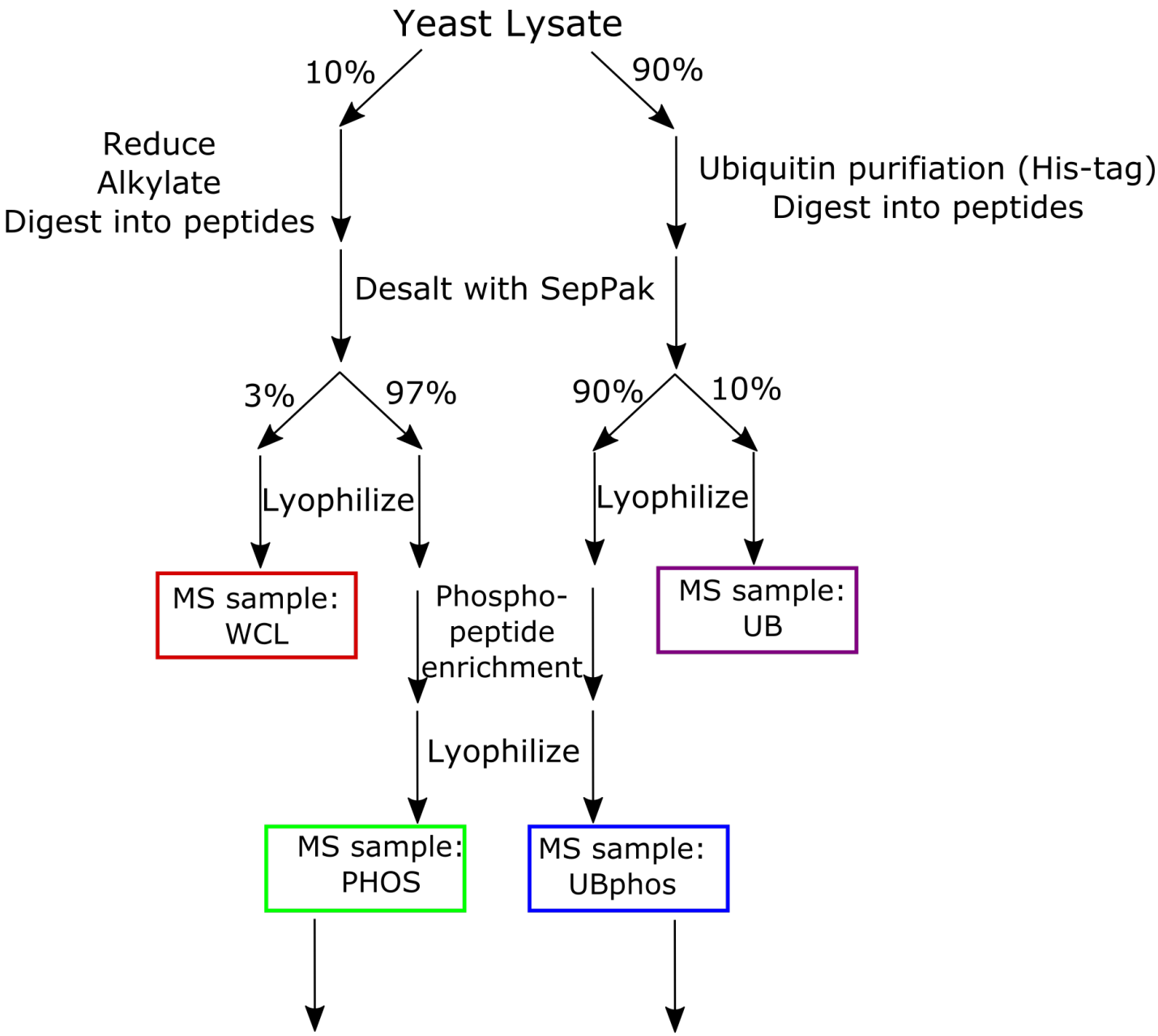
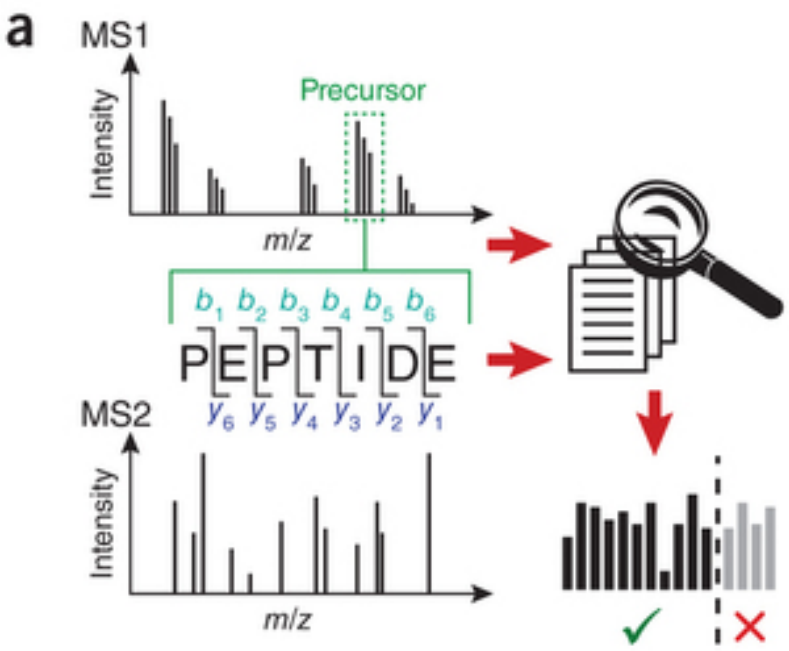
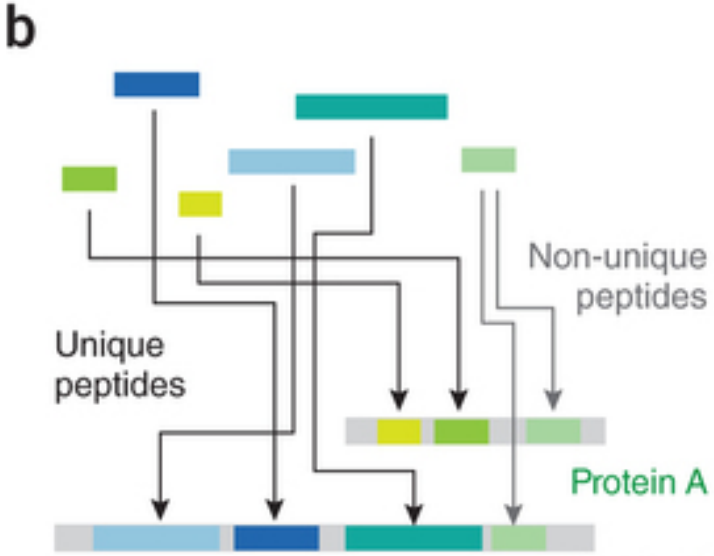
Mass spec data analysis

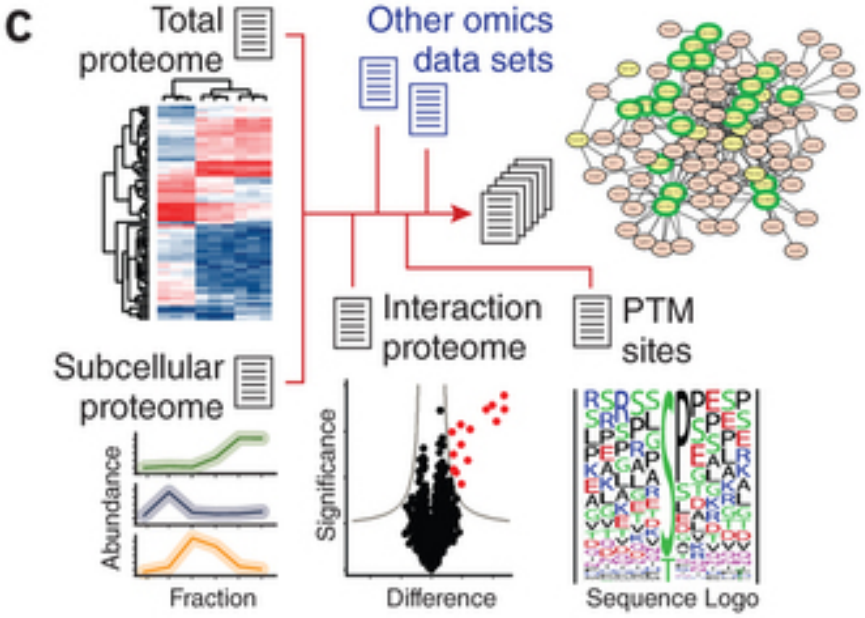
- Live MS stream
- RAW data file
- Proteomics data analysis pipeline steps
- Data Analysis
  - Assumptions
  - Approaches
- Output files
  - What do different columns mean?
  - What columns are useful?


