

# Data Independent Acquisition Strategies for Quantitative Proteomics: The Challenges of Scaling Up to Meet Demand

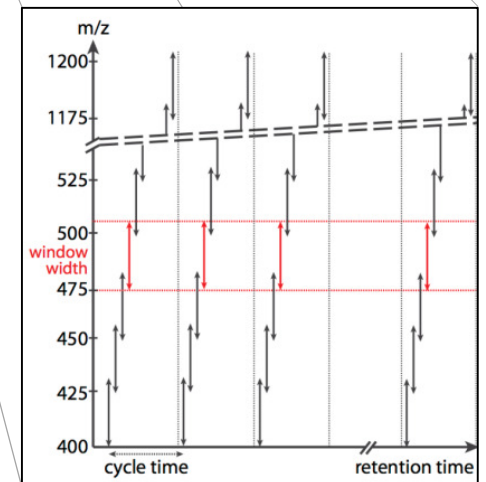
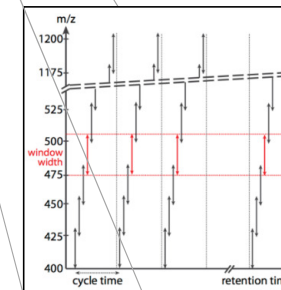
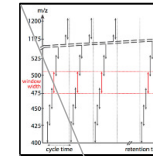
Data Independent Acquisition Interest Group

Presiding: Ben Collins (ETH Zurich), Isabell Bludau (ETH Zurich)

65th ASMS Conference on Mass Spectrometry and Allied Topics

June 7th, 2017

Room: 235 – 238 -- Indiana Convention Center, Indianapolis, IN



Real time poll

[bit.ly/2sEPrpu](https://bit.ly/2sEPrpu)

# The breadth of DIA methodology is increasing

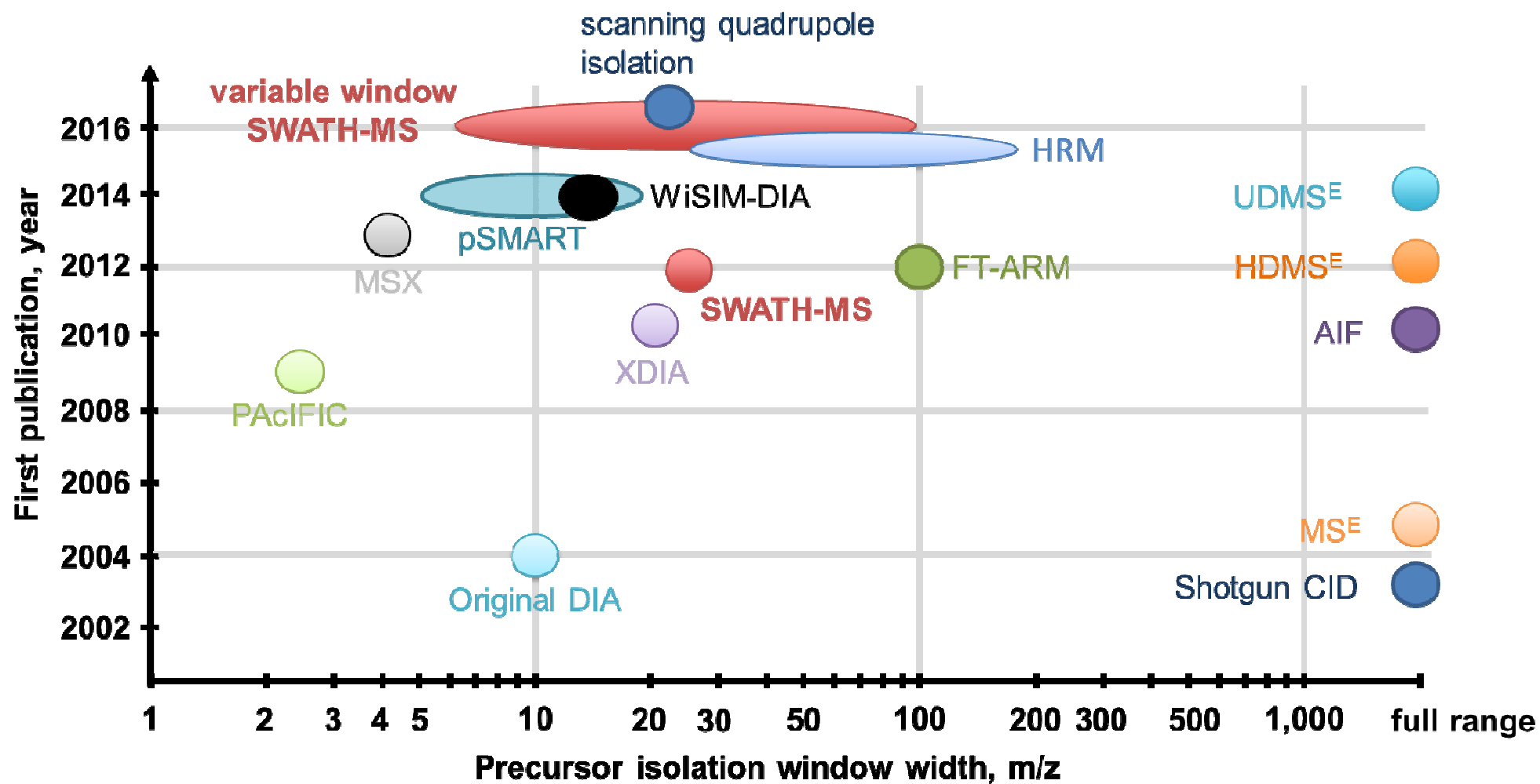
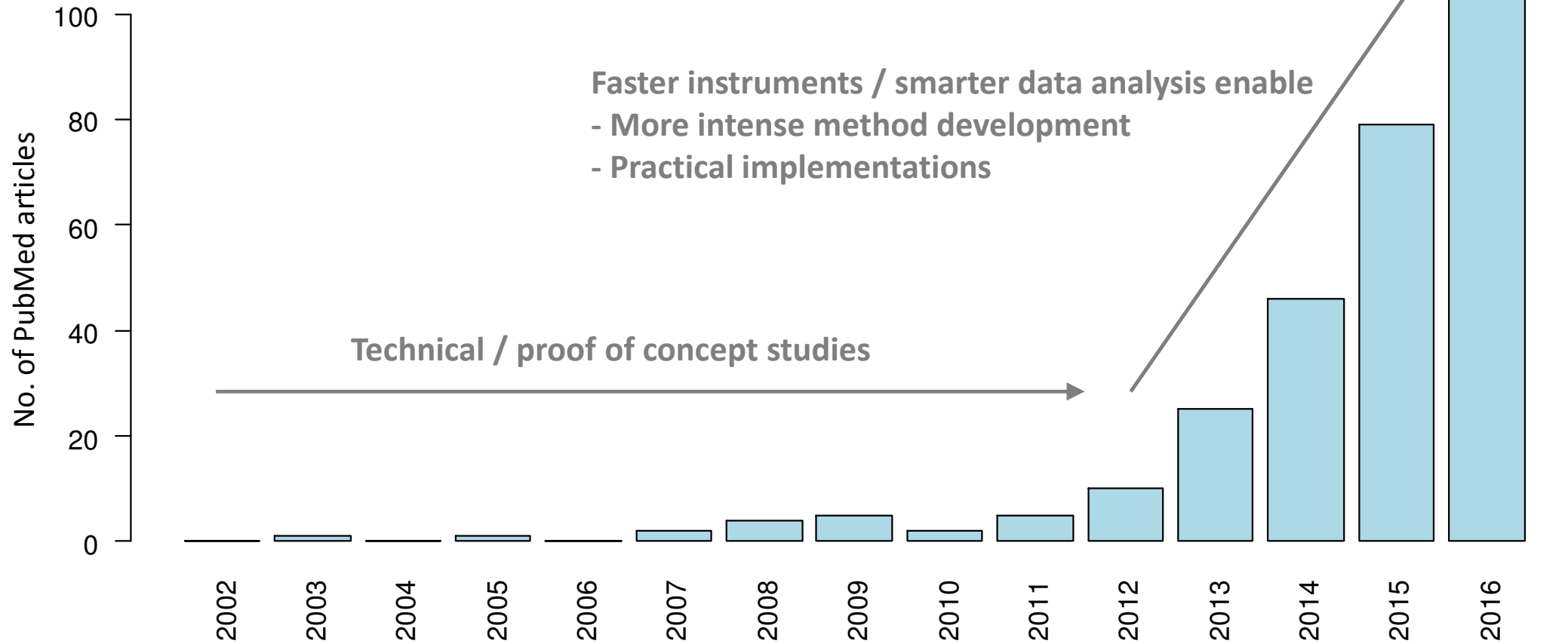


Figure from Tina Ludwig – SWATH-MS tutorial (in prep)

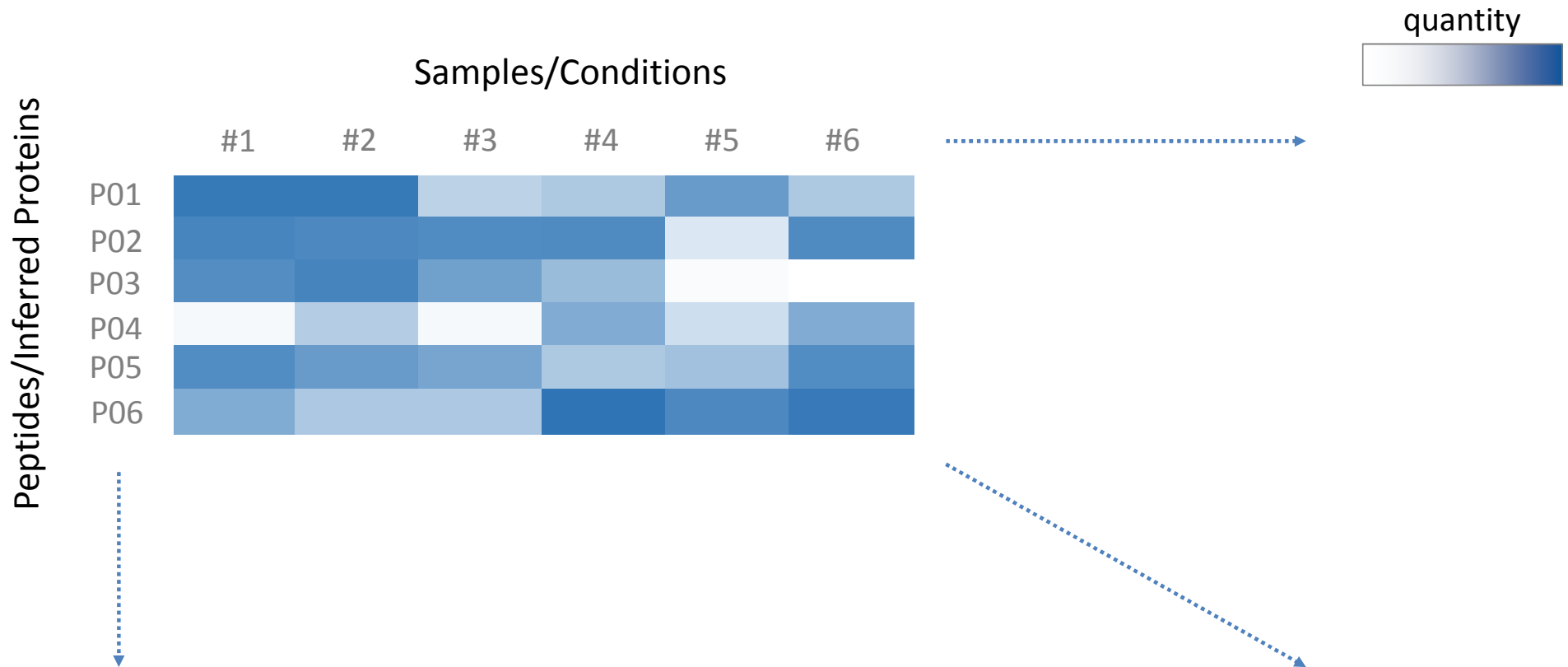
# The development and application of DIA to life science research is increasing

PubMed query:

((proteom\* OR protein) OR peptide) AND ("data independent acquisition" OR SWATH)

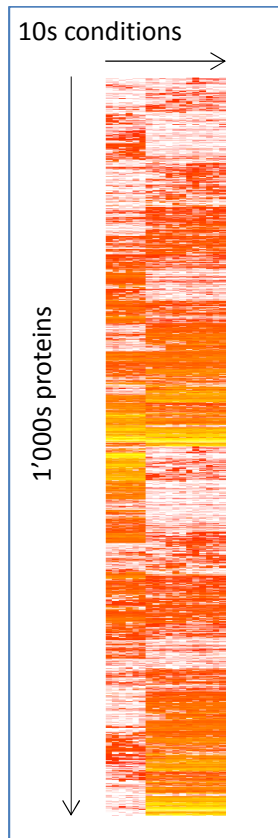


# Scaling up – in which dimensions?

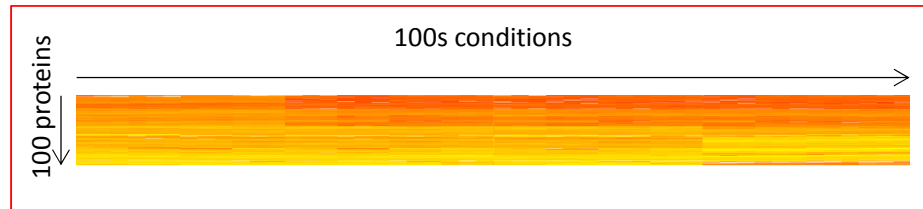


# The primary goal of DIA is data completeness

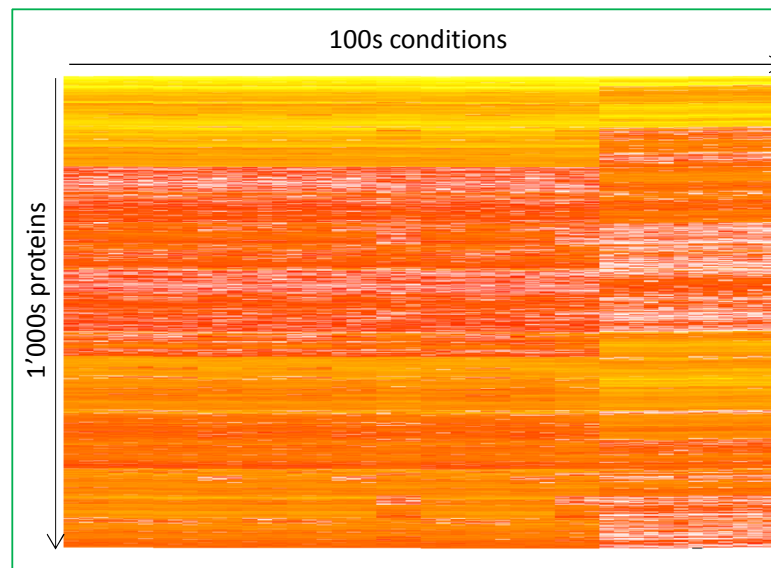
## Historical problem of DDA based approaches



## Targeted proteomics (SRM/PRM)



## DIA



Broad/deep coverage

Flexibility in targets

# Scaling up DIA – no. of samples/conditions

## Population-based analysis (personalized med., etc)

## Large-scale knock-out screens

Published online: February 4, 2015

Article

Quantitative variability of 342 plasma proteins in a human twin population

Yansheng Liu<sup>1,\*,†</sup>, Alfonso Buil<sup>2,†</sup>, Ben C Collins<sup>1,†</sup>,  
Olga Vitek<sup>3</sup>, Jeppe Mouritsen<sup>1</sup>, Genevieve Lachy<sup>1</sup>,  
Ruedi Aebersold<sup>1,5,\*\*</sup>

TRANSPARENT PROCESS OPEN ACCESS

molecular systems biology

bioRxiv beta

Precise label-free quantitative proteomes in high-throughput by microLC and data-independent SWATH acquisition

Michael Müller<sup>1,2</sup>

**Issues - scaling up samples?**

- Error control (FDR)
- Data completeness
- Quantitative precision/accuracy
- Quality control

**Workshop**

✓  
✓  
✓ ✗  
✗

RESEARCH ARTICLE

PROTEOMICS

Systems proteomics of mitochondria function

Evan G. Williams,<sup>1\*</sup> Yibo Wu,<sup>2\*</sup> Pooja Jha,<sup>1</sup> Sébastien Dubuis,<sup>2</sup> Peter Blattmann,<sup>2</sup>  
Carmen A. Argmann,<sup>3</sup> Sander M. Houten,<sup>3</sup> Tiffany Amariuta,<sup>1</sup> Witold Wolski,<sup>2</sup>  
Nicola Zamboni,<sup>2</sup> Ruedi Aebersold,<sup>2,4,†</sup> Johan Auwerx<sup>1,†</sup>

232 human

Genetic as

386 mouse liver samples

sample numbers

# Scaling up DIA – no. of peptides/inferred proteins



focused

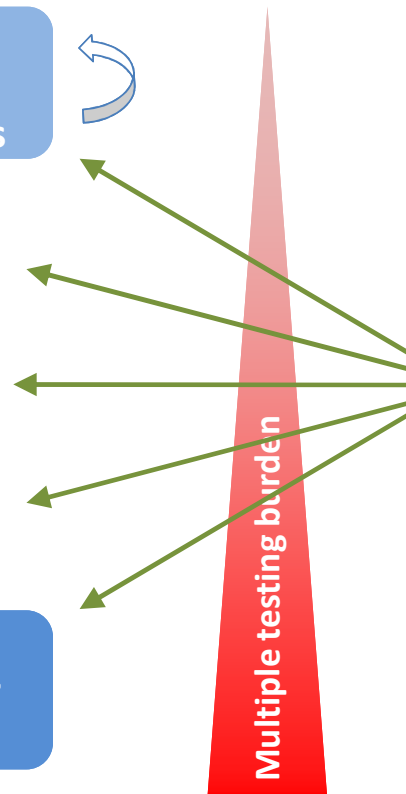


DAI 'targeting' scale

SRM style  
quantification  
with known targets

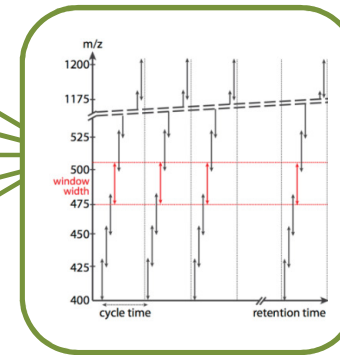


No hypothesis --  
Quantify whatever  
we can



Multiple testing burden

Statistics



Flexibility

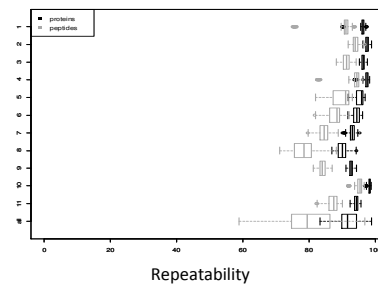
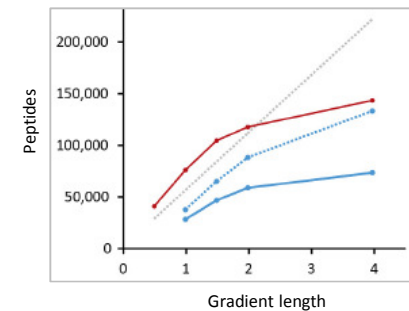
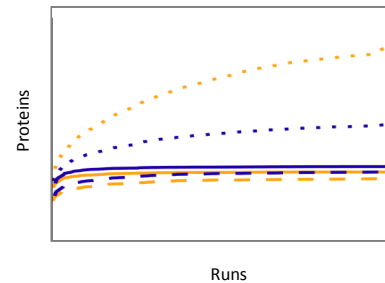
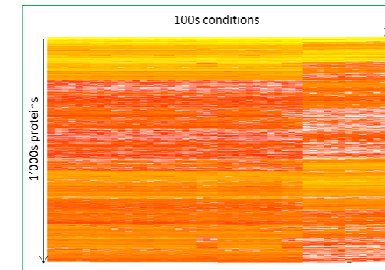
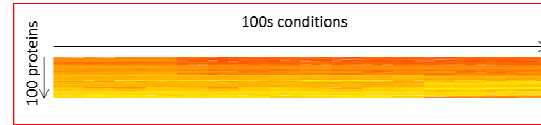


global



# Topics for discussion

1. Mike MacCoss (Univ. of Washington)
  - DIA as targeted proteomics
2. Alexey Nesvizhskii (Univ. of Michigan)
  - DIA as discovery proteomics
3. Isabel Bludau (ETH Zurich)
  - Error rate control at various levels
4. Lukas Reiter (Biognosys)
  - Depth of proteome coverage
5. Ben Collins (ETH Zurich)
  - Repeatability and data completeness (optional - depending on time)



# The goal is a community discussion!

- If you have a question or comment
  1. Raise your hand
  2. Shout
  3. Throw something
  4. Use the ASMS app
- Let's answer Question 1 in the poll

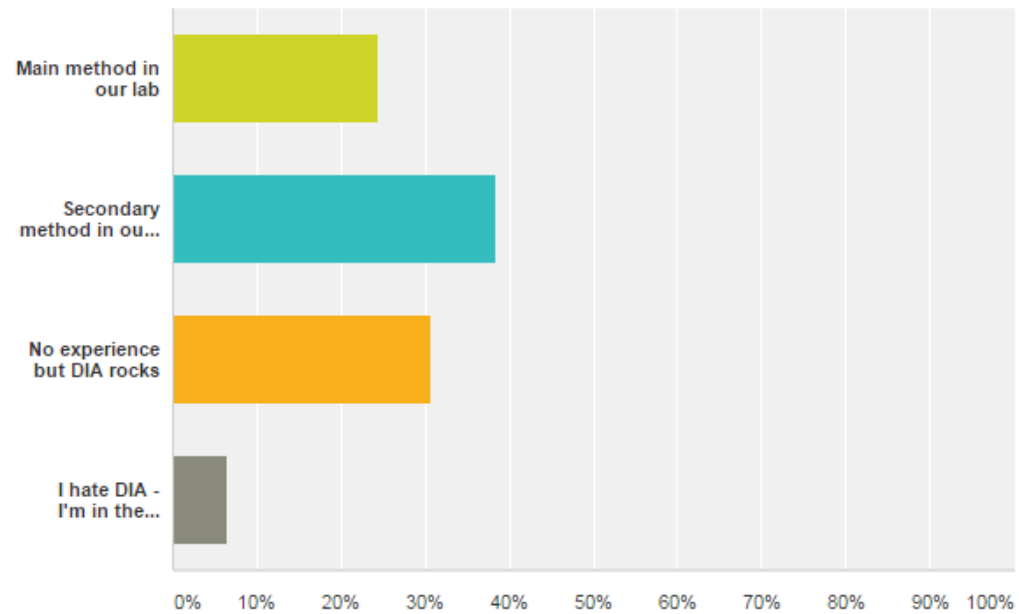
# Survey question 1

## 1. What's your experience with DIA?

- Main method in our lab
- Secondary method in our lab
- No experience but DIA rocks
- I hate DIA - I'm in the wrong room!

