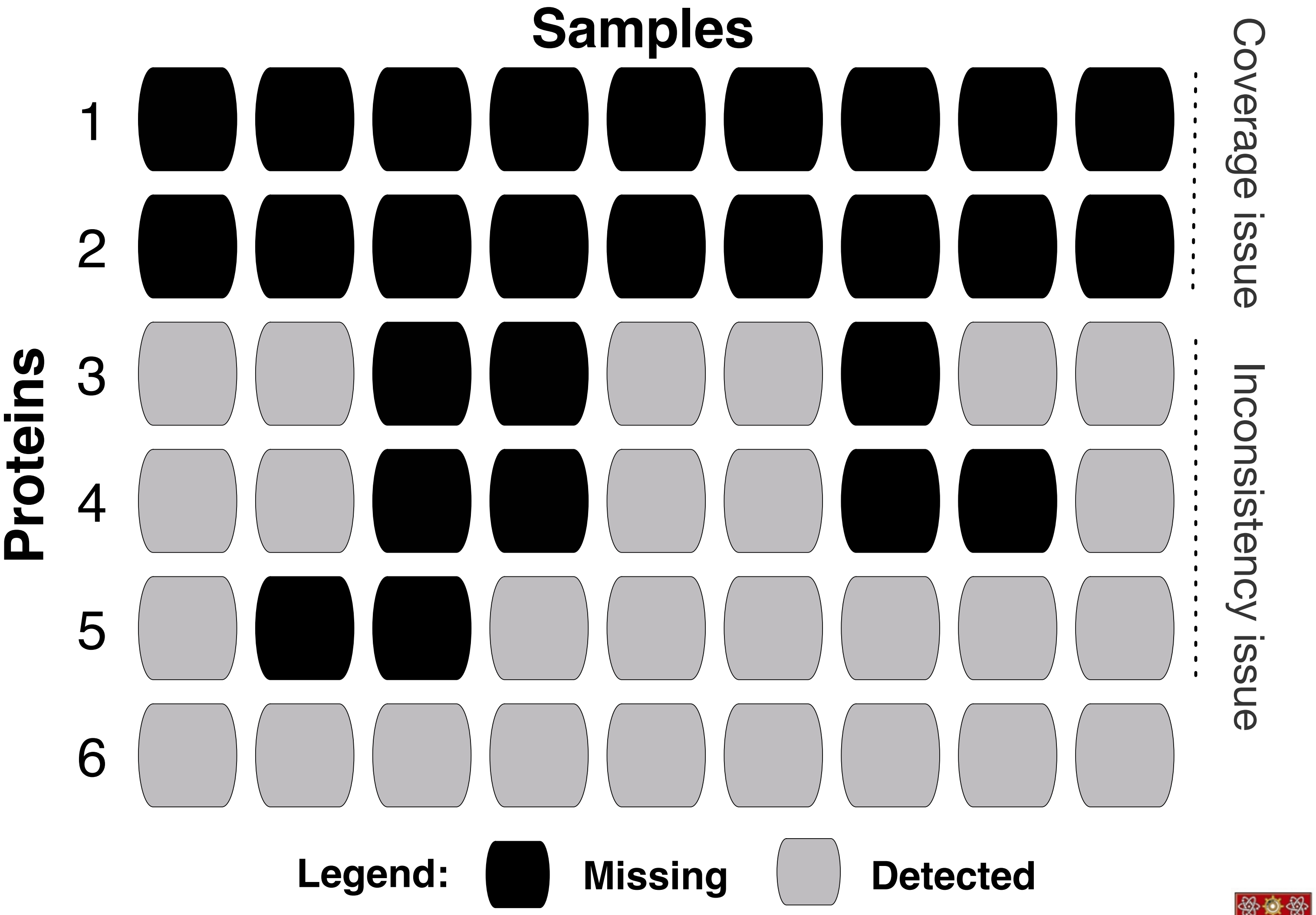
Missing proteins (Tiers 1 to 5)

PE Tier	Inclusion criteria	Percentage of proteome against 20,055 proteins	Notes
1	Evidence at proteome level	~82.0% (16,518)	*At least 2 unique non-overlapped peptides at least 9 amino acid residues long
2	Evidence at transcript level only	~11.5% (2,290)	*The transcript must be confidently detected, but no corresponding protein evidence
3	Homology inference only	~3.0% (565)	*Inferred homologues without protein or transcript support
4	Predicted	~0.5% (94)	*Predicted coding sequence, without homology, transcript or protein support
5	Dubious	~3.0% (588)	*The sequence may not fully meet the criteria for a predicted coding sequence *Uncertainty over the veracity of the coding sequence (i.e., we do not know the sequence is correct) *Some studies do not consider PE5 as MPs

