Peptide-level robust ridge regression modeling improves both sensitivity and specificity in quantitative proteomics

Ludger Goeminne

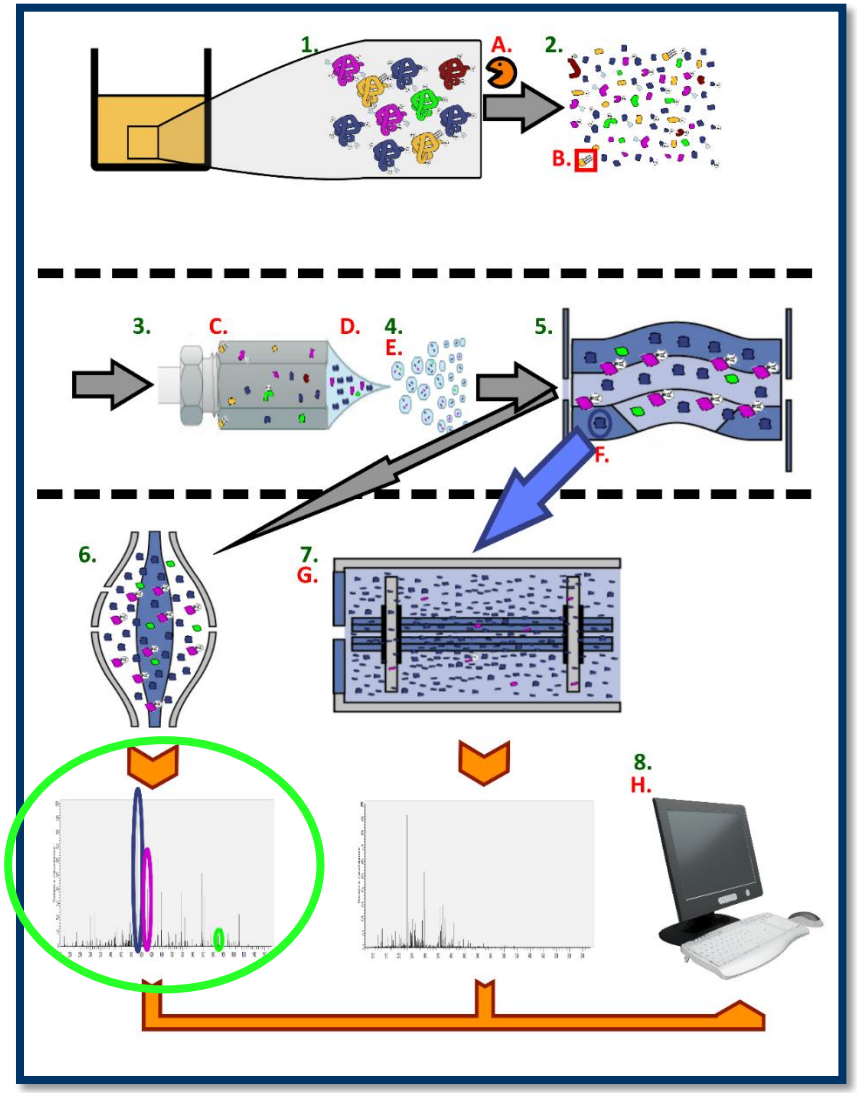
Promotors:

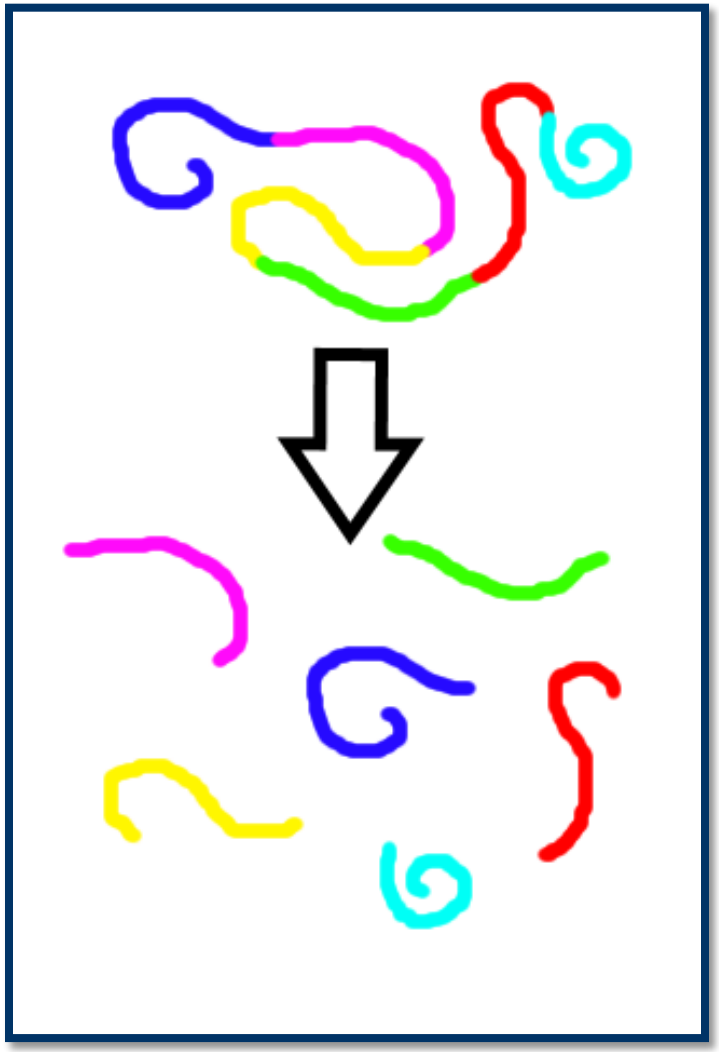
Lieven Clement

Kris Gevaert

Klaas Vandepoele

A story about **relative protein quantification** in **label-free shotgun proteomics**





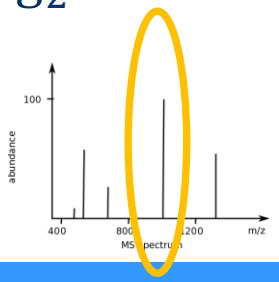
Outline

- Problem: reliable differential quantitation
- Solution: peptide-based models
- Improving solution via:
  1. Shrinkage estimation
  2. Borrowing information across proteins
  3. Weighing down outliers
- Leads to:
  1. Better fold change estimates
  2. Better sensitivity and specificity
- Conclusions: all of the above
- Acknowledgements

Problem: how to do differential quantification?

For each protein:

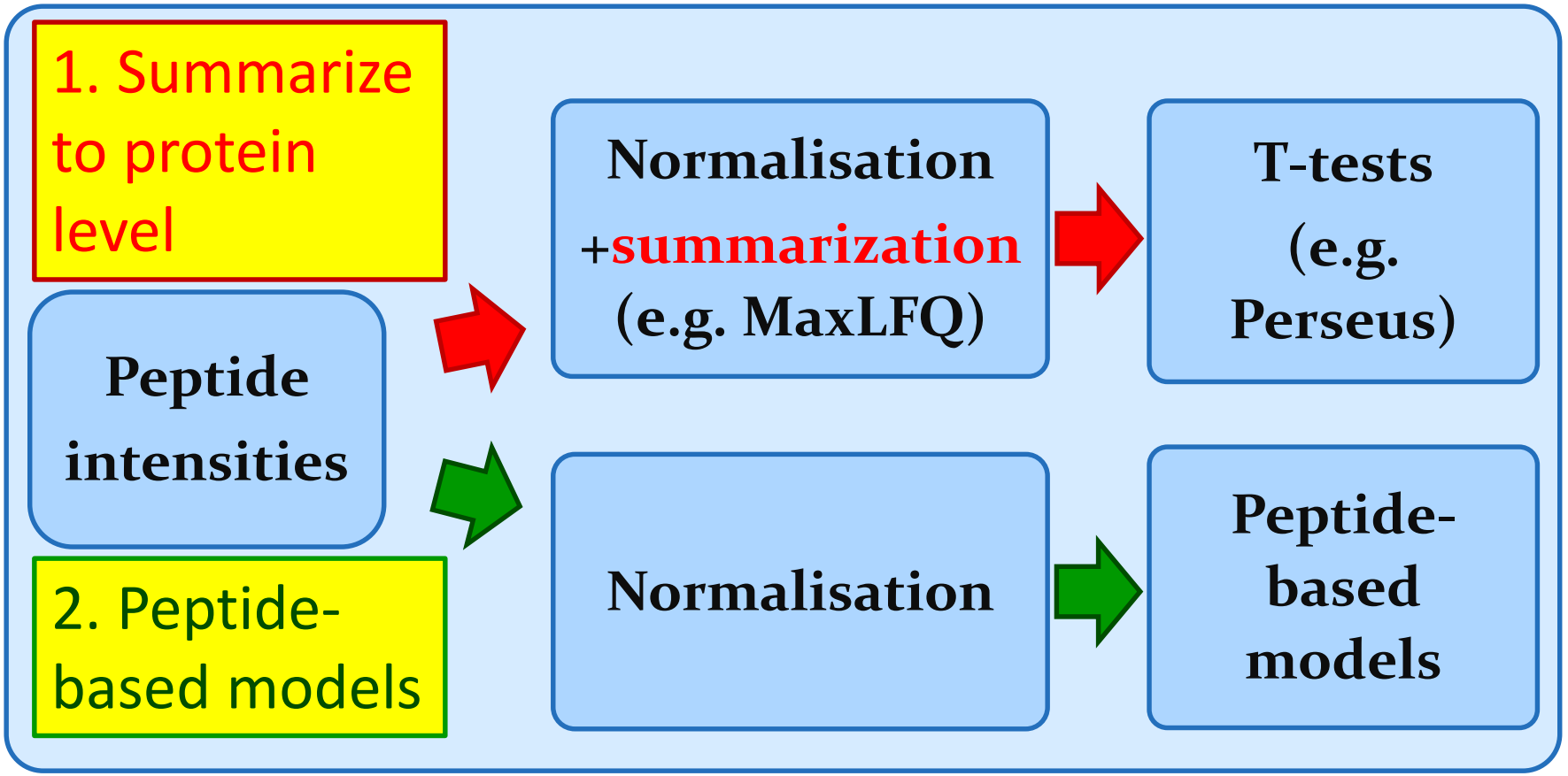
$\log_2 MS\ intensity$ → $treatment + peptide + repeat$



