SCIENCE MEETS LIFE

MSqRob: analysis of label-free proteomics data in an R/Shiny environment

Ludger Goeminne

11/01/2017

Promotors:

Lieven Clement

Kris Gevaert

VIB

UNIVERSITEIT GENT

FACULTY OF SCIENCES

FACULTY OF MEDICINE AND HEALTH SCIENCES

AGENTSCHAP INNOVEREN & ONDERNEMEN

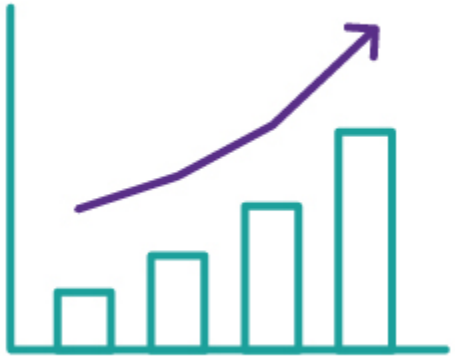
BIG N2N

Bioinformatics Institute Ghent

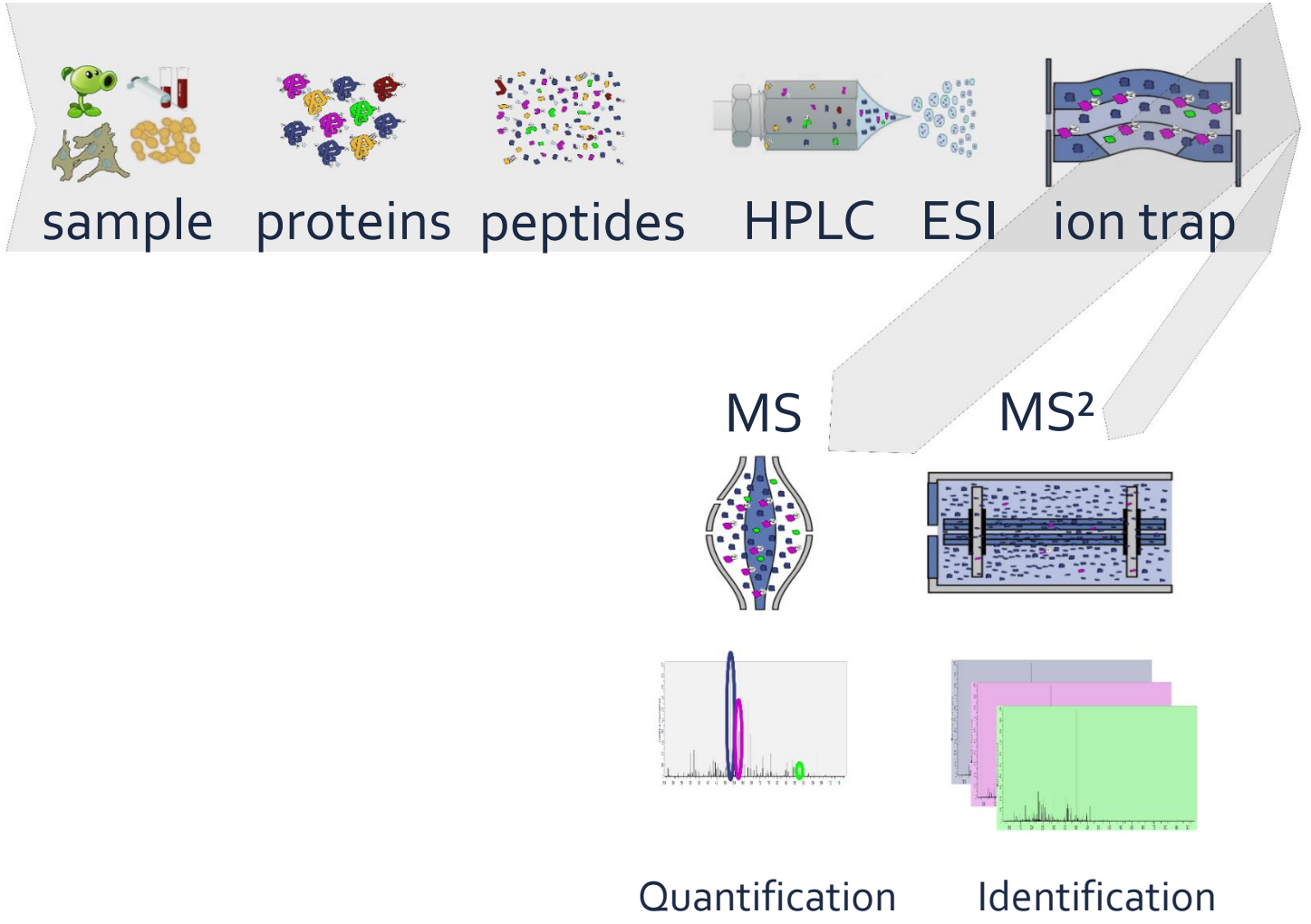
from nucleotides to networks

How often have you been stuck in the data analysis part?

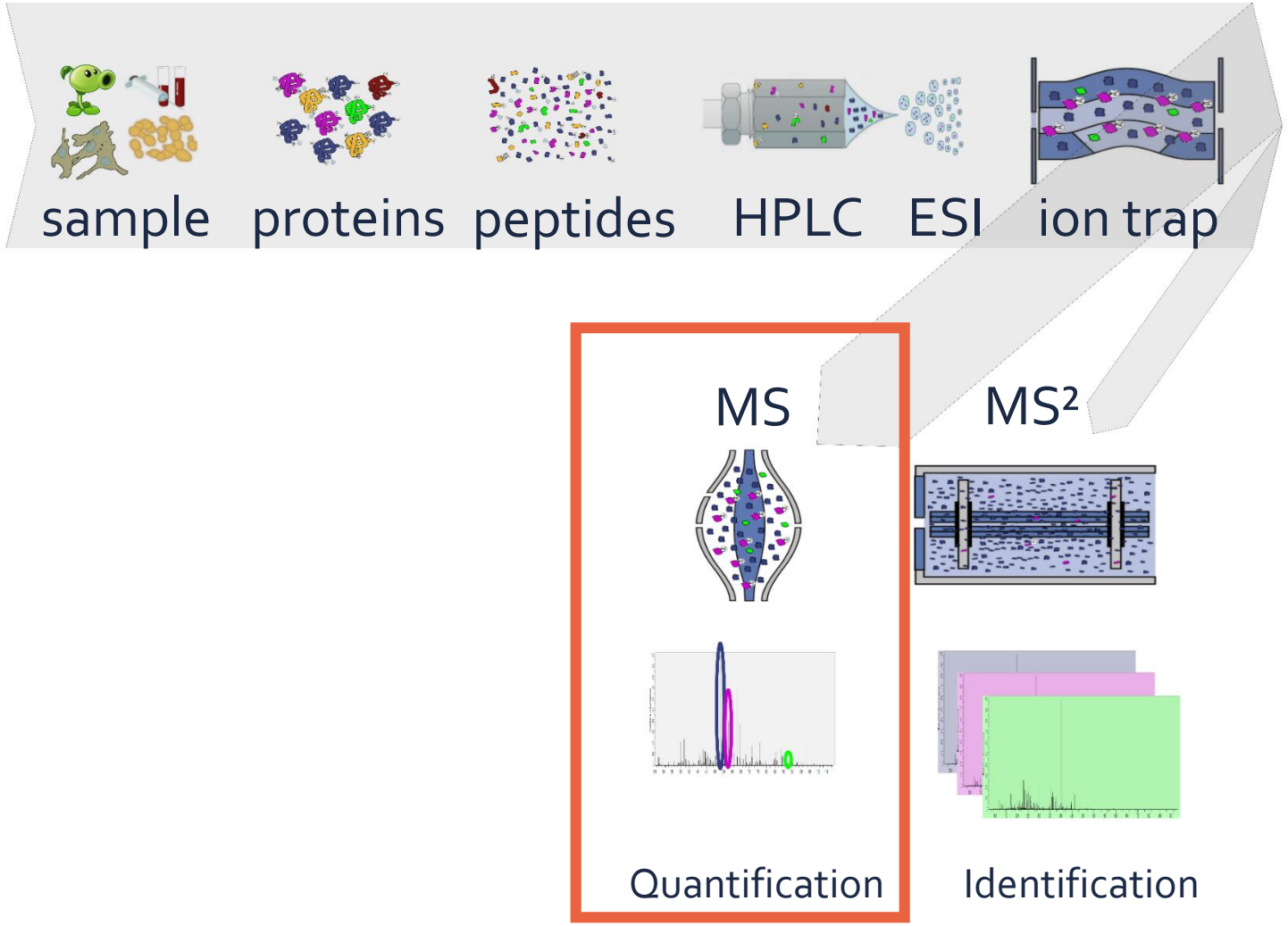
??????????