Missing proteins and relations to coverage, consistency problems

- An MP may be one of the following
  - sequence is known but hard to detect,
  - sequence is known but never detected in MS
  - sequence is not known but evidence exists for it e.g. via gene prediction or in raw spectra.
- The “missing-protein problem” (MPP)---viz. the difficulty in detecting certain proteins despite transcript or theoretical evidence---should more rightfully be considered a narrow manifestation of the more general coverage (the inability to survey the entire proteome) and consistency (the inability to consistently detect a protein) problems

Coverage and consistency



Proteins	Samples								
1	■	■	■	■	■	■	■	■	■
2	■	■	■	■	■	■	■	■	■
3	□	□	■	■	□	□	■	□	□
4	□	□	■	■	□	□	■	■	□
5	□	■	■	□	□	□	□	□	□
6	□	□	□	□	□	□	□	□	□

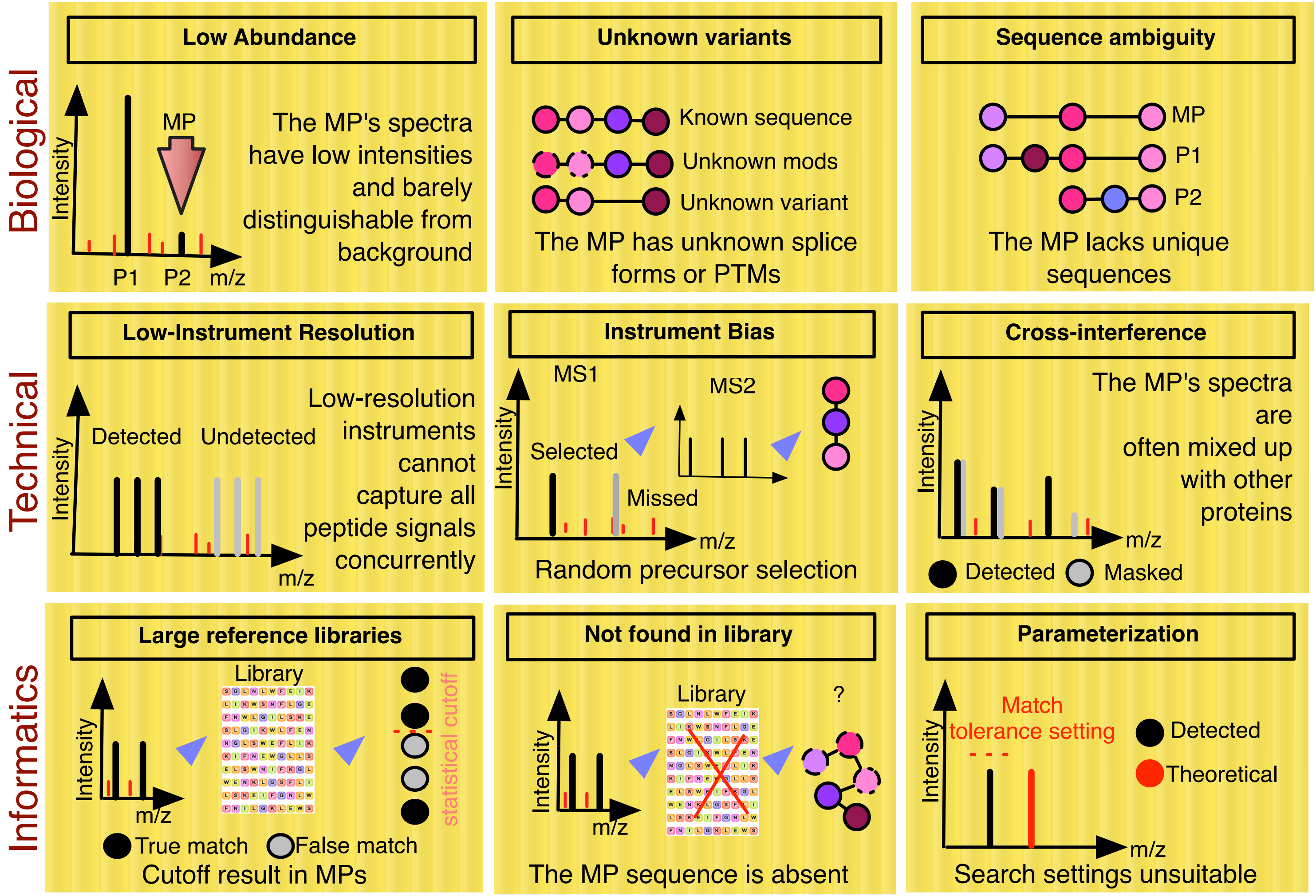
Coverage issue (proteins 1–2)

