Statistical Analysis with MSstats2

Meena Choi

Purdue University

2013.3.10

Overview

1. R packages : MSstats2 and SRMstats
2. Default analysis of a label-based SRM experiment (Human Plasma : Ovarian Cancer)
    1. Whole conceptual analysis
    2. How to analyze in R
    3. How to analyze in Skyline
3. A study of the importance of the quality of peaks
4. Another example of a label-free SRM (Rat plasma)
    1. Normalization
    2. A study of poor quality or inconsistent peptides

MSstats2 and SRMstats

Label-free shotgun MS

Label-free & label-based SRM

Label-free & label-based LC-MS & SRM

What we can do :

1. Test proteins for differential abundance
2. Quantify proteins in biological samples
3. Design of experiment

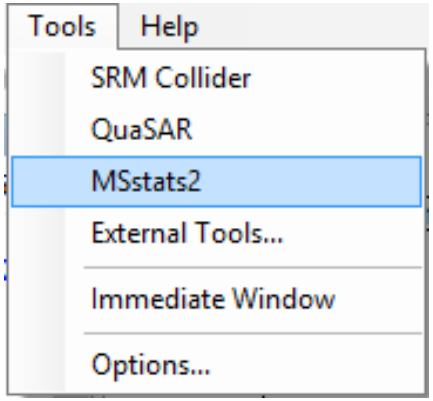
- Download : http://www.stat.purdue.edu/~ovitek/Software.html

  (MSstats2 is under evaluation for Bioconductor and integrated with Skyline.)

- Contact : Meena Choi (choi67@purdue.edu),

  Ching-Yun(Veavi) Chang(chang54@purdue.edu)

MSstats2 with Skyline



- Use as an external tool
- Automatically run the functions for
  - Preprocessing the data
  - Comparing between group
  - Calculating the sample size
  - Drawing the plots related with
- For the beginner of R, we can do statistical analysis with default options through Skyline easily.

- Use R-based platform if you want the detailed options for all functions such as,
  - Normalization
  - Detailed options for all plots
  - The number of peptides, transitions or power calculation
  - Quantification for sample
- With R-based platform, we can take advantage of options and modify the data easily.

How to start

- Required package
  - gplots, lme4, lattice, limma, marray
  - Need to install the required package once. Then they will be loaded automatically with MSstats2 or SRMstats.

- Installation
  - Select ‘packages’ in toolbar and then ‘Install package(s)’ in dropdown option.
  - Or use ‘install. packages’ function. (see R script example)

- See tutorial document for all detailed of running SRMstats and analysis through Skyline.

Two Example Datasets

Human Plasma : Ovarian Cancer

- OV (66) vs Control (15)
- No technical replicate
- Total 81 injections (Runs)
- Labeled SRM
- Good quality

Rat Plasma : Risk of heart disease

- High salt (7) vs. Low salt (7)
- 3 Technical replicates
- Total 42 injections (Runs)
- Label-free SRM
- Truncated peaks

What we can do with MSstats2?

- Ovarian Cancer data : label-based SRM
  - Initial data processing and visualization ('profile plot', 'QC plot')
  - Group comparison with several options and 'volcano plots'
  - Sample size calculation plot
  - The effect of poor quality features for statistical analysis
  - Label-based vs. Label-free SRM analysis

- Rat data : label-free SRM
  - The effect of poor quality features and inconsistent peptides

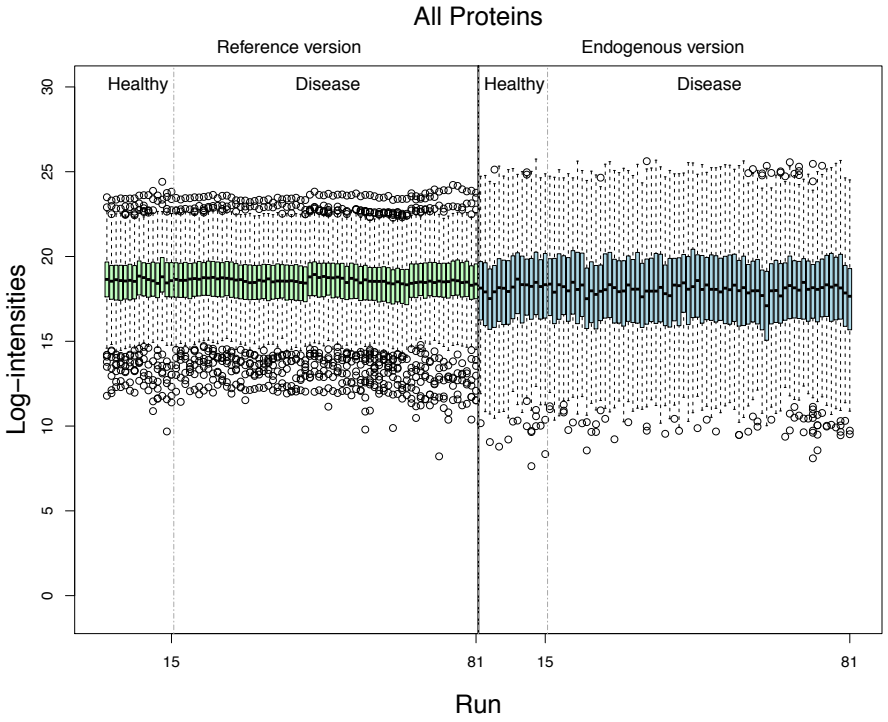
Overview

1. R packages : MSstats2 and SRMstats
2. Default analysis of a label-based SRM experiment (Human Plasma : Ovarian Cancer)
    1. Whole conceptual analysis
    2. How to analyze in R
    3. How to analyze in Skyline
3. A study of the importance of the quality of peaks
4. Another example of a label-free SRM (Rat plasma)
    1. Normalization
    2. A study of poor quality or inconsistent peptides