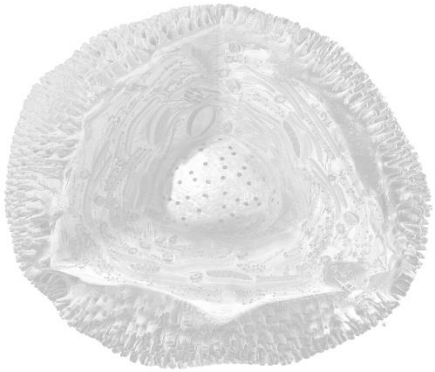
MS-based proteomics identifies many thousands of peptides

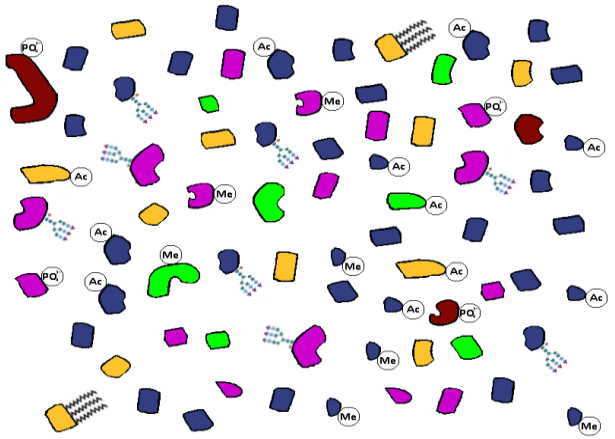


MS-based proteomics identifies many thousands of peptides

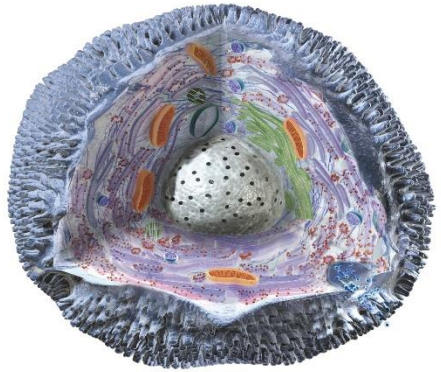


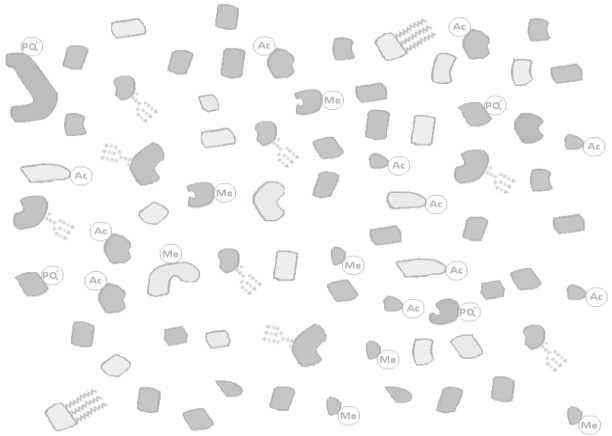
MS-based proteomics identifies peptides…





But we need protein-level information





Statistics and proteomics join forces

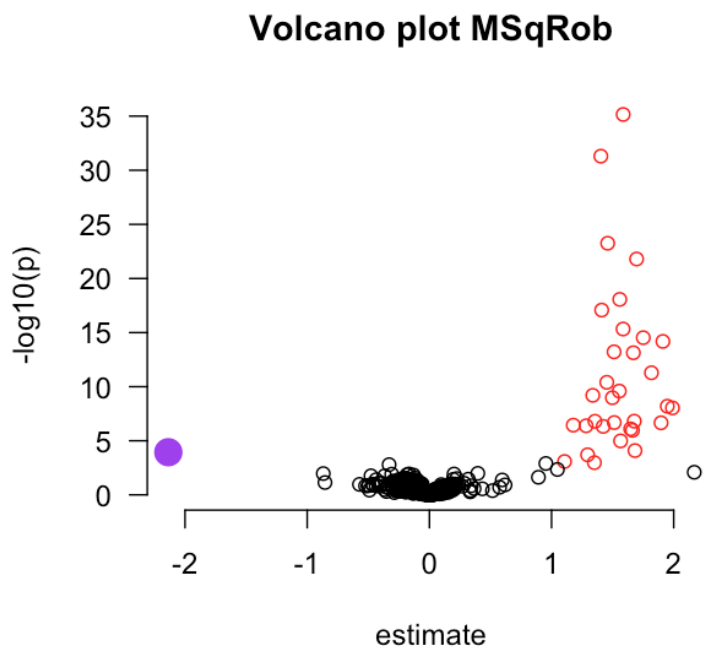
Lieven Clement: statistics

Kris Gevaert: proteomics

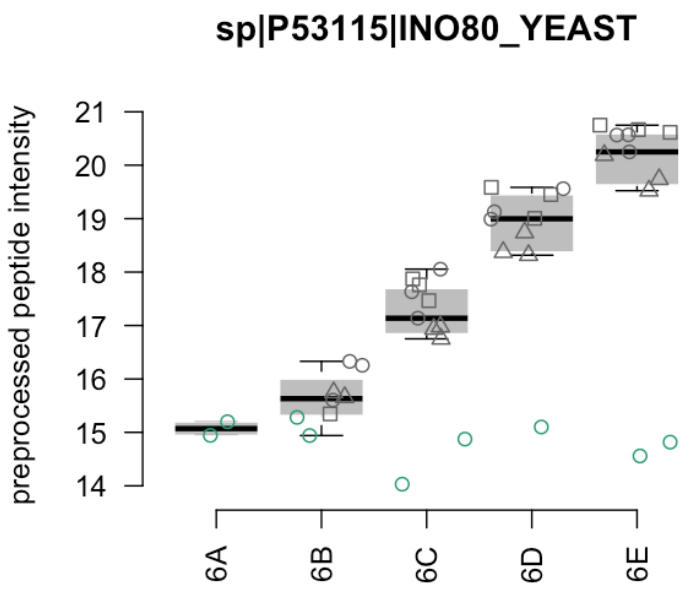
MSqRob
can solve your data-analysis problems

```
            Contrast 1
condition6A         -1
condition6B          1
```