Inconsistency issue (proteins 3–5)

Legend: ■ Missing □ Detected

Goh and Wong. Advanced bioinformatics methods for practical applications of proteomics. Briefings in Bioinformatics, 2017

Why do proteins go missing?



Biological

Low Abundance

The MP's spectra have low intensities and barely distinguishable from background

Unknown variants

The MP has unknown splice forms or PTMs

Sequence ambiguity

The MP lacks unique sequences

Technical

Low-Instrument Resolution

Low-resolution instruments cannot capture all peptide signals concurrently

Instrument Bias

Random precursor selection

Cross-interference

The MP's spectra are often mixed up with other proteins

Informatics

Large reference libraries

Cutoff result in MPs

Not found in library

The MP sequence is absent

Parameterization

Search settings unsuitable

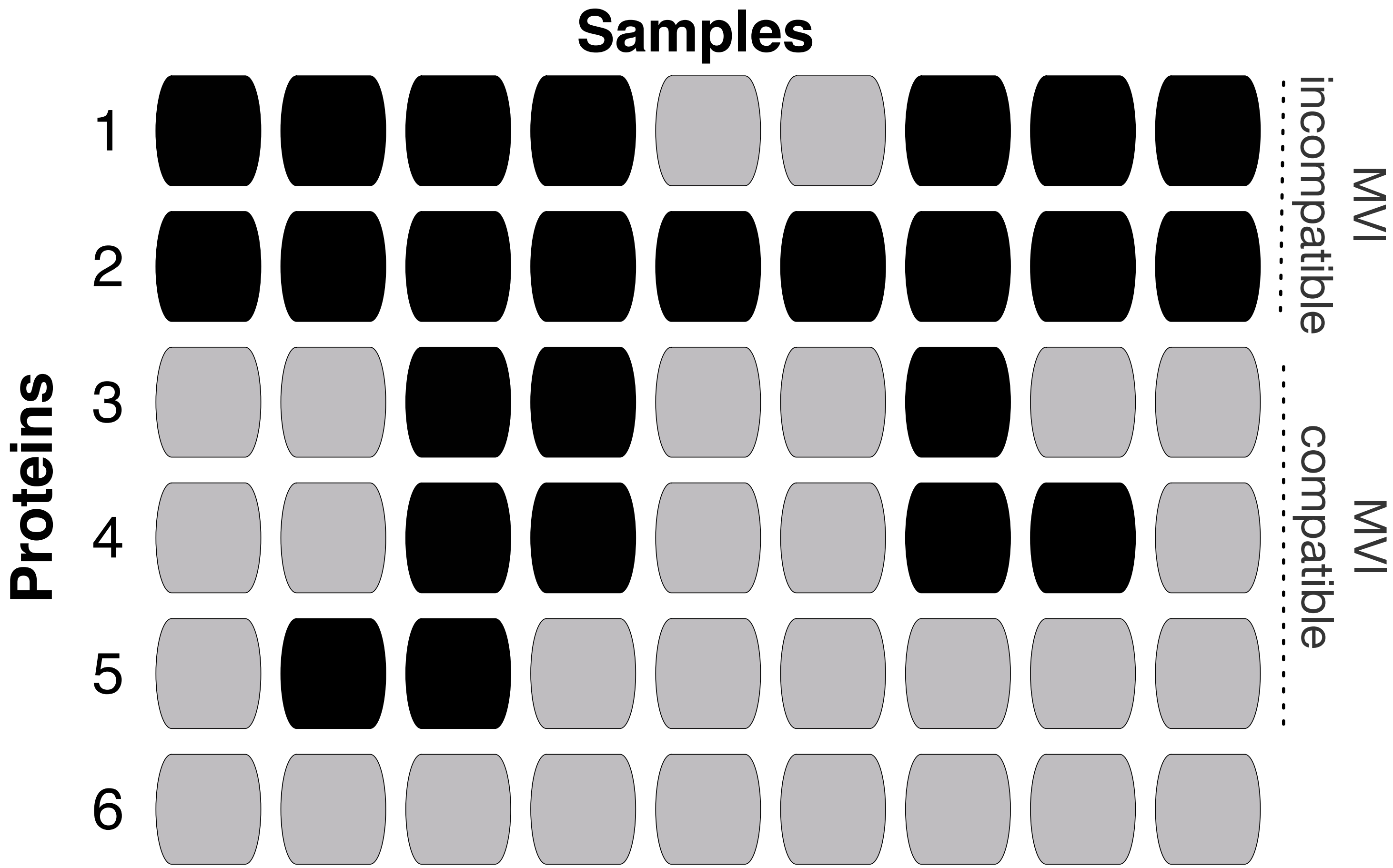
Zhou et al. Understanding missing proteins: A functional perspective. Drug Discovery Today, 2018