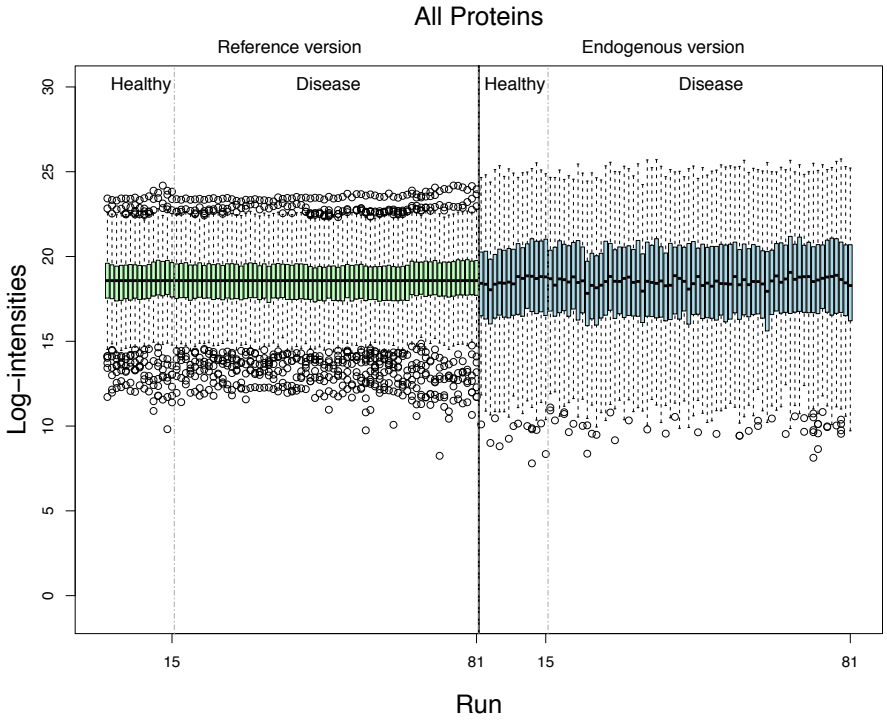
Quality Control plot

- Show the systematic bias between MS runs
- Constant Normalization

Before Normalization



After Normalization



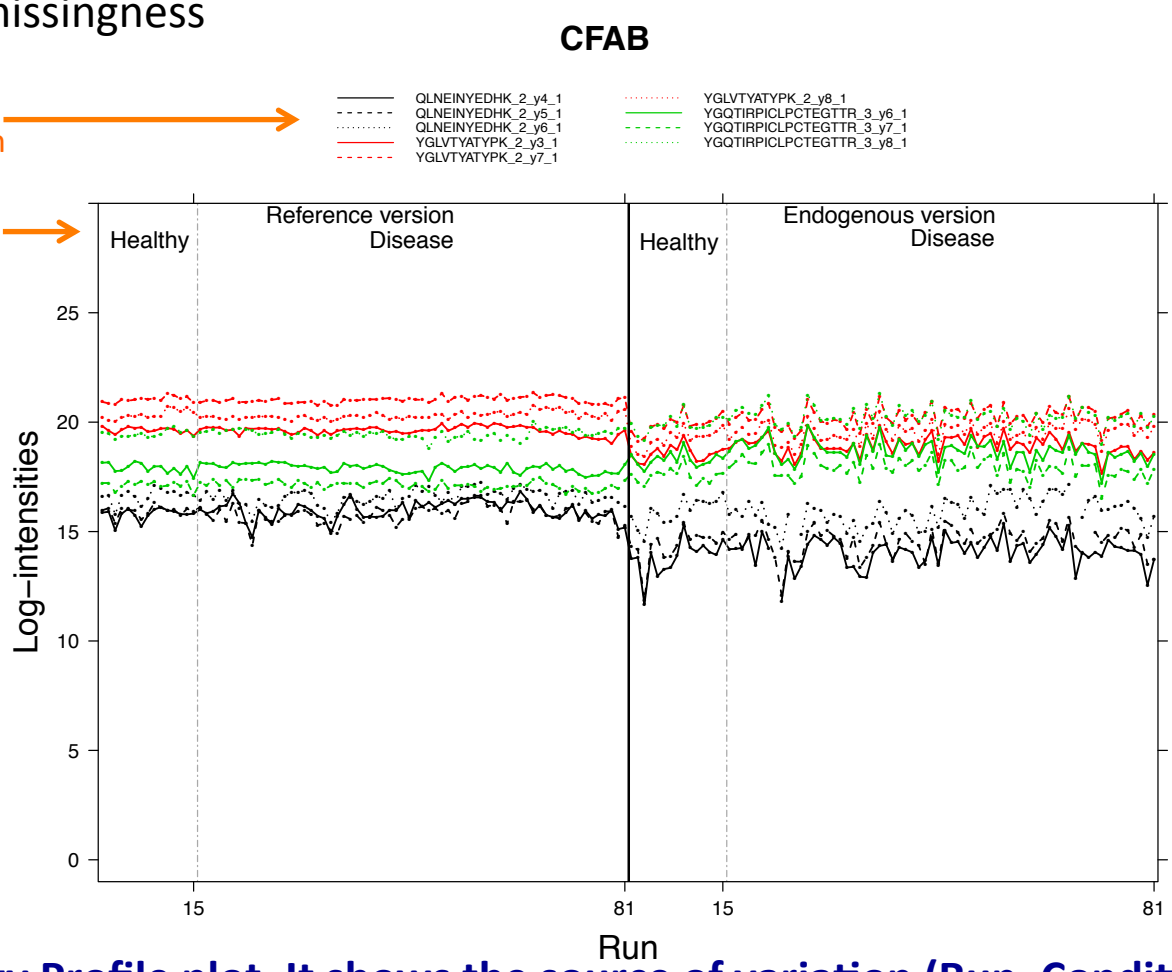
After normalization, the reference signals for all proteins are stable across MS runs.

Profile plot

- Show the potential source of variation, such as Run, Transition, Condition
- Check missingness