Yeast Lysate
10%
90%
Reduce
Alkylate
Digest into peptides
Ubiquitin purifiation (His-tag)
Digest into peptides
Desalt with SepPak
3%
97%
90%
10%
Lyophilize
Lyophilize
MS sample:
WCL
Phospho-
peptide
enrichment
MS sample:
UB
Lyophilize
MS sample:
PHOS
MS sample:
UBphos

Data analysis pipeline







Peptide fragment fingerprinting (PFF)

protein sequence database

in silico digest

YSFVATAER

HETSINGK

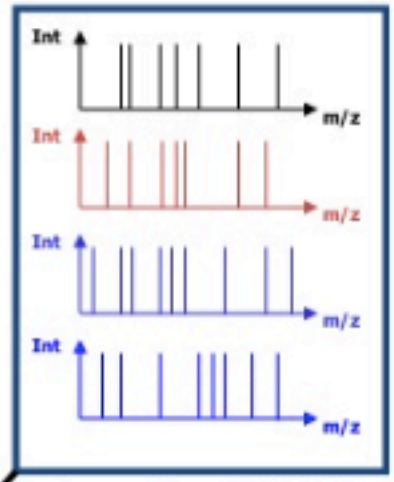
MILQEESTVYYR

SEFASTPINK

...

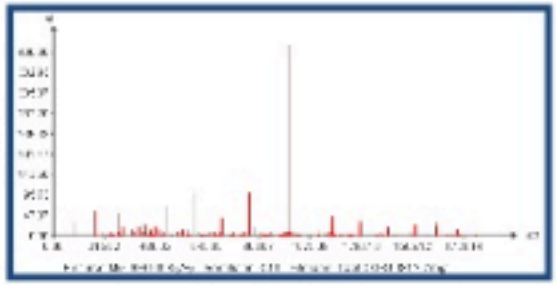
peptide sequences

in silico MS/MS



theoretical MS/MS spectra

in silico matching



experimental MS/MS spectrum

1) YSFVATAER 34
2) YSFVSAIR 12
3) FFLIGGGGK 2

peptide scores

The Results: Distinguishing Right from Wrong

In large proteomics data sets (for which manual data inspection is impossible), how can we distinguish between correct and incorrect peptide assignments?

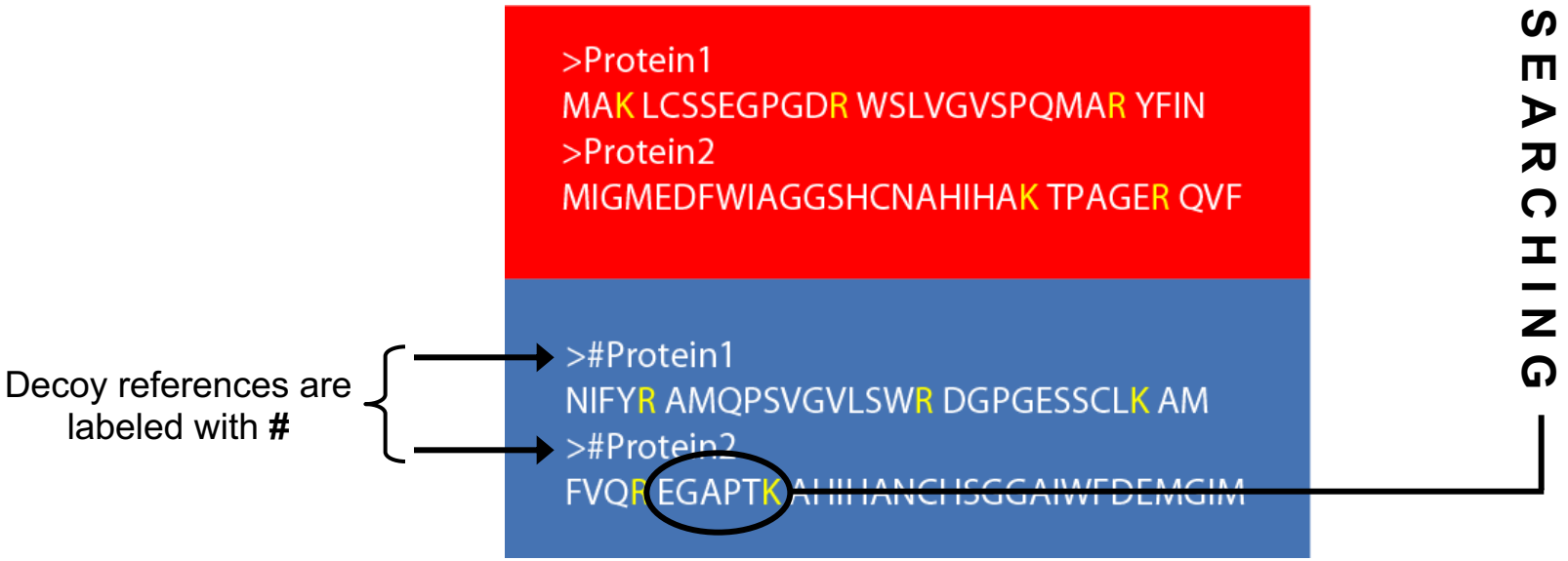
Use "decoy" sequences to distract non-peptidic, non-uniquely matchable, or otherwise unmatchable spectra into a search space that is known *a priori* to be incorrect

Use the frequency of "decoy" sequences among total sequences to estimate the overall frequency of wrong answers (False Positive Rate)

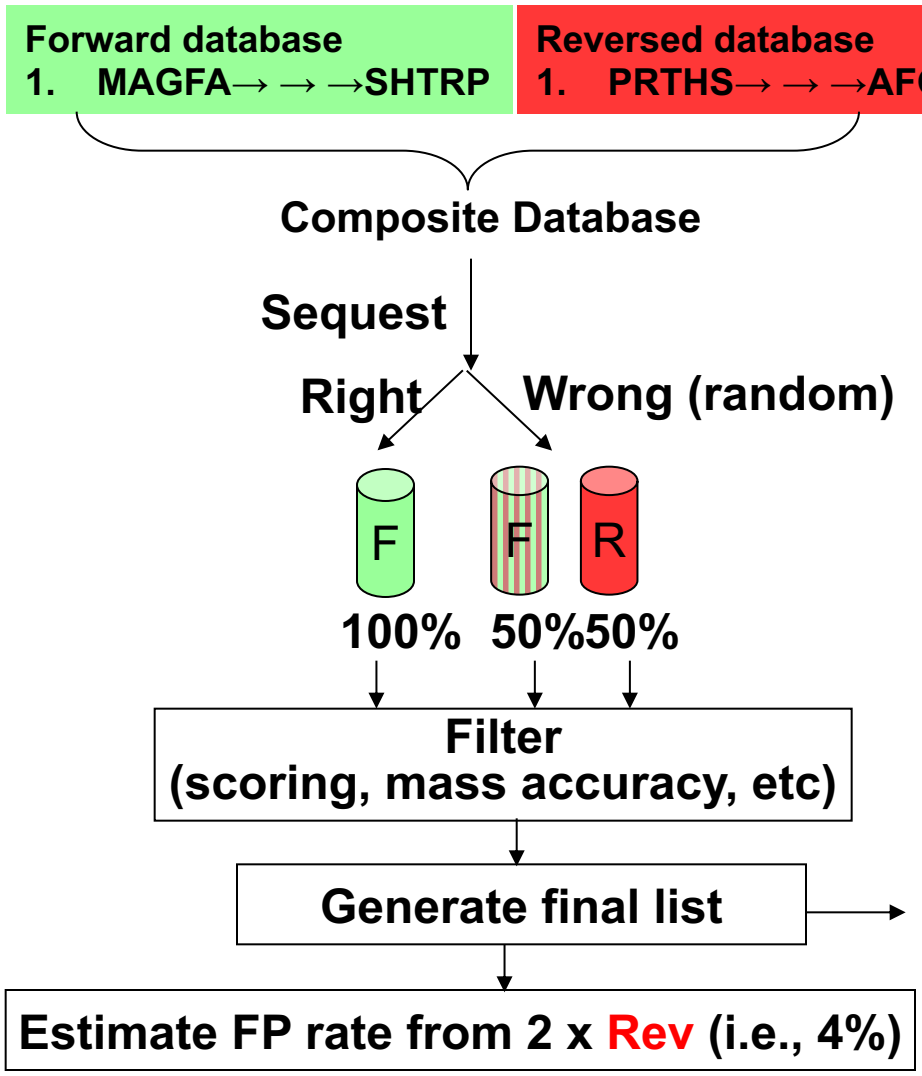
Adjust filtering criteria to achieve a ~ 1% False Positive Rate