## What's your experience with DIA?

Answered: 78 Skipped: 0



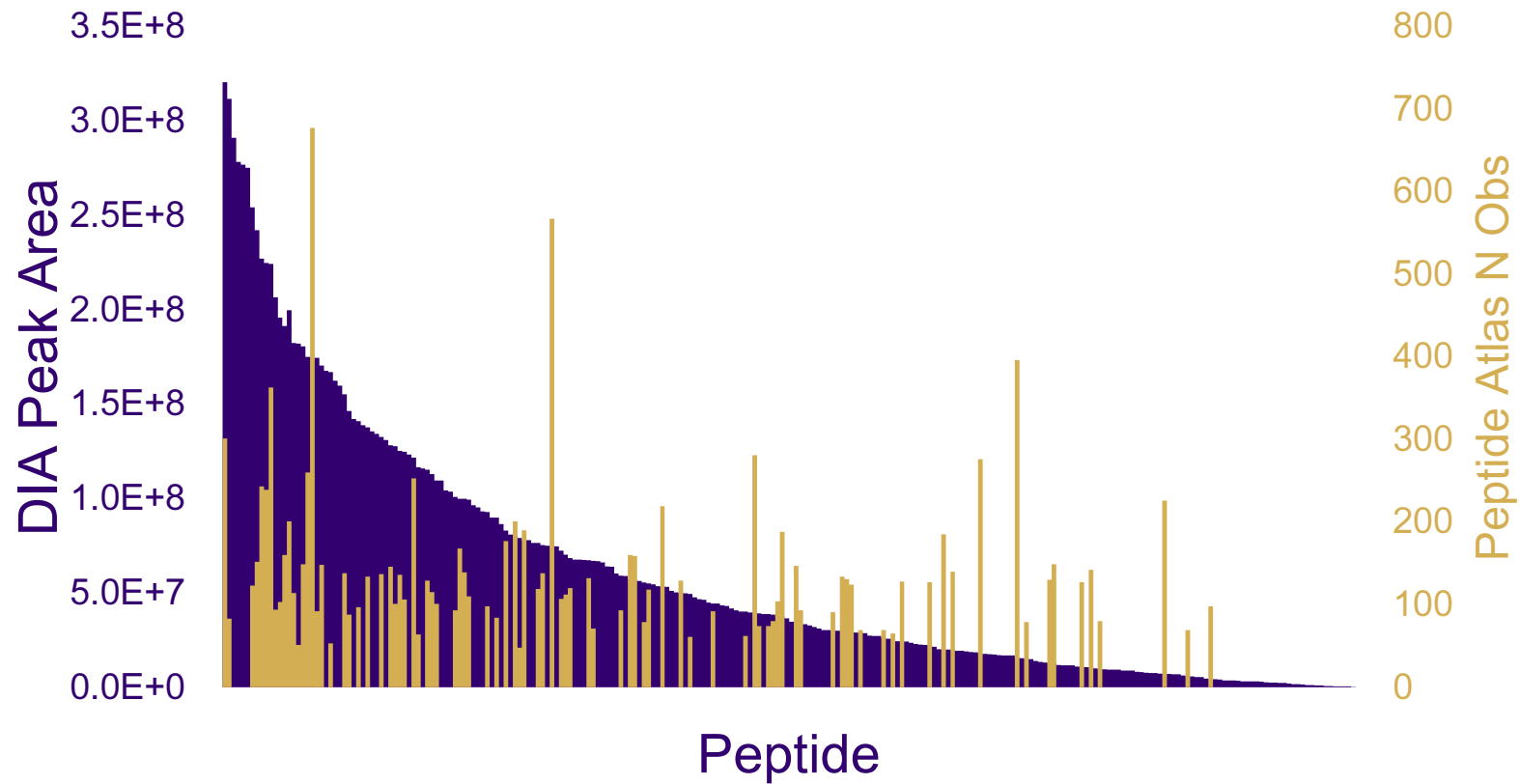
Answer Choices	Responses
▼ Main method in our lab	24.36% 19
▼ Secondary method in our lab	38.46% 30
▼ No experience but DIA rocks	30.77% 24
▼ I hate DIA - I'm in the wrong room!	6.41% 5
Total	78

Mike

# Improving DIA Assay Design using Lessons Learned from Targeted Assay Development

# Apolipoprotein B100

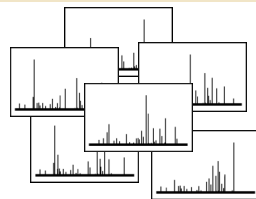
## DIA vs. PeptideAtlas



# Pecan: Detecting Peptides Directly from DIA Data

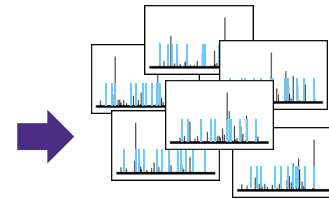
## Spectrum-centric analysis

What peptides best explain the data?



MS/MS spectra

example tools:  
SEQUEST, MASCOT ... etc.



Peptide spectrum matches  
(PSMs)

p-value  
q-value

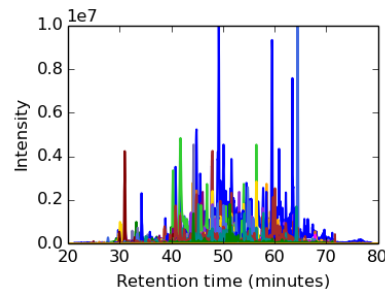
## Peptide-centric analysis

Which peptides are detected in our data?

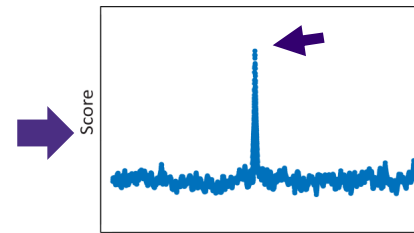
Peptides of interest



example tools:  
OpenSWATH, Skyline



Extracted MS/MS data

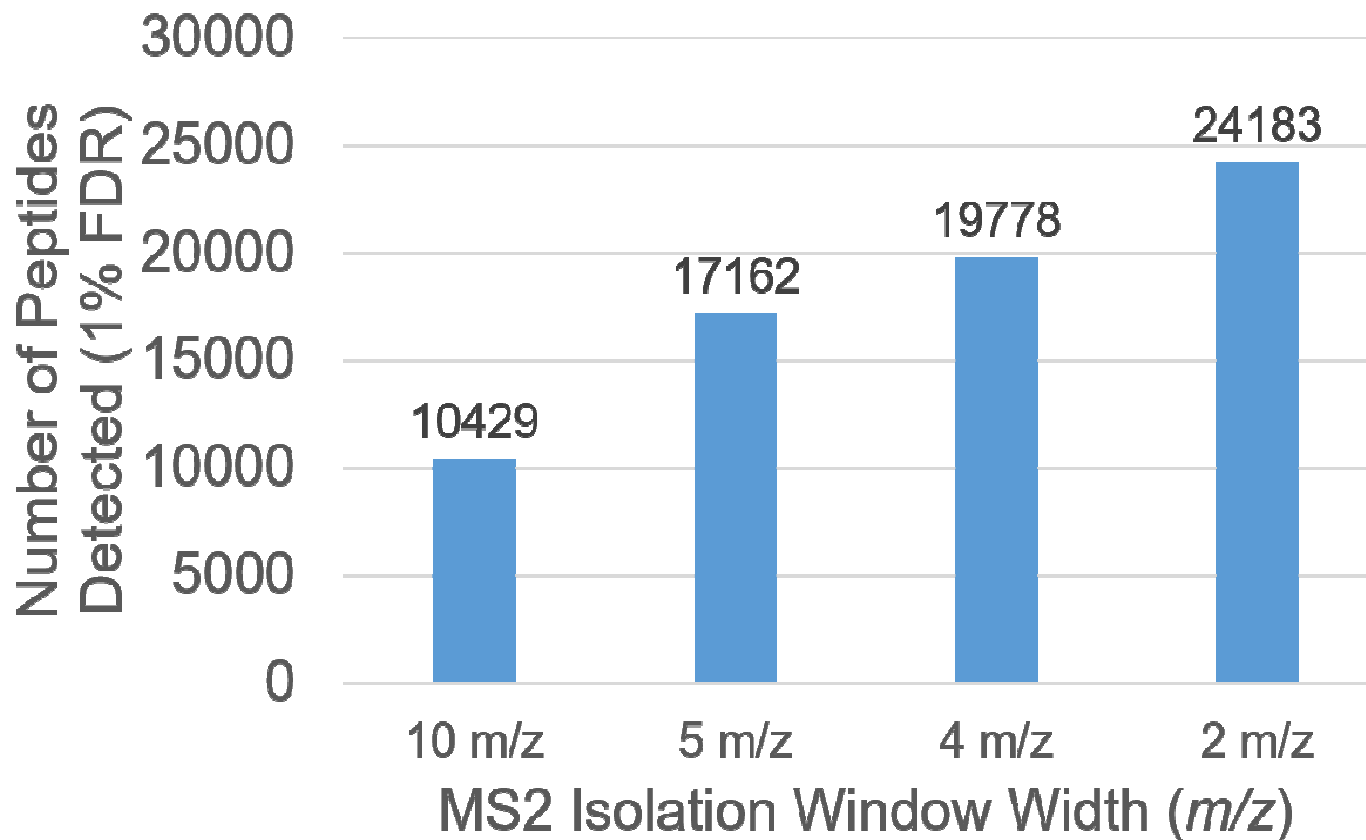


Evidence for detection

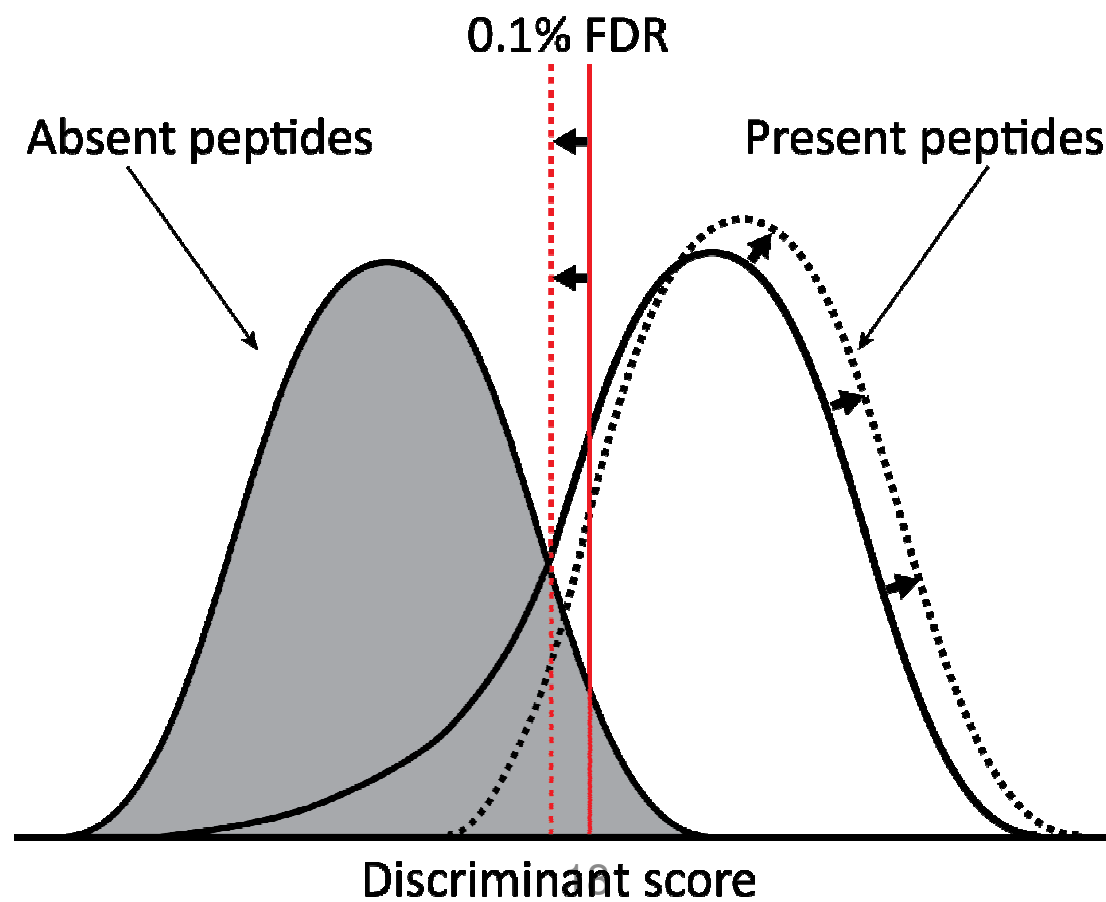
p-value  
q-value



## Problem #1: Detection of Peptides in DIA Data is Inversely Proportional to Isolation Width

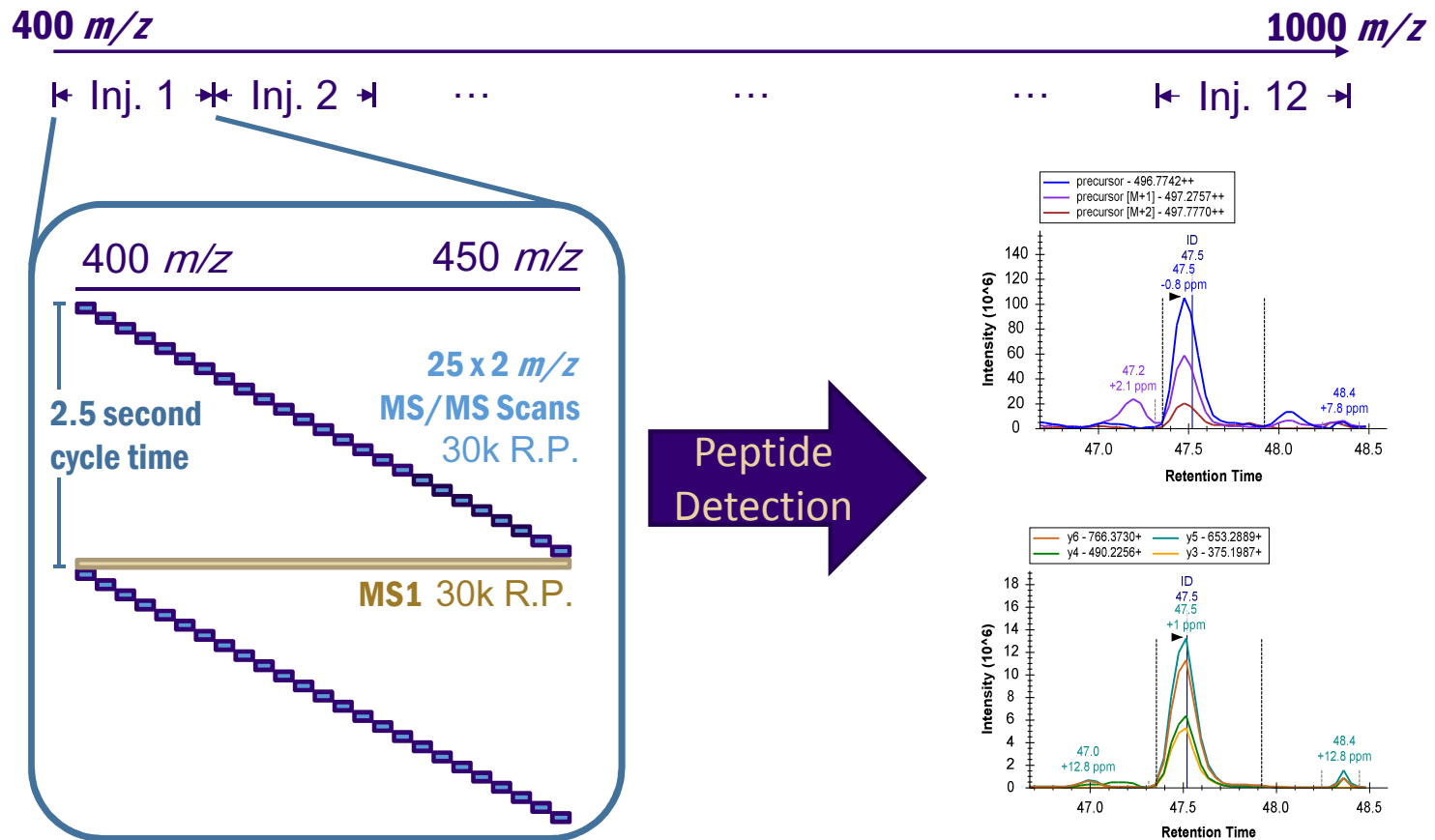


# Narrow Isolation Windows Improves the Score Discrimination Between Absent and Present Peptides



# Picking Peptides Directly from DIA Data

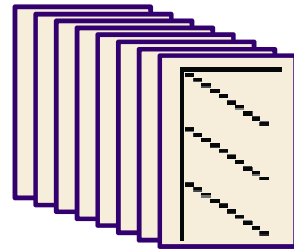
12 LC-MS/MS Runs: ~1  $\mu$ L plasma



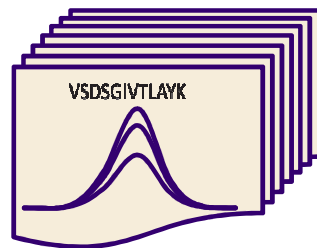
# Chromatogram ~~Spectral Library~~ Workflow

Build a Chromatogram Library

12 DIA Runs  
**2  $m/z$  Isolation**  
Pooled Sample

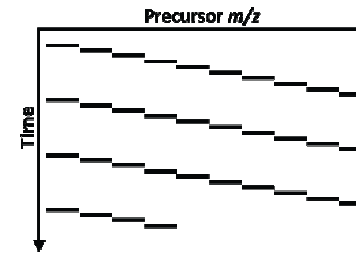


Pecan Peptide  
Detection

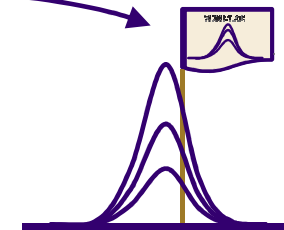


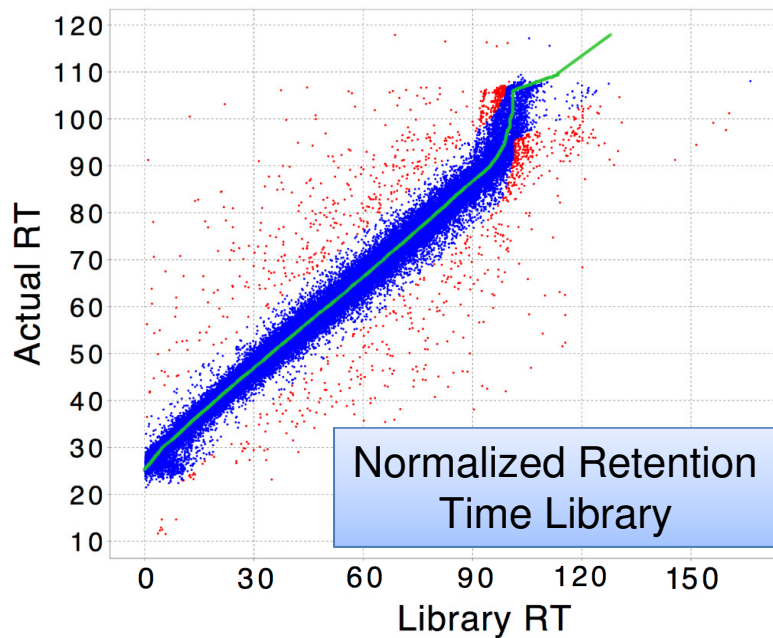
Acquire DIA Data

1 DIA Run  
**20  $m/z$  Isolation**  
/per sample

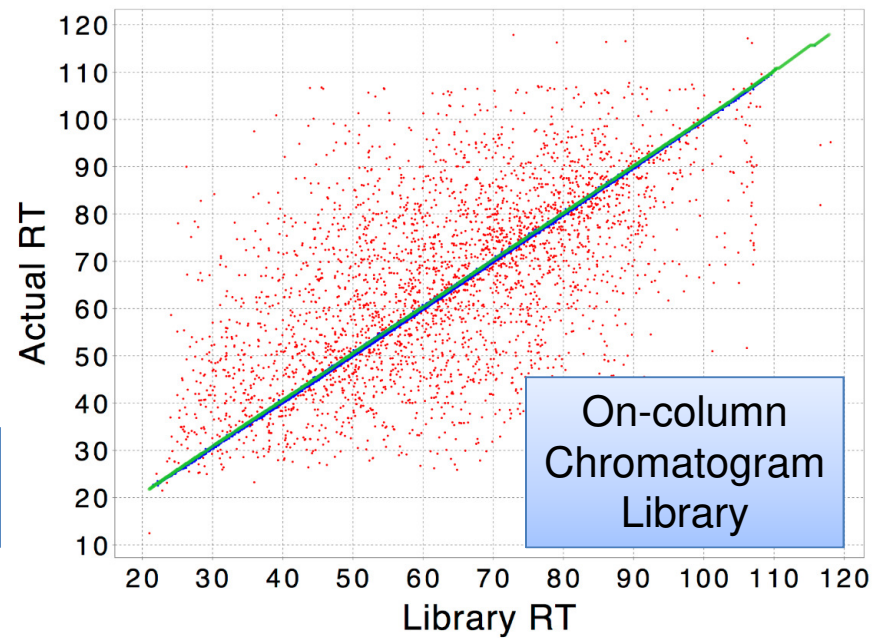


Chromatogram Extraction  
Quantification

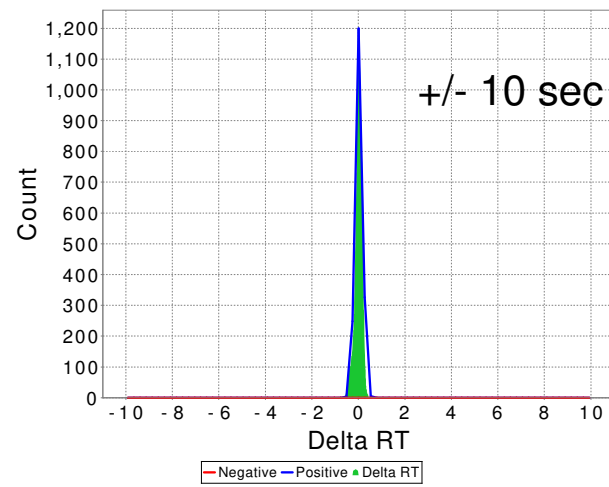
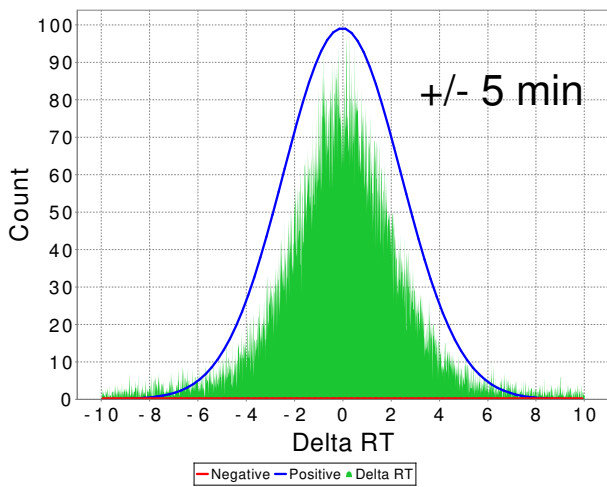