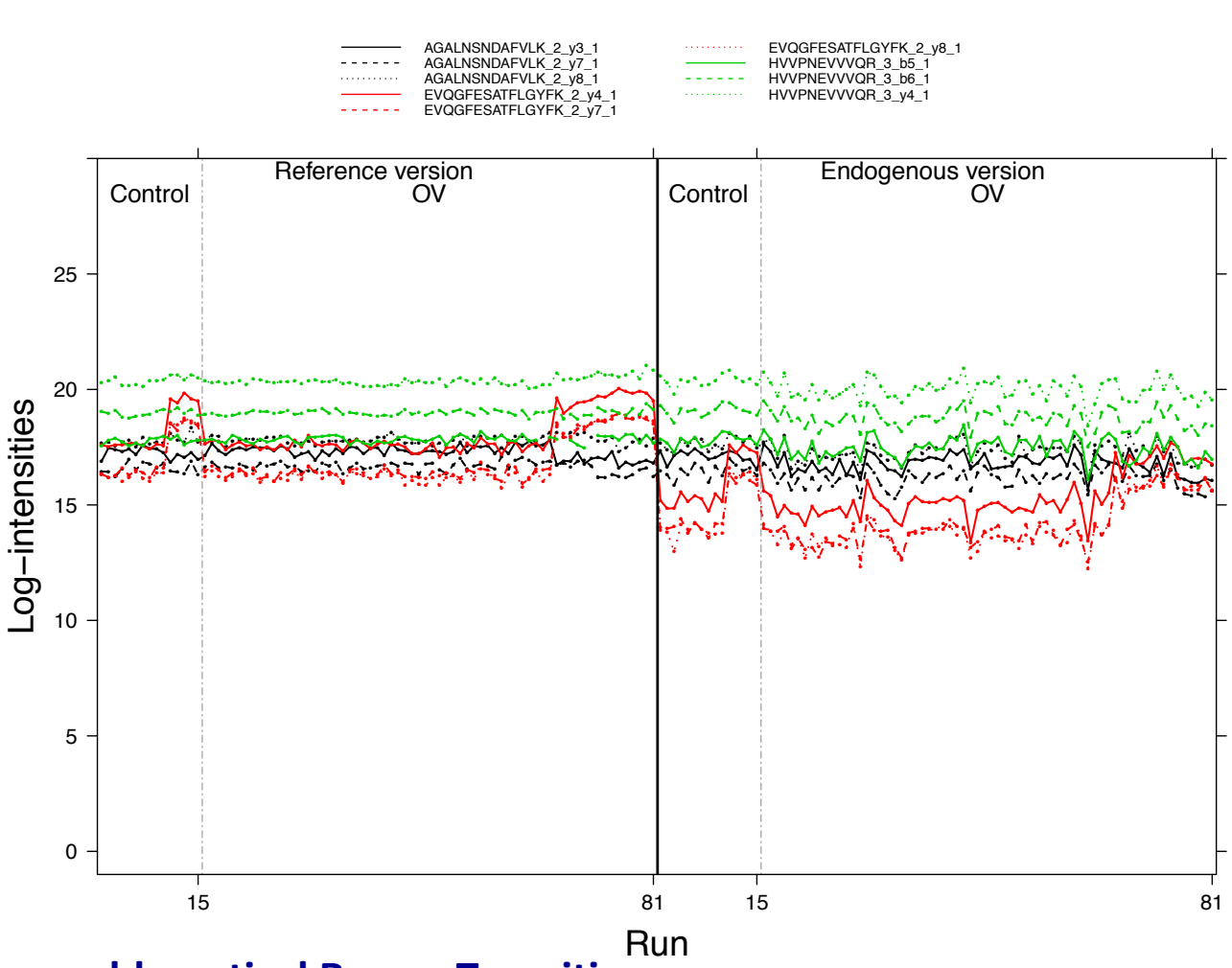
Color : peptide
Linetype : transition

Reference or not, Condition



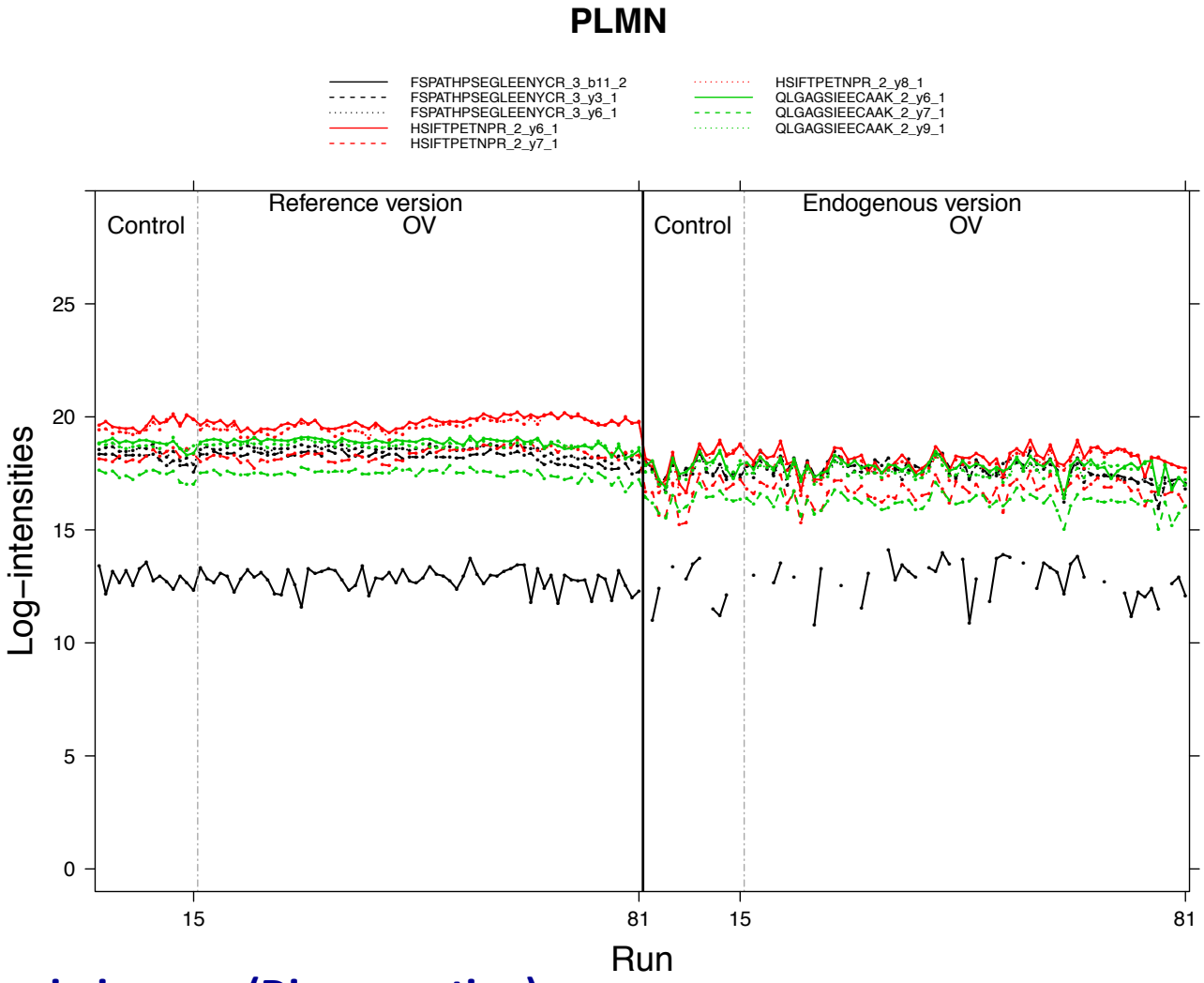
Good quality Profile plot. It shows the source of variation (Run, Condition, Transition)

Profile plot



Detect the problematical Run or Transition

Profile plot



Show the missingness (Disconnection)

QC plot and Profile plot

- Can detect the source of variation :
  - Run, Subject, Feature, Condition
- Can find any problematic observation :
  - Outliers, missingness
- Show how the normalization works.