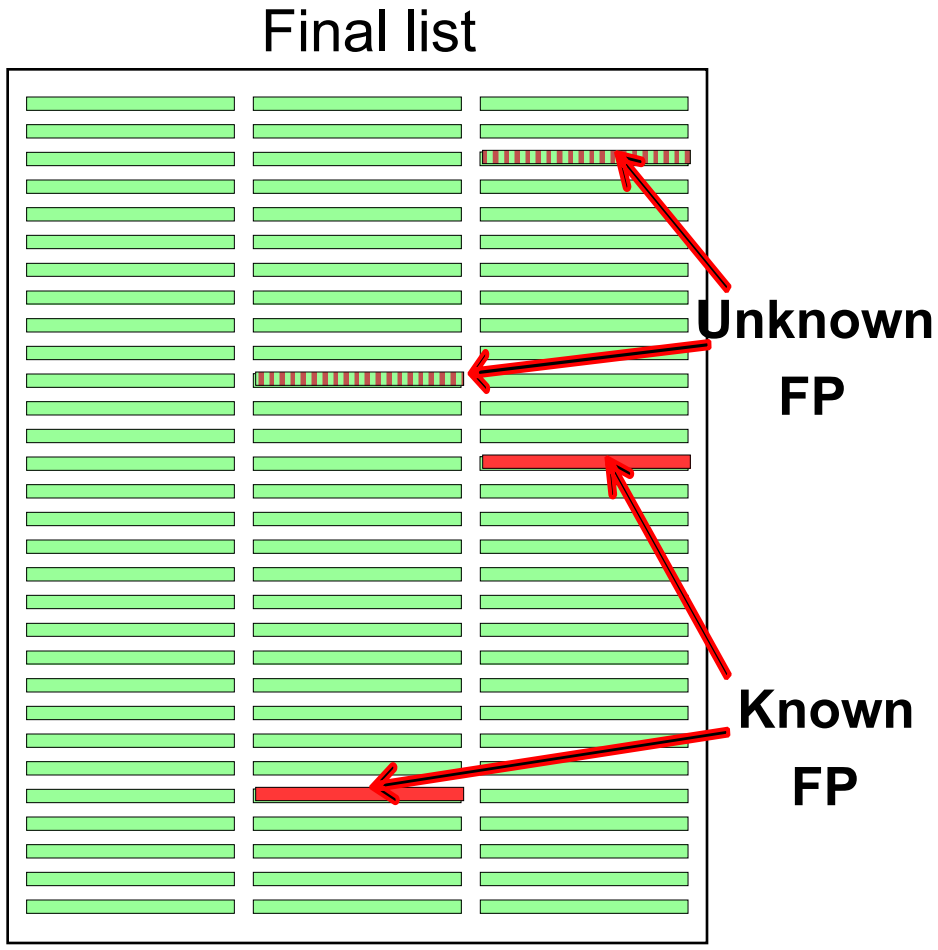
Decoy Sequences? A "Reversed" Database!

We generate decoy sequences by reversing each protein sequence in a given database, such that the resultant *in silico* digest contains nonsense peptides, then append the reversed database to the end of the forward database