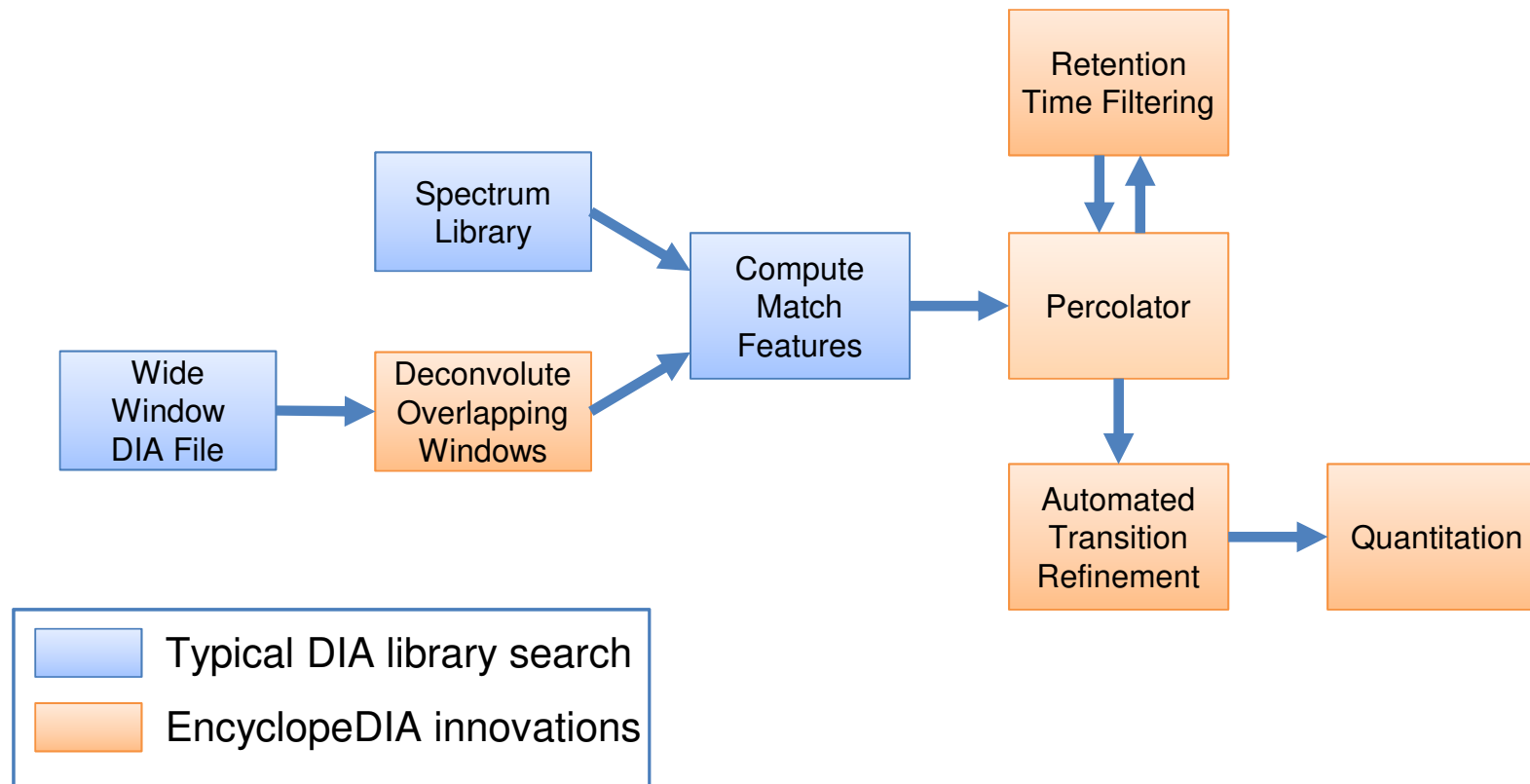
Retention Time Fit · Data Used In Fit · Data Removed From Fit



Retention Time Fit · Data Used In Fit · Data Removed From Fit

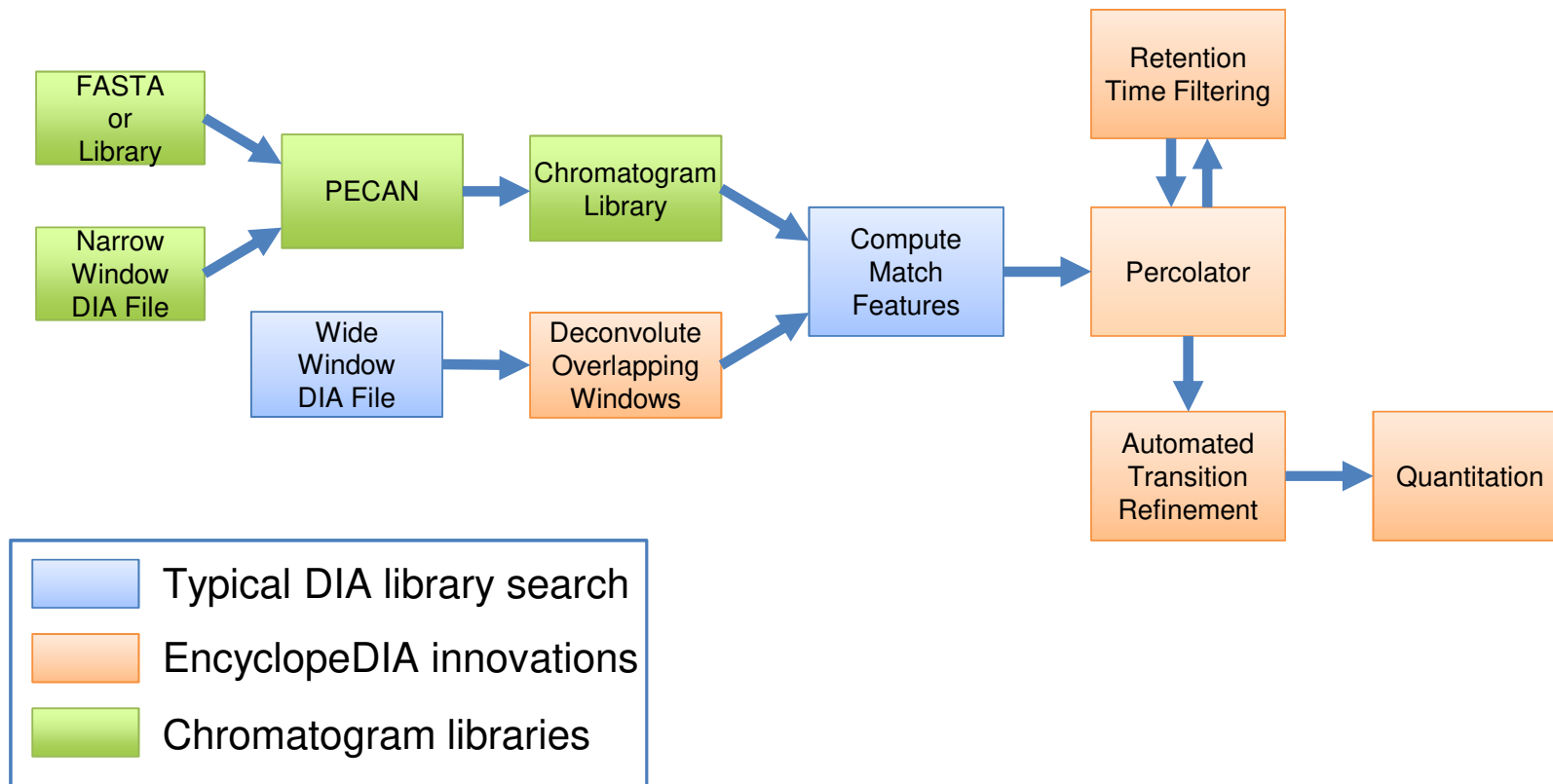


# EncyclopeDIA workflow

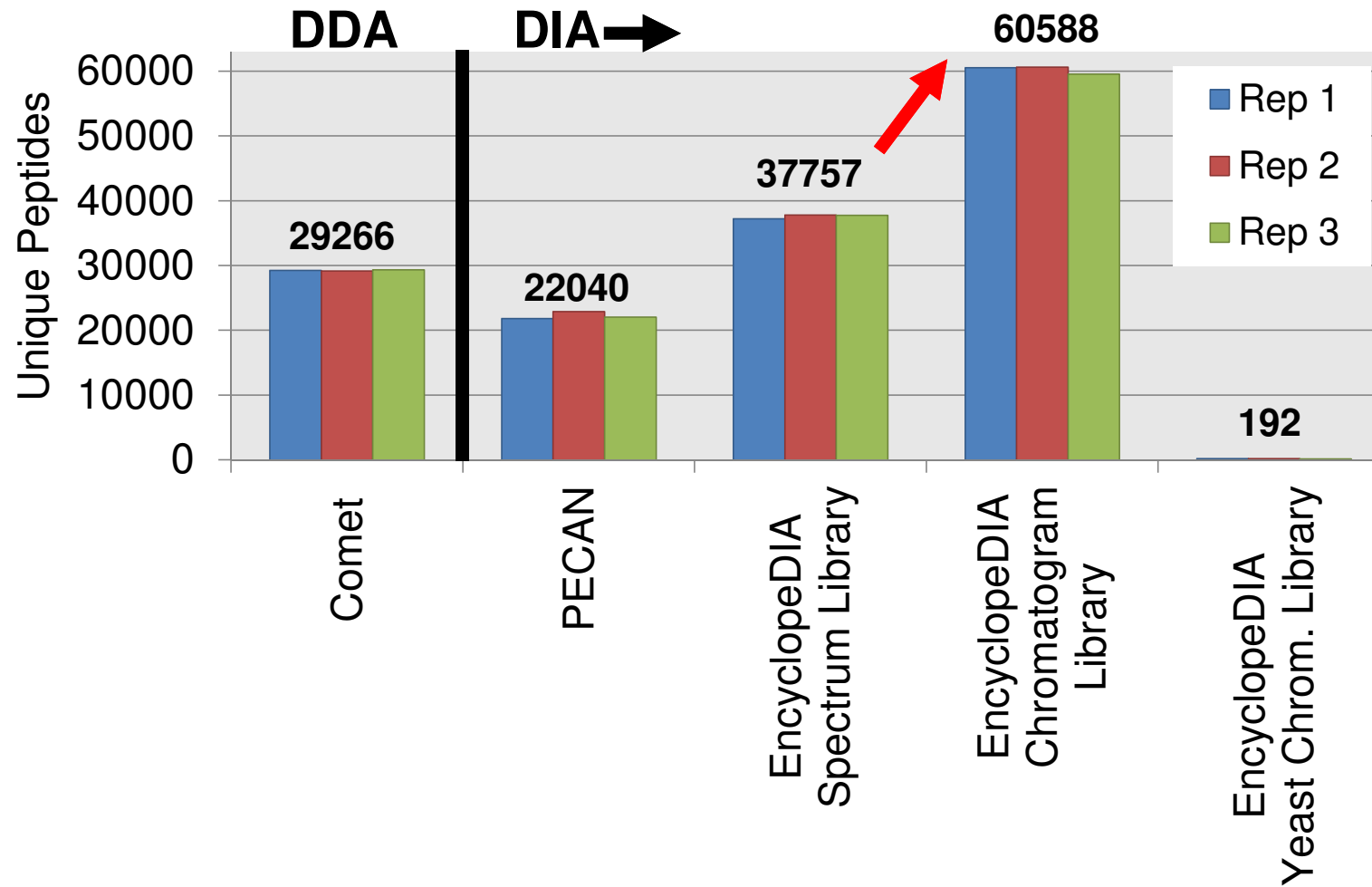


Percolator from Käll L et al, Nat Methods. 2007 Nov;4(11):923-5.

# EncyclopeDIA workflow



# Chromatogram libraries are significantly more powerful than spectrum libraries

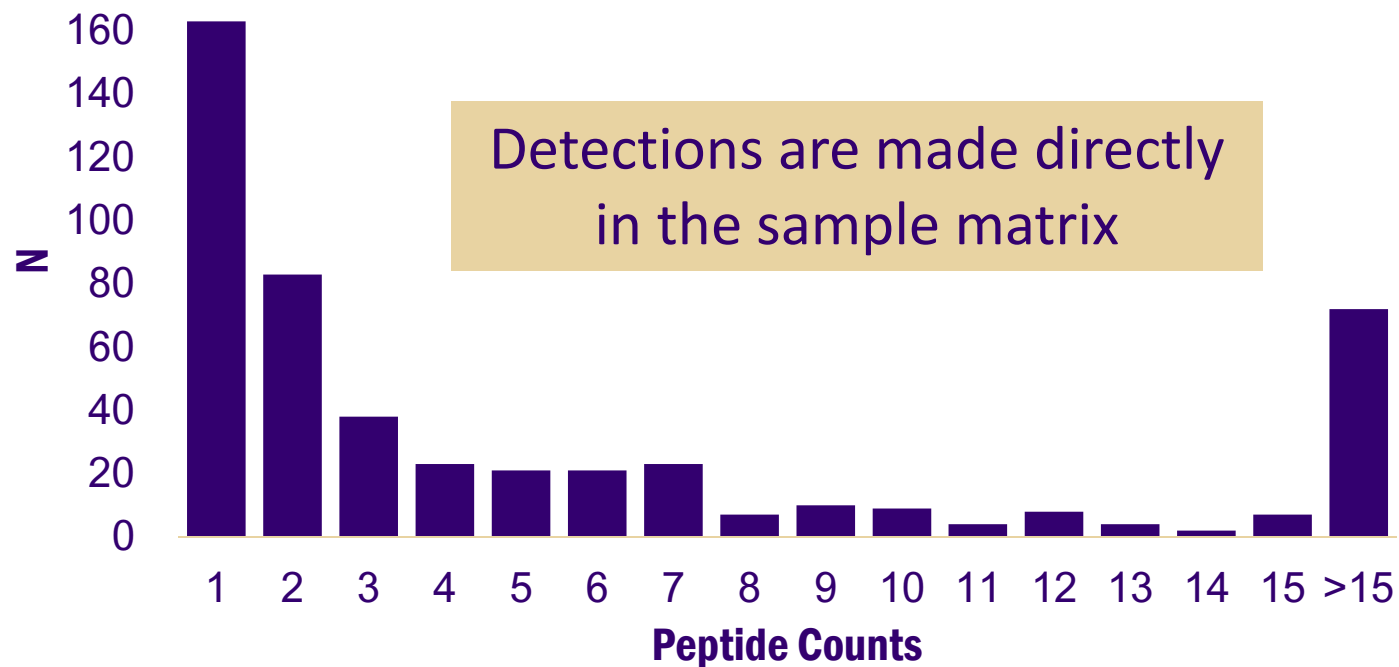




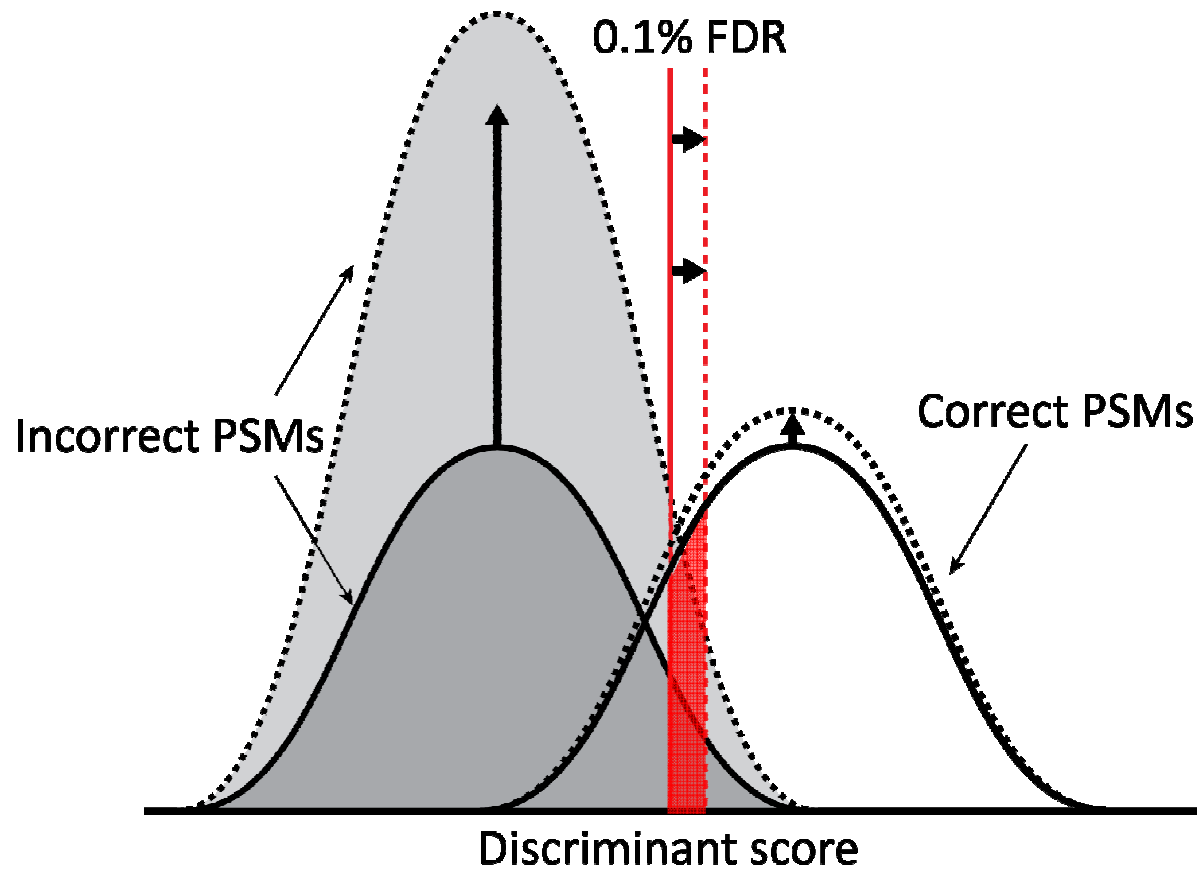
# Human Plasma Chromatogram Library

4,244 unique peptides | 495 proteins

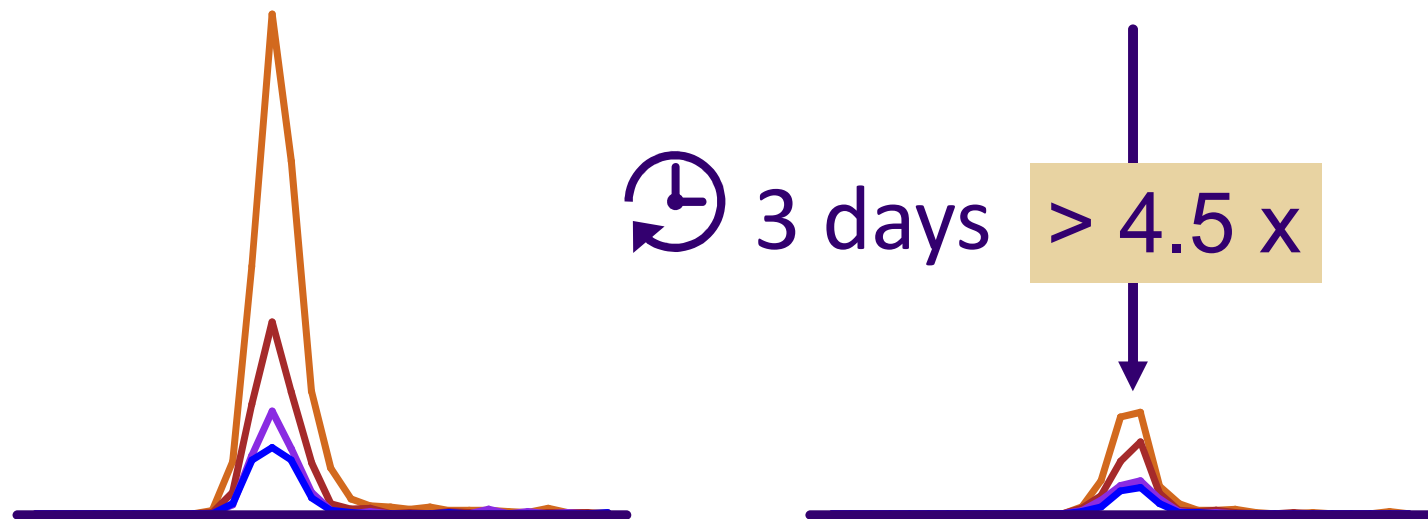
3,880 peptides mapping to one protein



## Lesson 2: Don't Use a Large Library Just Because You Can



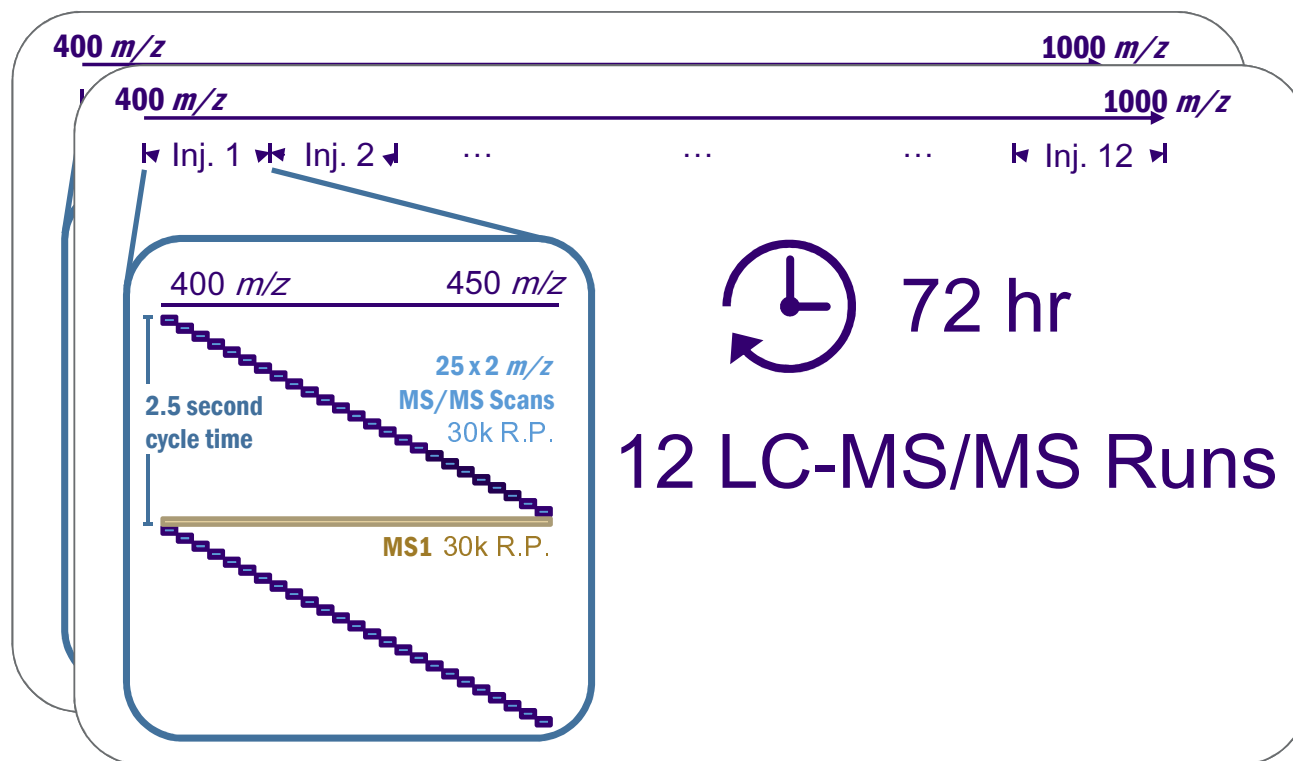
## Lesson 3: Not all Peptides are Stable