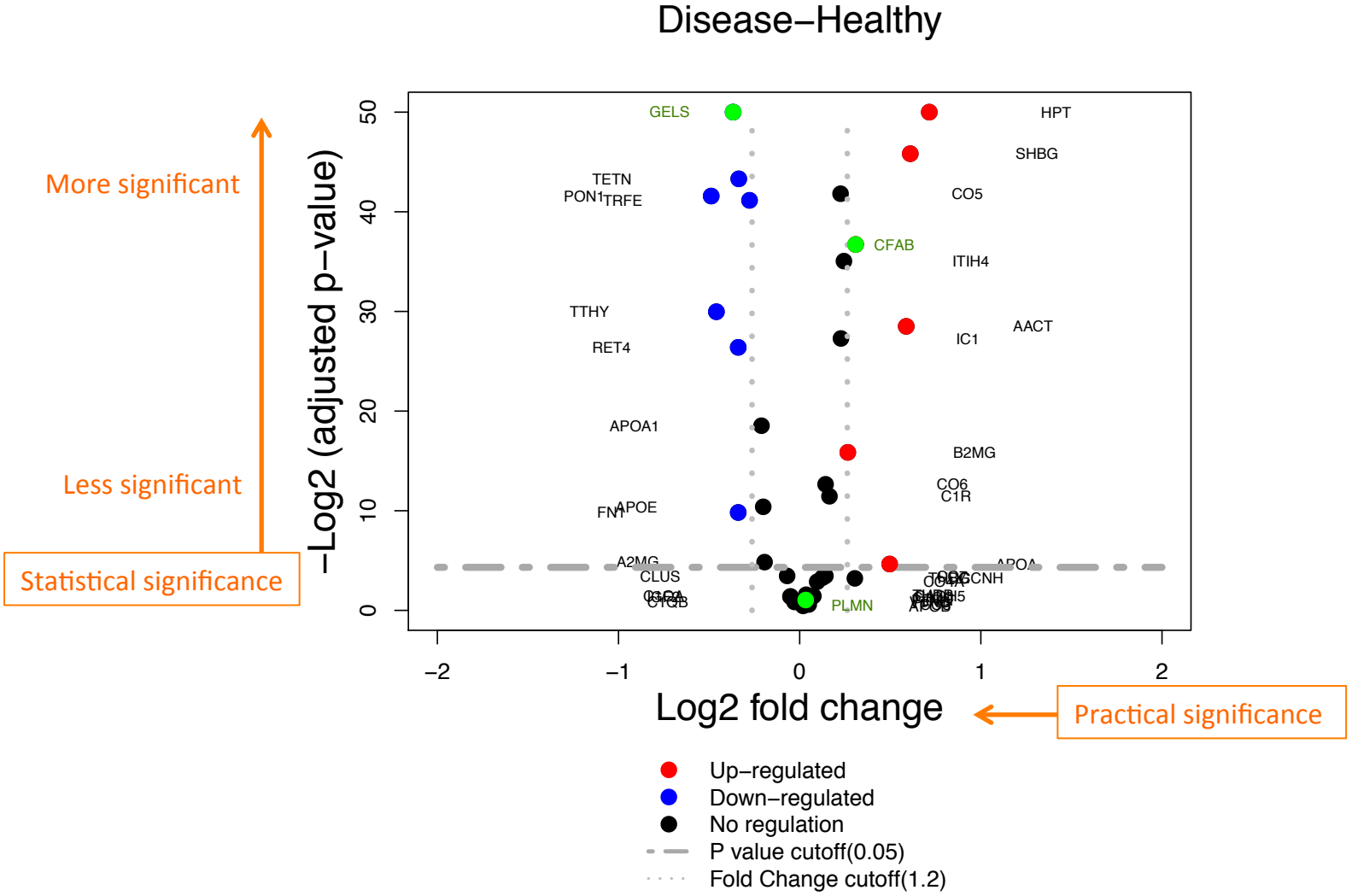
Group Comparison

Comparison : **Disease – Healthy** (Ovarian Cancer – Control)

- with default option for model (random Run, fixed Subject)
- output with FDR<0.05, Fold Change cutoff=1.2

Protein	Label	log2FC	SE	Tvalue	DF	pvalue	adj.pvalue	
GELS	OV-Control	-0.3671696	0.03828988	-9.589208	627	0.000000e+00	0.000000e+00	Significant
HPT	OV-Control	0.7155773	0.06299801	11.358729	869	0.000000e+00	0.000000e+00	
SHBG	OV-Control	0.6106531	0.06883696	8.871005	158	1.332268e-15	1.598721e-14	
TETN	OV-Control	-0.3360433	0.04176910	-8.045260	393	1.021405e-14	9.192647e-14	
PON1	OV-Control	-0.4882415	0.05897174	-8.279245	157	5.062617e-14	3.037570e-13	
TRFE	OV-Control	-0.2766892	0.03620692	-7.641889	632	7.971401e-14	4.099578e-13	
CFAB	OV-Control	0.3090737	0.04303860	7.181314	628	1.959766e-12	8.818946e-12	Significant
TTHY	OV-Control	-0.4595182	0.07071673	-6.498013	369	2.639720e-10	9.502990e-10	
AACT	OV-Control	0.5881425	0.08990597	6.541751	158	8.028156e-10	2.627396e-09	
RET4	OV-Control	-0.3389700	0.05631861	-6.018792	386	4.089053e-09	1.132353e-08	
APOA1	OV-Control	-0.2108136	0.04271180	-4.935723	626	1.025170e-06	2.636151e-06	
BZMG	OV-Control	0.2655696	0.05824220	4.559745	363	7.012014e-06	1.682883e-05	
APOE	OV-Control	-0.2012671	0.05623494	-3.579040	629	3.714142e-04	7.428285e-04	
FN1	OV-Control	-0.3391011	0.09657341	-3.511330	157	5.822912e-04	1.103289e-03	
AZMG	OV-Control	-0.1938152	0.08199489	-2.363747	153	1.934731e-02	3.482516e-02	
APOA	OV-Control	0.4966726	0.21668170	2.292176	157	2.322386e-02	3.981232e-02	
PLMN	OV-Control	0.03325885	0.04229023	0.7864428	603	4.319171e-01	4.859067e-01	Not significant