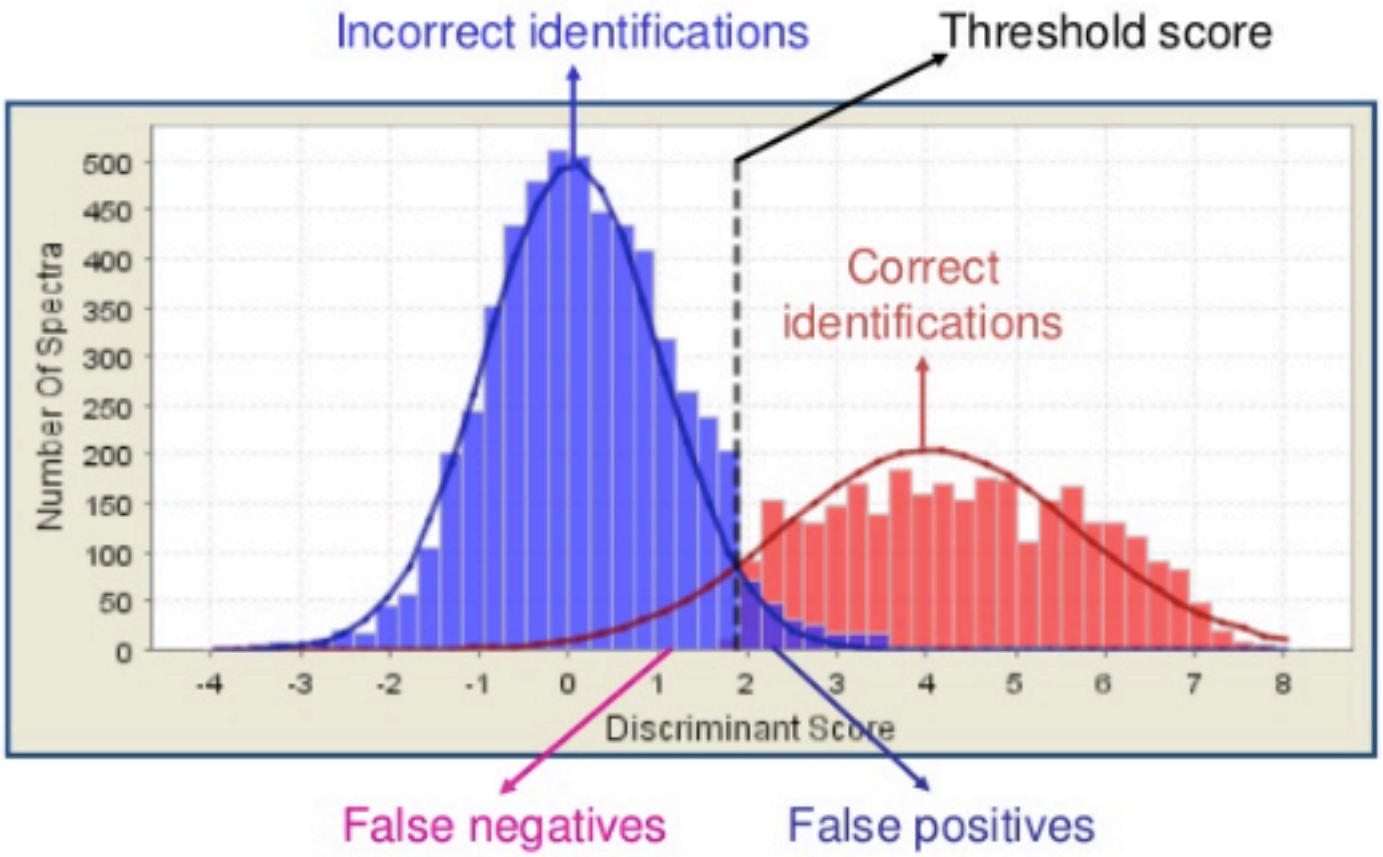


Target/Decoy Database Searching



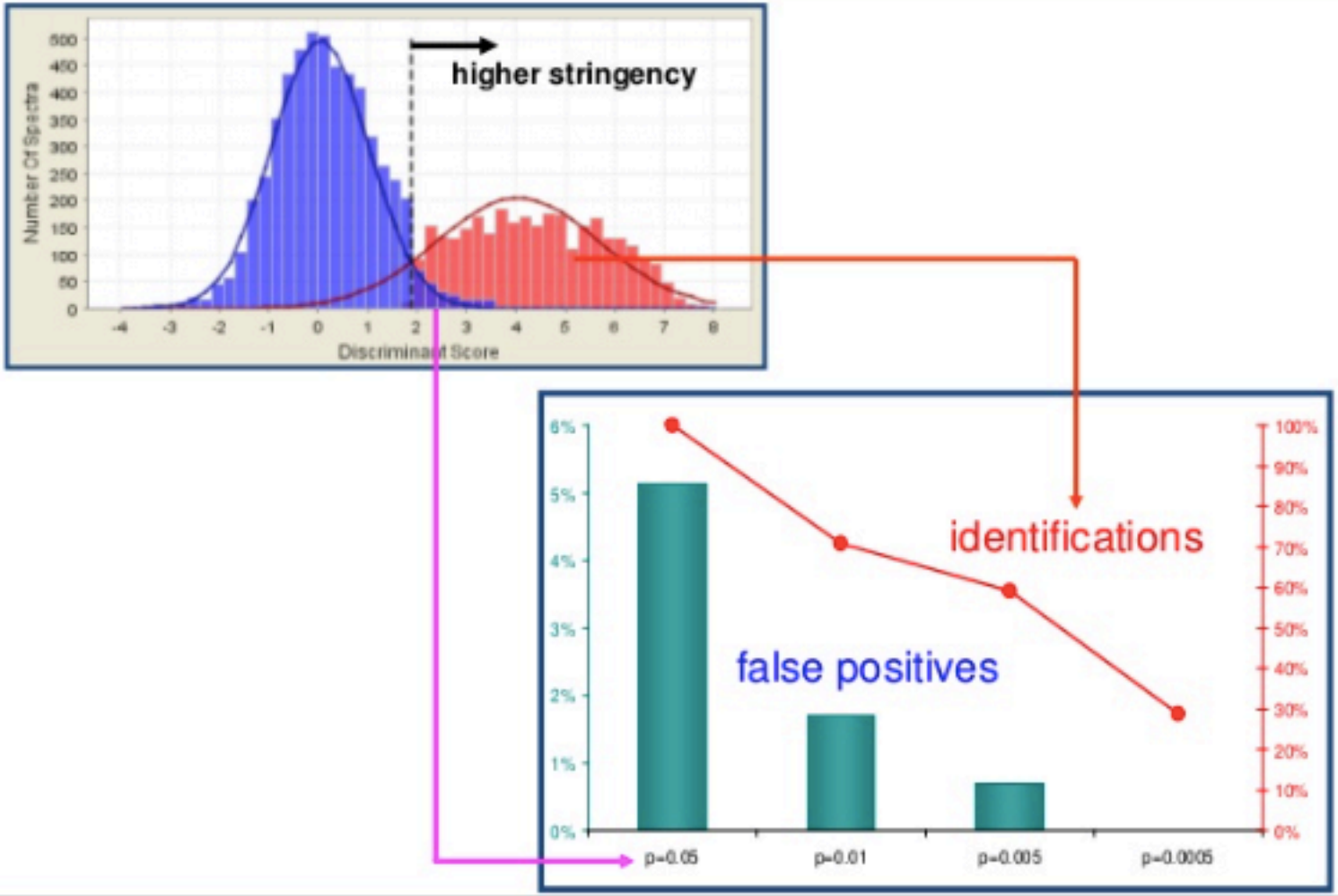


Overall concept of scores and cut-offs

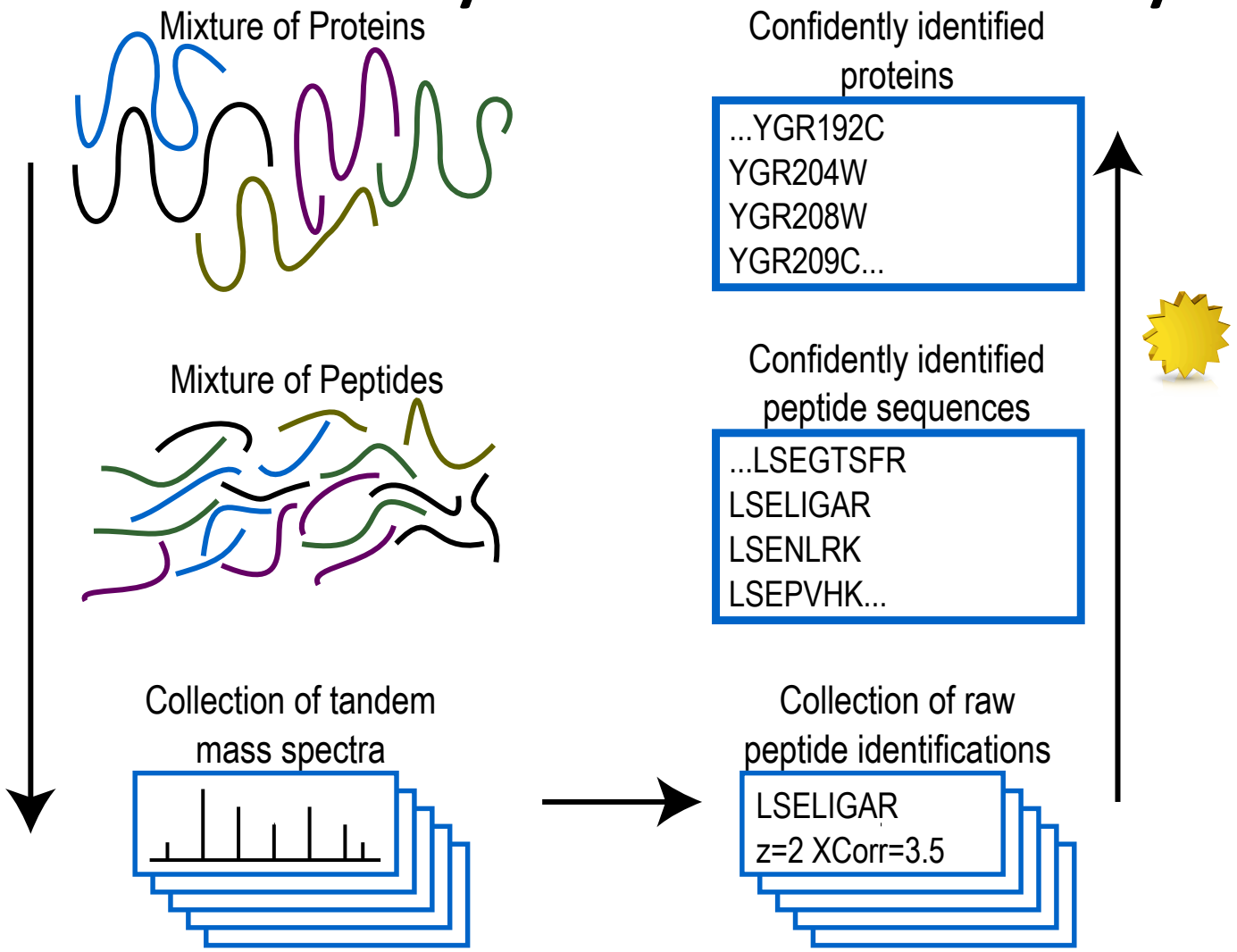


***Adapted from*: www.proteomesoftware.com – Wiki pages**

Playing with probabilistic cut-off scores



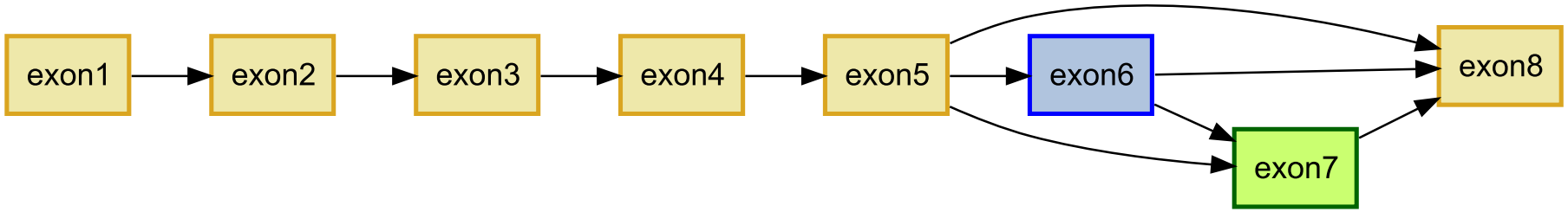
Disassembly and reassembly



After AI Nesvizhskii, *Mol Cell Proteomics* (2005) 4: 1419-40.

Protein isoforms

- A single gene may give rise to many transcripts that overlap for one or more exons.
- When isoforms are listed as separate proteins in the FASTA, a peptide may match a shared or distinctive part of a protein sequence.
- VEGF incorporates eight exons, where either 6 or 7, both, or neither may be incorporated.