**Apolipoprotein B100 | TTLTAFGFASADLIEIGLEGK @ 4°C**

**Jim Bollinger | ASMS 2014**

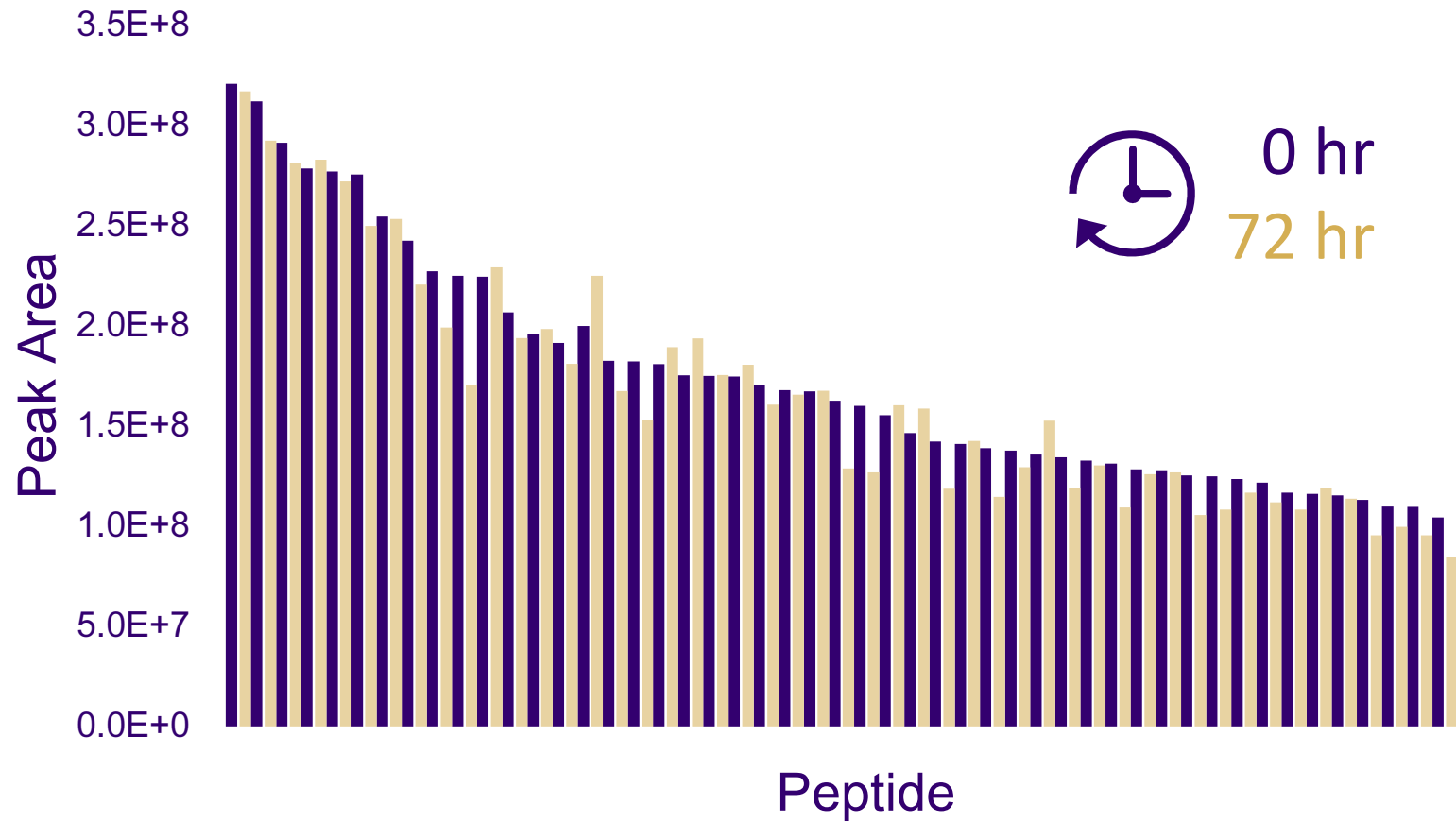
# Assessing Peptide Stability with DIA



# Peptide Stability

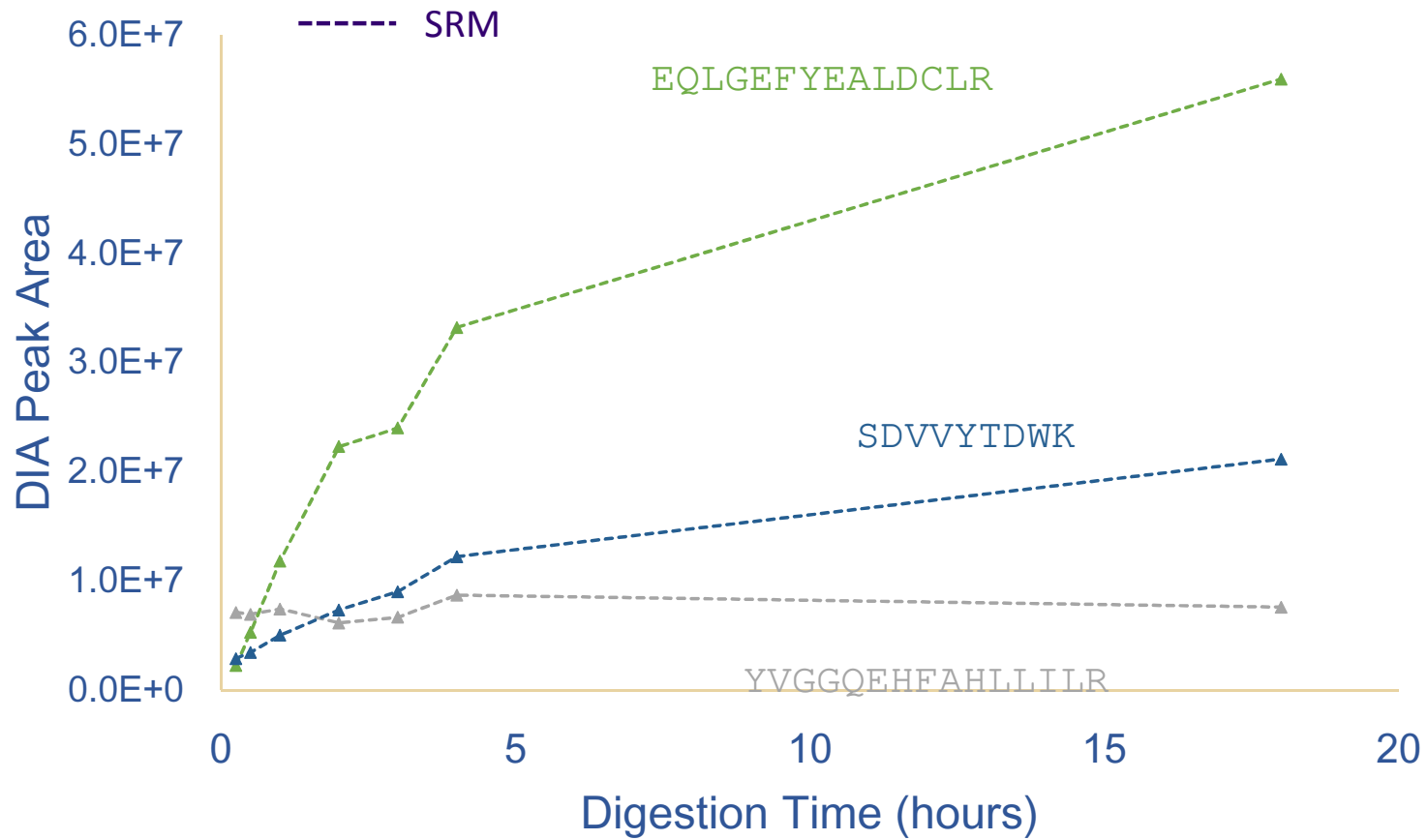
## Apolipoprotein B100

\*top 50 peptides

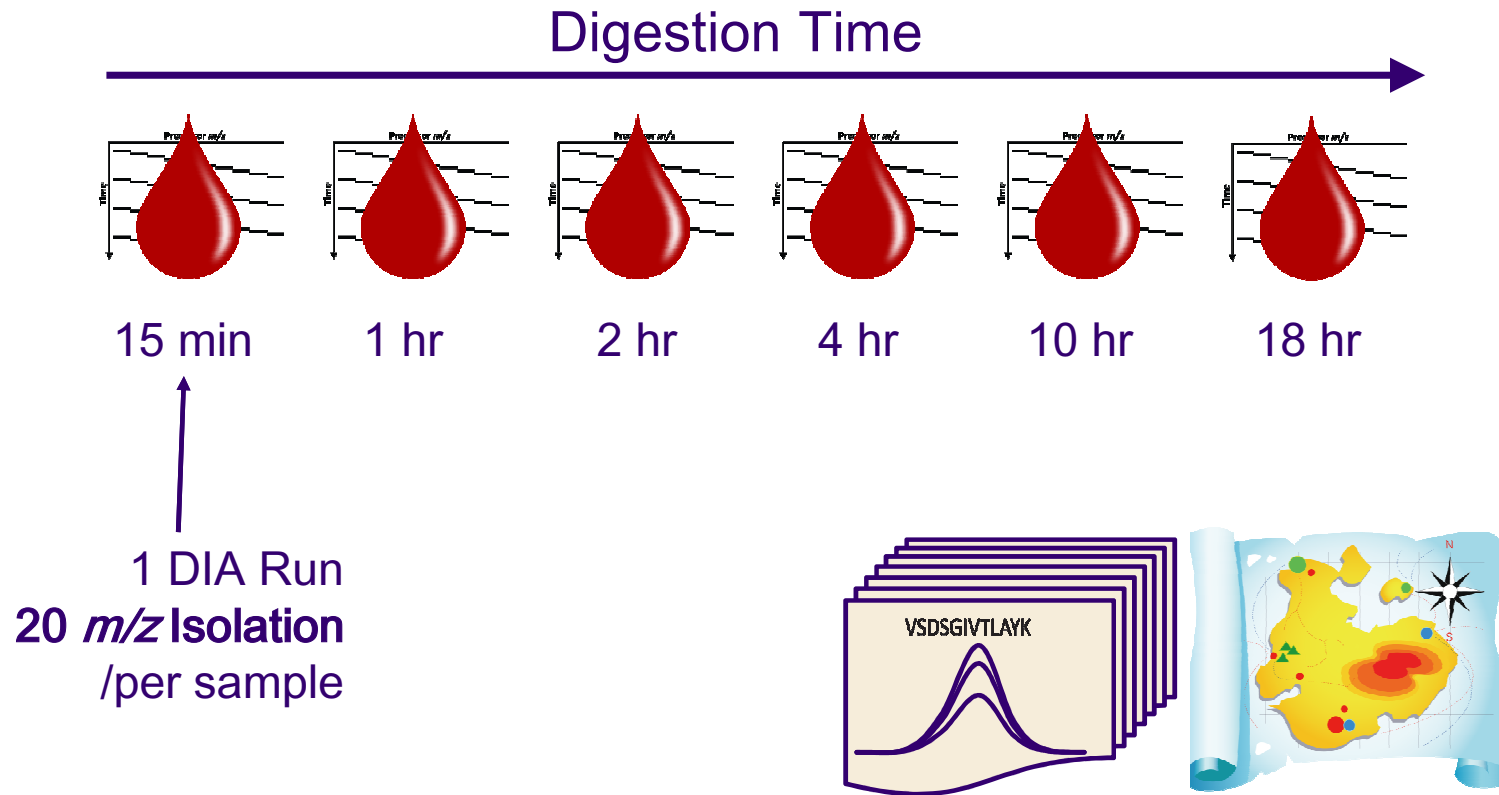


# Lesson 4: Know your Digestion

## Alpha-1-acid glycoprotein

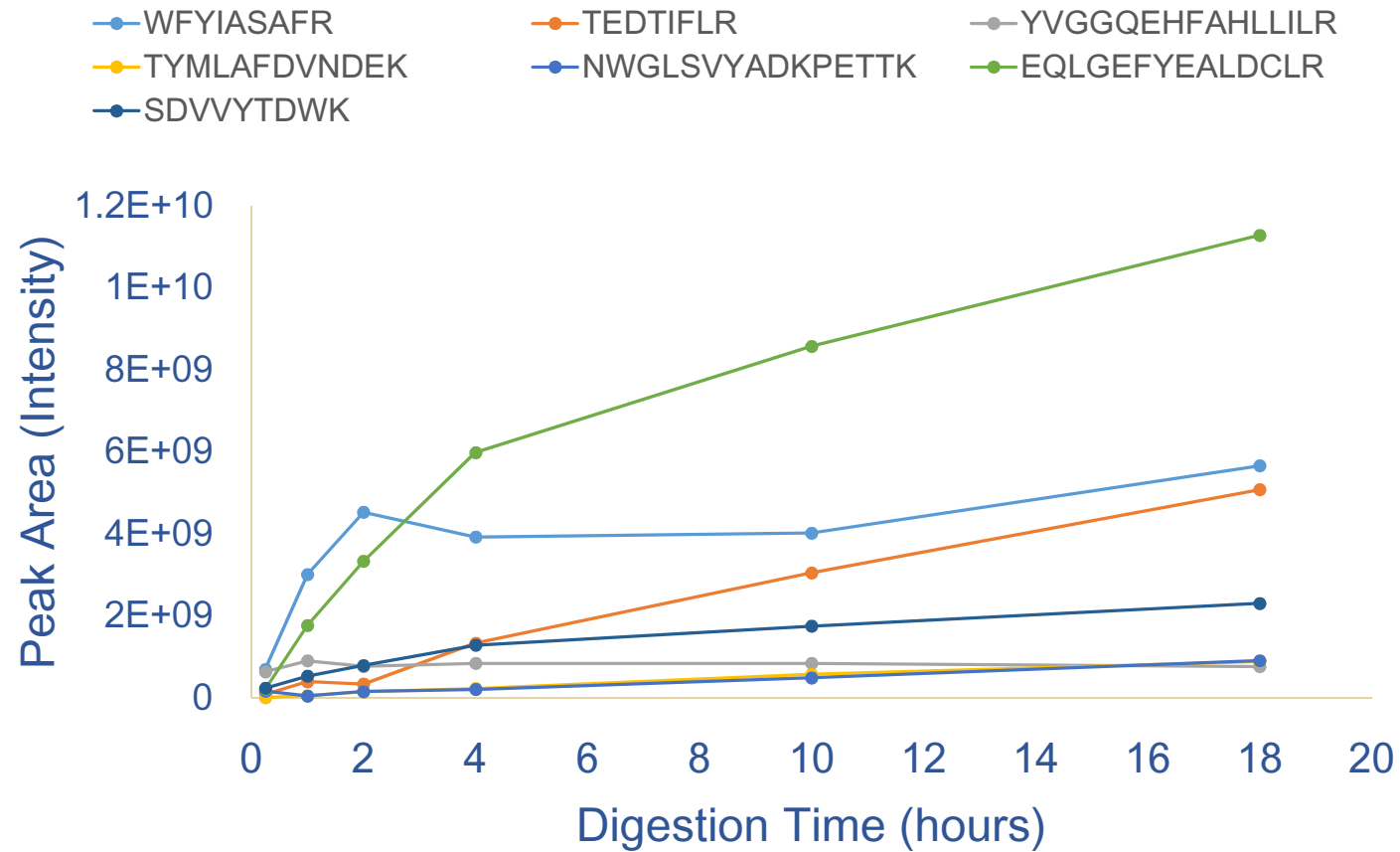


# Digestion Time Course by DIA



# Digestion Time Course (DIA)

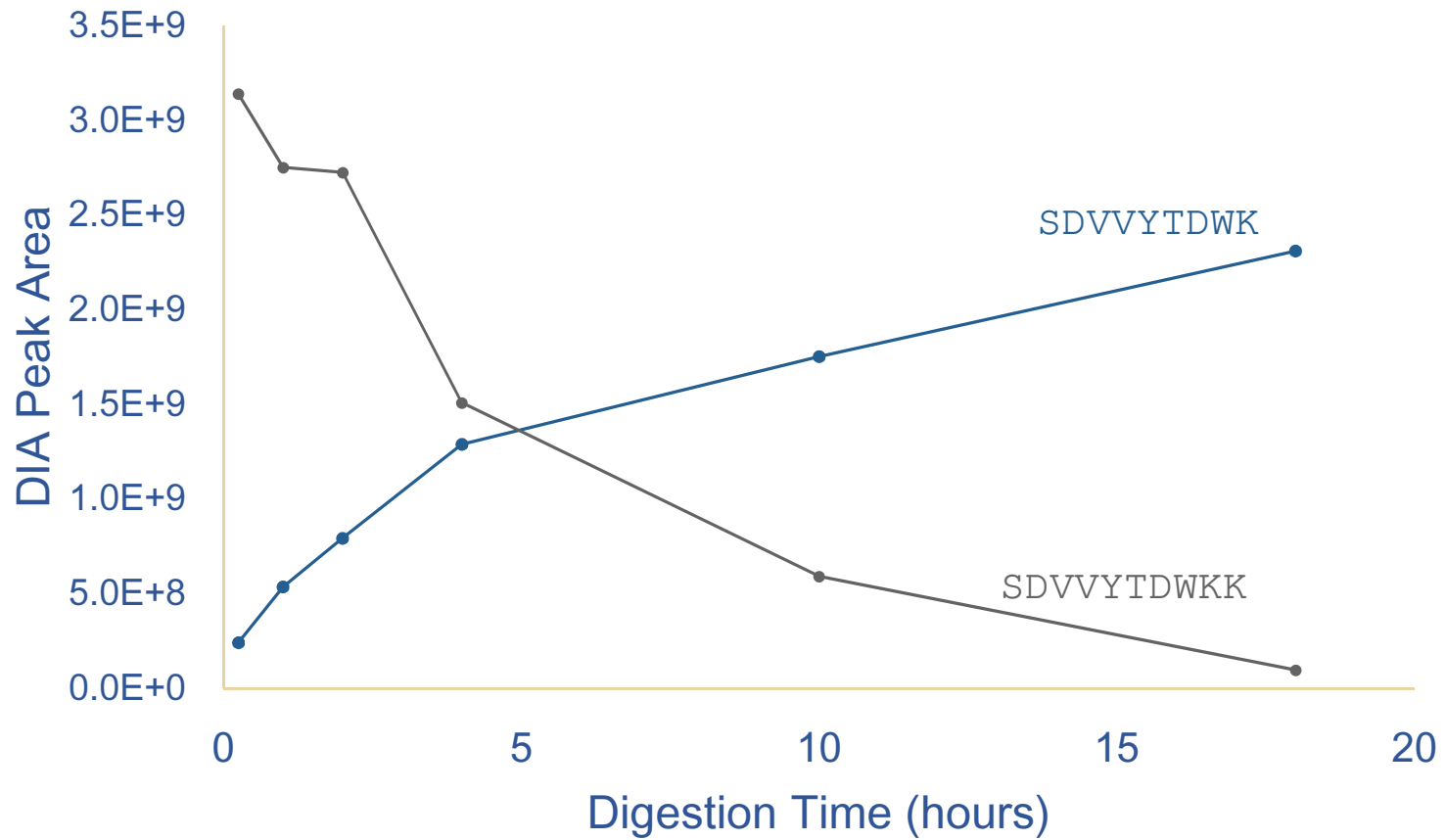
## Alpha-1-acid glycoprotein



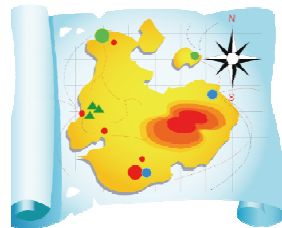
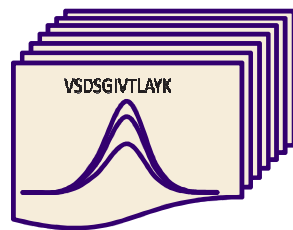
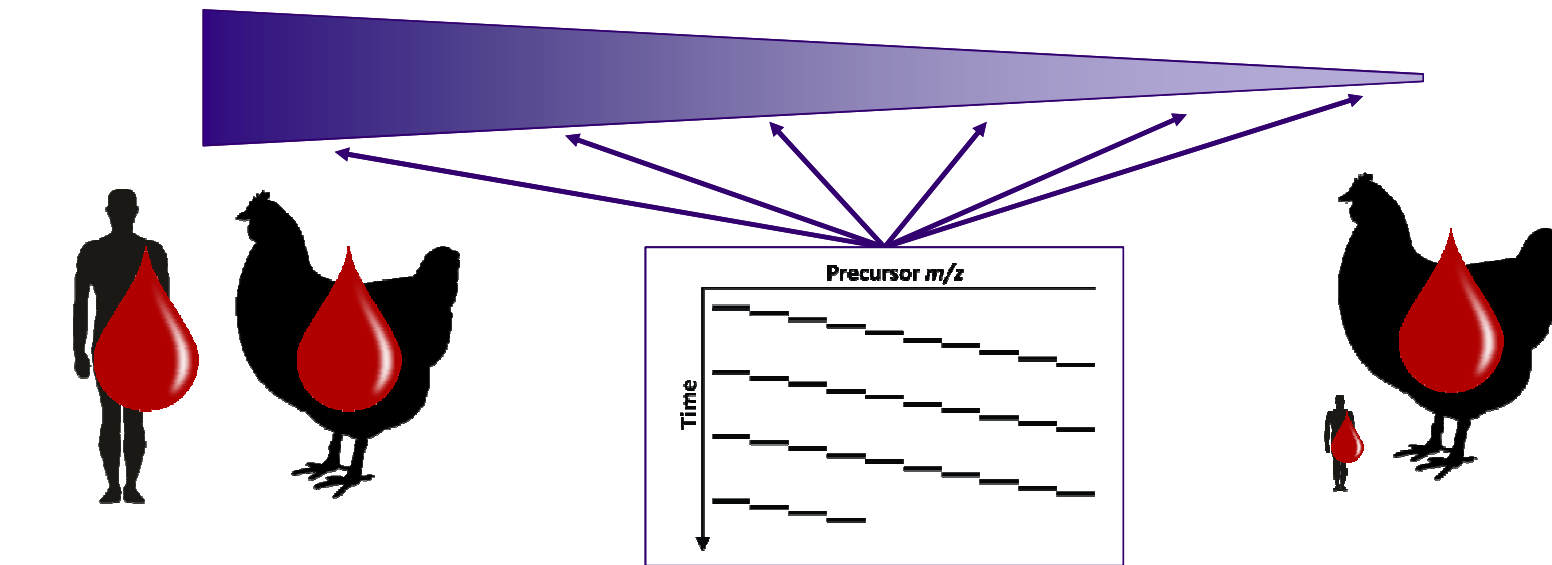