Volcano Plot

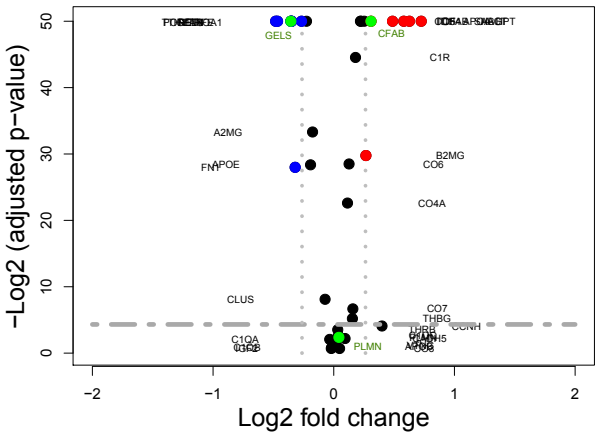


Random Run, Fixed Subject, FDR<0.05 and FC cutoff=1.2

Group Comparison with different options

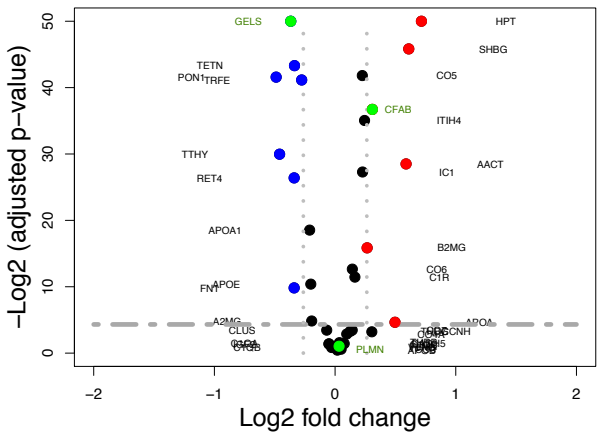
- Scope of biological replication : fixed (“restricted”) / random (“expanded”)
- Scope of technical MS run replication : fixed (“restricted”) / random (“expanded”)
- Interference : contain interference transitions, need additional model interaction

Fixed Run, Fixed Subject



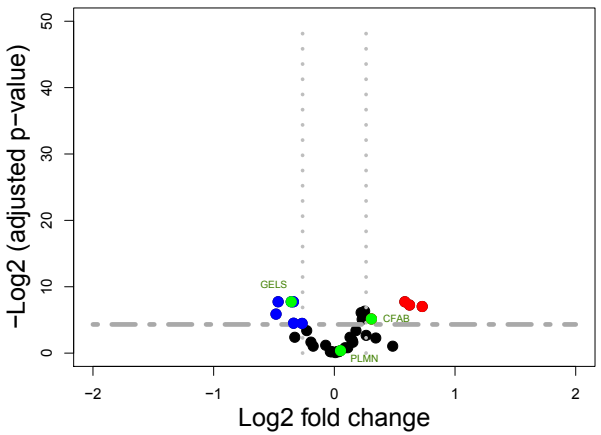
Top 5	log2FC	SE	Adj p-value
HPT	0.7249	0.0179	<0.0001
TRFE	-0.2668	0.0074	<0.0001
PON1	-0.4658	0.0180	<0.0001
SHBG	0.5803	0.0239	<0.0001
TTHY	-0.4813	0.0285	<0.0001

Random Run, Fixed Subject



Top 5	log2FC	SE	Adj p-value
GELS	-0.3672	0.0383	<0.0001
HPT	0.7156	0.0630	<0.0001
SHBG	0.6107	0.0688	<0.0001
TETN	-0.3360	0.0418	<0.0001
PON1	-0.4882	0.0590	<0.0001

Random Run, Random Subject



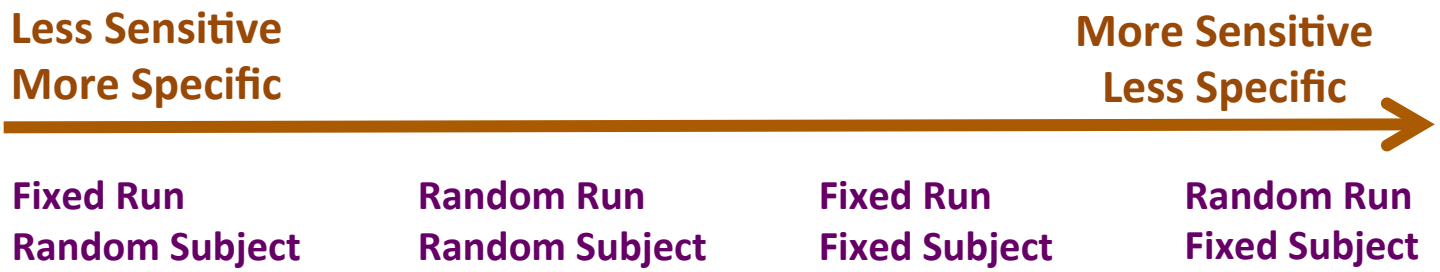
Top 5	log2FC	SE	Adj p-value
PON1	-0.4658	0.1249	0.0047
TETN	-0.3394	0.0915	0.0047
SHBG	0.5845	0.1589	0.0047
GELS	-0.3573	0.0987	0.0047
AACT	0.6232	0.1808	0.0066

Summary for comparison

- Three Proteins from Profile plots

Protein	Fixed Run, Fixed Subject			Random Run, Fixed Subject			Random Run, Random Subject		
	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value
CFAB	0.3087	0.0248	<0.0001	0.3091	0.0430	<0.0001	0.3071	0.1125	0.0282
GELS	-0.3557	0.0214	<0.0001	-0.3672	0.0383	<0.0001	-0.3573	0.0987	0.0047
PLMN	0.0427	0.0299	0.1923	0.0333	0.0423	0.4859	0.0508	0.0966	0.7718

Conclusion



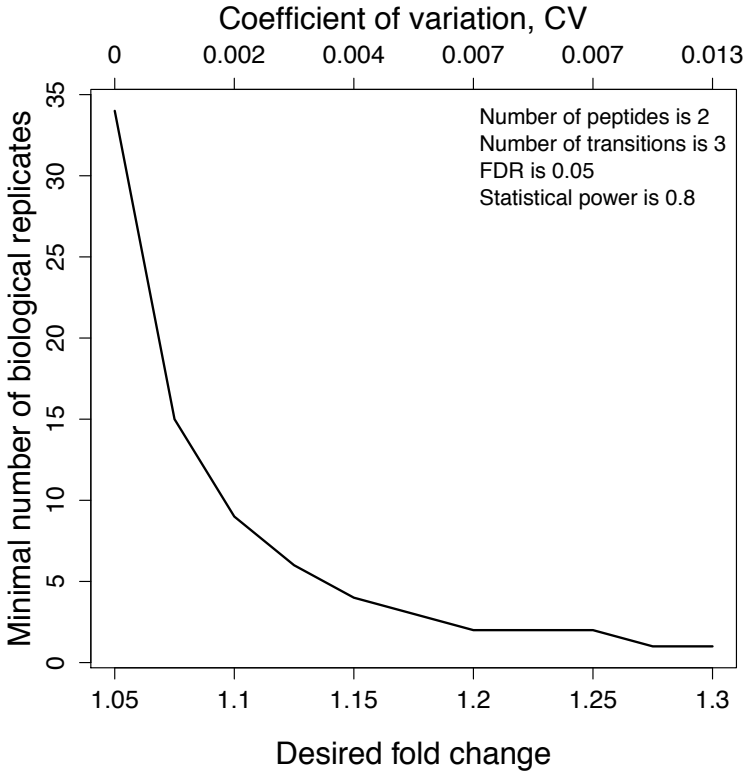
- The choice of the model should depend on the desired scope of biological conclusions, and not on the sensitivity/specificity.