What's so tricky about assembly?

- The proteins containing each identified peptide are reported by identifying algorithm.
- A maximal list of proteins includes each that contains any observed peptide.
- Some proteins are indistinguishable on the basis of observed peptides.
- Remaining proteins may overlap with others in their observed peptides.

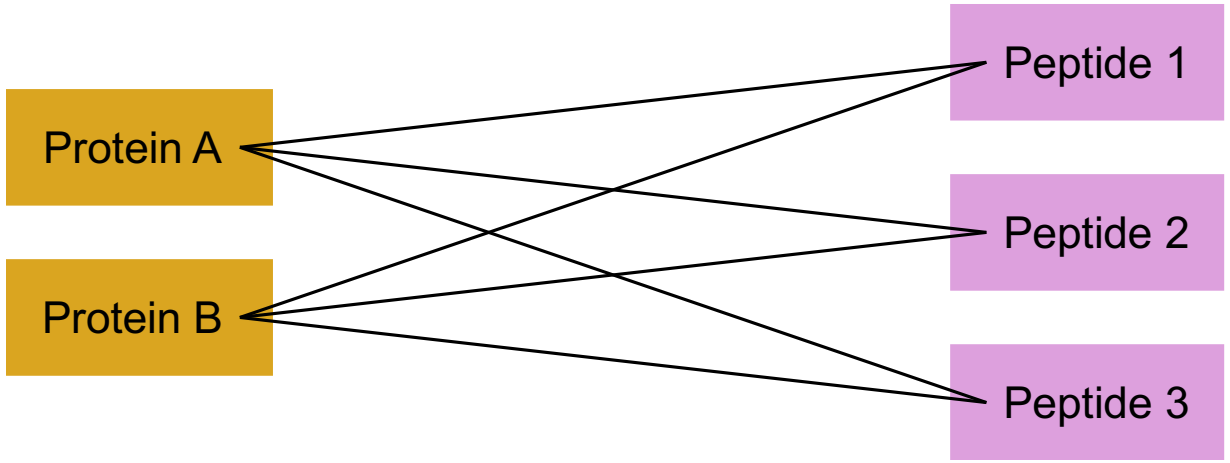
Parsimony

- *noun*: “economy of explanation in conformity with Occam's razor”
  - Merriam Webster OnLine

- “*Plurality ought never be posed without necessity.*”
  - William of Occam

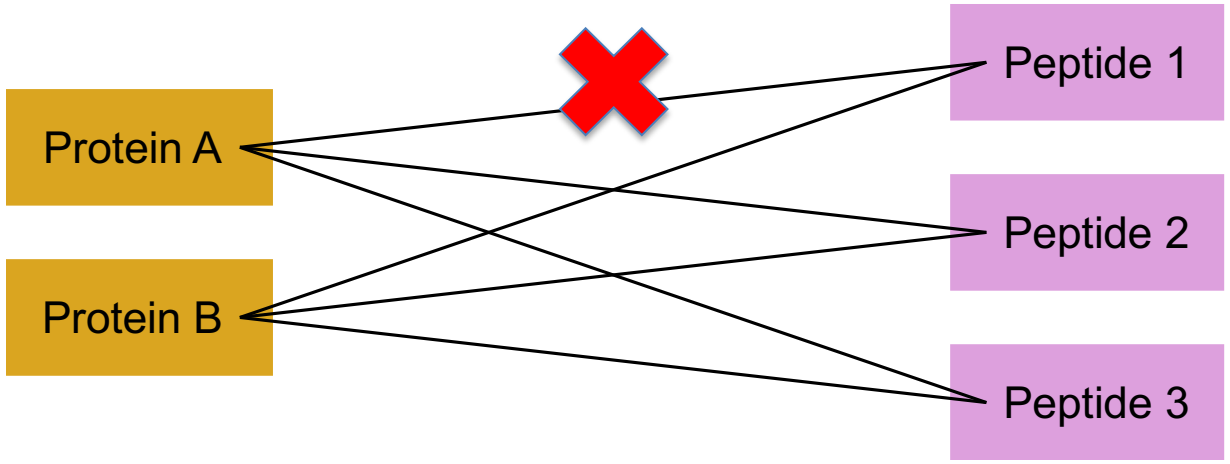
Indiscernible / equivalent proteins

- Occur when two proteins are equally good at explaining a set of peptides.
- Counted as a single *protein group* in most software packages.



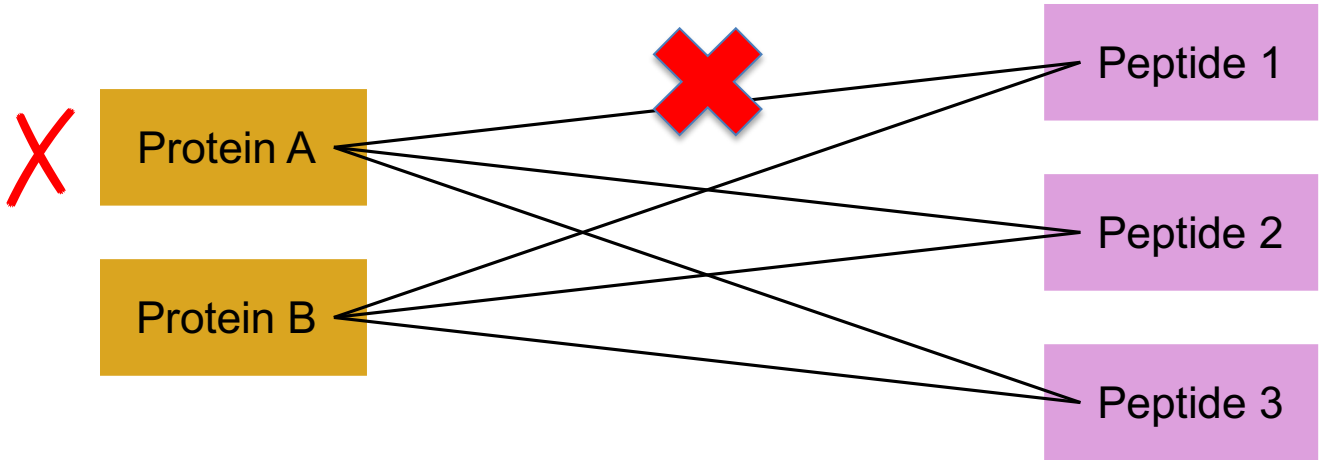
Indiscernible / equivalent proteins

- Occur when two proteins are equally good at explaining a set of peptides.
- Counted as a single *protein group* in most software packages.