22.6464	Treatment 1	Peptide A	Rep 1
17.85773	Treatment 1	Peptide B	Rep 1
15.4947	Treatment 1	Peptide C	Rep 2
14.02125	Treatment 1	Peptide D	Rep 2
18.0965	Treatment 2	Peptide A	Rep 3
14.59100	Treatment 2	Peptide B	Rep 3
14.2959	Treatment 2	Peptide C	Rep 3

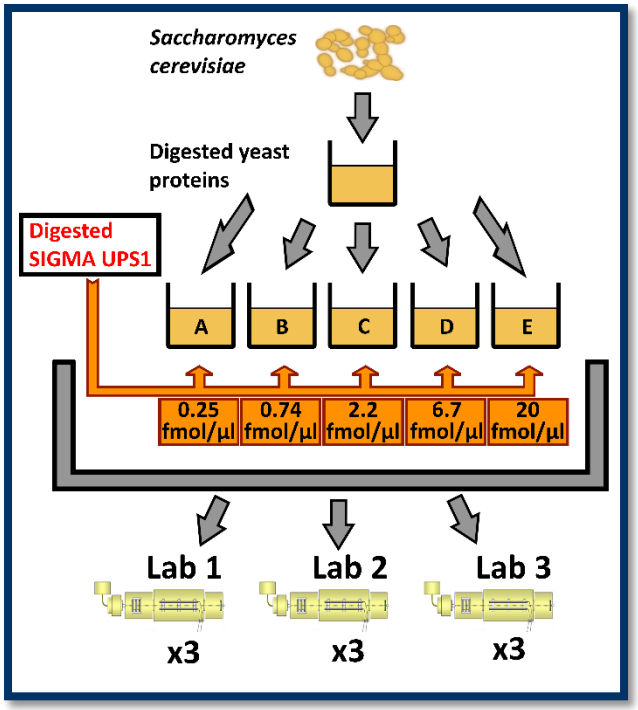
Problem: how to do differential quantification?

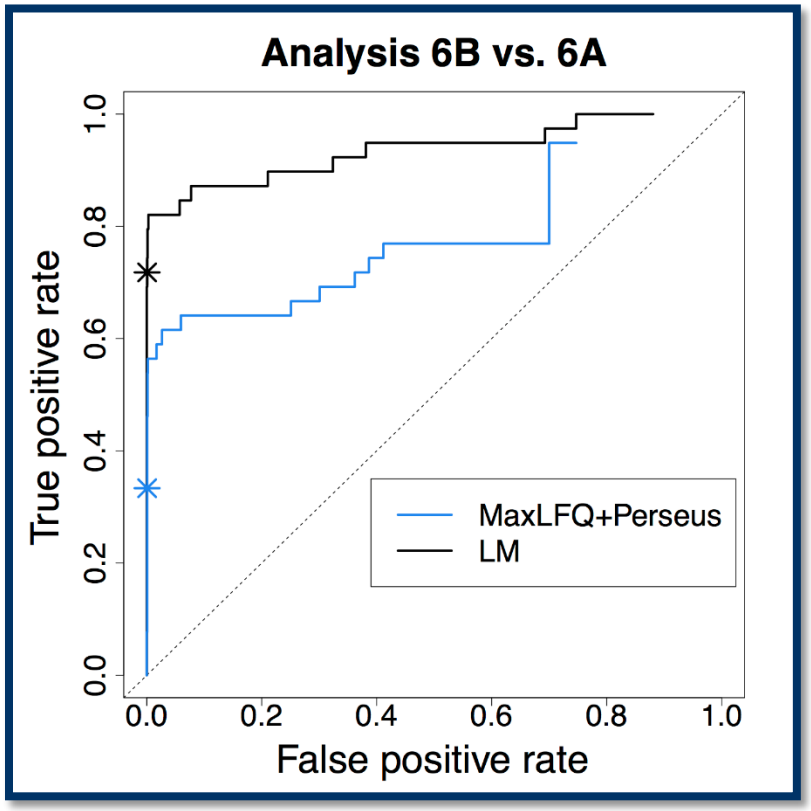
Solution: 2 main ways:



Peptide-based models are superior

Spike-in: 48 human proteins in yeast proteome





Daly, et al. (2008), *Journal of Proteome Research*, 7, (3), 1209-1217.

Clough et al. (2009, *Journal of Proteome Research*, 8, (11), 5275-5284.