Volcano plot



Detail plot



Analysis of label-free proteomics data with MSqRob

Performance: why it works so well

Features: how you can use MSqRob

Analysis of label-free proteomics data with MSqRob

Performance: why it works so well

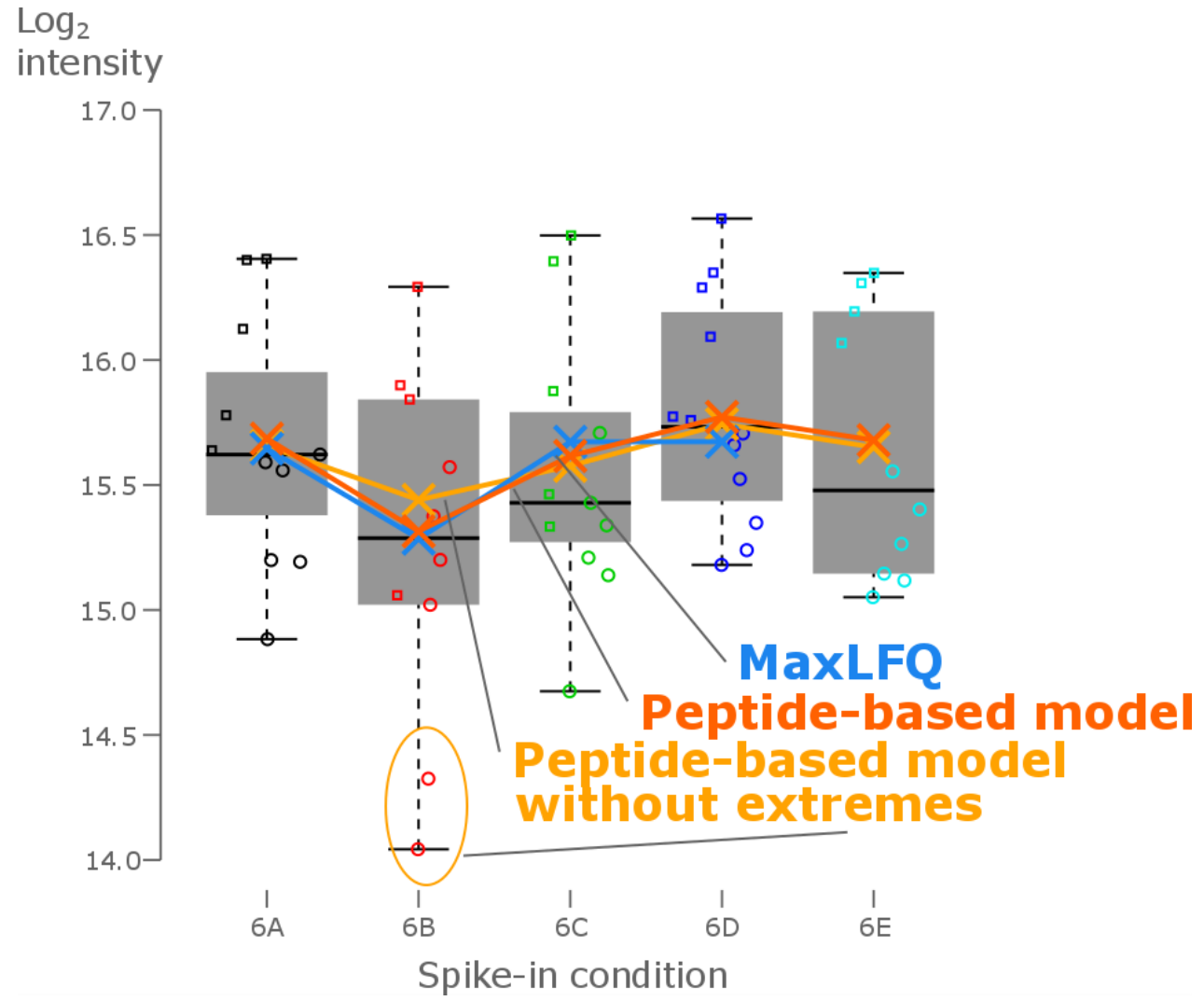
Features: how you can use MSqRob

Peptide-based models work better than summarization-based approaches

Identification	Preprocessing Normalization (+ summarization)	Inference Modeling + significance testing	Performance Mean relative pAUC
	MaxLFQ	t-test	68 %
	Quantile normalization No summarization	Peptide-based model	96 %

(Goeminne *et al.*, 2015, JPR)

Existing methods suffer from overfitting, unstable variances and outliers



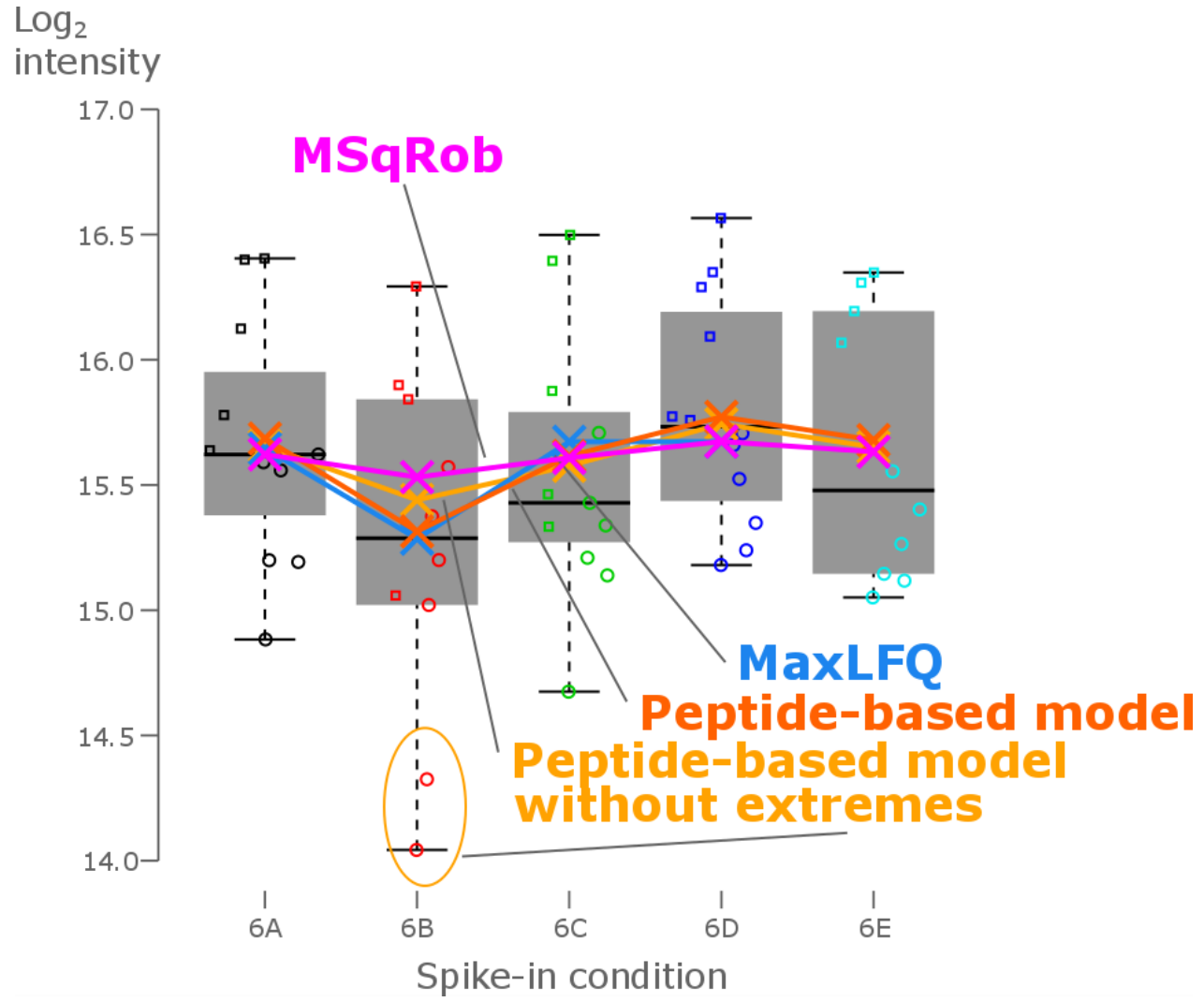
MSqRob adds 3 improvements to peptide-based models