NANYANG TECHNOLOGICAL UNIVERSITY

Missing value imputation (MVI)

- A few strategies:
  - We fill in 0s or a fixed value based on the average of all protein expression
  - For each missing value per protein, we fill in the average value based on all observed values for that same protein
  - We estimate the missing value based on proteins that are known to be correlated

Missing value imputation (MVI)



Limitation: It mostly only works well for inconsistency issue

NANYANG TECHNOLOGICAL UNIVERSITY

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

MVI methods really don’t work very well

CV(RMSE) ~ 20% at 75th percentile

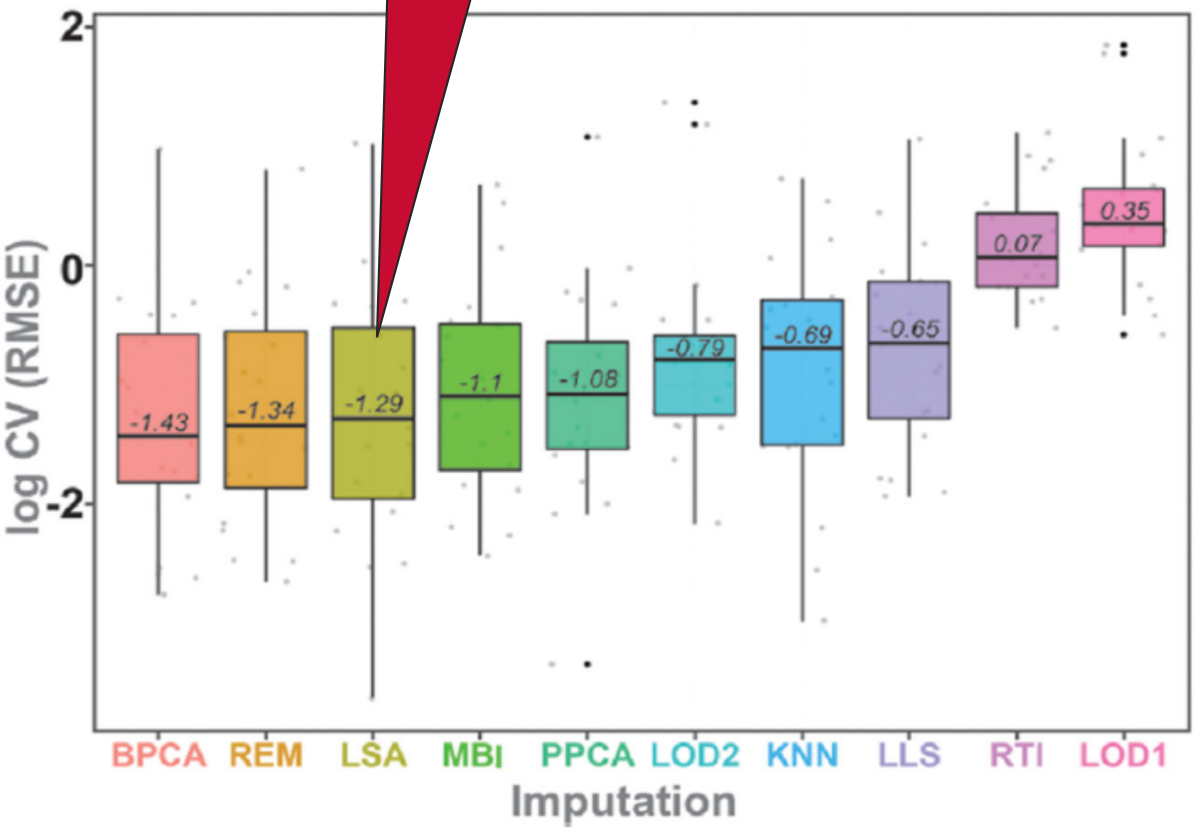


Figure 2.
Boxplot of the average $\log_{10}$ CV(RMSE) for the imputed dilution series datasets (Table 1) at the (A) peptide and (B) protein levels. The lower line represents the 25th percentile, the upper line of the box represents the 75th percentile, and the inner line corresponds to the median $\log_{10}$ CV(RMSE).

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

High abundance has lower % of MPs. However, low abundance is not a solely explanation. The MPs are equally distributed across the horizontal median.

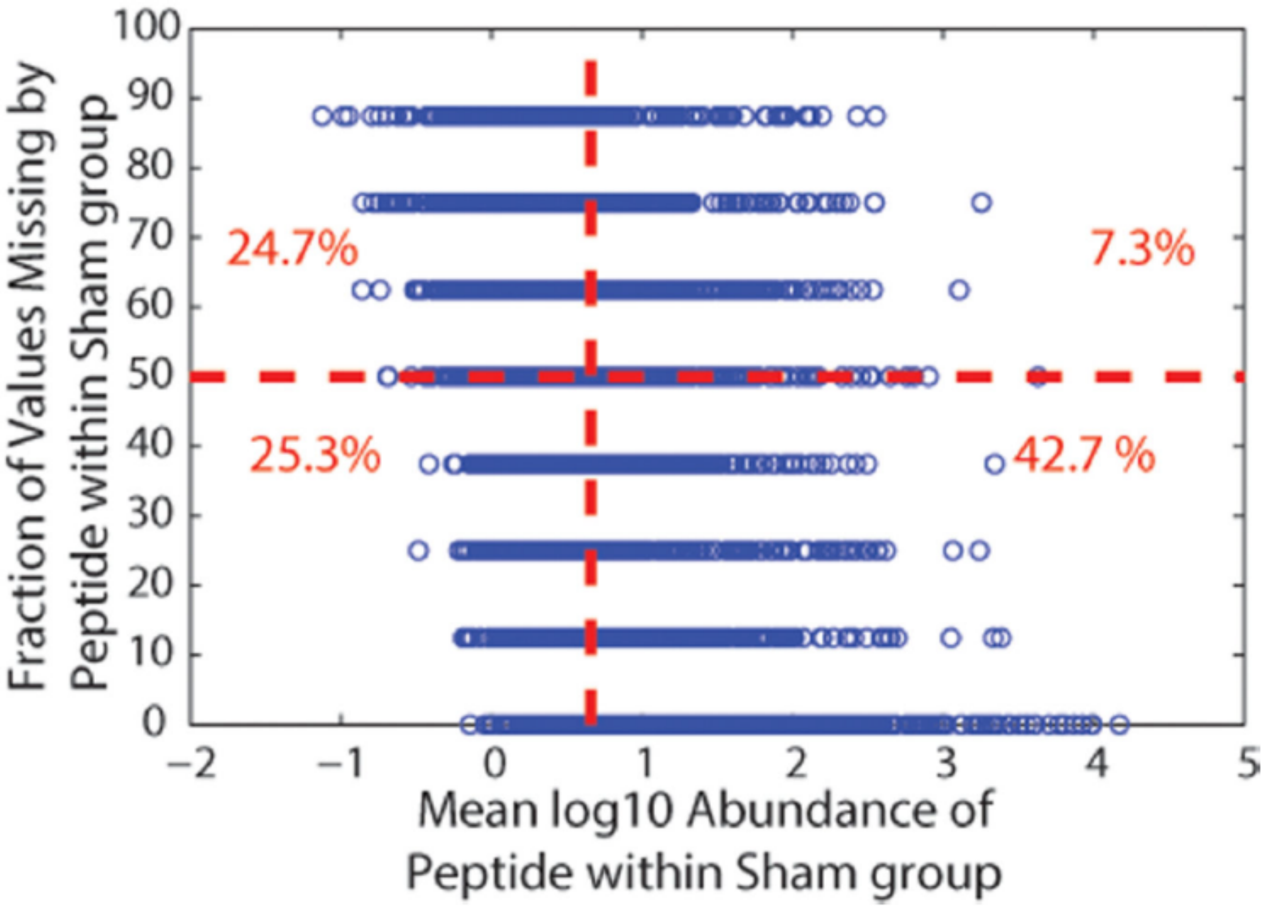


Figure 1.