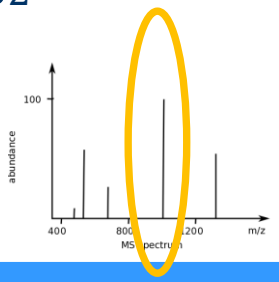
Karpievitch et al. (2009), *Bioinformatics*, 25, (16), 2028-2034.

Goeminne et al. (2015), *Journal of Proteome Research* , 14, (6), 2457-2465.

Peptide-based models

A model **for each protein:**

$$\log_2 MS\ intensity \sim intercept + treatment + peptide + repeat + error$$

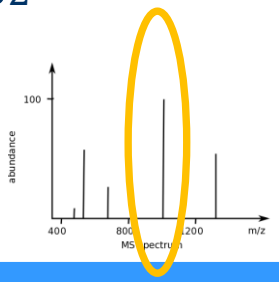


22.6464	Intercept	Treatment 1	Peptide A	Rep 1	Error 1
17.85773	Intercept	Treatment 1	Peptide B	Rep 1	Error 2
15.4947	Intercept	Treatment 1	Peptide C	Rep 2	Error 3
14.02125	Intercept	Treatment 1	Peptide D	Rep 2	Error 4
18.0965	Intercept	Treatment 2	Peptide A	Rep 3	Error 5
14.59100	Intercept	Treatment 2	Peptide B	Rep 3	Error 6
14.2959	Intercept	Treatment 2	Peptide C	Rep 3	Error 7

Peptide-based models

A model **for each protein:**

$$\log_2 MS\ intensity \sim intercept + treatment + peptide + repeat + error$$

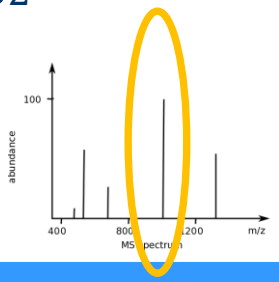


22.6464	16	1.5	4.5	0.5	0.1464
17.85773	16	1.5	-0.2	0.5	0.05773
15.4947	16	1.5	-1	-0.7	-0.3053
14.02125	16	1.5	-2	-0.7	-0.77875
18.0965	16	-1.5	4.5	-0.3	-0.6035
14.59100	16	-1.5	-0.2	-0.3	0.59100
14.2959	16	-1.5	-1	-0.3	0.0959

Peptide-based models

A model **for each protein:**

$$\log_2 MS\ intensity \sim intercept + treatment + peptide + repeat + error$$

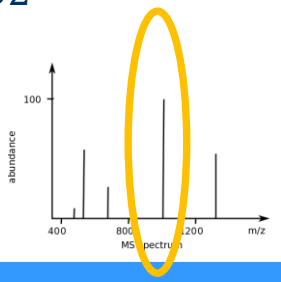


22.6464	16	1.5	4.5	0.5	0.1464
17.85773	16	1.5	-0.2	0.5	0.05773
15.4947	16	1.5	-1	-0.7	-0.3053
14.02125	16	1.5	-2	-0.7	-0.77875
18.0965	16	-1.5	4.5	-0.3	-0.6035
14.59100	16	-1.5	-0.2	-0.3	0.59100
14.2959	16	-1.5	-1	-0.3	0.0959