Shrinkage estimation

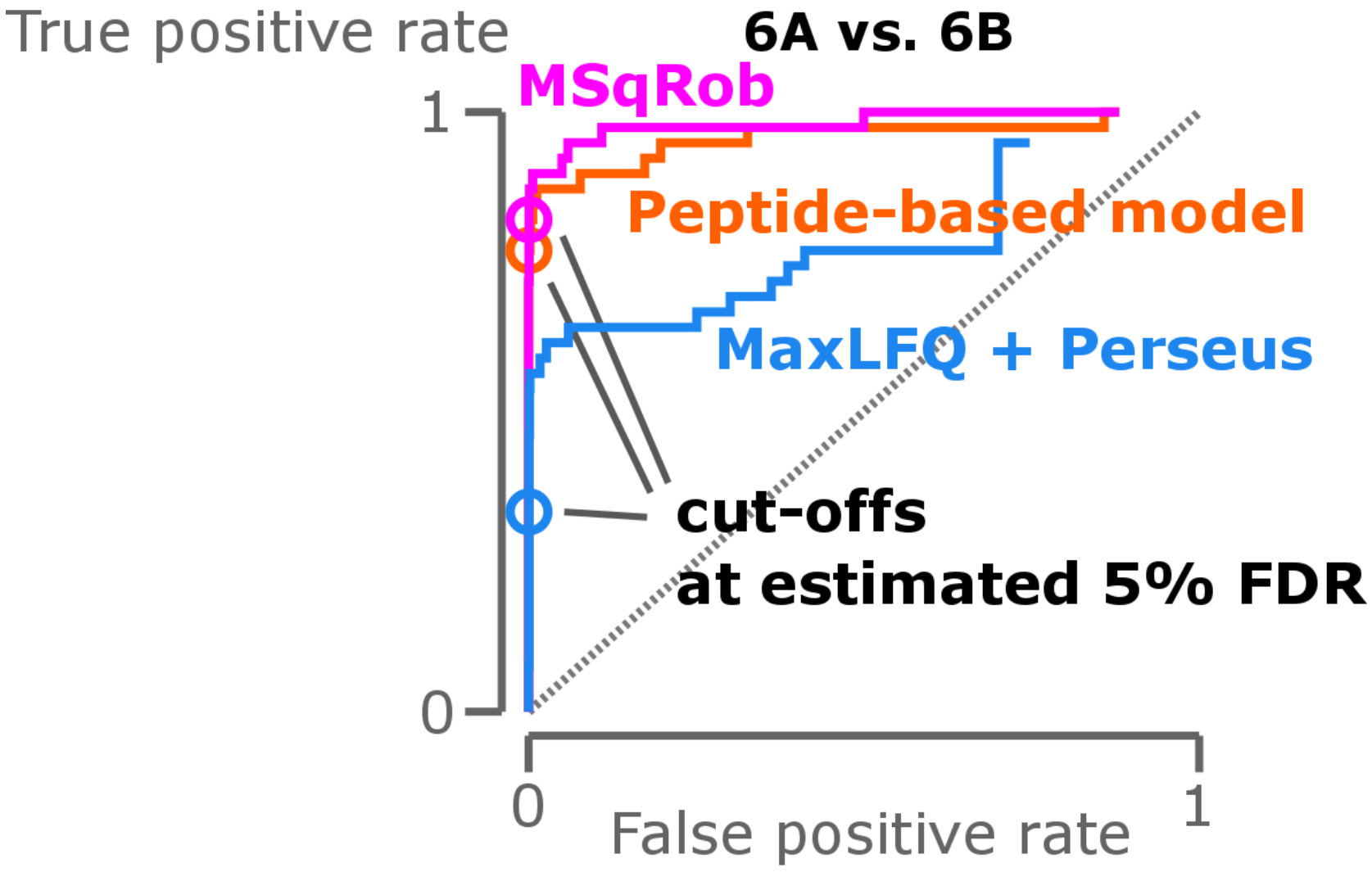
Borrowing information across proteins

Weighing down outliers

MSqRob provides more stable estimates



MSqRob outperforms other methods



Analysis of label-free proteomics data with MSqRob

Performance: why it works so well

Features: how you can use MSqRob

MSqRob handles data in a natural way

Fixed effects: genotype,treatment

Random effects: peptide, biological repeat, mass spec run

=> MSqRob can handle **complex designs**

Import from MaxQuant's peptides.txt

Project Name

project

Specify the location where your output will be saved

Browse... /Users/lgoeminn/Doc

Upload complete

Specify the location of your experimental annotation file

Browse... annotation.xlsx

Upload complete

Specify the location of your peptides.txt file

Browse... peptides.txt

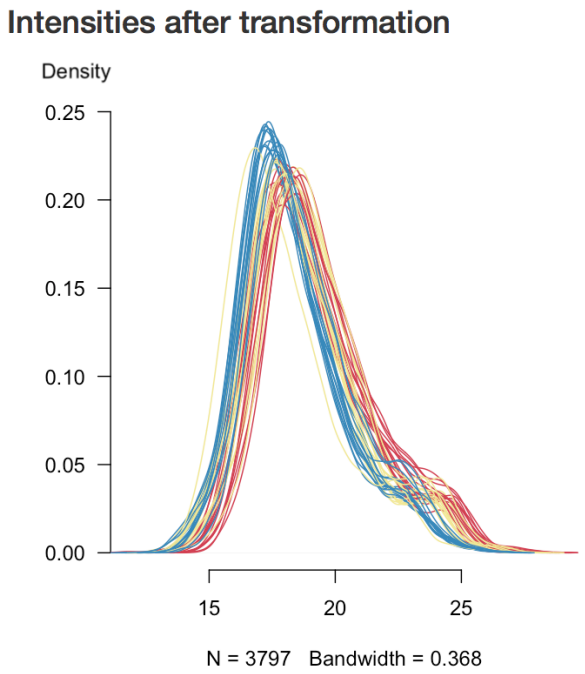
Upload complete

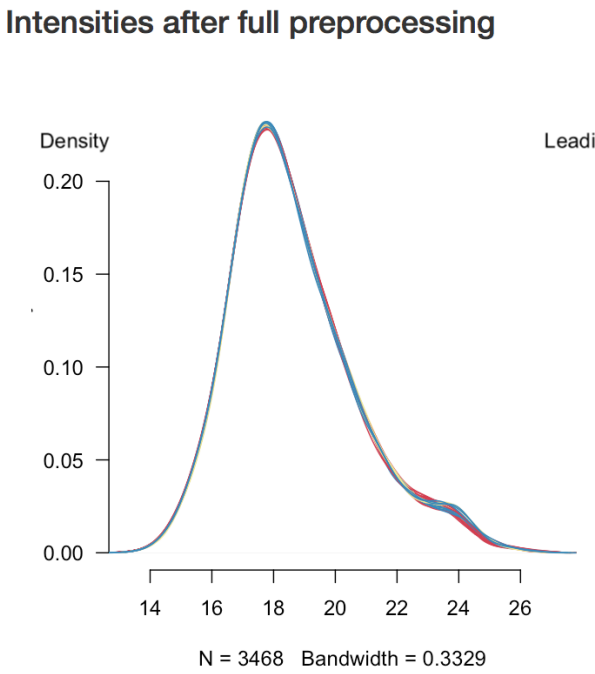
Add an annotation file

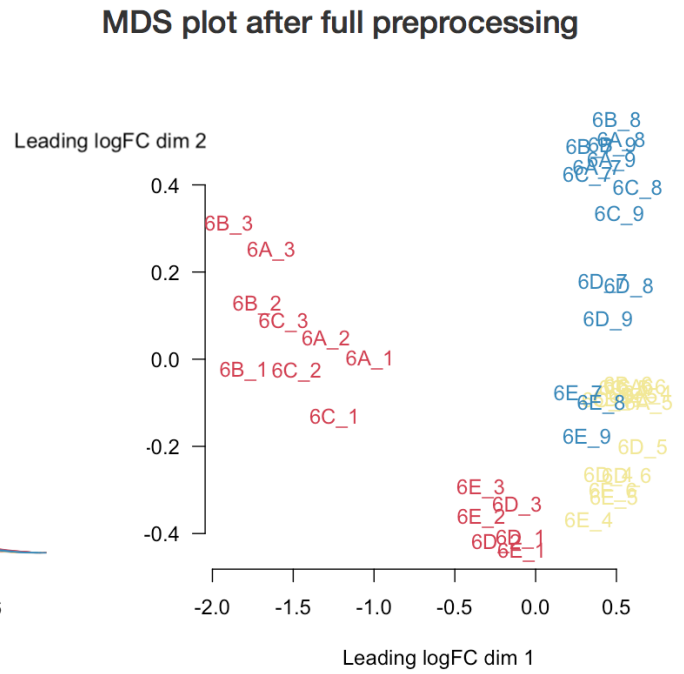
A1 | run