# Investigating Digestion Kinetics

## Alpha-1-acid glycoprotein



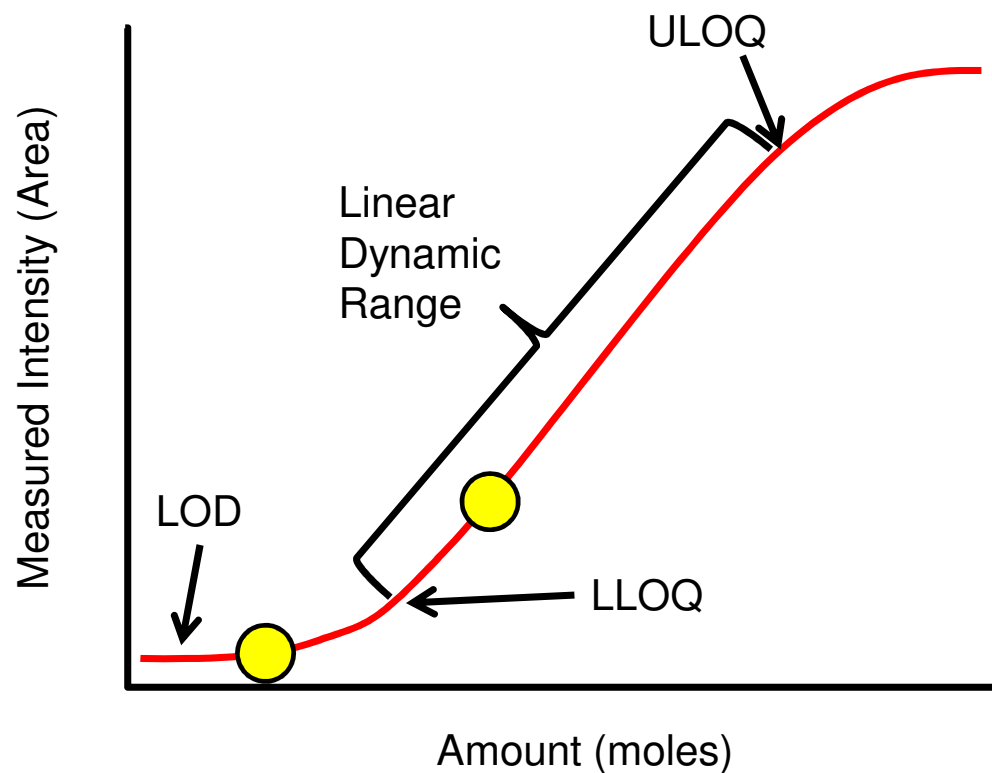
# Lesson 5: Evaluate Linearity and LoQ



6 LC-MS/MS Runs  
(not including replicates)

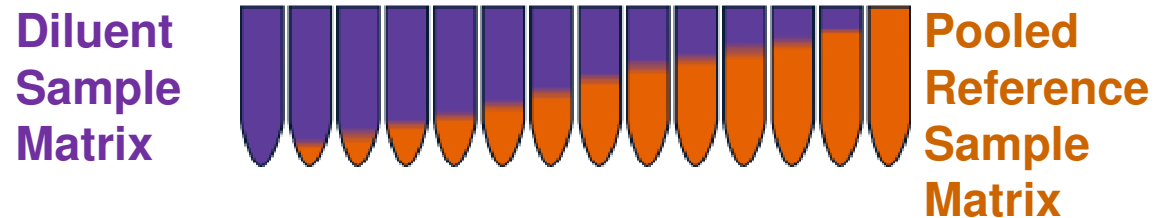
Grant RP, Hoofnagle AN, Clinical  
Chemistry 2014

# Are the measurements quantitative or just differential?



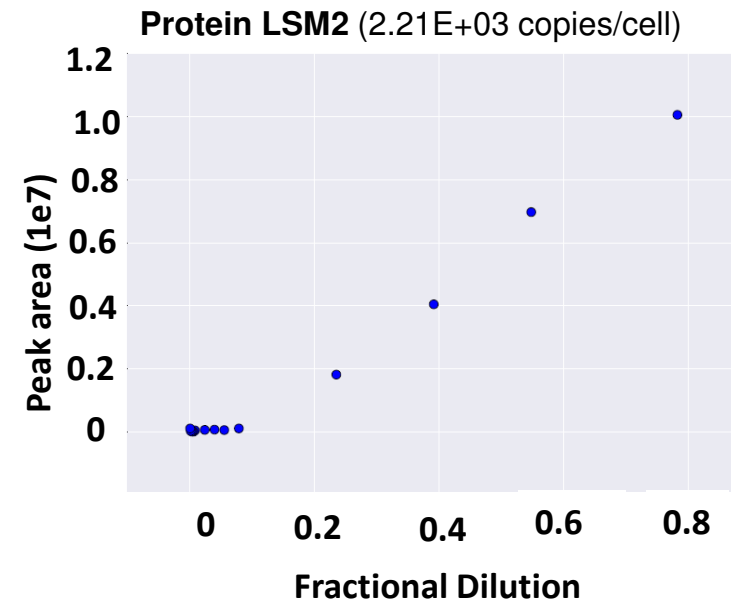
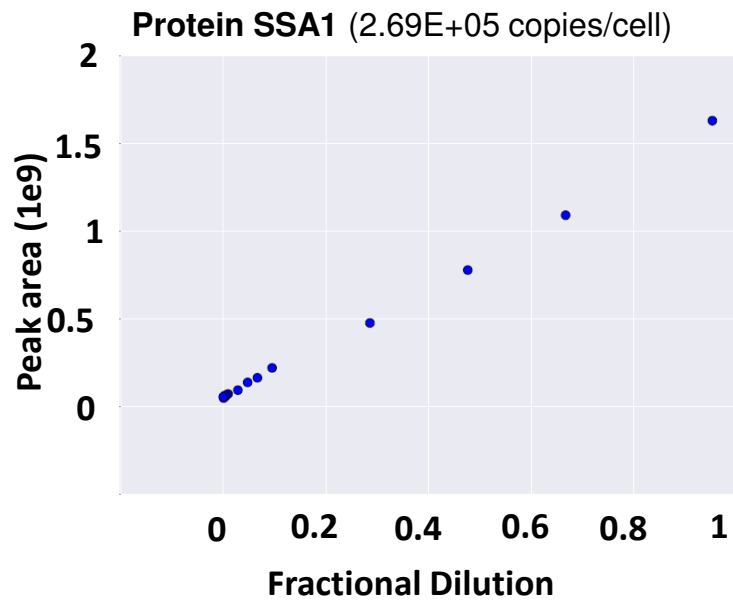
LOD: Limit of Detection  
LLOQ: Lower Limit of Quantitation  
ULOQ: Upper Limit of Quantitation

# Method to measure both LOQ and Linearity

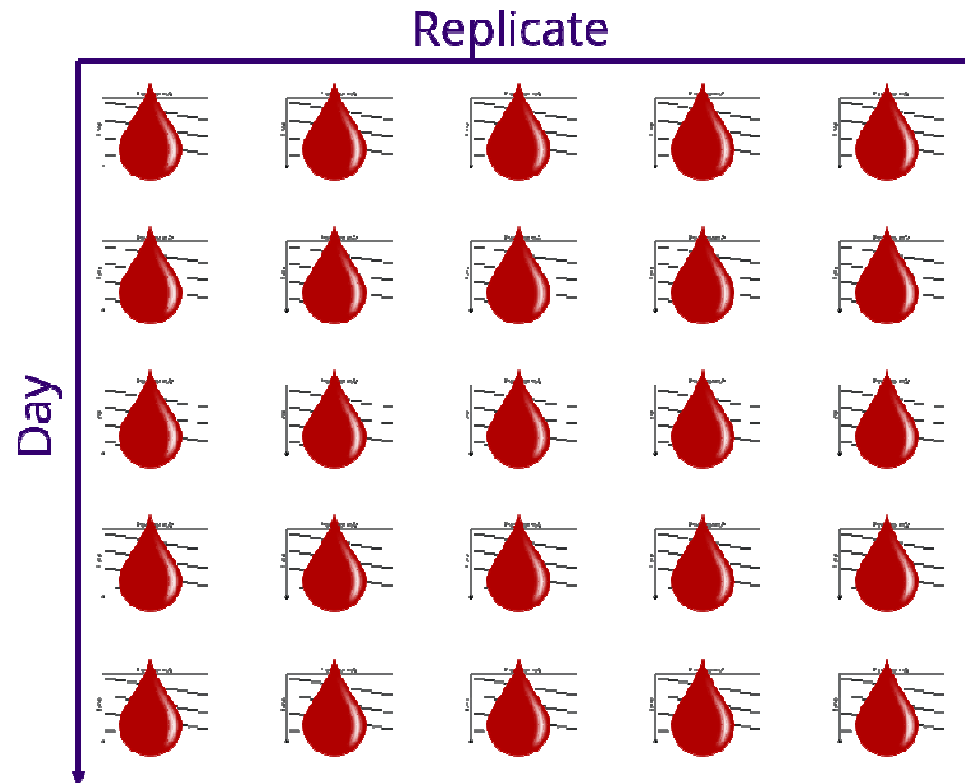


- Possible Samples to Use as a Diluent Matrix
  - Stable isotope labeled version of the matrix.
    - $^{15}\text{N}$  or SILAC labeled cells
  - A diverged species
    - For human plasma we use chicken plasma.

# Reference Yeast BY4742 Diluted in $^{15}\text{N}$ Yeast (S288c)



# Assess Reproducibility (5x5)



# DIA Assay Workflow

---

Peptide Detection and Stability  
24 x 2 hr LC-MS/MS

Digestion Time Course  
8 x 2 hr LC-MS/MS

Linearity / LoQ Assessment  
10 x 2 hr LC-MS/MS

Reproducibility  
25 x 2 hr LC-MS/MS

**67 LC-MS/MS Runs**

Alexey





# DIA

---

Is it a shotgun proteomics method?

*Yes, DIA as not “less shotgun” than DDA*

Is it a discovery proteomics method?

*Yes, DIA is a untargeted data acquisition method. It is even “less targeted” than DDA*

# Lessons from History

---

High throughput proteomics methods that rely on the previously generated (proteomics) data have not been very successful

- Reference databases of 1D and 2D SDS Page gels (SWISS-2DPAGE database)
- AMT (accurate mass and time) approach
- Spectral library searching as replacement for database search

Flicker - V0.67-beta \$Date: 2004/02/23 14:05:49 \$  
File Edit View Landmark Transform Quantify Help

Flicker (OFF) Flicker is off  
 Click to access DB  
 Allow transforms  
 Sequential transforms  
Flashing is OFF  
Done

ZoomMag: 1.00x

swiss2D page : P02647

1) Match spot  
2) Make map active  
3) Click on spot  
4) Putative ID pops up

PLASMA\_HUMAN\_id.gif (230, 302) 20677 tot.gif

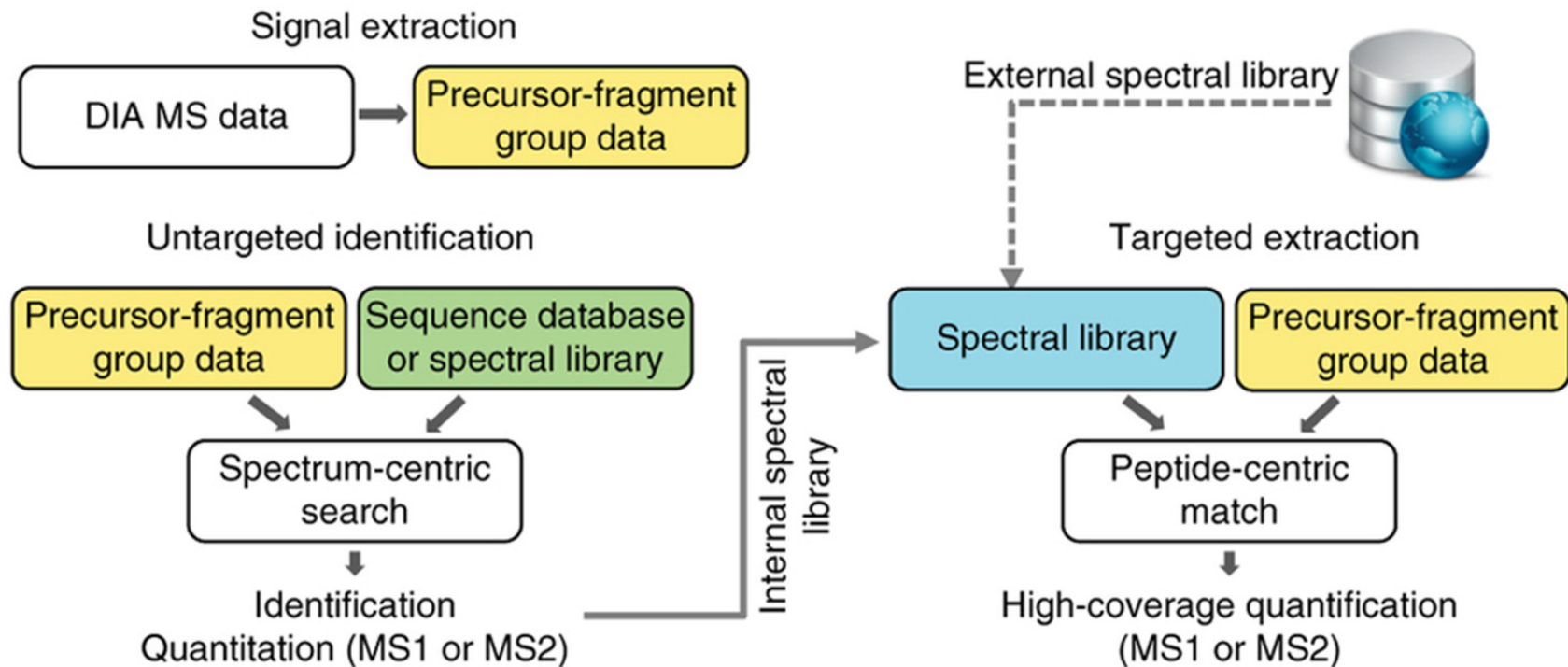
Delay 0.30Sec

Delay 0.30Sec

Information about the entry  
Name and origin of the protein  
Description  
Gene name(s)  
Tissue  
Taxonomy  
References

(1) MAJIDIG OF GEL  
MDCDB=9114204; PubM=128649 DDCS; EUPALS; ESI-level found  
Nicolson D F., Prager E., Fogart H., Barck A., Keller F., Pasqual C., Sanchez J.-C.,  
Tost J.-D., Rydquist B., Varga E., Aguiar D., Hughes G.J.

# Hybrid (Direct+ Targeted) Strategy

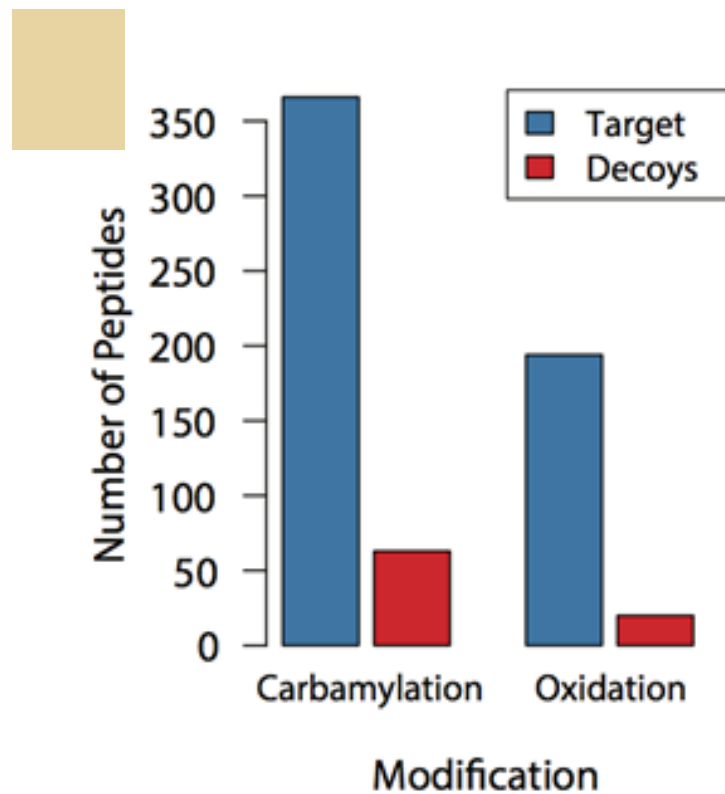


C.C. Tsou *et al.* DIA-Umpire: comprehensive computational framework for data independent acquisition proteomics ***Nature Methods***, 2015



# Violation of the Target-Decoy Assumption in Closed Searches

---



- ▶ Selected spectra corresponding to common modifications identified in open search and examined their identifications in closed search
- ▶ Under target-decoy assumptions, these spectra should all be incorrect and match equally to target and decoy sequences
- ▶ Target-decoy assumption is violated: 6X difference for carbamylation, 9X for oxidation

## Clothes And Body Parts

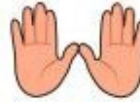
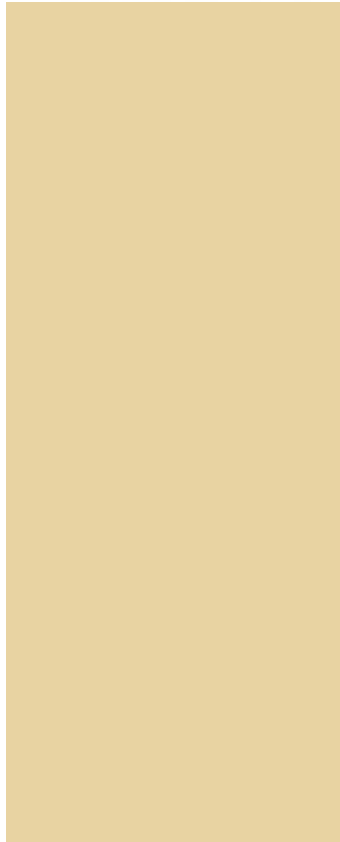
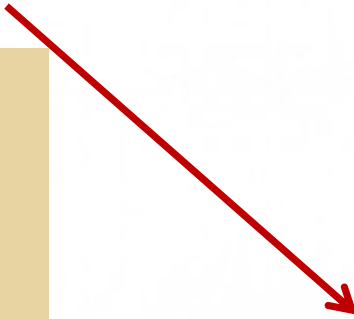
Match the images on the left to their corresponding images on the right.





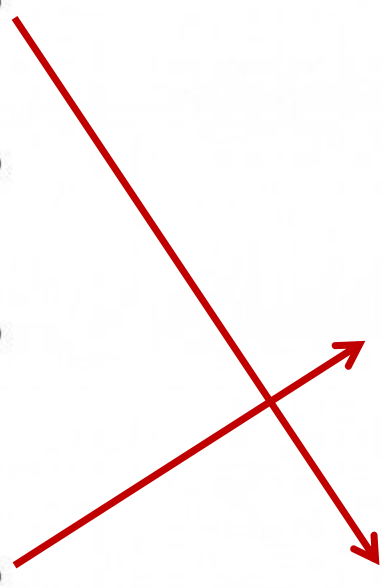
## Clothes And Body Parts

Match the images on the left to their corresponding images on the right.



## Clothes And Body Parts

Match the images on the left to their corresponding images on the right.



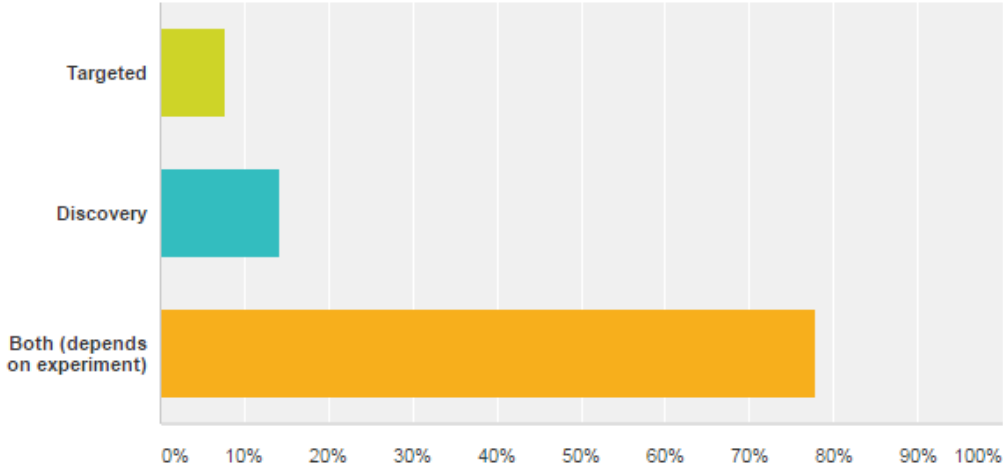
# Survey question 2

2. In which mode should DIA data be analyzed?

- Targeted
- Discovery
- Both (depends on experiment)

# In which mode should DIA data be analyzed?

Answered: 77 Skipped: 1



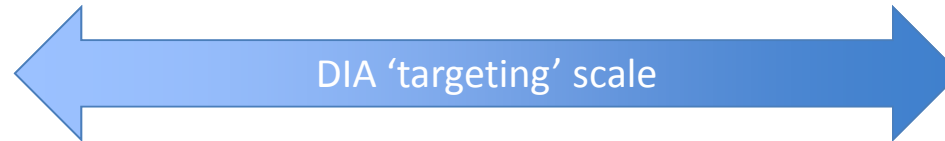
Answer Choices	Responses
▼ Targeted	7.79% 6
▼ Discovery	14.29% 11
▼ Both (depends on experiment)	77.92% 60
Total	77

Isabell

# Scaling up DIA – Error rate control



focused



global

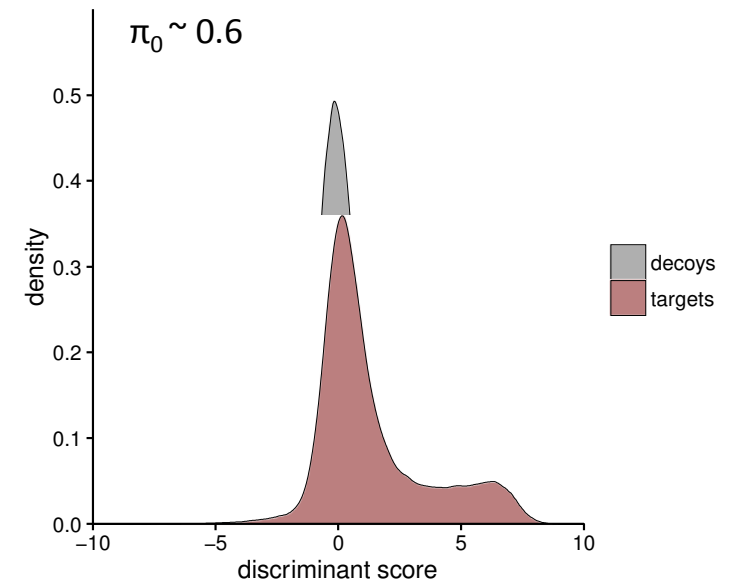
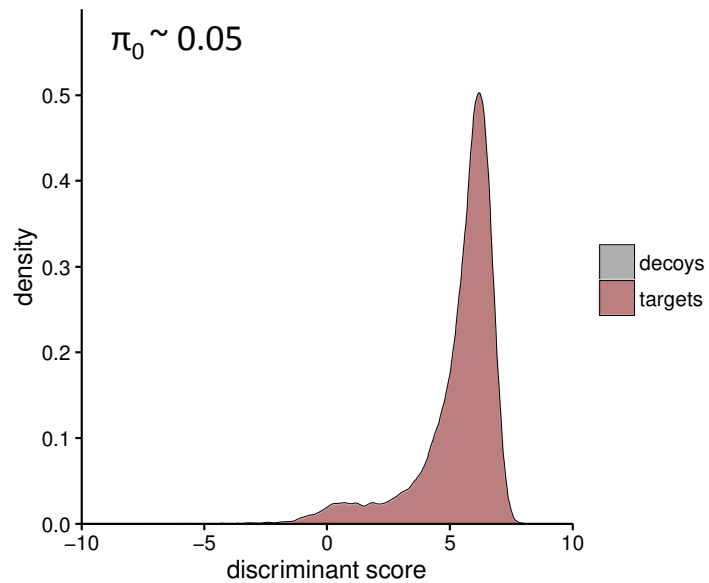