Peptide-based models

A model **for each protein:**

$$\log_2 MS\ intensity \sim intercept + treatment + peptide + repeat + error$$



22.6464
17.85773
15.4947
14.02125
18.0965
14.59100
14.2959

1.5
1.5
1.5
1.5
-1.5
-1.5
-1.5

1.5

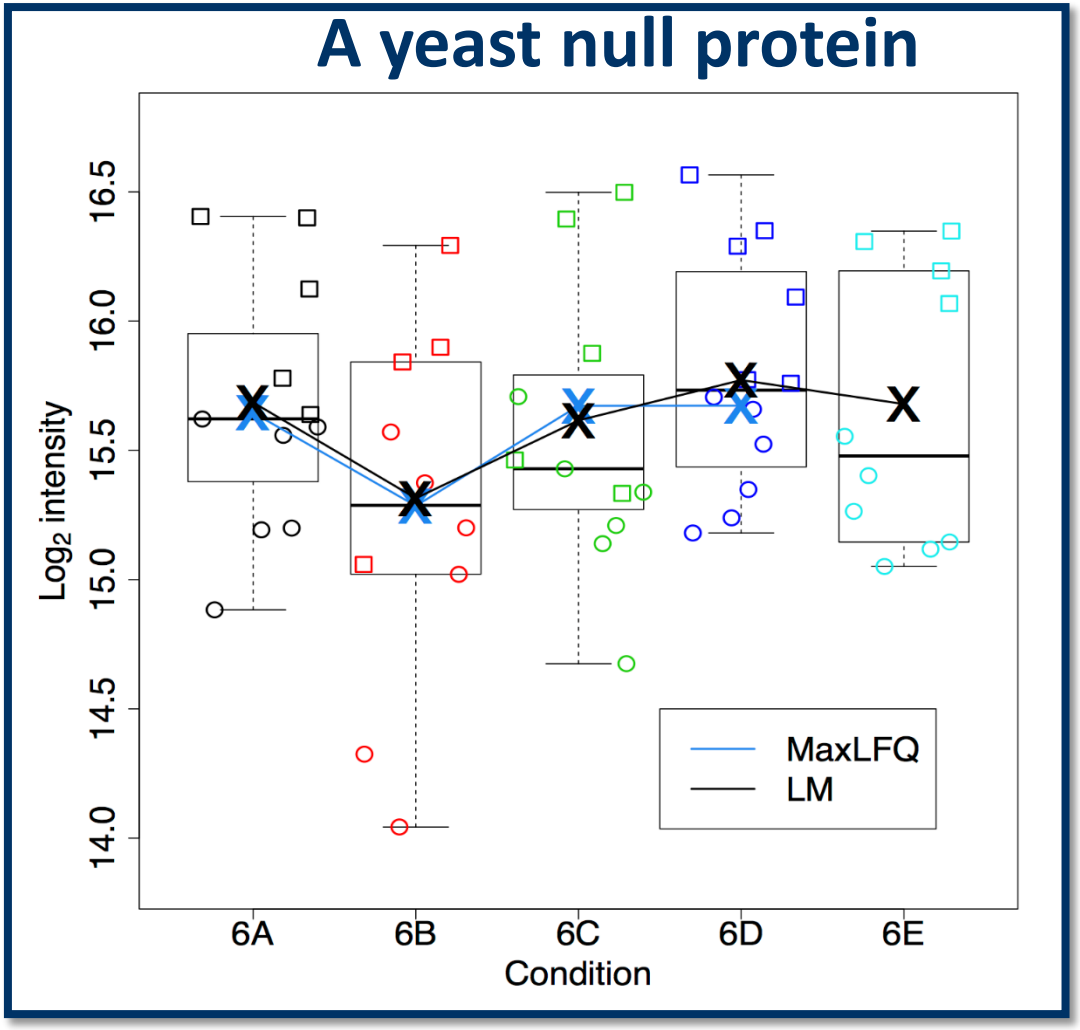
-1.5

-

3

Still some issues…

1. Unstable DA estimates
2. Unstable variance estimates
3. Outliers



Structure of my presentation

- Problem: reliable differential quantitation
- Solution: peptide-based models
- Improving solution via:
  1. Shrinkage estimation
  2. Borrowing information across proteins
  3. Weighing down outliers
- Leads to:
  1. Better fold change estimates
  2. Better sensitivity and specificity
- Conclusions: all of the above
- Acknowledgements

How can we improve upons existing peptide-based models?

Problem	**Solution**
1. Unstable DA estimates	1. **Shrinkage** estimation
2. Unstable variance estimates	2. **Borrow information** across proteins
3. Outliers	3. **Weigh down** outlying peptides

This will lead to:

1. Better fold change estimates
2. Better ranking

1. **Shrinkage estimation**

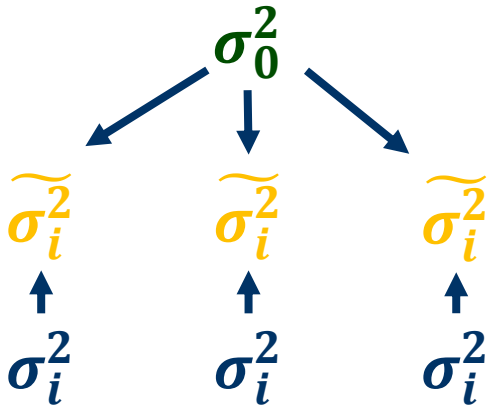
E.g. **ridge regression**: minimize the following loss function:

$$\sum (y - X\hat{\beta})^2 + \lambda_{treat} \sum \hat{\beta}^2_{treat} + \lambda_{pep} \sum \hat{\beta}^2_{pep} + \lambda_{instr} \sum \hat{\beta}^2_{instr}$$

- Penalty on the effect sizes: **shrinkage toward 0**
- **Biased** but (much) more **stable** estimator
- **Sparse data: shrinkage ↗**
- λs: via cross-validation or link with mixed models

2. **Borrow information** across proteins: **Empirical Bayes variance estimation**

Data decides!



- Stabilizes variance estimates
- Get rid of proteins with low fold changes and low variance caused by data sparsity

More details (limma paper):

Smyth (2004), Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3, Article3.