	A	B	C
1	**run**	**condition**	**lab**
2	6A_1	6A	LTQ-Orbitrap_86
3	6A_2	6A	LTQ-Orbitrap_86
4	6A_3	6A	LTQ-Orbitrap_86
5	6A_4	6A	LTQ-OrbitrapO_65
6	6A_5	6A	LTQ-OrbitrapO_65
7	6A_6	6A	LTQ-OrbitrapO_65
8	6A_7	6A	LTQ-OrbitrapW_56
9	6A_8	6A	LTQ-OrbitrapW_56
10	6A_9	6A	LTQ-OrbitrapW_56
11	6B_1	6B	LTQ-Orbitrap_86
12	6B_2	6B	LTQ-Orbitrap_86
13	6B_3	6B	LTQ-Orbitrap_86
14	6B_4	6B	LTQ-OrbitrapO_65
15	6B_5	6B	LTQ-OrbitrapO_65
16	6B_6	6B	LTQ-OrbitrapO_65
17	6B_7	6B	LTQ-OrbitrapW_56
18	6B_8	6B	LTQ-OrbitrapW_56
19	6B_9	6B	LTQ-OrbitrapW_56
20	6C_1	6C	LTQ-Orbitrap_86
21	6C_2	6C	LTQ-Orbitrap_86
22	6C_3	6C	LTQ-Orbitrap_86
23	6C_4	6C	LTQ-OrbitrapO_65
24	6C_5	6C	LTQ-OrbitrapO_65
25	6C_6	6C	LTQ-OrbitrapO_65
26	6C_7	6C	LTQ-OrbitrapW_56
27	6C_8	6C	LTQ-OrbitrapW_56
28	6C_9	6C	LTQ-OrbitrapW_56

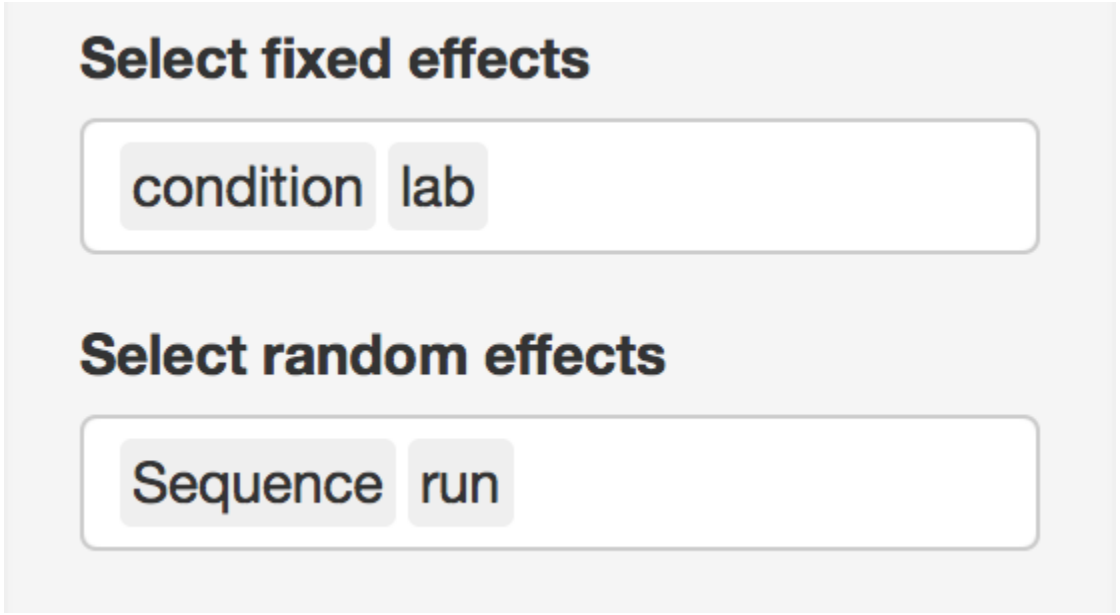
Preprocess your data



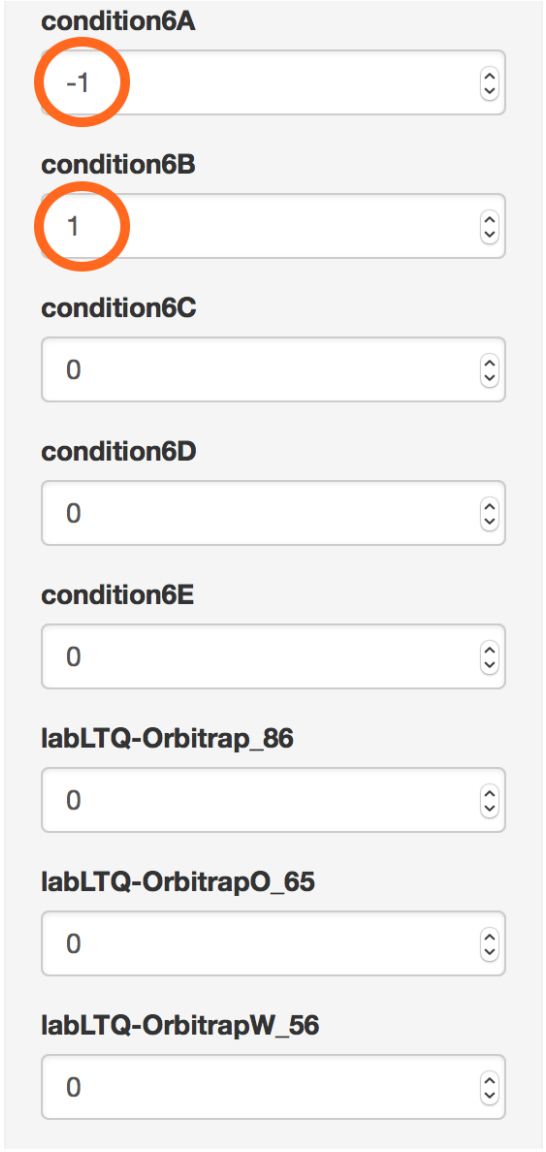




Select fixed and random effects



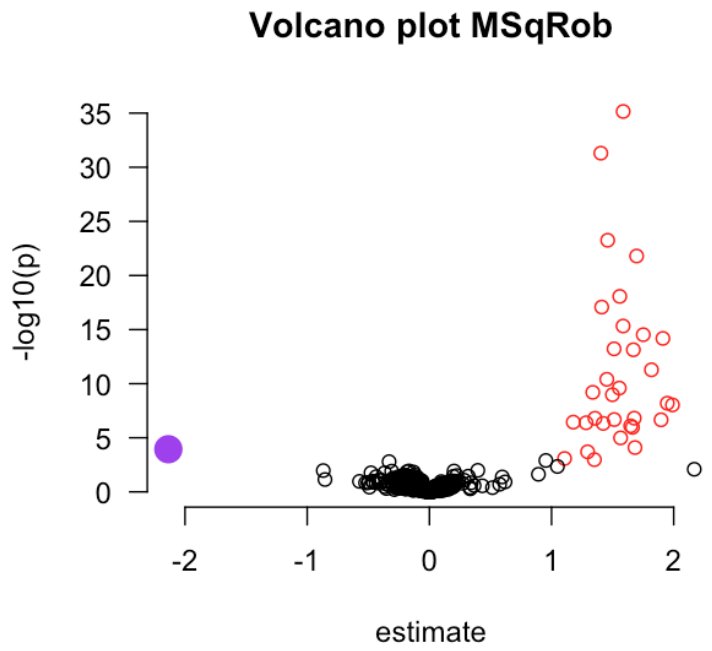
Test the appropriate research hypotheses