## Sample-specific spectral library

- Built from same samples as DIA data
- Majority of peptides/proteins detectable

## Comprehensive deep spectral library

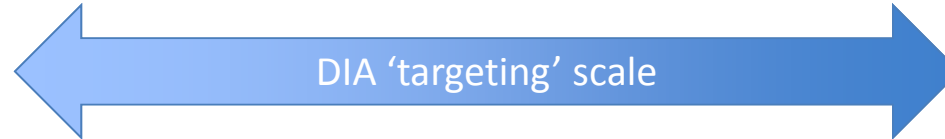
- Built from multiple different samples
- Majority of peptides/proteins NOT detectable



# Scaling up DIA – Error rate control



focused



global



## Sample-specific spectral library

- Built from same samples as DIA data
- Majority of peptides/proteins detectable

Protein-level error can be handled on library generation level



## Comprehensive deep spectral library

- Built from multiple different samples
- Majority of peptides/proteins NOT detectable

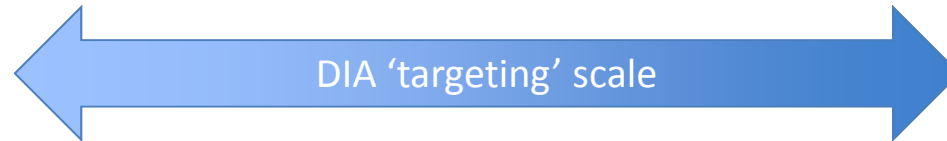
Protein-level error accumulates if not carefully controlled



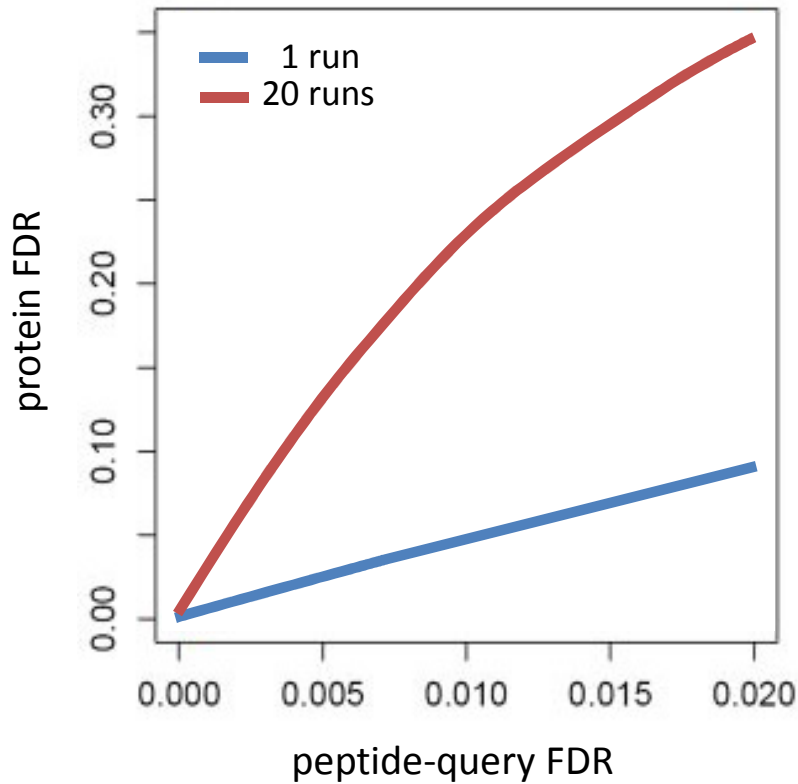
# Scaling up DIA – Error rate control



focused



global



## Comprehensive deep spectral library

- Built from multiple different samples
- Majority of peptides/proteins NOT detectable

Protein-level error accumulates if not carefully controlled





# Scaling up DIA – Error rate control

- 1. Control error rate on protein level:**

Take best peptide peak group per protein for FDR / q-value estimation on protein level

- 2. Control error rate globally across all samples within a study: Protein master list**

Take best peptide peak group per protein across all samples in a study to generate a protein master list at 1% FDR

Extended version of PyProphet: <https://github.com/PyProphet>

Rosenberger & Bludau *et al.* (submitted)

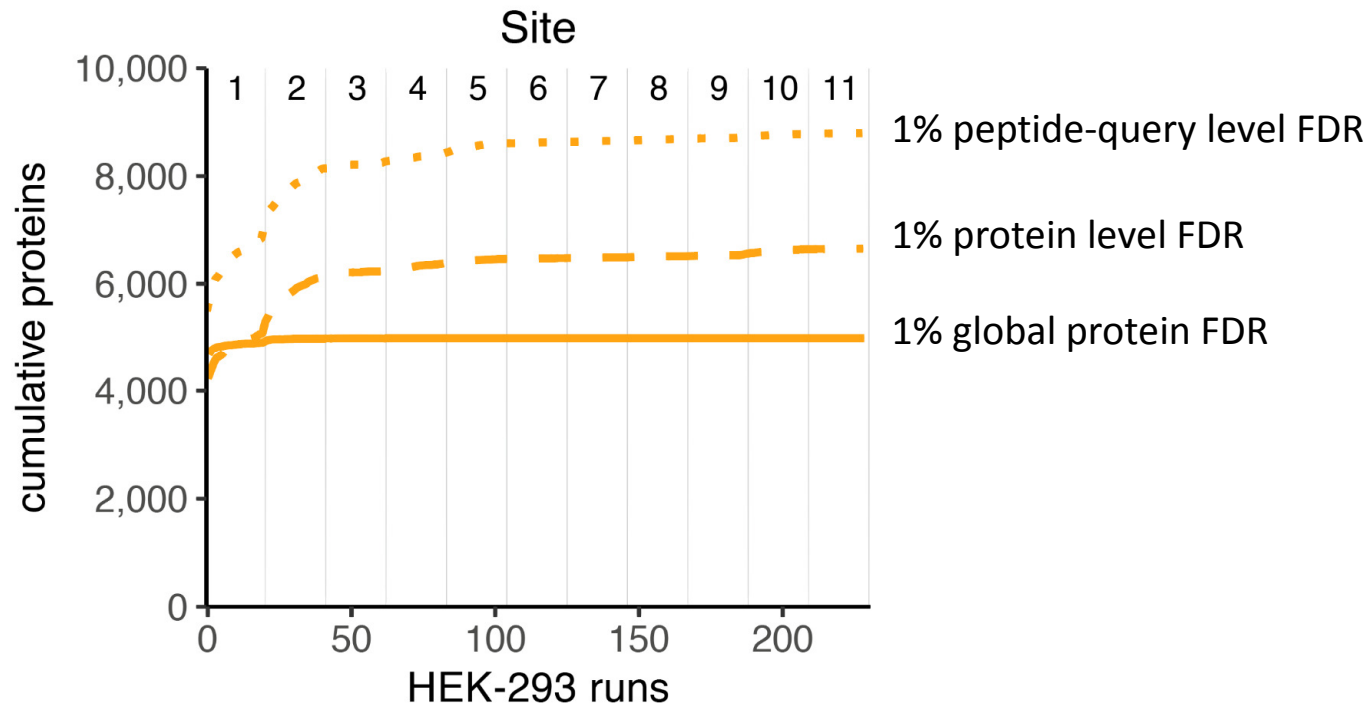
# Scaling up DIA – Error rate control

## DIA data:

- Inter-laboratory study:  
229 DIA measurements of same HEK-293 cell lysate  
Collins et al. (2017)

## Spectral library:

- Combined assay library (CAL):  
331 DDA injections of different human tissues and cell types including HEK293  
Rosenberger, G. et al. (2014)



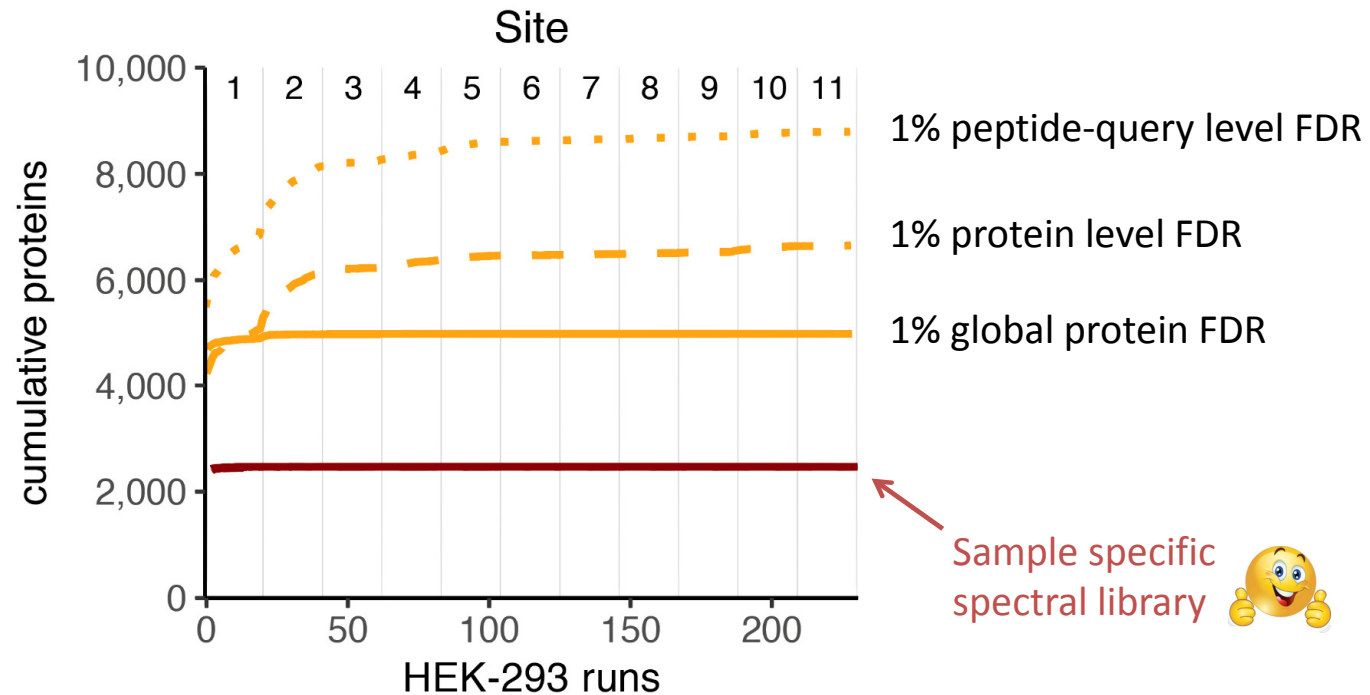
# Scaling up DIA – Error rate control

## DIA data:

- Inter-laboratory study:  
229 DIA measurements of same HEK-293 cell lysate  
Collins et al. (2017)

## Spectral library:

- Combined assay library (CAL):  
331 DDA injections of different human tissues and cell types including HEK293  
Rosenberger, G. et al. (2014)



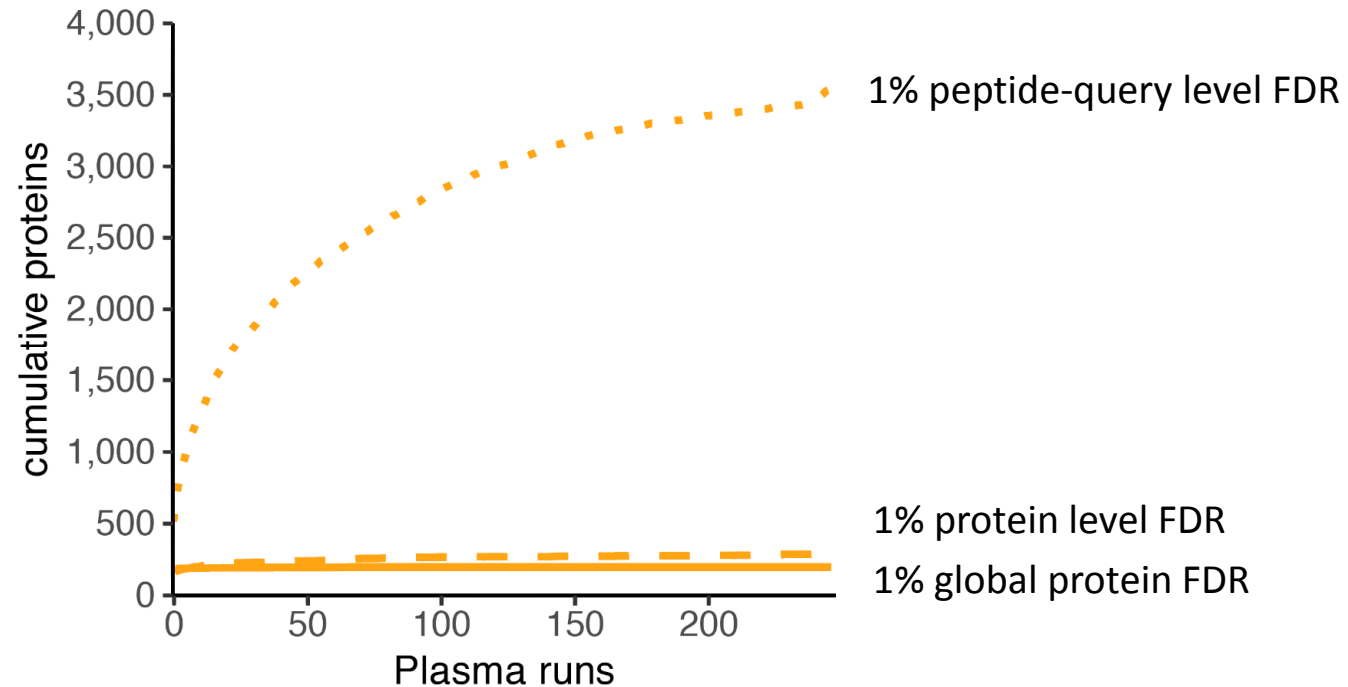
# Scaling up DIA – Error rate control

## DIA data:

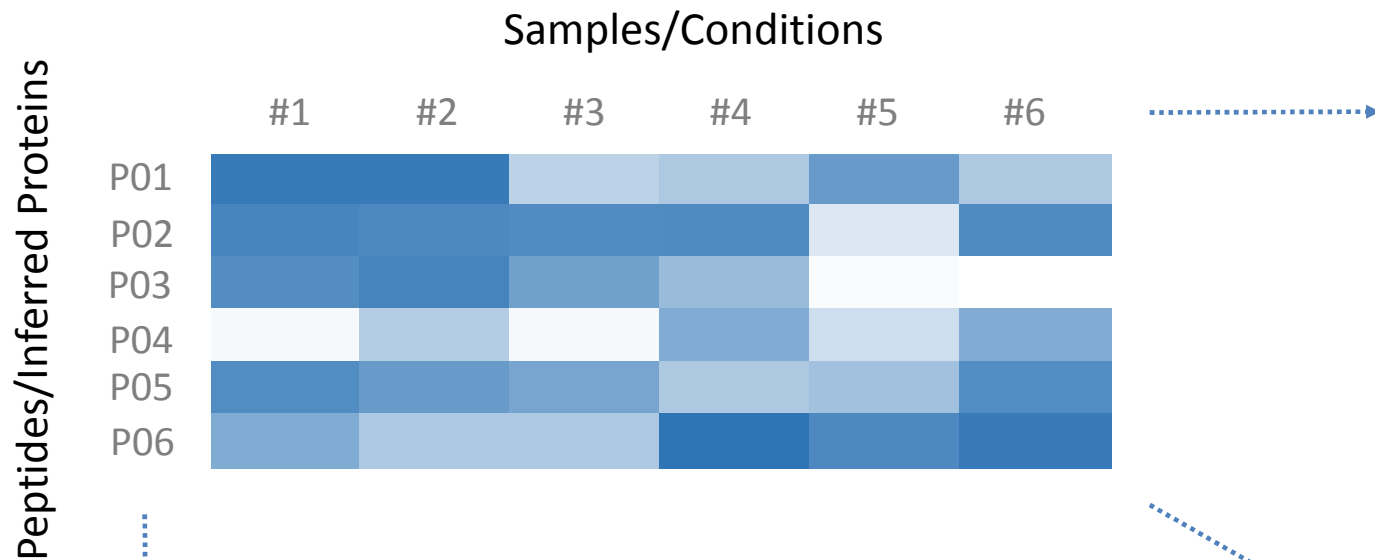
- Blood plasma dataset with 246 samples  
Liu et al. (2015)

## Spectral library:

- Combined assay library (CAL):  
331 DDA injections of different human  
tissues and cell types including HEK293  
Rosenberger, G. *et al.* (2014)



# Scaling up DIA – Error rate control



For large comprehensive  
spectral libraries:  
Use protein-level FDR

&

Context matters!  
Use global protein master list  
constraint to avoid error  
accumulation across samples!

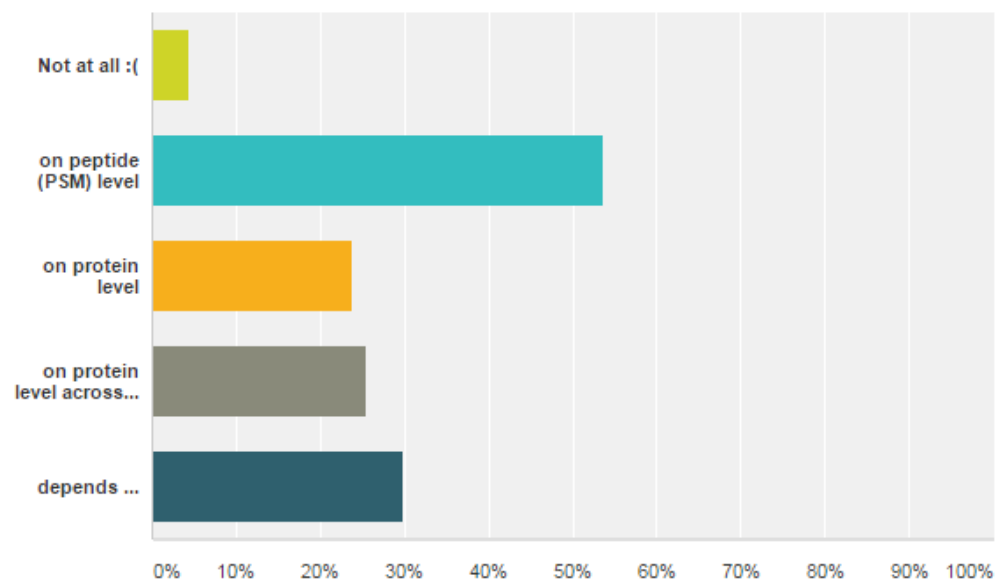
# Survey question 3

## 3. At which level do you control your FDR?

- Not at all :(
- on peptide (PSM) level
- on protein level
- on protein level across the entire dataset
- depends ...

## At which level do you control your FDR?

Answered: 67 Skipped: 11



Answer Choices	Responses
▼ Not at all :(	4.48% 3
▼ on peptide (PSM) level	53.73% 36
▼ on protein level	23.88% 16
▼ on protein level across the entire dataset	25.37% 17
▼ depends ...	29.85% 20
Total Respondents: 67	

Lukas



ASMS 2017

**DIA Workshop –  
“Depth of Proteome Coverage”**

Lukas Reiter, Biognosys

# Depth of Proteome Coverage

Why having a large proteome coverage?

- Discovery
  - E.g. drug target deconvolution with Limited proteolysis (LiP) \*
- Low abundant wish list proteins combined with discovery
- Multi OMICS -> increase overlap with other data sets

How can the proteome coverage be increased?

- Sample (prep)
- **Chromatography**
- Instrumentation
- **DIA Method**
- **Spectral library**
- **Precision iRT**

*Leuenerger et al. Cell-wide analysis... Science (2017)*

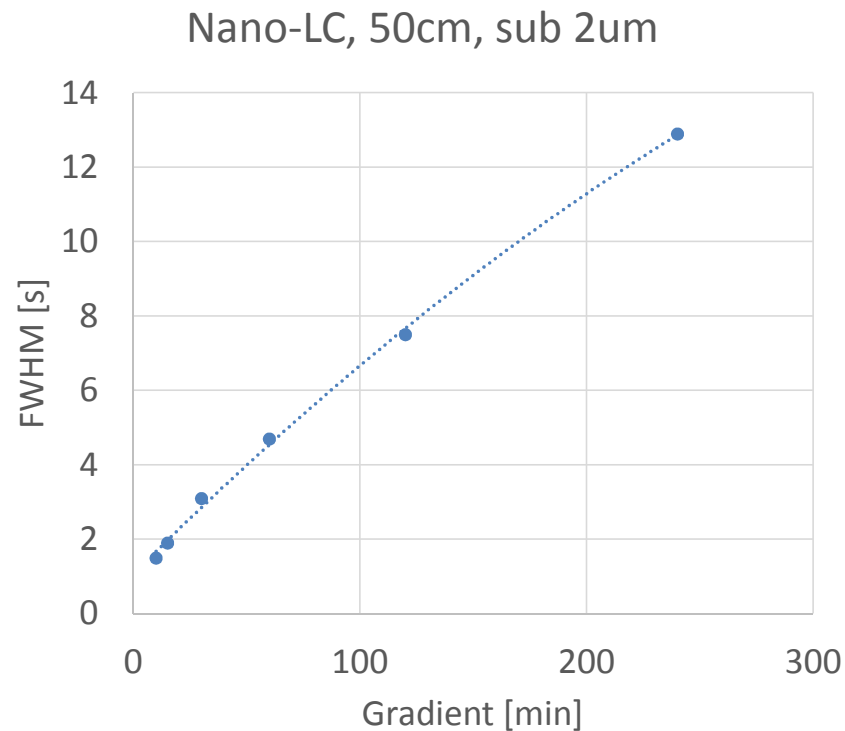
# Chromatography

How to get a high peak capacity?