Sample size calculation

The number of biological replicates $= J \geq \left(\frac{4\sigma^2}{KL} - \frac{2\sigma^2(1-w)}{KL}\right)\left(\frac{Z_{1-\beta} + Z_{1-\alpha/2}}{\Delta}\right)^2$

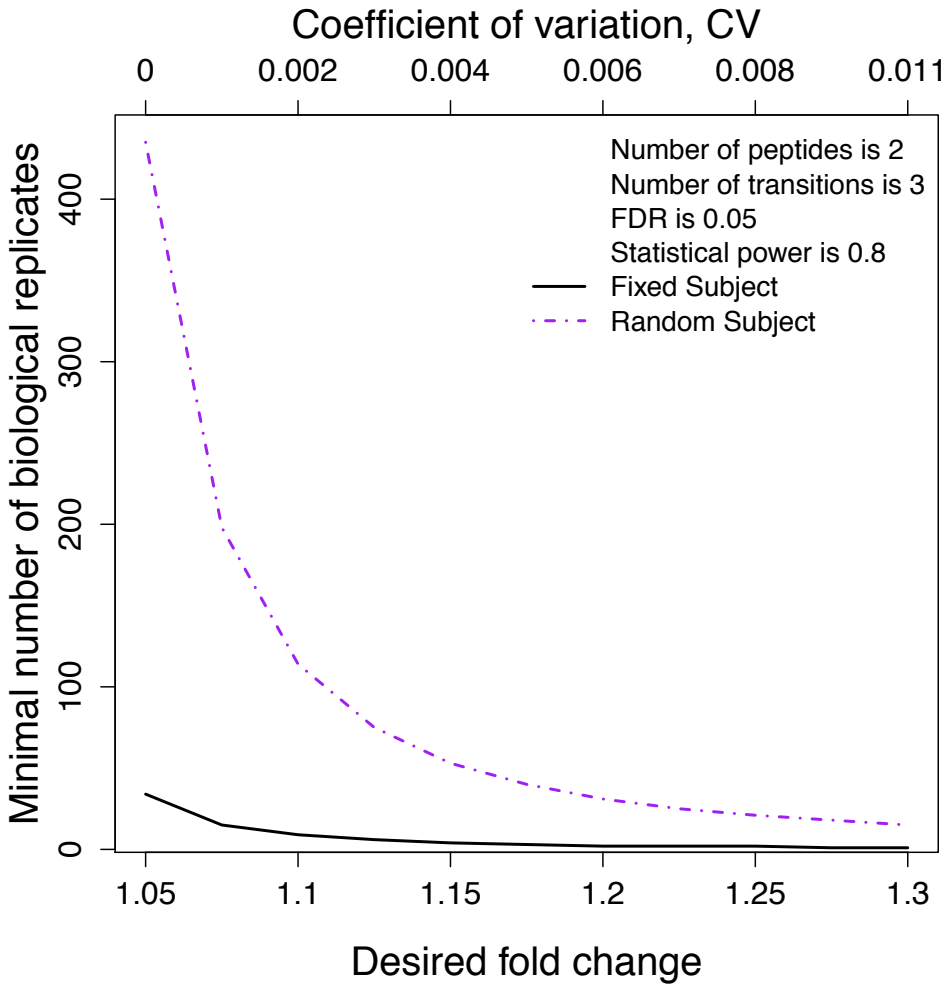
- Use the current dataset for variance estimation : with fixed Subject or random Subject
- Also calculate
  - The number of peptide per protein
  - The number of transition per peptide
  - power

desiredFC	numSample	numPep	numTran	FDR	power	CV
1.050	34	2	3	0.05	0.8	0.000
1.075	15	2	3	0.05	0.8	0.001
1.100	9	2	3	0.05	0.8	0.002
1.125	6	2	3	0.05	0.8	0.003
1.150	4	2	3	0.05	0.8	0.004
1.175	3	2	3	0.05	0.8	0.005
1.200	2	2	3	0.05	0.8	0.007
1.225	2	2	3	0.05	0.8	0.007
1.250	2	2	3	0.05	0.8	0.007
1.275	1	2	3	0.05	0.8	0.013
1.300	1	2	3	0.05	0.8	0.013



Output and plot of Sample size calculation with fixed Subject

Sample size calculation



Need more number of biological replicate with random subject for model

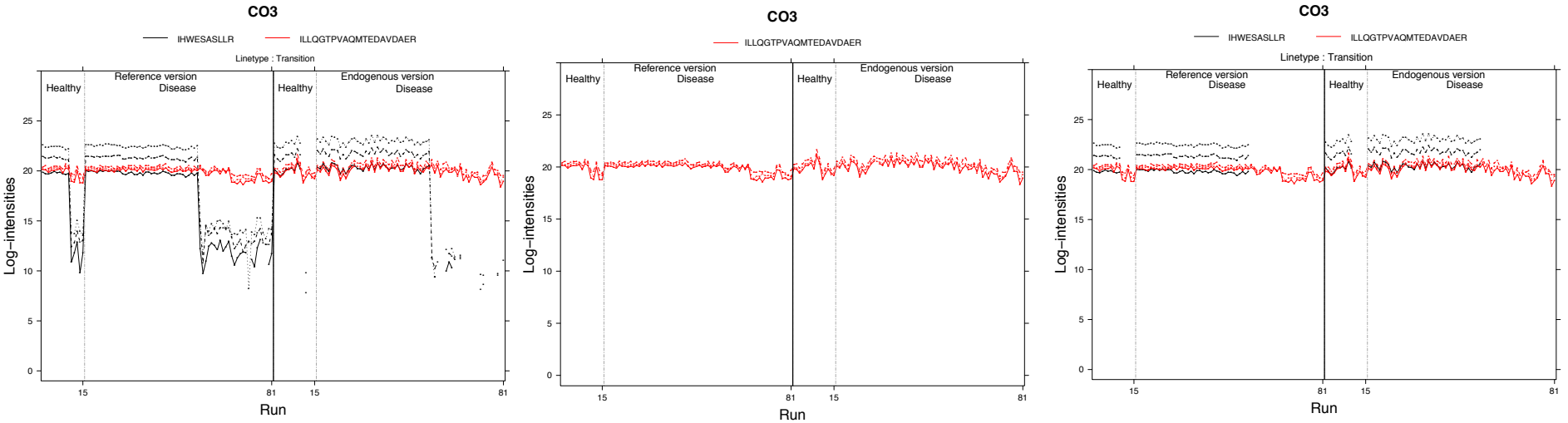
Overview

1. R packages : MSstats2 and SRMstats
2. Default analysis of a label-based SRM experiment (Human Plasma : Ovarian Cancer)
    1. Whole conceptual analysis
    2. How to analyze in R
    3. How to analyze in Skyline
3. A study of the importance of the quality of peaks
4. Another example of a label-free SRM (Rat plasma)
    1. Normalization
    2. A study of poor quality or inconsistent peptides