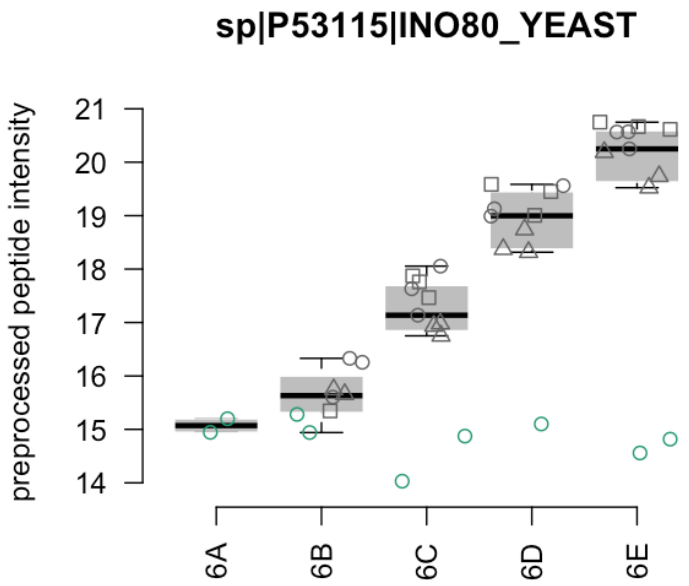
Inspect the results graphically

```
            Contrast 1
condition6A         -1
condition6B          1
```

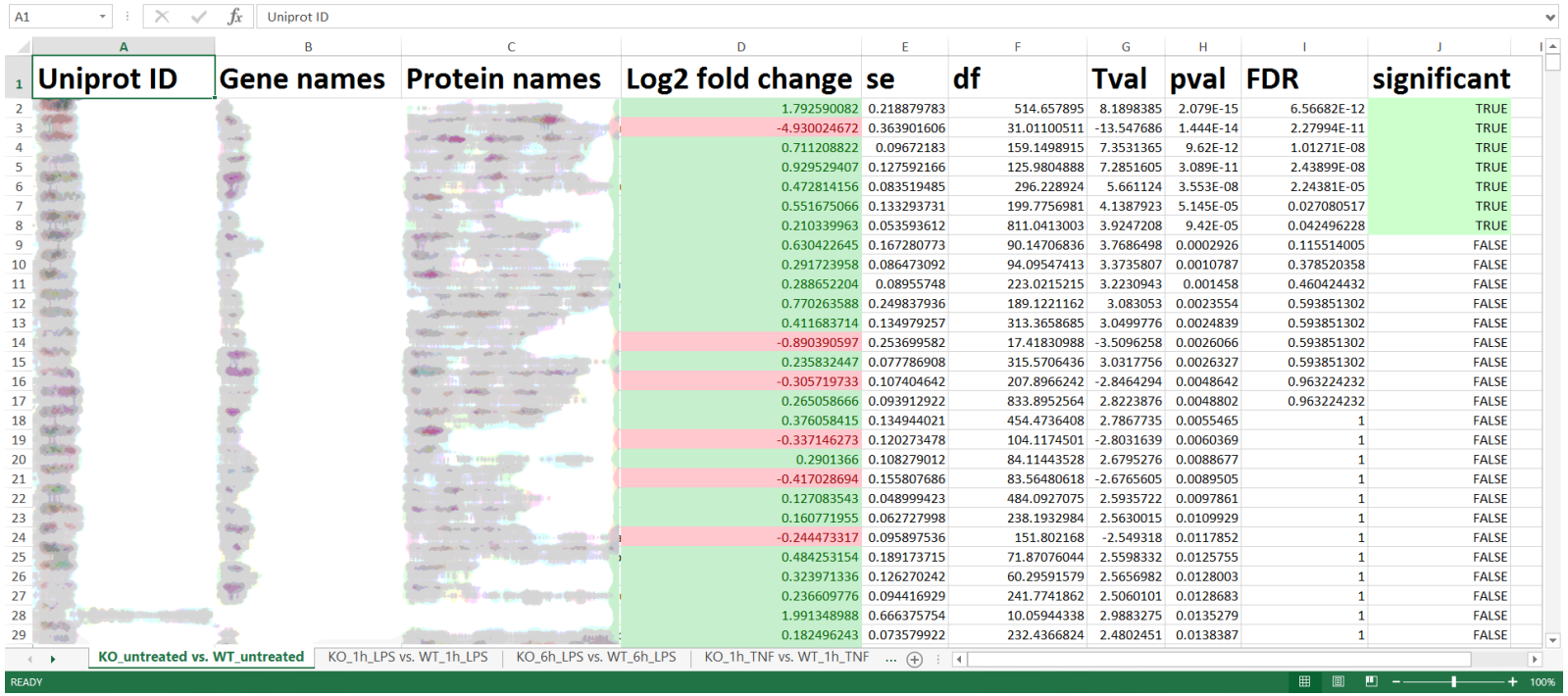
Volcano plot



Detail plot



Export the results to Excel



A1 | Uniprot ID

	A	B	C	D	E	F	G	H	I	J
1	Uniprot ID	Gene names	Protein names	Log2 fold change	se	df	Tval	pval	FDR	significant
2	[illegible]	[illegible]	[illegible]	1.792590082	0.218879783	514.657895	8.1898385	2.079E-15	6.56682E-12	TRUE
3	[illegible]	[illegible]	[illegible]	-4.930024672	0.363901606	31.01100511	-13.547686	1.444E-14	2.27994E-11	TRUE
4	[illegible]	[illegible]	[illegible]	0.711208822	0.09672183	159.1498915	7.3531365	9.62E-12	1.01271E-08	TRUE
5	[illegible]	[illegible]	[illegible]	0.929529407	0.127592166	125.9804888	7.2851605	3.089E-11	2.43899E-08	TRUE
6	[illegible]	[illegible]	[illegible]	0.472814156	0.083519485	296.228924	5.661124	3.553E-08	2.24381E-05	TRUE
7	[illegible]	[illegible]	[illegible]	0.551675066	0.133293731	199.7756981	4.1387923	5.145E-05	0.027080517	TRUE
8	[illegible]	[illegible]	[illegible]	0.210339963	0.053593612	811.0413003	3.9247208	9.42E-05	0.042496228	TRUE
9	[illegible]	[illegible]	[illegible]	0.630422645	0.167280773	90.14706836	3.7686498	0.0002926	0.115514005	FALSE
10	[illegible]	[illegible]	[illegible]	0.291723958	0.086473092	94.09547413	3.3735807	0.0010787	0.378520358	FALSE
11	[illegible]	[illegible]	[illegible]	0.288652204	0.08955748	223.0215215	3.2230943	0.001458	0.460424432	FALSE
12	[illegible]	[illegible]	[illegible]	0.770263588	0.249837936	189.1221162	3.083053	0.0023554	0.593851302	FALSE
13	[illegible]	[illegible]	[illegible]	0.411683714	0.134979257	313.3658685	3.0499776	0.0024839	0.593851302	FALSE
14	[illegible]	[illegible]	[illegible]	-0.890390597	0.253699582	17.41830988	-3.5096258	0.0026066	0.593851302	FALSE
15	[illegible]	[illegible]	[illegible]	0.235832447	0.077786908	315.5706436	3.0317756	0.0026327	0.593851302	FALSE
16	[illegible]	[illegible]	[illegible]	-0.305719733	0.107404642	207.8966242	-2.8464294	0.0048642	0.963224232	FALSE
17	[illegible]	[illegible]	[illegible]	0.265058666	0.093912922	833.8952564	2.8223876	0.0048802	0.963224232	FALSE
18	[illegible]	[illegible]	[illegible]	0.376058415	0.134944021	454.4736408	2.7867735	0.0055465	1	FALSE
19	[illegible]	[illegible]	[illegible]	-0.337146273	0.120273478	104.1174501	-2.8031639	0.0060369	1	FALSE
20	[illegible]	[illegible]	[illegible]	0.2901366	0.108279012	84.11443528	2.6795276	0.0088677	1	FALSE
21	[illegible]	[illegible]	[illegible]	-0.417028694	0.155807686	83.56480618	-2.6765605	0.0089505	1	FALSE
22	[illegible]	[illegible]	[illegible]	0.127083543	0.048999423	484.0927075	2.5935722	0.0097861	1	FALSE
23	[illegible]	[illegible]	[illegible]	0.160771955	0.062727998	238.1932984	2.5630015	0.0109929	1	FALSE
24	[illegible]	[illegible]	[illegible]	-0.244473317	0.095897536	151.802168	-2.549318	0.0117852	1	FALSE
25	[illegible]	[illegible]	[illegible]	0.484253154	0.189173715	71.87076044	2.5598332	0.0125755	1	FALSE
26	[illegible]	[illegible]	[illegible]	0.323971336	0.126270242	60.29591579	2.5656982	0.0128003	1	FALSE
27	[illegible]	[illegible]	[illegible]	0.236609776	0.094416929	241.7741862	2.5060101	0.0128683	1	FALSE
28	[illegible]	[illegible]	[illegible]	1.991348988	0.666375754	10.05944338	2.9883275	0.0135279	1	FALSE
29	[illegible]	[illegible]	[illegible]	0.182496243	0.073579922	232.4366824	2.4802451	0.0138387	1	FALSE