- UHPLC
- Nano-LC
- 75um ID long columns
- Sub 2um beads
- Long gradients
- Low dead volumes

**1m column, 4h gradient**

- Peak capacity > 700



FWHM: measured as median for all peptides identified in a HeLa digest

Peak capacity:  $1 + g / (\text{FWHM} * 1.7)$

# DIA Method

## Step by step optimization

### 1) Data points per peak

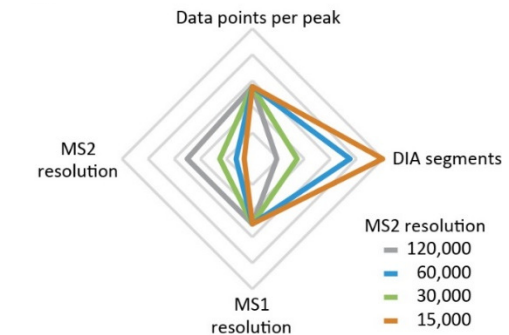
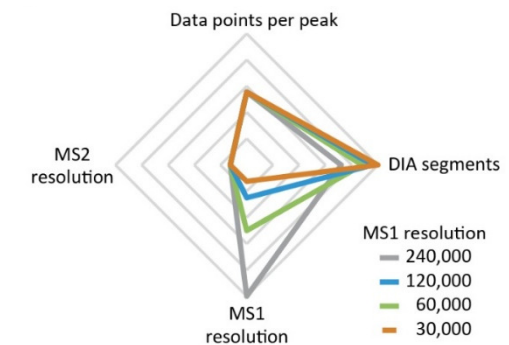
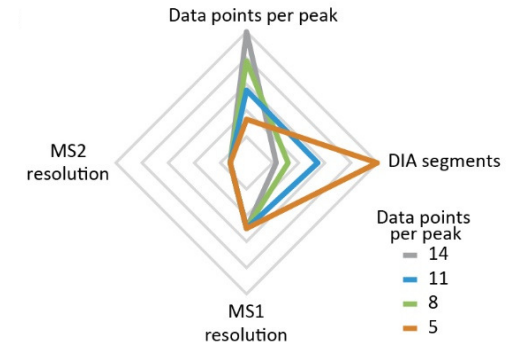
- Generate a scouting method which over samples the peaks
- Scale the number of MS2 segments from the scouting method to result in 5, 8, 11 and 14 data points per peak
- Pick the best method

### 2) MS1 resolution

- Vary the MS1 resolution from 30'000 to 240'000 (balance the MS2 segments to keep data points per peak constant)
- Pick the best method

### 3) MS2 resolution

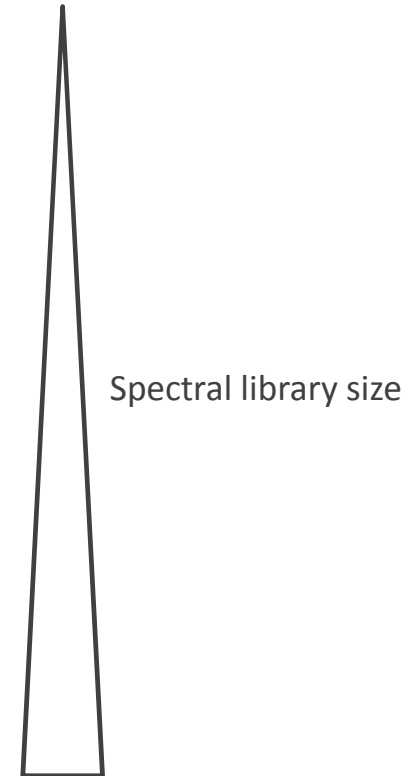
- Vary the MS2 resolution from 15'000 to 120'000 (balance the MS2 segments to keep data points per peak constant)
- Pick the best method



# Spectral Library

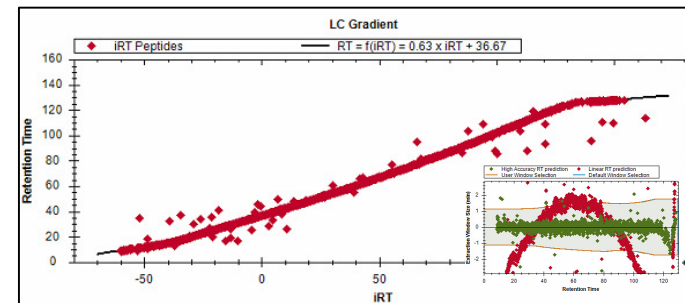
## Evolution over time for Biognosys

- Replicate injections and DDA on project samples
- Mild fractionation
  - 6 high pH reversed phased fractions
- Deep fractionation
  - Two condition pools
  - Pooled micro fractions from UHPLC
  - 10 fractions each
- ...
- Resource spectral libraries or from large sets of synthetic peptides



# Precision iRT

- Extends the indexed retention time (iRT) concept
- Allows very precise targeting in retention time dimension
- Dynamically adapts
- XIC windows of 1-3% of gradient length can be achieved
  - Especially when using spectral libraries acquired on exactly the same setup



# Depth of Proteome Coverage

## *Some example data*

- Setup
  - Deep project-specific spectral library (MaxQuant & Spectronaut)
  - Analyzed with Spectronaut
  - Peptide and protein FDR 1%
  - 4h gradients, 1m column
  - HEK-293 sample
- Single run results
  - 7'060 protein groups, 154'643 precursors
  - Median XIC width 8.5 min
  - Peak capacity 710, median FWHM 13 s
- Technical triplicates
  - 6'534 proteins with CVs < 20%
  - 123'700 precursors with CVs < 20%
  - Data completeness for precursors: 91%

# Survey Question 4

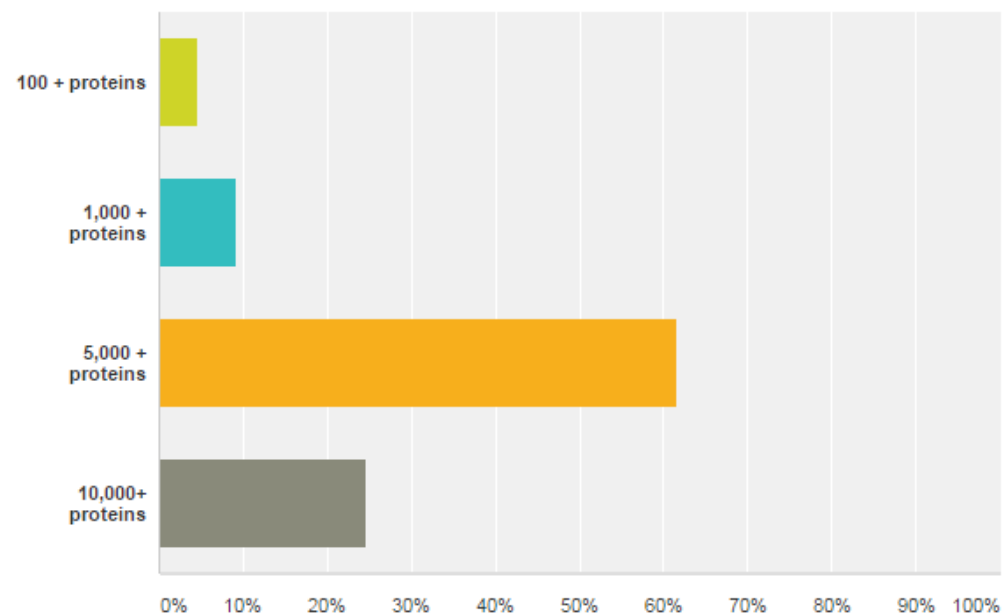
4. What do you consider a deep proteome coverage?

- 100 + proteins
- 1,000 + proteins
- 5,000 + proteins
- 10,000+ proteins



## What do you consider a deep proteome coverage?

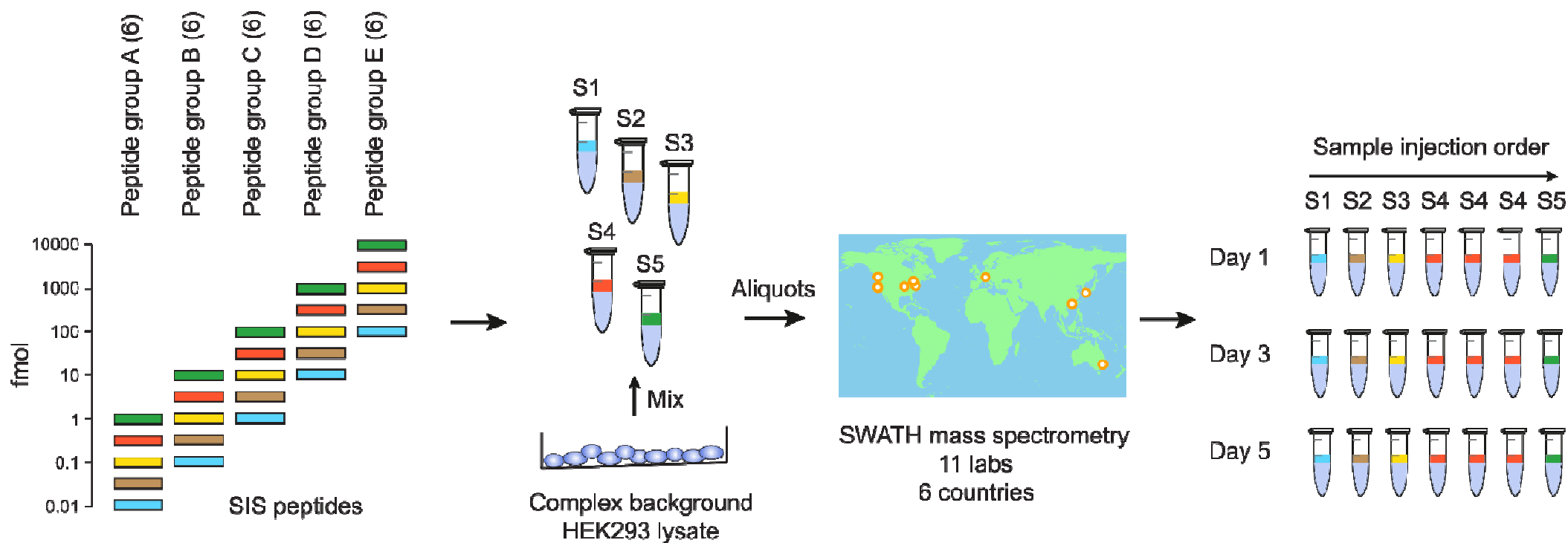
Answered: 65 Skipped: 13



Answer Choices	Responses
▼ 100 + proteins	4.62% 3
▼ 1,000 + proteins	9.23% 6
▼ 5,000 + proteins	61.54% 40
▼ 10,000+ proteins	24.62% 16
Total	65

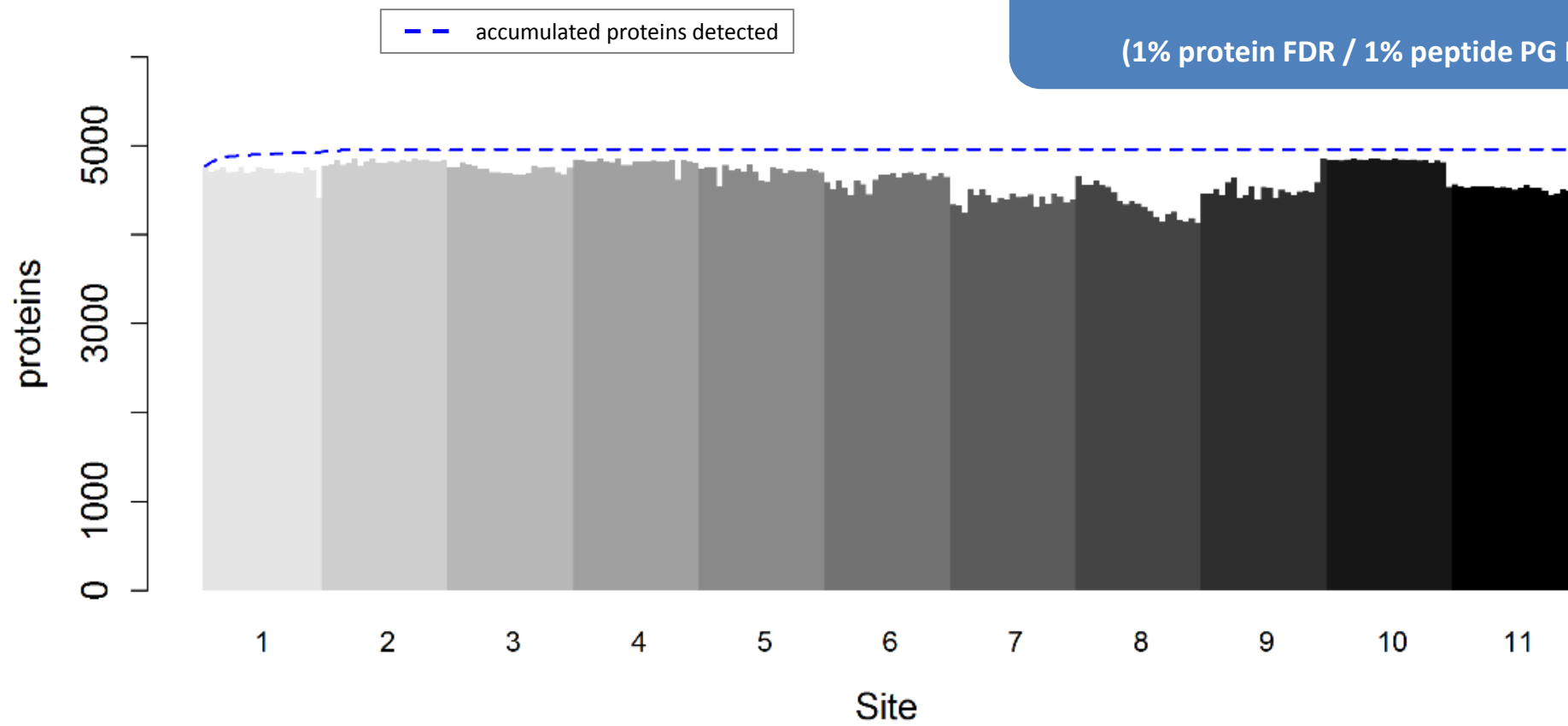
Ben

# Study design -- Inter-lab SWATH-MS



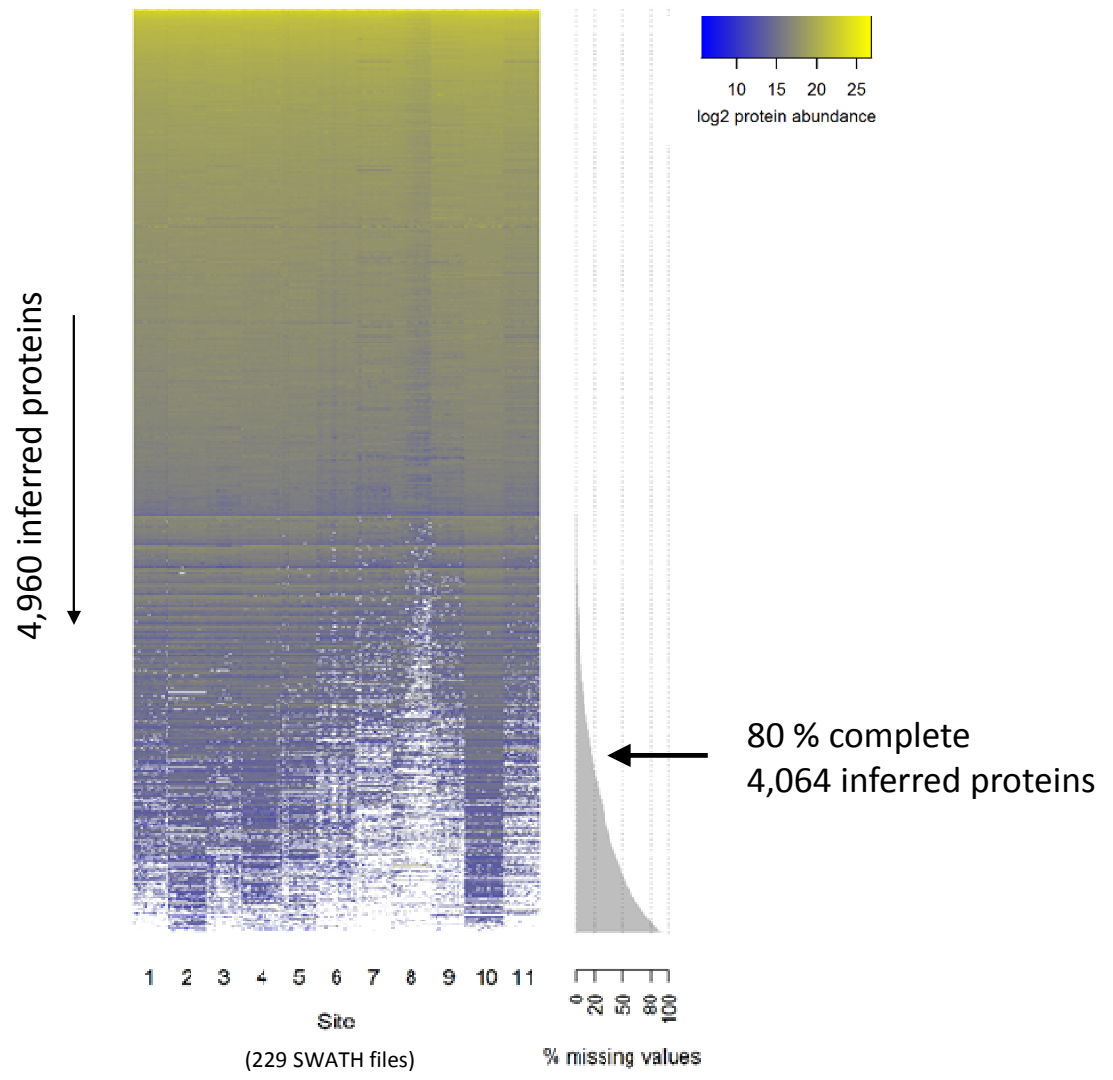
## Peptide/protein detection rates (HEK293 lysate)

Total -- 4,960 proteins / 39,928 peptide PGs  
Site median – 4,691 / 34,286 PGs  
(1% protein FDR / 1% peptide PG FDR)



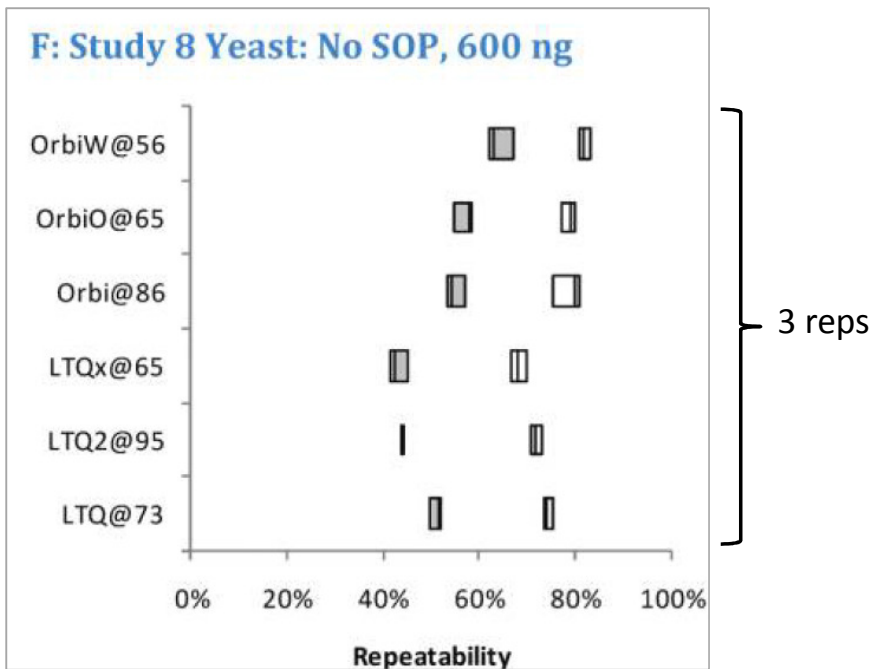
# Data completeness

No alignment  
No ID propagation  
between runs

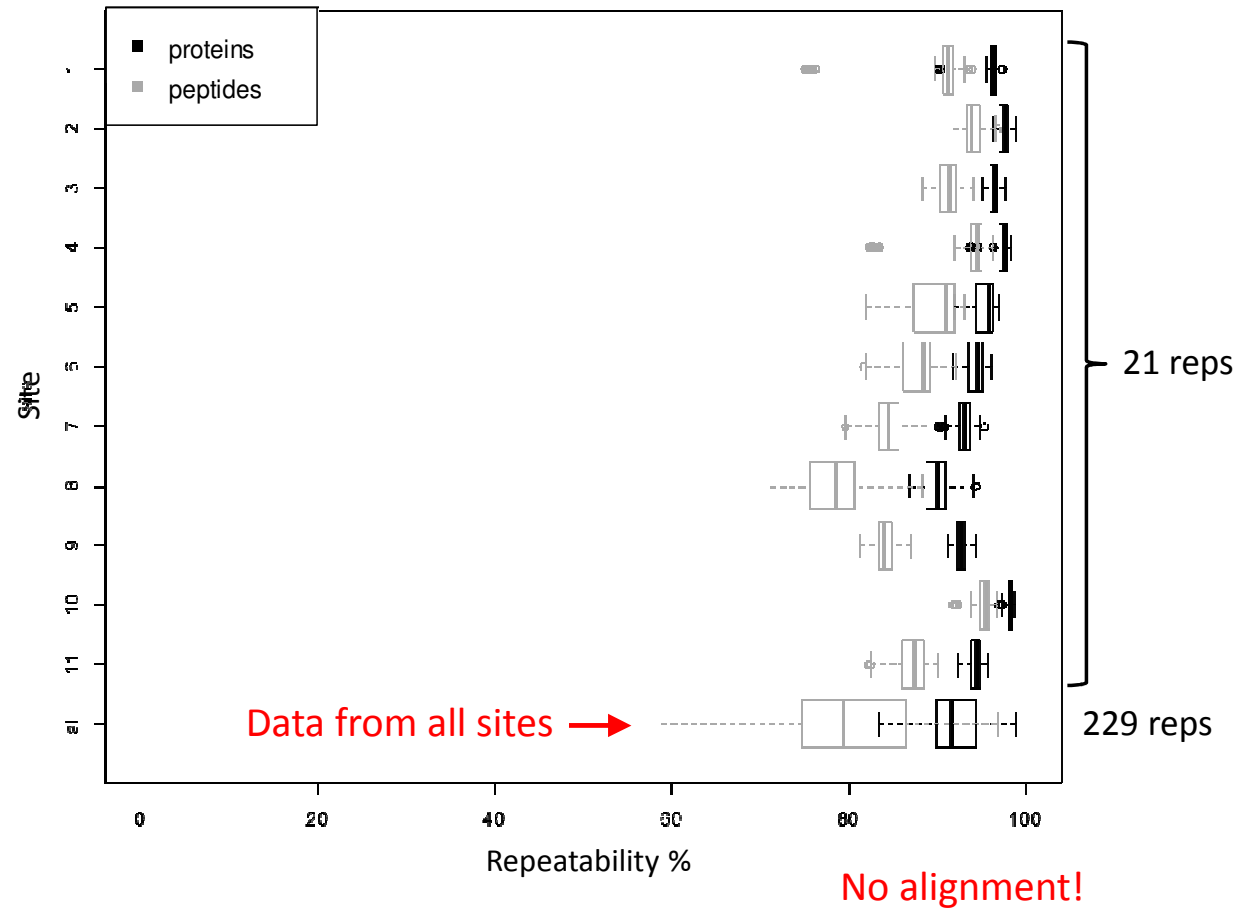


# Repeatability of identification

Tabb et al. – DDA (2010)