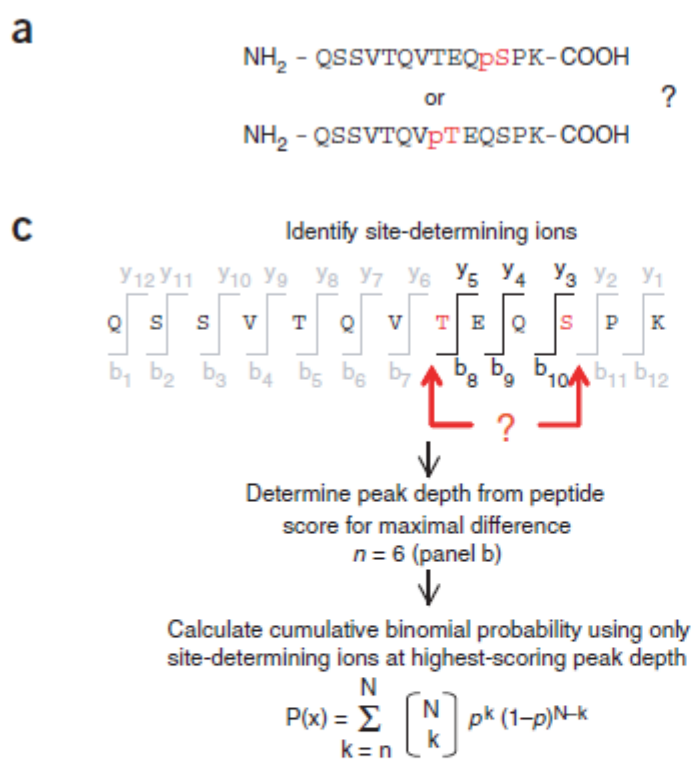


Indiscernible / equivalent proteins

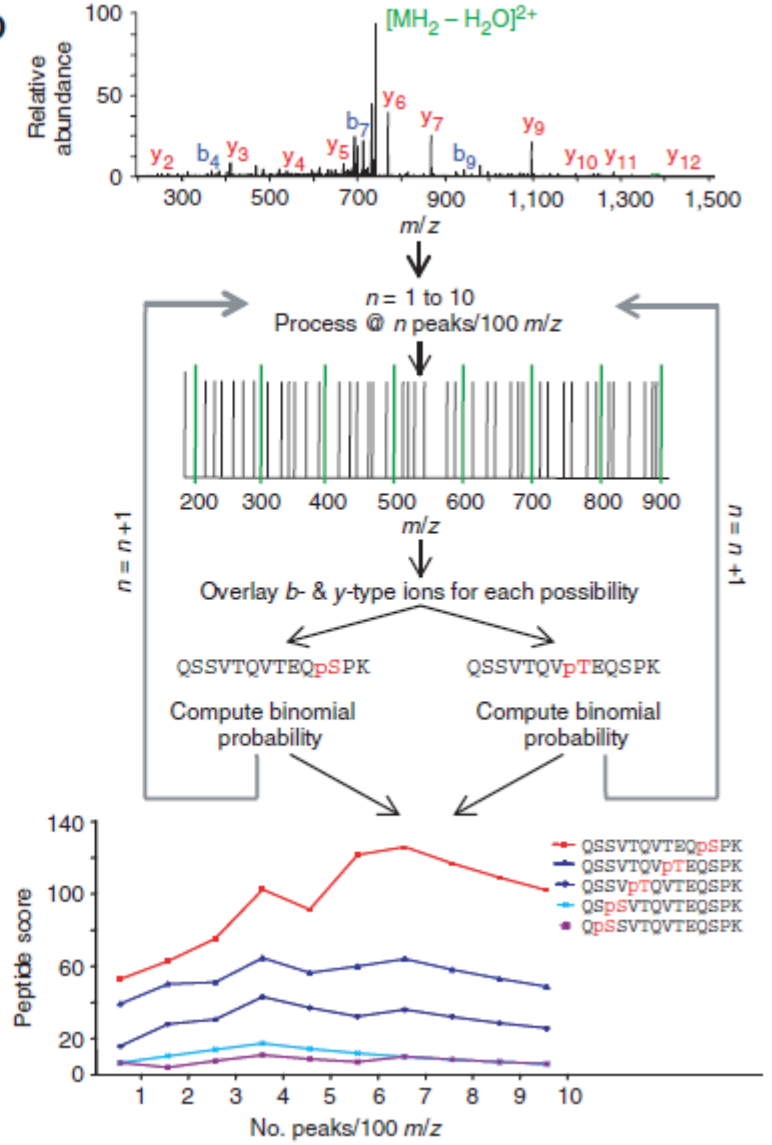
- Occur when two proteins are equally good at explaining a set of peptides.
- Counted as a single *protein group* in most software packages.



phosphorylation site localization

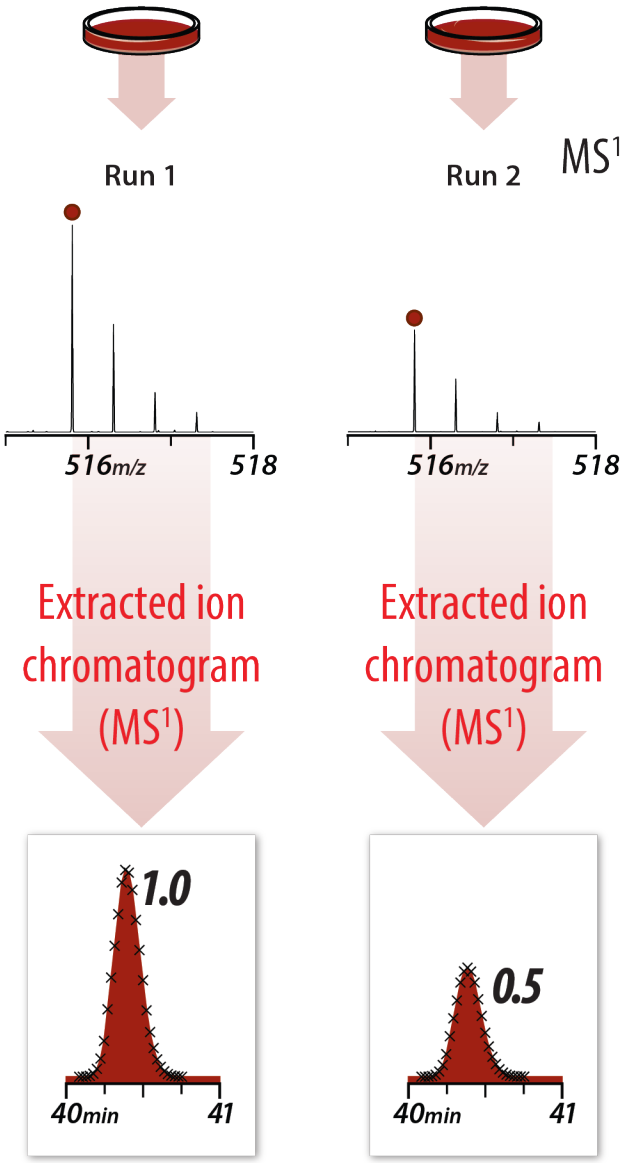


Phosphopeptide	QSSVTQVTE QpS PK	QSSVTQ VpTEQSP K
Trials (*N*)	6 (y_3, y_4, y_5, b_8, b_9, b_{10})	6 (y_3, y_4, y_5, b_8, b_9, b_{10})
Successes (*n*)	5 (y_3, y_4, y_5, b_9, b_{10})	0
p (6 peaks / 100 *m*/*z*)	0.06	0.06
P	0.0000044	1.0
Score [–10 × log(*P*)]	53.57	0
Ascore = ambiguity score (difference of the top two candidates)	53.57 – 0 = 53.57	