KO_untreated vs. WT_untreated | KO_1h_LPS vs. WT_1h_LPS | KO_6h_LPS vs. WT_6h_LPS | KO_1h_TNF vs. WT_1h_TNF ...

READY 100%

Download MSqRob from GitHub

https://github.com/ludgergoeminne/MSqRob

Goeminne, L.J.E., Gevaert, K. and Clement, L. **Peptide-level robust regression improves estimation, sensitivity and specificity in data-dependent quantitative label-free shotgun proteomics**. Molecular and Cellular Proteomics 15(2), pp 567-668.

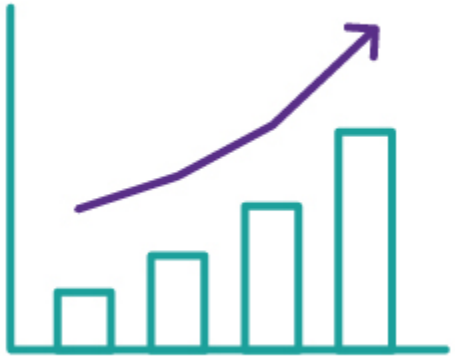
Analysis of label-free proteomics data with MSqRob

Performance: why it works so well

Features: how you can use MSqRob

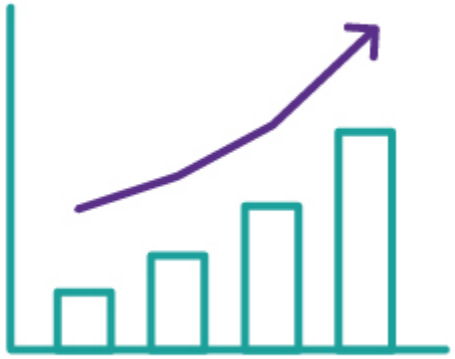
How often have you been stuck in the data analysis part?

??????????



MSqRob can solve your problem

MSqRob



Visit our posters!

ludger.goeminne@vib-ugent.be

Visit our posters!

ludger.goeminne@vib-ugent.be

MSqRob: analysis of label-free proteomics data in an R/Shiny environment

Ludger Goeminne[1,2,3,4*], Kris Gevaert[1,2,4], Lieven Clement[3,4]

[1]Medical Biotechnology Center, VIB, Ghent, Belgium; [2]Department of Biochemistry; [3]Department of Applied Mathematics, Computer Science and Statistics; [4]Bioinformatics Institute Gent. BIG-N2N, Ghent University, Belgium. *Contact: ludger.goeminne@vib-ugent.be

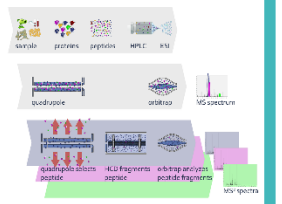
In MS-based proteomics, proteins are not completely covered and peptides that are identified in one sample are often missing in other samples. Common workflows adopt software tools that have graphical user interfaces, but are often based on less sensitive protein level abundance values and/or provide inefficient or even inappropriate statistical inference.

MSqRob is an R package that accounts for peptide-specific effects as well as differences in the number of peptide identifications. It copes with overfitting, unstable variances and outliers by three modular extensions: (1) ridge regression, (2) empirical Bayes variance estimation and (3) M-estimation. MSqRob provides state-of-the-art statistical inference for label-free proteomics experiments with simple and complex designs: MSqRob can cope with multifactorial, block, repeated measures and time series designs, which cannot be analyzed properly in existing proteomics data analysis software. The Shiny graphical user interface for MSqRob is very user-friendly and requires no statistical programming experience.

Goeminne, L.J.E., Gevaert, K. and Clement, L. Molecular and Cellular Proteomics 15(2), pp 567-668.

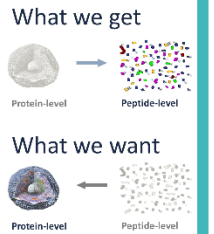
Download MSqRob: https://github.com/ludgergoeminne/MSqRob

Proteomics workflow



Proteins are extracted from a sample, digested into peptides, separated on a column and ionized. An MS spectrum represents a peptide's abundance, MS^2 spectra allow for peptide identifications.

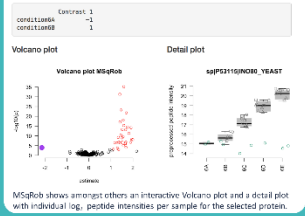
Problem



MSqRob workflow

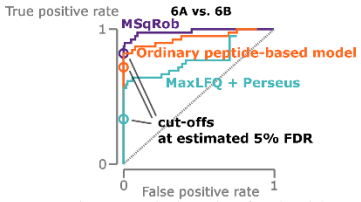
- Import data (MaxQuant or Mascot)
- Preprocessing (e.g. $\log_2$ transformation, normalization, ...)
- Robust peptide-based regression
- Test contrasts of interest (FDR-corrected p-values)

Results



MSqRob shows amongst others an interactive Volcano plot and a detail plot with individual $\log_2$ peptide intensities per sample for the selected protein.

ROC curves



Receiver operating characteristic curves show that MSqRob outperforms other methods.

Conclusions

- MSqRob is user-friendly
- MSqRob provides stable fold change estimates
- MSqRob provides has a better sensitivity and specificity

Analyzing repeated measures designs in label-free proteomics with MSqRob (MCP 2016 15(2):657-68.)

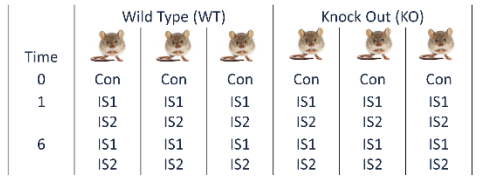
Lieven Clement[3,4,*], Ludger Goeminne[1,2,3,4], Emmy Van Quickelberghe[1,2,4] & Kris Gevaert[1,2,4]

[1]Medical Biotechnology Center, VIB, Ghent, Belgium; [2]Department of Biochemistry; [3]Department of Applied Mathematics, Computer Science and Statistics; [4]Bioinformatics institute Ghent, BIG -N2N, Ghent University, Belgium. *Contact: lieven.clement@UGent.be

Background: In repeated measures designs different observations are obtained on the same experimental unit (EU), which increases statistical power for within subject treatment effects because the between-subject variability can be eliminated from the estimation. Data of the same EU, however, are typically more similar than data between EUs. Most existing workflows cannot address experiments with complex designs and correlation, resulting in a power loss when assessing treatment effects within EU (e.g. compound effects) and improper error rate control for effects between EU (e.g. KO vs WT).