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.

How about we use the idea of "guilt-by-association?"

- 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 but not E.

⇒ 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 (recover missing proteins)

⇒ 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

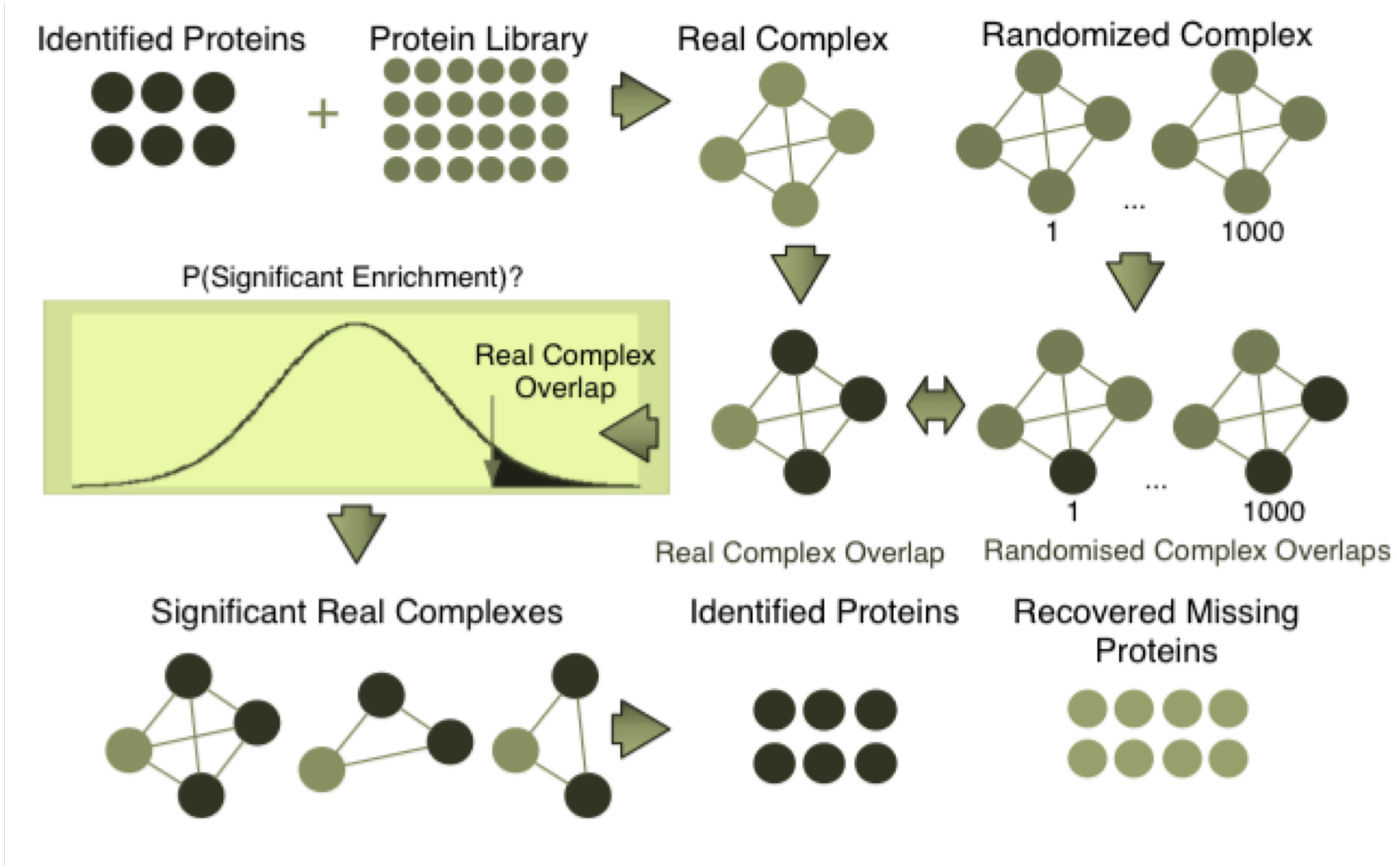
⇒removing noise

Goh and Wong. Integrating networks and proteomics: moving forward. Trends in Biotechnology, 2016