Overview

1. R packages : MSstats2 and SRMstats
2. Default analysis of a label-based SRM experiment (Human Plasma : Ovarian Cancer)
    1. Whole conceptual analysis
    2. How to analyze in R
    3. How to analyze in Skyline
3. A study of the importance of the quality of peaks
4. Another example of a label-free SRM (Rat plasma)
    1. Normalization
    2. A study of poor quality or inconsistent peptides

Poor quality feature, unrecognized : CO3



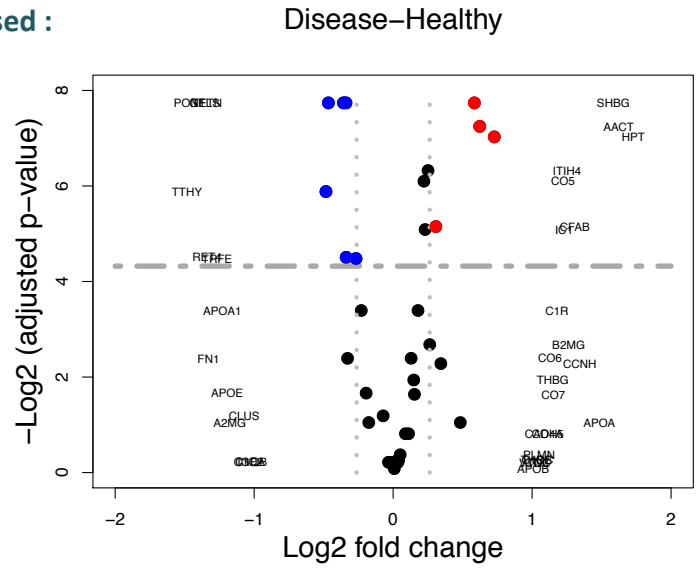
		With Features		Remove the feature entirely		Replace with missing values		No Interaction	
	Options	log2FC	Adj P-value	log2FC	Adj P-value	log2FC	Adj P-value	log2FC	Adj P-value
Labeed-based	Fixed Run Fixed Subject	0.0485 (0.0968)	0.6166	0.15 (0.0206)	<0.0001	0.1775 (0.0156)	<0.0001	0.1587 (0.028)	<0.0001
	Random Run Fixed Subject	0.0502 (0.1088)	0.6635	0.1539 (0.0460)	0.0018	0.1976 (0.0375)	<0.0001	0.1811 (0.0398)	<0.0001
	Random Run Random Subject	0.0342 (0.1669)	0.8621	0.1496 (0.0936)	0.2054	0.1819 (0.0925)	0.1185	0.1612 (0.0944)	0.1729
Label-free	Fixed Subject	-0.4280 (0.2991)	0.1841	0.1514 (0.0139)	<0.0001	0.2060 (0.0175)	<0.0001	0.1864 (0.0166)	<0.0001
	Random Subject	-0.4227 (0.5514)	0.6471	0.1513 (0.1573)	0.555	0.2059 (0.1463)	0.2840	0.1864 (0.1462)	0.3532

Summary for poor quality

- With poor quality features,
  - fold change is quite different. Also the conclusion is different.
- In this case, remove the feature entirely, or replace missing values get similar result because there are other good features.
- Replace vs no interaction : not much different because the number of missing values are reasonable. However, the number of missing values are large, it can be affected.

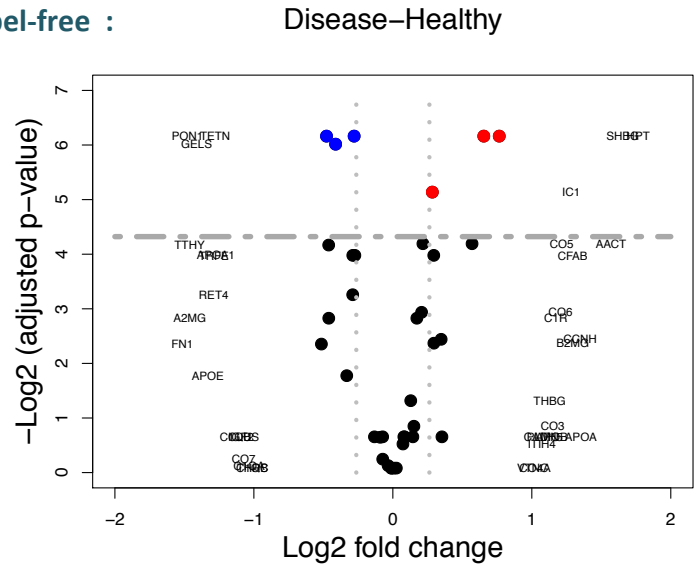
Labeled vs Label-free : Comparison with random Subject

Label-based :



Top 7	log2FC	SE	Adj p-value
GELS	-0.3576	0.0988	0.0047
PON1	-0.4660	0.1250	0.0047
SHBG	0.5840	0.1589	0.0047
TETN	-0.3397	0.0917	0.0047
AACT	0.6626	0.1808	0.0066
HPT	0.7278	0.2177	0.0076
TTHY	-0.4834	0.1642	0.0171

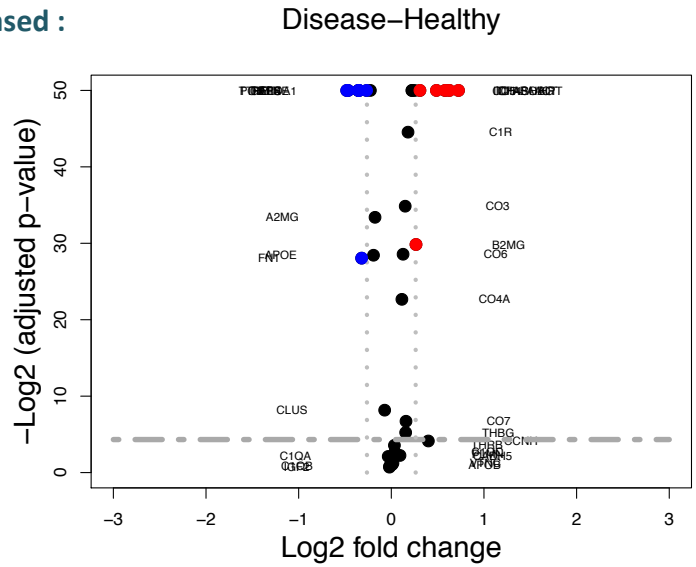
Label-free :