Repeated Measures Design

Baseline control, early and late responses on inflammatory stimuli (IS)



MSqRob workflow

- Import data (MaxQuant or Mascot)
- Preprocessing (e.g. $\log_2$ transformation, normalization, ...)
- Robust peptide-based regression
- Test contrasts of interest (FDR-corrected p-values)

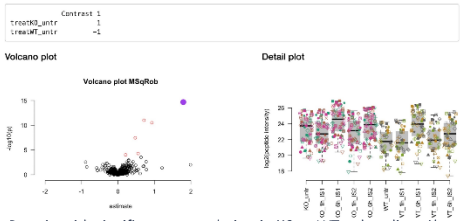
Model

$$\log_2 \text{Intensity} = G \times IS \times T + (1|\text{mouse}) + (1|\text{run}) + (1|\text{Peptide}) + \varepsilon$$

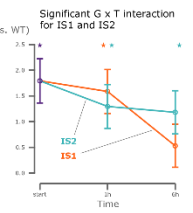
$\log_2$ peptide intensity modeled by genotype (G), Inflammatory Stimulus (IS) & time (T) main effects & interactions + random effects for mouse, run and peptide to address correlation. Normal error.

Results

MSqRob GUI Output



Additional Plot



Protein with significant upregulation in KO vs WT at baseline, 1h upon treatment with IS1 stimulus and 1h and 6h with IS2, and a significant interaction, i.e. the upregulation in KO decreases over time.

Conclusion

- Powerful analysis of complex designs with fixed and random effects
- Robust estimation and shrinkage of fixed effects
- Data exploration and visualisation
- **Download MSqRob package:** https://github.com/ludgergoeminne/MSqRob

Visit our posters!

ludger.goeminne@vib-ugent.be

Lieven Clement

Kris Gevaert

VIB
science meets life

UNIVERSITEIT GENT

BIG N2N
Bioinformatics Institute Ghent
from nucleotides to networks

AGENTSCHAP INNOVEREN & ONDERNEMEN

Vlaanderen
is ondernemen

MSqRob: analysis of label-free proteomics data in an R/Shiny environment

Ludger Goeminne[1,2,3,4*], Kris Gevaert[1,2,4], Lieven Clement[3,4]
[1]Medical Biotechnology Center, VIB, Ghent, Belgium; [2]Department of Biochemistry; [3]Department of Applied Mathematics, Computer Science and Statistics; [4]Bioinformatics Institute Gent. BIG-N2N, Ghent University, Belgium. *Contact: ludger.goeminne@vib-ugent.be

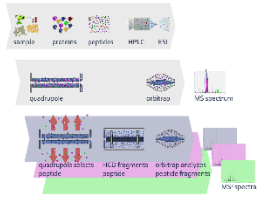
In MS-based proteomics, proteins are not completely covered and peptides that are identified in one sample are often missing in other samples. Common workflows adopt software tools that have graphical user interfaces, but are often based on less sensitive protein level abundance values and/or provide inefficient or even inappropriate statistical inference.

MSqRob is an R package that accounts for peptide-specific effects as well as differences in the number of peptide identifications. It copes with overfitting, unstable variances and outliers by three modular extensions: (1) ridge regression, (2) empirical Bayes variance estimation and (3) M-estimation. MSqRob provides state-of-the-art statistical inference for label-free proteomics experiments with simple and complex designs: MSqRob can cope with multifactorial, block, repeated measures and time series designs, which cannot be analyzed properly in existing proteomics data analysis software. The Shiny graphical user interface for MSqRob is very user-friendly and requires no statistical programming experience.

Goeminne, L.J.E., Gevaert, K. and Clement, L. Molecular and Cellular Proteomics 15(2), pp 567-668.

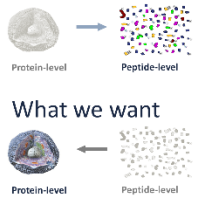
Download MSqRob: https://github.com/ludgergoeminne/MSqRob

Proteomics workflow



Proteins are extracted from a sample, digested into peptides, separated on a column and ionized. An MS spectrum represents a peptide's abundance. MS^2 spectra allow for peptide identifications.

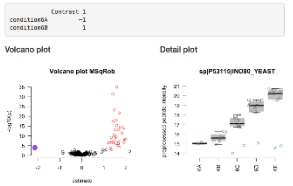
Problem



MSqRob workflow

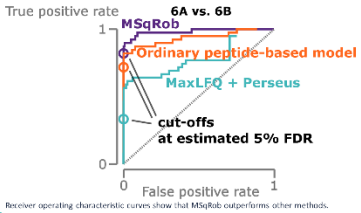
- Import data (MaxQuant or Mascot)
- Preprocessing (e.g. $\log_2$ transformation, normalization, ...)
- Robust peptide-based regression
- Test contrasts of interest (FDR-corrected p-values)

Results



MSqRob shows amongst others an interactive Volcano plot and a detail plot with individual $\log_2$ peptide intensities per sample for the selected protein.

ROC curves



Receiver operating characteristic curves show that MSqRob outperforms other methods.

Conclusions

- MSqRob is user-friendly
- MSqRob provides stable fold change estimates
- MSqRob provides has a better sensitivity and specificity

Analyzing repeated measures designs in label-free proteomics with MSqRob (MCP 2016 15(2):657-68.)

Lieven Clement[3,4,*], Ludger Goeminne[1,2,3,4], Emmy Van Quickelberghe[1,2,4] & Kris Gevaert[1,2,4]
[1]Medical Biotechnology Center, VIB, Ghent, Belgium; [2]Department of Biochemistry; [3]Department of Applied Mathematics, Computer Science and Statistics; [4]Bioinformatics institute Ghent, BIG -N2N, Ghent University, Belgium. *Contact: lieven.clement@UGent.be

Background: In repeated measures designs different observations are obtained on the same experimental unit (EU), which increases statistical power for within subject treatment effects because the between-subject variability can be eliminated from the estimation. Data of the same EU, however, are typically more similar than data between EUs. Most existing workflows cannot address experiments with complex designs and correlation, resulting in a power loss when assessing treatment effects within EU (e.g. compound effects) and improper error rate control for effects between EU (e.g. KO vs WT).

Repeated Measures Design

Baseline control, early and late responses on inflammatory stimuli (IS)

	Wild Type (WT)			Knock Out (KO)		
Time						
0	Con	Con	Con	Con	Con	Con
1	IS1	IS1	IS1	IS1	IS1	IS1
	IS2	IS2	IS2	IS2	IS2	IS2
6	IS1	IS1	IS1	IS1	IS1	IS1
	IS2	IS2	IS2	IS2	IS2	IS2

MSqRob workflow

- Import data (MaxQuant or Mascot)
- Preprocessing (e.g. $\log_2$ transformation, normalization, ...)
- Robust peptide-based regression
- Test contrasts of interest (FDR-corrected p-values)

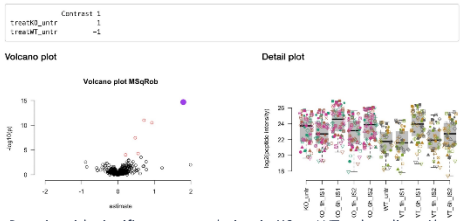
Model

$$\log_2 \text{Intensity} = G \times IS \times T + (1|\text{mouse}) + (1|\text{run}) + (1|\text{Peptide}) + \varepsilon$$

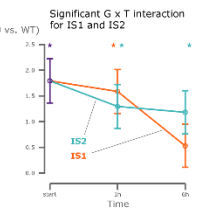
$\log_2$ peptide intensity modeled by genotype (G), Inflammatory Stimulus (IS) & time (T) main effects & interactions + random effects for mouse, run and peptide to address correlation. Normal error.

Results

MSqRob GUI Output



Additional Plot



Protein with significant upregulation in KO vs WT at baseline, 1h upon treatment with IS1 stimulus and 1h and 6h with IS2, and a significant interaction, i.e. the upregulation in KO decreases over time.

Conclusion

- Powerful analysis of complex designs with fixed and random effects
- Robust estimation and shrinkage of fixed effects
- Data exploration and visualisation
- **Download MSqRob package:** https://github.com/ludgergoeminne/MSqRob