3. **Weigh down outlying peptides**

E.g. **M estimation with Huber weights**

Minimize the following loss function:

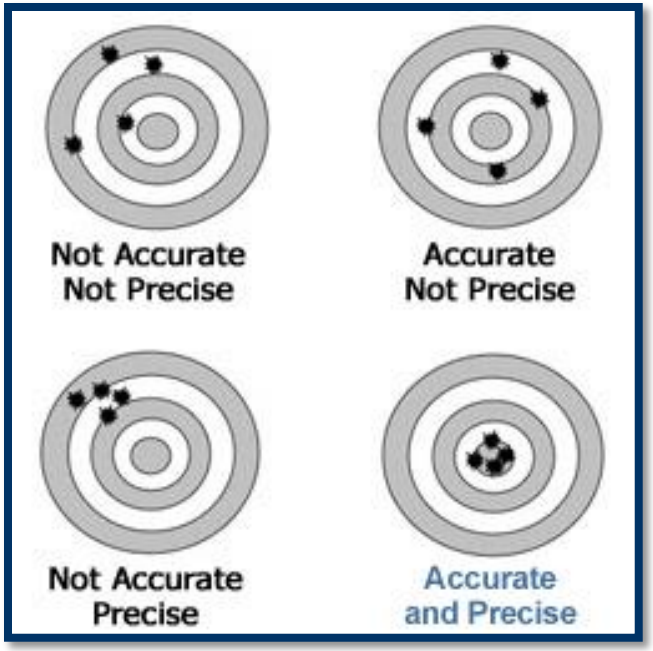
$$\sum \boldsymbol{w}\left(y - X\hat{\beta}\right)^2 + \lambda_{treat}\sum \hat{\beta}^2_{treat} + \lambda_{pep}\sum \hat{\beta}^2_{pep} + \lambda_{instr}\sum \hat{\beta}^2_{instr}$$

- Weigh down outlying observations

Results!