Goh and Wong. Design principles for clinical network-based proteomics. Drug Discovery Today, 2016

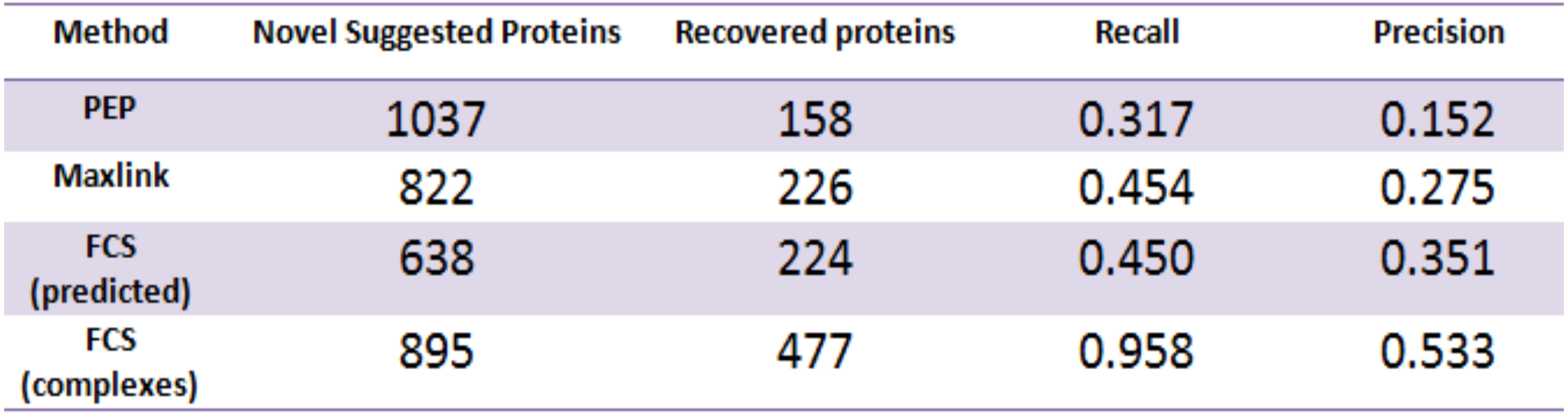
How about we use the idea of "guilt-by-association?"

The functional class scoring (FCS) algorithm



Goh et al. Comparative network-based recovery analysis and proteomic profiling of neurological changes in valporic acid-treated mice. *JPR*, 2013

Does context really work?



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

But we can’t rank the individual proteins simply based on p-values. Can we do better? This is a story for another time. Or simply refer to https://www.comp.nus.edu.sg/~wongls/talks/wls-incob2017.pdf

Goh et al. Comparative network-based recovery analysis and proteomic profiling of neurological changes in valporic acid-treated mice. *JPR*, 2013

What have we learnt?

- Getting good quality PSMs requires consideration of a large number of factors
- The p-value, FDR and PEP are used as statistical approaches for different purposes
- There are 3 strategies for creating decoy libraries in FDR estimation
- Proteomics is plagued with coverage and consistency issues, requiring various rescue analysis

You should be able to

- Describe the various factors affecting PSM quality
- Describe p-values, FDR and PEP
- Describe and evaluate the various decoy library generation strategies (sequence reversal, sequence randomization) for FDR estimation
- Describe coverage and consistency issues in proteomics

Readings

- Elias and Gygi. Target-Decoy Search Strategy for Mass Spectrometry-Based Proteomics. Methods Mol Biol. 604: 55–71, 2010.
- 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
- Goh and Wong. Advanced bioinformatics methods for practical applications of proteomics. Briefings in Bioinformatics, 2017 (https://doi.org/10.1093/bib/bbx128)
- Käll et al. Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. JPR. 7(1):29-34, 2008

Readings

- Goh and Wong. Integrating networks and proteomics: moving forward. Trends in Biotechnology, 34(12):951-959, 2016
- Goh and Wong. Design principles for clinical network-based proteomics. Drug Discovery Today, 21(7):1130-1138, 2016
- Goh and Wong. Dealing with confounders in omics analysis. Trends in Biotechnology, S0167-7799(18)30047-7, 2018
- Goh and Wong. Advanced bioinformatics methods for practical applications in proteomics. Briefings in Bioinformatics, 2017.
- Zhou et al. Understanding missing proteins: A functional perspective. Drug Discovery Today, 23(3):644--651, March 2018.