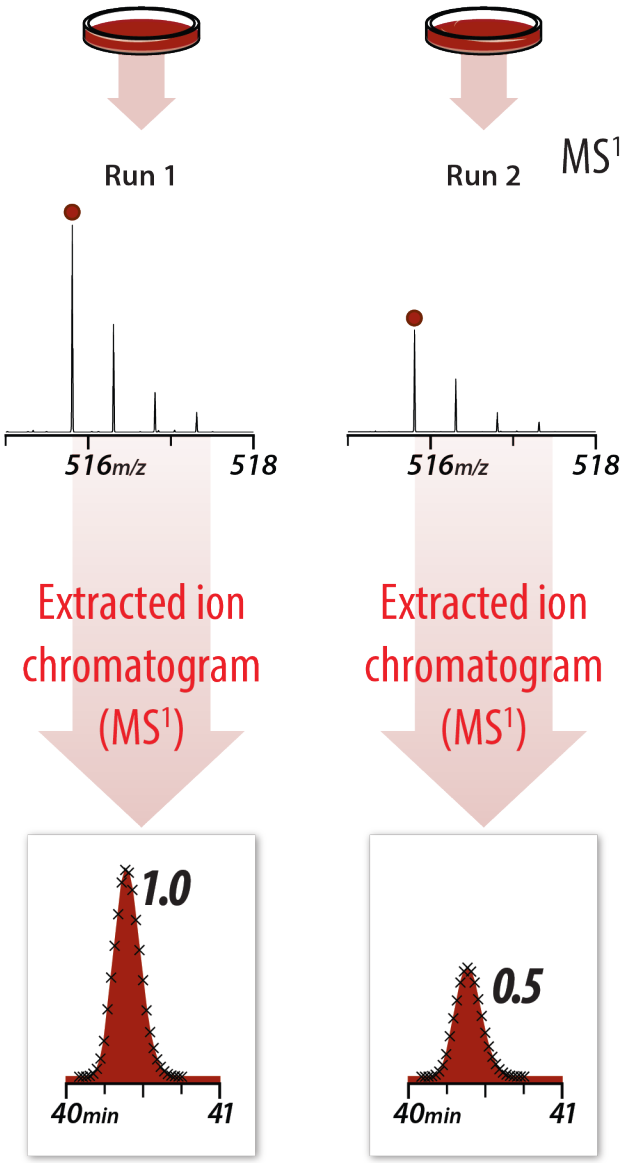


Beausoleil et al., *Nat. Biotechnol,* 2006

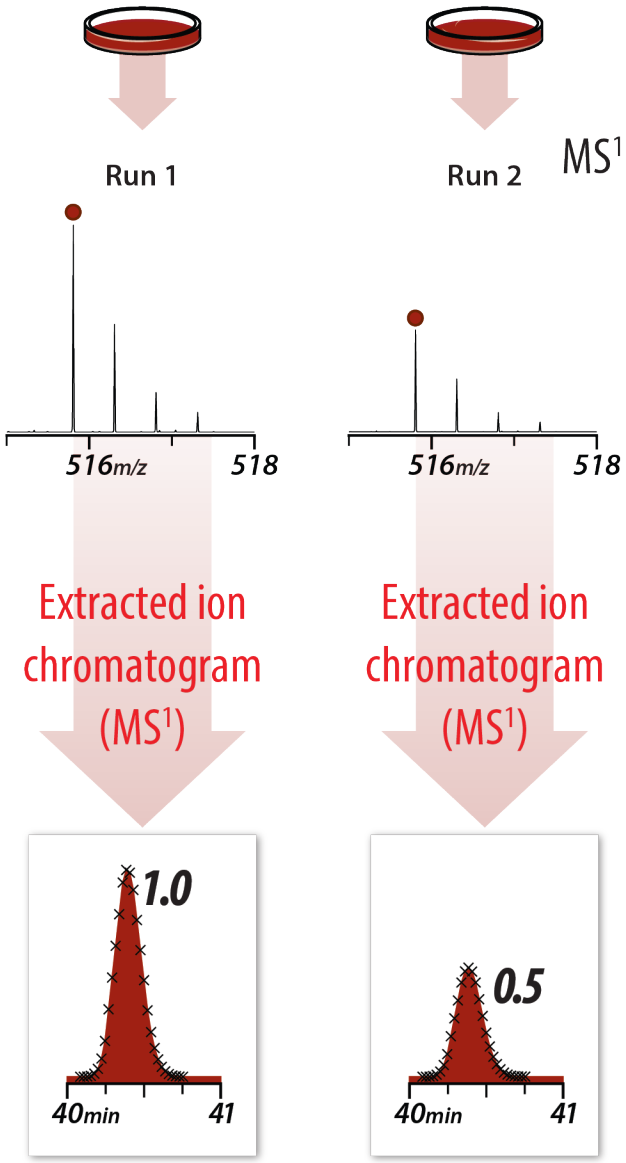
Label free quantitation AUC

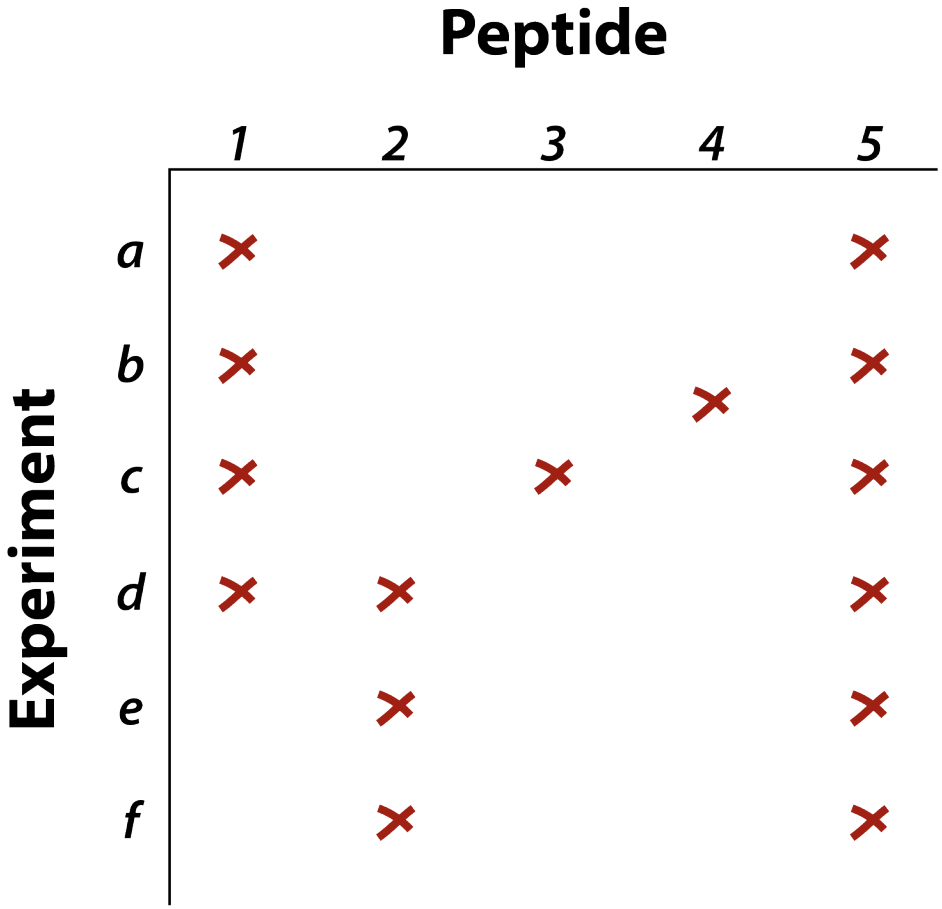


Label free quantitation AUC



Label free quantitation AUC





Data results:

Good news: every team had at least 1 nice looking phosphorylation run

Bad news: The Ubiquitin phosphorylation analysis didn't go so well.

Box Color Code:

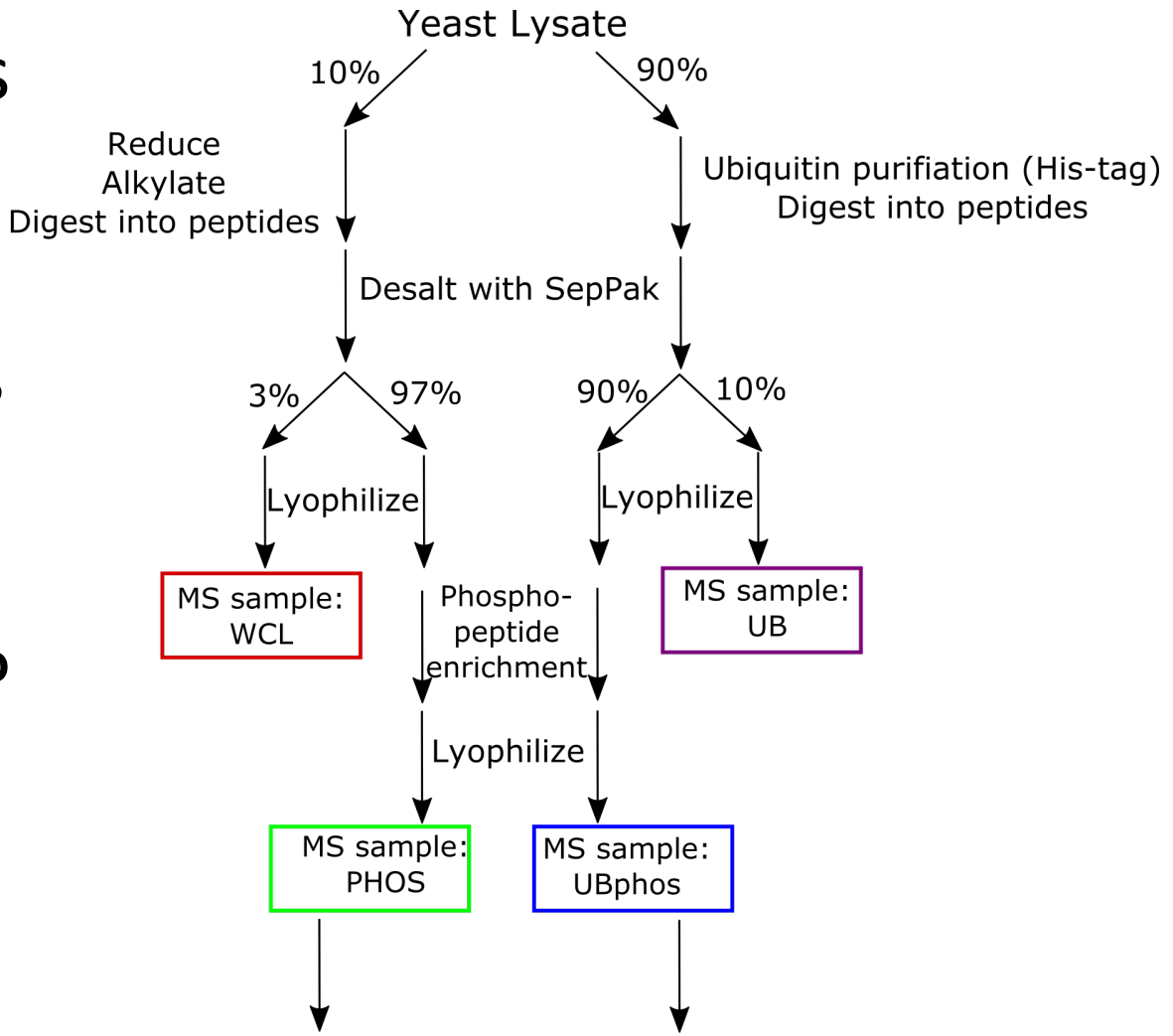
Green = Good signal, normal looking run

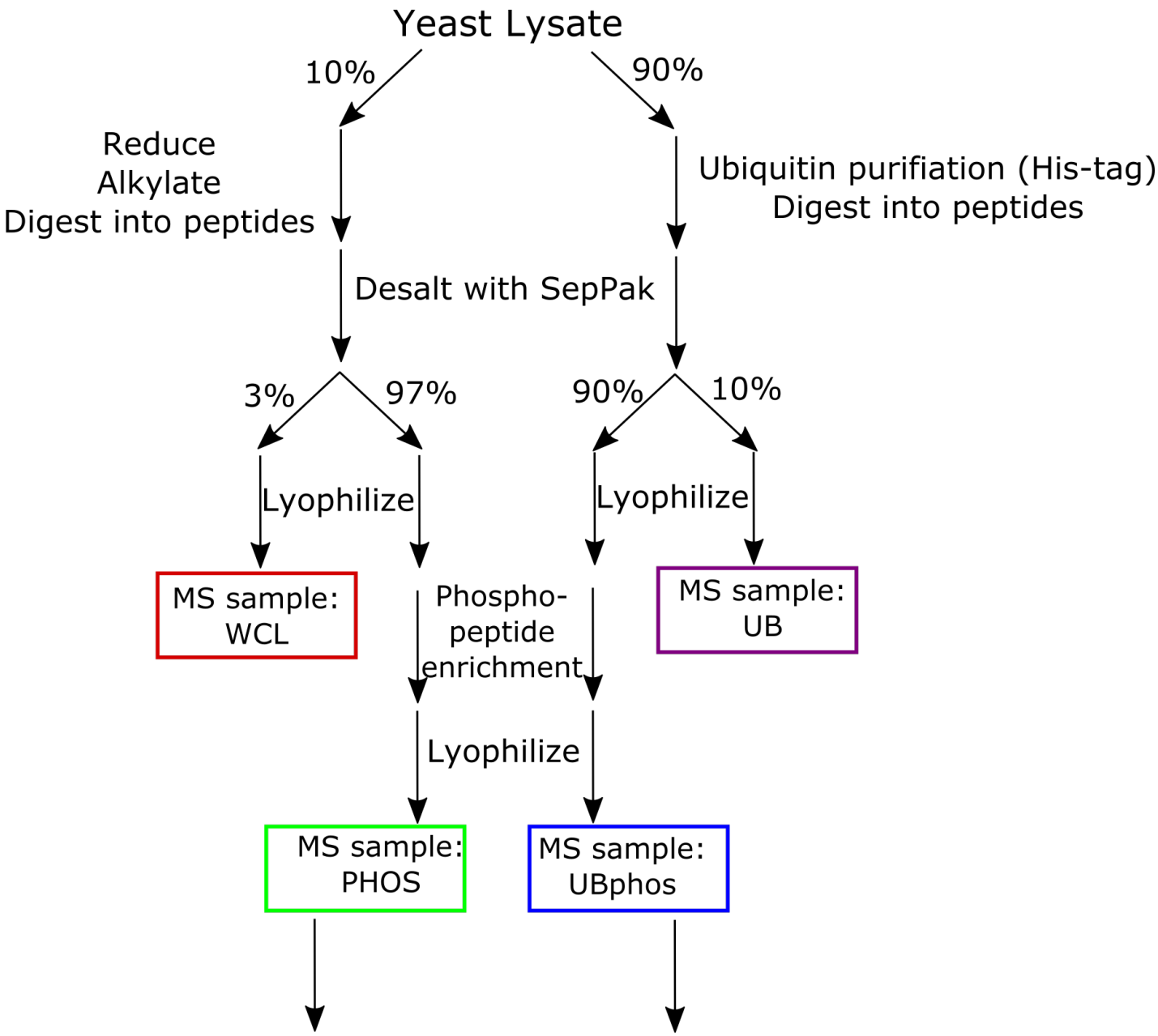
Red = Very low or no signal

Team	Condition	Ub	WCL	Ub Phos	WCL phos
APUBSCRAWL	Swe1 KO	QE20150928-51	QE20150928-53	QE20150928-92	QE20150928-52
APUBSCRAWL	Cerulenin	QE20150928-56	QE20150928-58	QE20150928-55	QE20150928-57
Control (David Mavor)	WT	QE20150928-102	QE20150928-104	QE20150928-101	QE20150928-103
EtOH	Kin3 KO	QE20150928-23	QE20150928-24	QE20150928-82	QE20150928-83
EtOH	menadione	QE20150928-95	QE20150928-97	QE20150928-94	QE20150928-96
ONION	Atg1 KO	QE20150928-78	QE20150928-80	QE20150928-77	QE20150928-79
ONION	rapamycin	QE20150928-71	QE20150928-73	QE20150928-70	QE20150928-72
PYND	Alk KO	QE20150928-37	QE20150928-38	QE20150928-89	QE20150928-90
PYND	5-fluorocytosine	QE20150928-40	QE20150928-41	QE20150928-85	QE20150928-86
SHMOO	CaCl2	QE20150928-34	QE20150928-35	not run, causing MS problems	not run, causing MS problems
SHMOO	CMK1 KO	QE20150928-43	QE20150928-46	QE20150928-44	QE20150928-45
WHANGEE	Tpk1 KO	QE20150928-61	QE20150928-63	QE20150928-60	QE20150928-62
WHANGEE	Tunicamycin	QE20150928-66	QE20150928-68	QE20150928-65	QE20150928-67

Data assumptions

- What assumptions can we make about your individual samples and the differences between samples?




Yeast Lysate
10%
90%
Reduce
Alkylate
Digest into peptides
Ubiquitin purifiation (His-tag)
Digest into peptides
Desalt with SepPak
3%
97%
90%
10%
Lyophilize
Lyophilize
MS sample:
WCL
Phospho-
peptide
enrichment
MS sample:
UB
Lyophilize
MS sample:
PHOS
MS sample:
UBphos

Data analysis approaches?

- What types of things can we look for?
  - Do you have good data?
  - Do you need/want to filter some things out first?
  - How do we decide what is a difference between control and KO/chemical?
  - What are you looking for from your different samples?
  - What biological attributes can we consider?
  - Sharing & Comparing between groups HIGHLY encouraged!

Files you have:

- Summary file: higher level comparison of runs
  - MS/MS identified column give you a quick metric of data quality
- Experimental design file: simple text file saying which mass spec data file goes with what experiment
- Evidence file: lots of detailed info. Don't plan on using it.
- Peptides:
- Protein Groups:
- Phospho(STY)Sites: Very useful.

What columns are most useful?

- Phospho(STY) file:
  - Protein: systematic name
  - Fasta header: will have the common gene name and protein description
  - Position within protein: phosphorylation site position
  - Localization probability: confidence of phosphorylation site assignment. Above 0.75 is high confidence. Is a general one for entire dataset, or an experiment specific score.
  - Intensity “Blank experiment”: abundance in that experiment. Don’t use the ones with “__#”
  - Id: unique identifier for each row
  - Evidence ID: if you want to map to evidence file
  - Reverse and contaminants

What columns are most useful?

- Protein Groups file:
  - Protein IDs
  - Fasta headers
  - Number of proteins – number of proteins within the protein ID group
  - Mol. Weight [kDa]
  - Sequence length: our could easily pull from fasta file
  - Intensity “Blank experiment”: abundance in that experiment. Don’t use the ones with “__#”
  - Reverse and contaminants
  - Id: unique identifier for each row