1. Better fold change estimates

-> More accurate and more precise

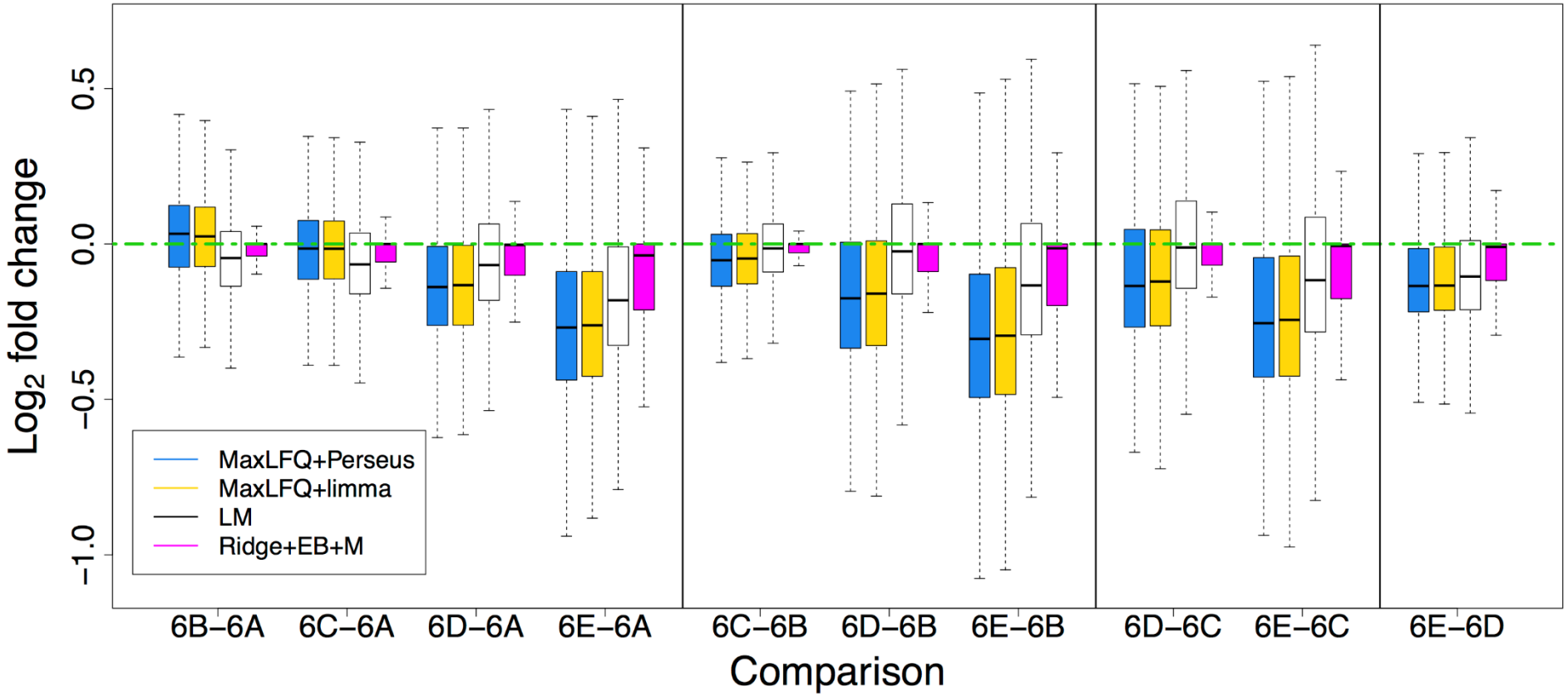


2. Better specificity and sensitivity

-> Improved ranking

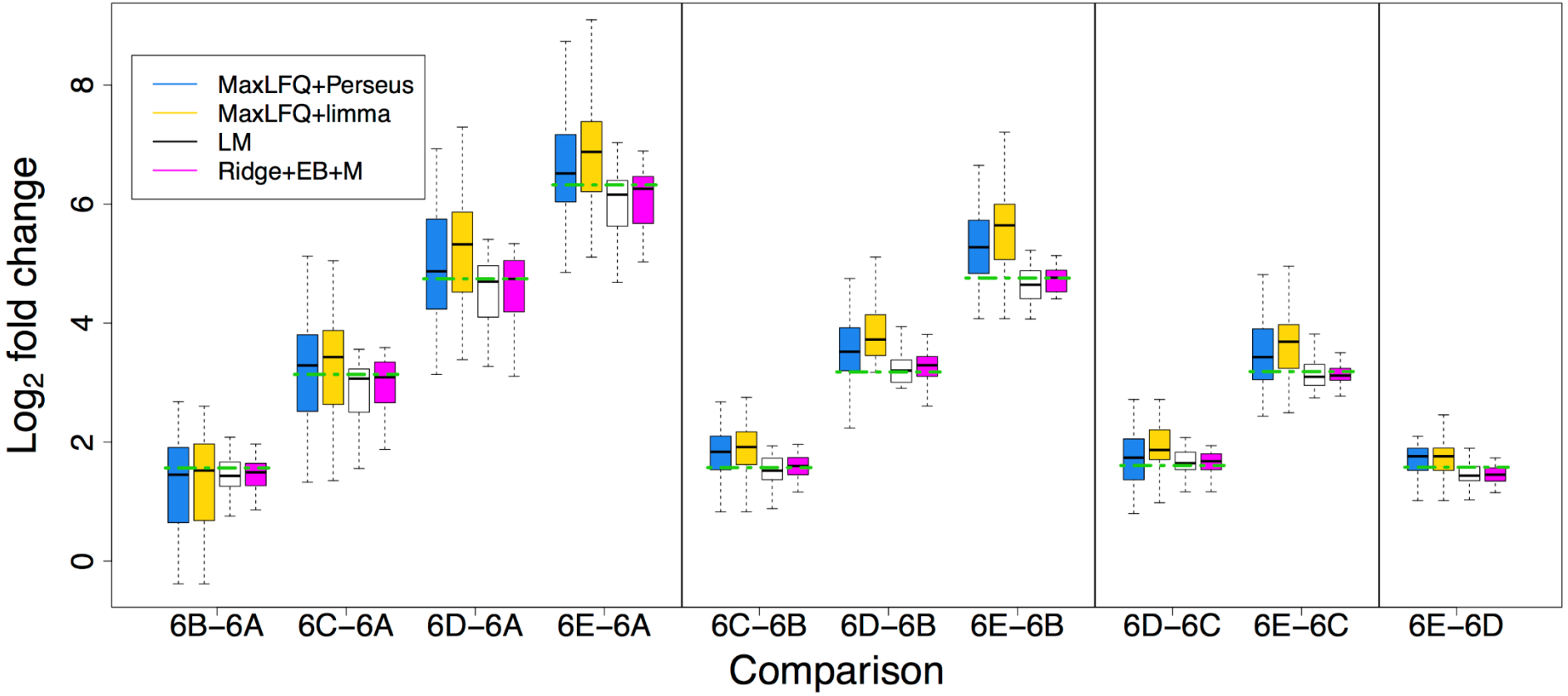
1. Better fold change estimates

For **null proteins**

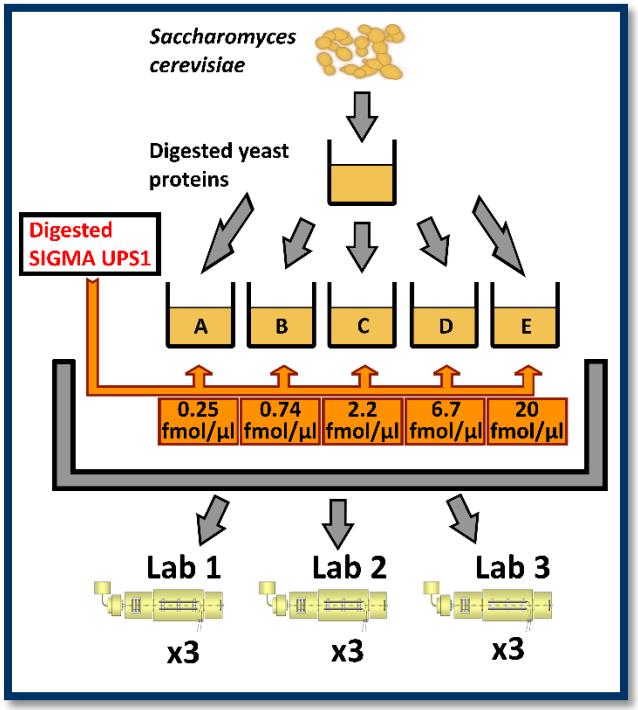


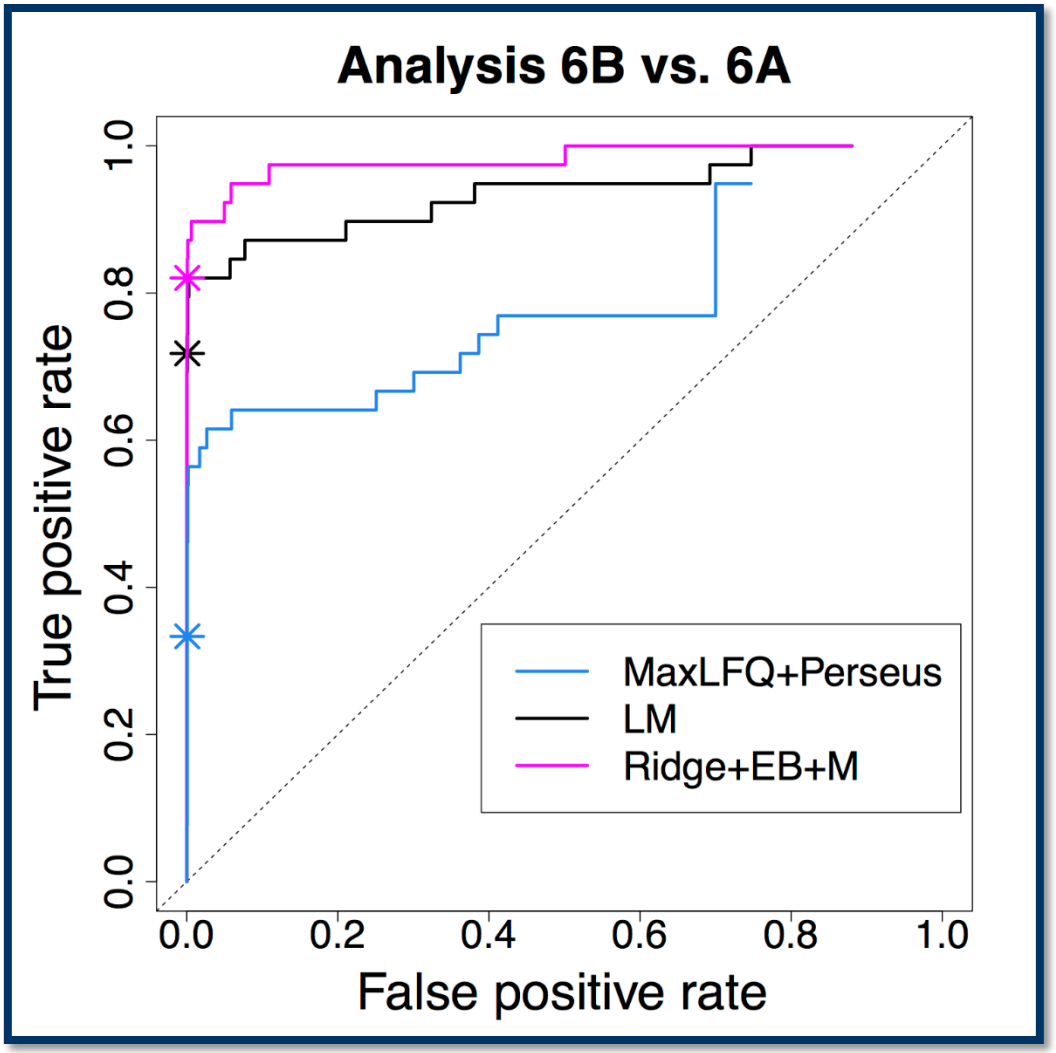
1. Better fold change estimates

For **differentially abundant proteins**



2. Better sensitivity and specificity





Conclusions

1. Use **peptide-based** models
2. Our peptide-based model uses:
    1. **Shrinkage estimation**
    2. **Empirical Bayes variance estimation**
    3. **Downweighing of outliers**
3. Advantages:
    1. **More stable fold change estimates**
    2. **Better sensititivity and specificity**

Papers

Goeminne et al. (2015), Summarization vs. Peptide-Based Models in Label-free Quantitative Proteomics: Performance, Pitfalls and Data Analysis Guidelines. *Journal of Proteome Research*.

Goeminne et al. (2015), Peptide-level robust ridge regression modeling with Empirical Bayes variance estimation and M estimation improves both sensitivity and specificity in quantitative label-free shotgun proteomics, *submitted*.

Acknowledgements: people

Lieven Clement
+ lab members

Kris Gevaert
+ lab members

Lennart Martens

Klaas Vandepoele

Andrea Argentini

Acknowledgements: organizations

Thank you for your attention!

Questions?