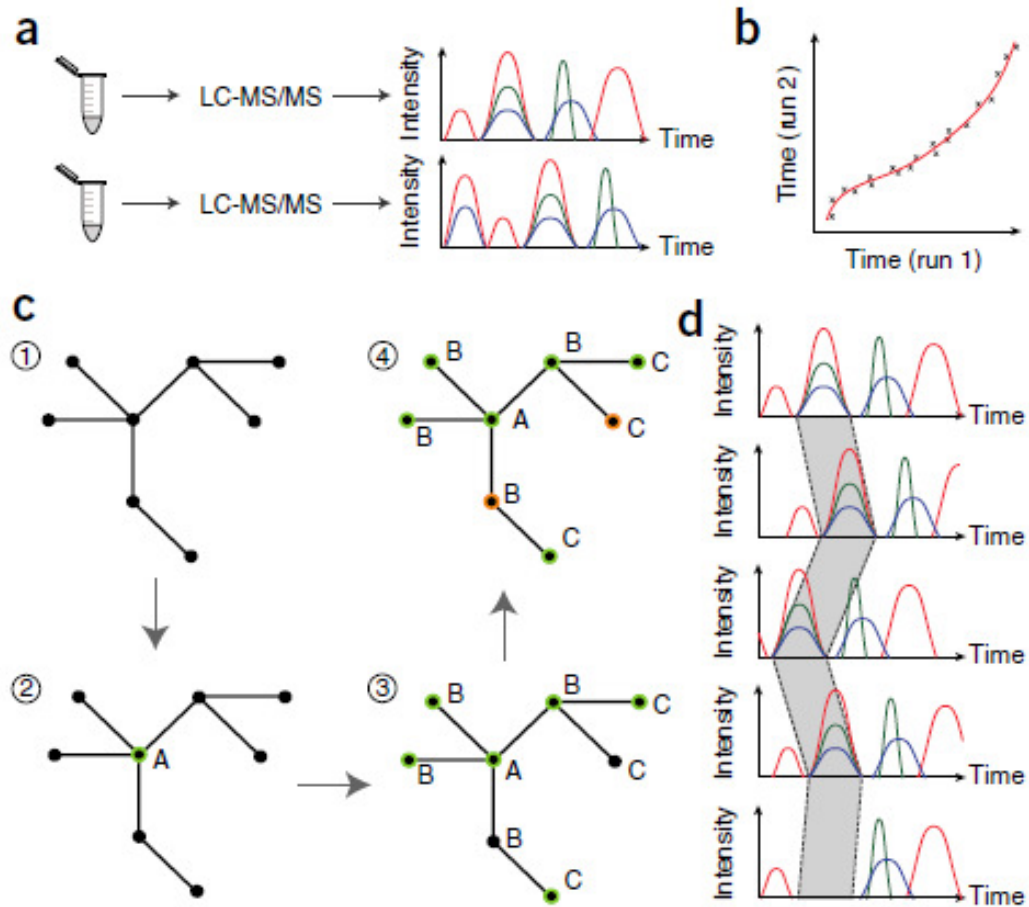


Inter-lab SWATH-MS



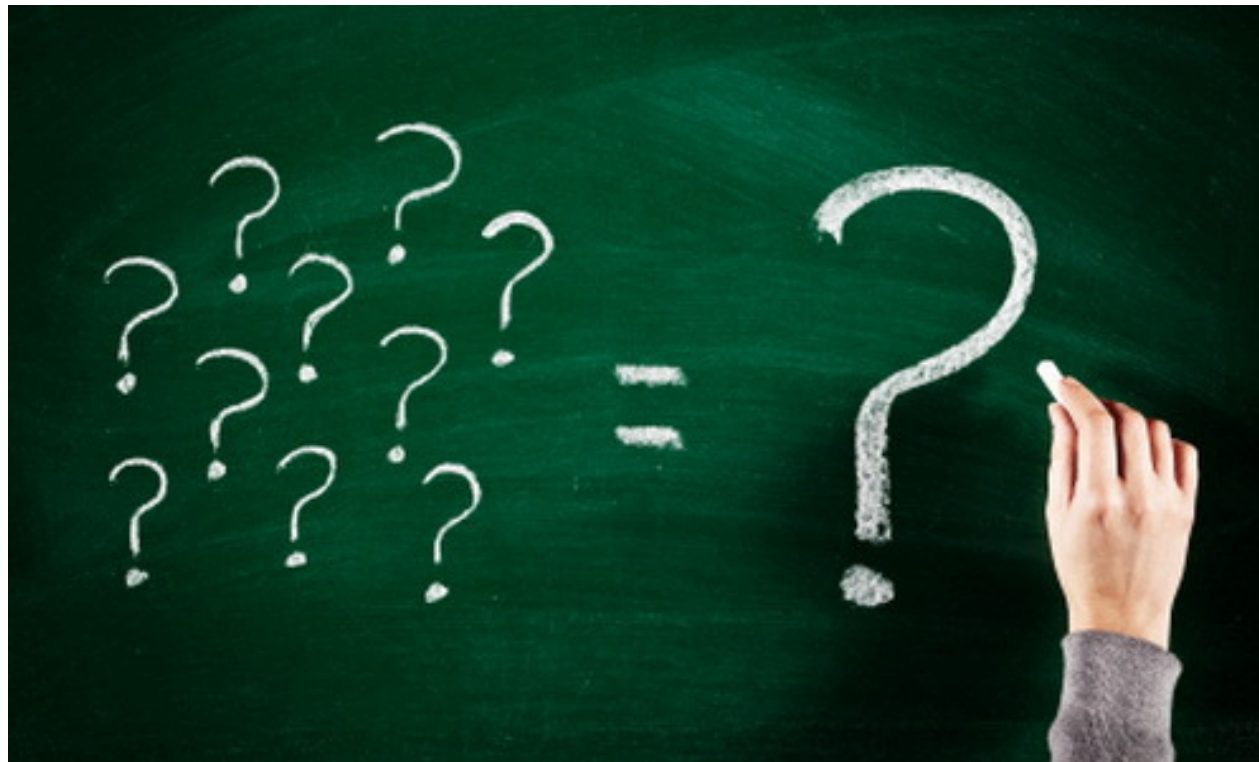
Tabb, D. L. *et al.* Repeatability and Reproducibility in Proteomic Identifications by Liquid Chromatography–Tandem Mass Spectrometry. *J. Proteome Res.* **9**, 761–776 (2010).

# Alignment can only improve completeness...



**TRIC**  
TTransfer of  
Identification  
Confidence

# Comments? Questions?

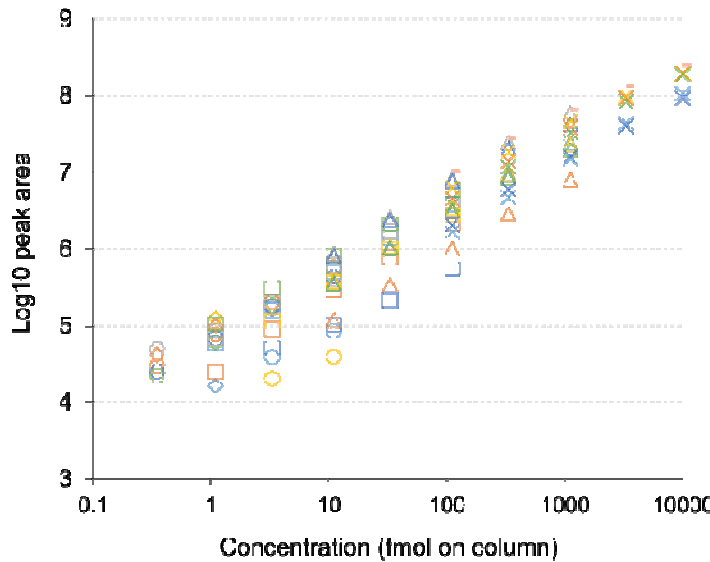




Linearity, dynamic range, and response differences

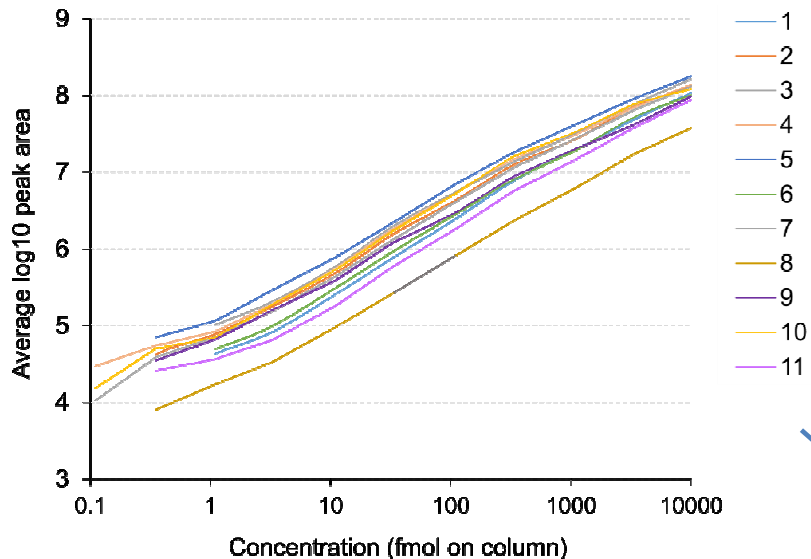
average

1 site



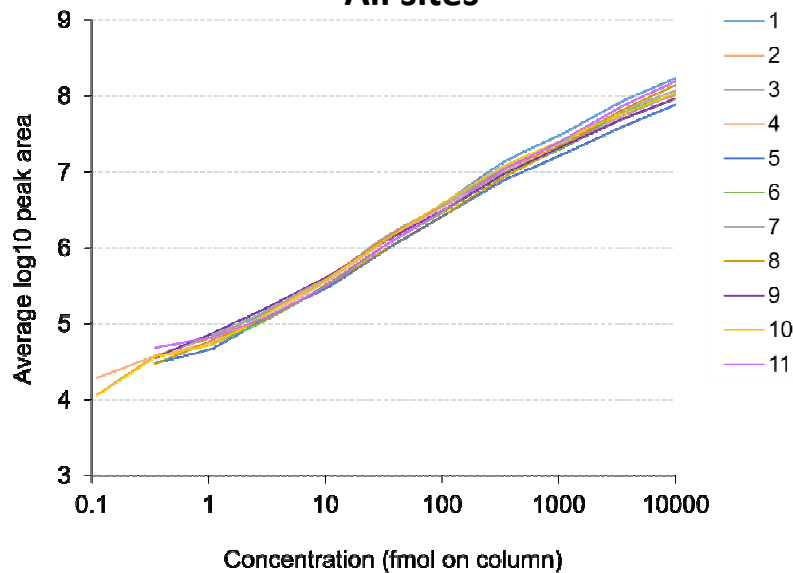
- ◆ AEN.LVE
- ◆ HNRPG.GFA
- ◆ SFPQ.FAQ
- ◆ SMD3.VAQ
- ◆ ZMAT5.APP
- ◆ ANRA2.SVQ
- ◆ DDX5.FV1
- ◆ ESRP2.EAS
- ◆ RUXF.CICAM.NN
- ◆ SFSWA.EAQ
- ◆ SMD1.EPV
- ◆ SNRPA.EVS
- ◆ APR.APA
- ◆ E2F7.LDF
- ◆ NCBP2.SDS
- ◆ ROA2.GCN
- ◆ SMD2.NNT
- ◆ SRSF7.AFS
- ◆ FL2D.YTD
- ◆ PHLA3.SGG
- ◆ SFB2.VGE
- ◆ SFR19.TPE
- ◆ SRSF3.AFG
- ◆ WDR83.DGQ
- ◆ AK17A.FDW
- ◆ HNHPU.SSG
- ◆ RBM5.GLP
- ◆ RBMX.LFI
- ◆ SRRM4.LGQ
- ◆ TRA2A.TGP

All sites

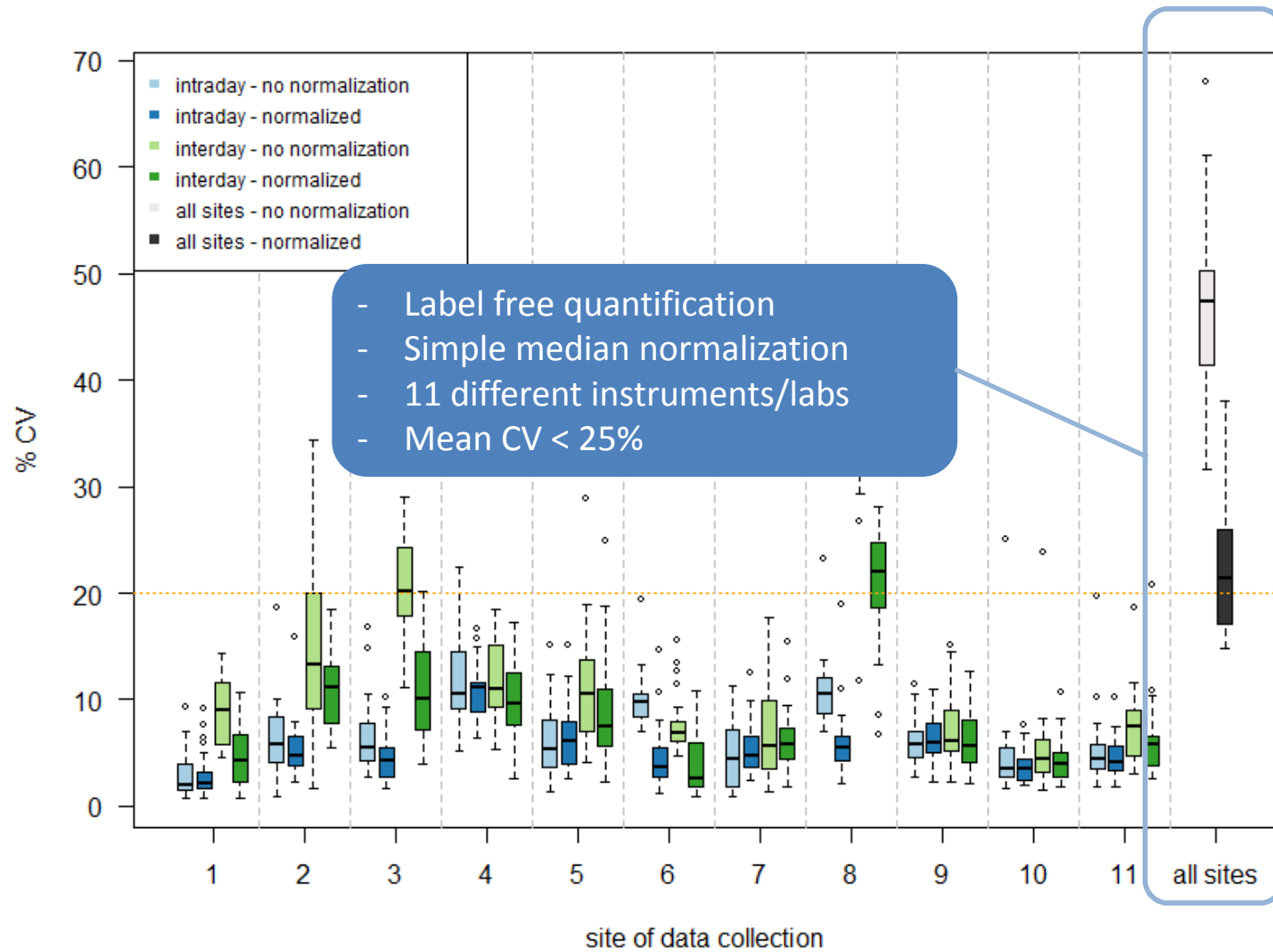


Normalize (based on HEK293 medians)

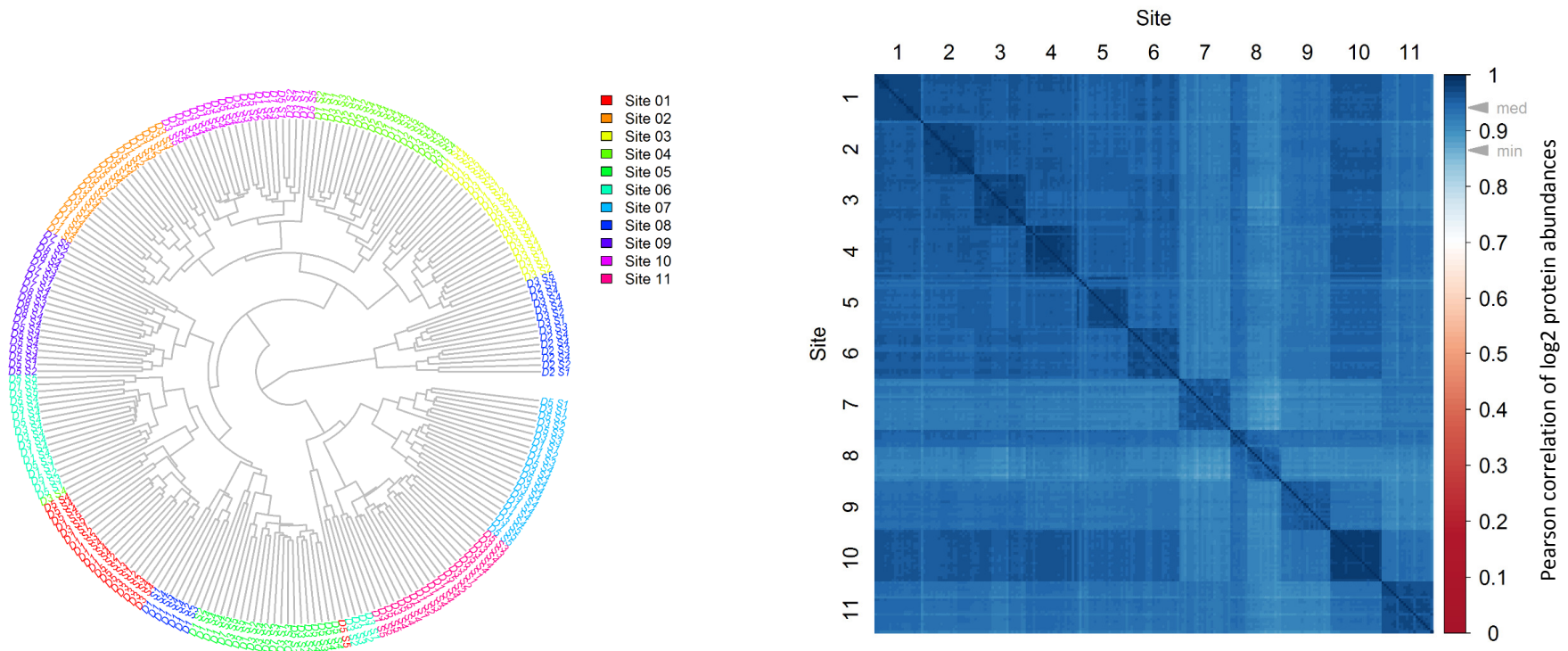
All sites



# Reproducibility (30 x SIL peptides)

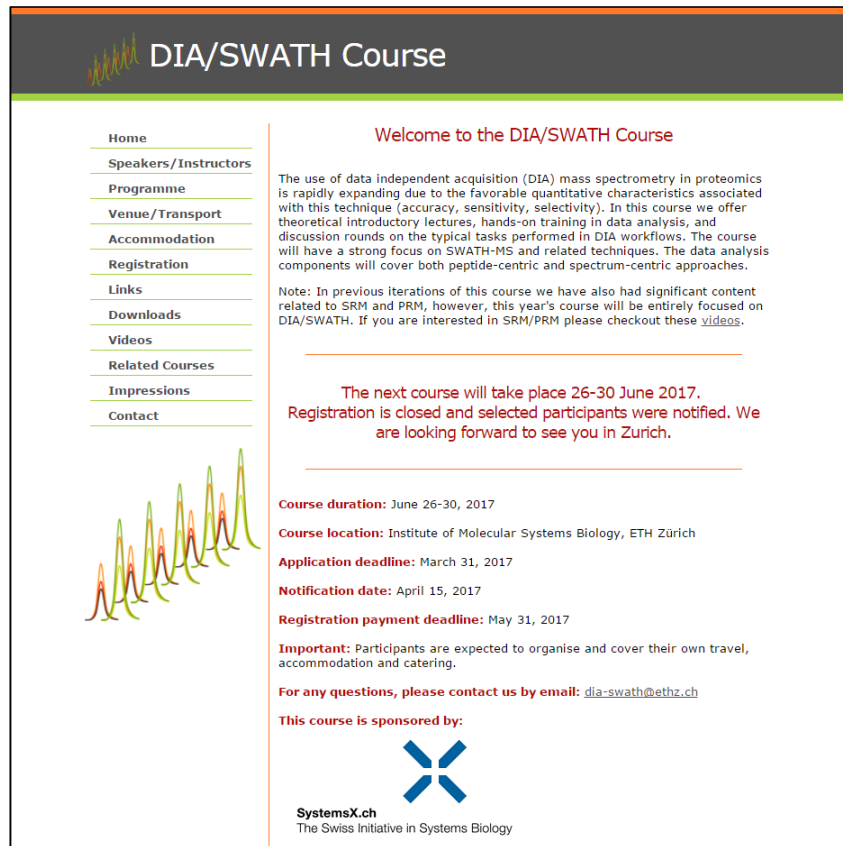


# Global similarity of quantitative protein abundance profiles (HEK293 lysate)



**Median Pearson correlation (overall) = 0.940**  
**Median Pearson correlation (within sites) = 0.971**  
**Minimum Pearson correlation = 0.868**

# If you want learn more about DIA/SWATH



The screenshot shows the homepage of the DIA/SWATH Course website. The header features a stylized chromatogram icon and the text "DIA/SWATH Course". A left-hand navigation menu lists various sections: Home, Speakers/Instructors, Programme, Venue/Transport, Accommodation, Registration, Links, Downloads, Videos, Related Courses, Impressions, and Contact. The main content area is titled "Welcome to the DIA/SWATH Course" and contains a detailed description of the course, a note about previous iterations, and a message stating that the next course will take place from June 26-30, 2017, with registration closed. Key dates and information are listed, including the course duration, location at the Institute of Molecular Systems Biology, ETH Zürich, application deadline of March 31, 2017, notification date of April 15, 2017, and registration payment deadline of May 31, 2017. An important note mentions that participants are responsible for their own travel and accommodation. Contact information is provided as [dia-swath@ethz.ch](mailto:dia-swath@ethz.ch). The course is sponsored by SystemsX.ch, The Swiss Initiative in Systems Biology, with a logo consisting of four blue arrows pointing towards the center.

**DIA/SWATH Course**

**Welcome to the DIA/SWATH Course**

The use of data independent acquisition (DIA) mass spectrometry in proteomics is rapidly expanding due to the favorable quantitative characteristics associated with this technique (accuracy, sensitivity, selectivity). In this course we offer theoretical introductory lectures, hands-on training in data analysis, and discussion rounds on the typical tasks performed in DIA workflows. The course will have a strong focus on SWATH-MS and related techniques. The data analysis components will cover both peptide-centric and spectrum-centric approaches.

Note: In previous iterations of this course we have also had significant content related to SRM and PRM, however, this year's course will be entirely focused on DIA/SWATH. If you are interested in SRM/PRM please checkout these [videos](#).


The next course will take place 26-30 June 2017.  
Registration is closed and selected participants were notified. We are looking forward to see you in Zurich.

**Course duration:** June 26-30, 2017  
**Course location:** Institute of Molecular Systems Biology, ETH Zürich  
**Application deadline:** March 31, 2017  
**Notification date:** April 15, 2017  
**Registration payment deadline:** May 31, 2017

**Important:** Participants are expected to organise and cover their own travel, accommodation and catering.

For any questions, please contact us by email: [dia-swath@ethz.ch](mailto:dia-swath@ethz.ch)

This course is sponsored by:

  
**SystemsX.ch**  
The Swiss Initiative in Systems Biology

[dia-swath-course.ethz.ch](http://dia-swath-course.ethz.ch)

Registration for this year is closed  
but lecture videos will be  
posted **late July 2017**

Thanks for participating!!

Ideas for discussion topics for next year to:

[collins@imsb.biol.ethz.ch](mailto:collins@imsb.biol.ethz.ch)

[bludau@imsb.biol.ethz.ch](mailto:bludau@imsb.biol.ethz.ch)