Top 6	log2FC	SE	Adj p-value
HPT	0.7656	0.2226	0.0139
PON1	-0.4772	0.1434	0.0139
SHBG	0.6544	0.1832	0.0139
TETN	-0.2785	0.0849	0.0139
GELS	-0.4117	0.1297	0.0155
IC1	0.2842	0.0978	0.0294

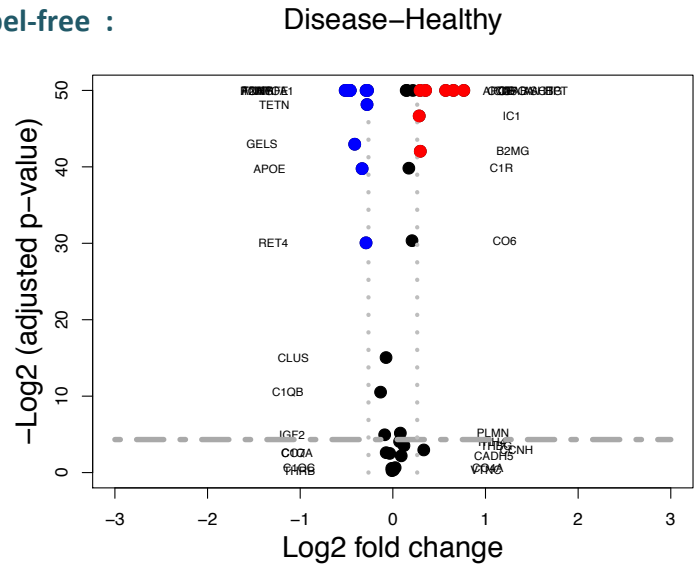
Labeled vs Label-free : Comparison with fixed Subject

Label-based :



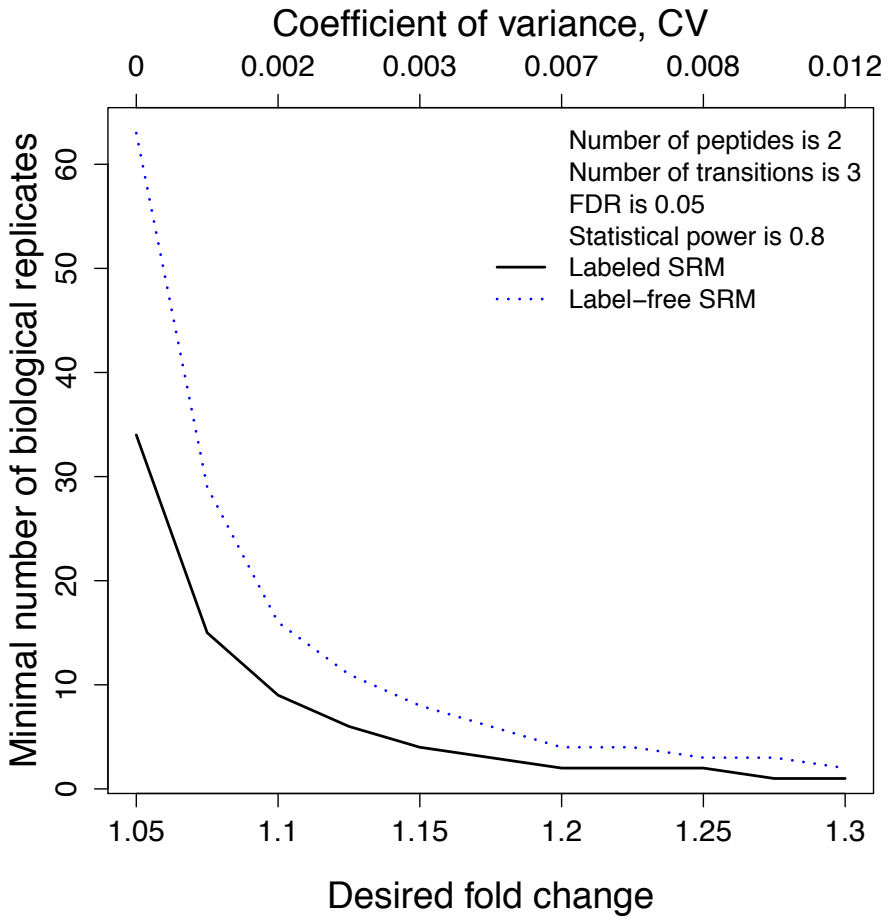
Top 7	log2FC	SE	Adj p-value
HPT	0.7249	0.0179	<0.0001
TRFE	-0.2668	0.0074	<0.0001
PON1	-0.4657	0.0180	<0.0001
SHBG	0.5803	0.0239	<0.0001
TTHY	-0.4813	0.0295	<0.0001
GELS	-0.3557	0.0214	<0.0001
RET4	-0.3483	0.0225	<0.0001

Label-free :



Top 7	log2FC	SE	Adj p-value
AACT	0.5701	0.0114	<0.0001
SHBG	0.6544	0.0209	<0.0001
A2MG	-0.4610	0.0180	<0.0001
PON1	-0.4773	0.0237	<0.0001
HPT	0.7656	0.0440	<0.0001
FN1	-0.5143	0.0402	<0.0001
APOA	0.3532	0.0280	<0.0001

Labeled vs Label-free : Sample Size



- We need the statistical model to make these plots : Here Fixed Subject used.
- The plots assume that the label-based and label-free have the same variance components, which may not be true.
- Label-free SRM need more sample for the same condition.
- Ideally would make separate pilot experiments with each technology.

Summary for Label-based and Label-free

- Comparison : conclusion is not different. However, SE is different.
- Sample Size Calculation : almost double of the number of sample size is need. Because SE for label-free is larger.

- The problem with the previous analysis is that we used references to peak picks, so it is only moderately representative of the real label-free analysis.

Overview

1. R packages : MSstats2 and SRMstats
2. Default analysis of a label-based SRM experiment (Human Plasma : Ovarian Cancer)
    1. Whole conceptual analysis
    2. How to analyze in R
    3. How to analyze in Skyline
3. A study of the importance of the quality of peaks
4. Another example of a label-free SRM (Rat plasma)
    1. A study of poor quality or inconsistent peptides
    2. Normalization

Example2 : Rat Plasma

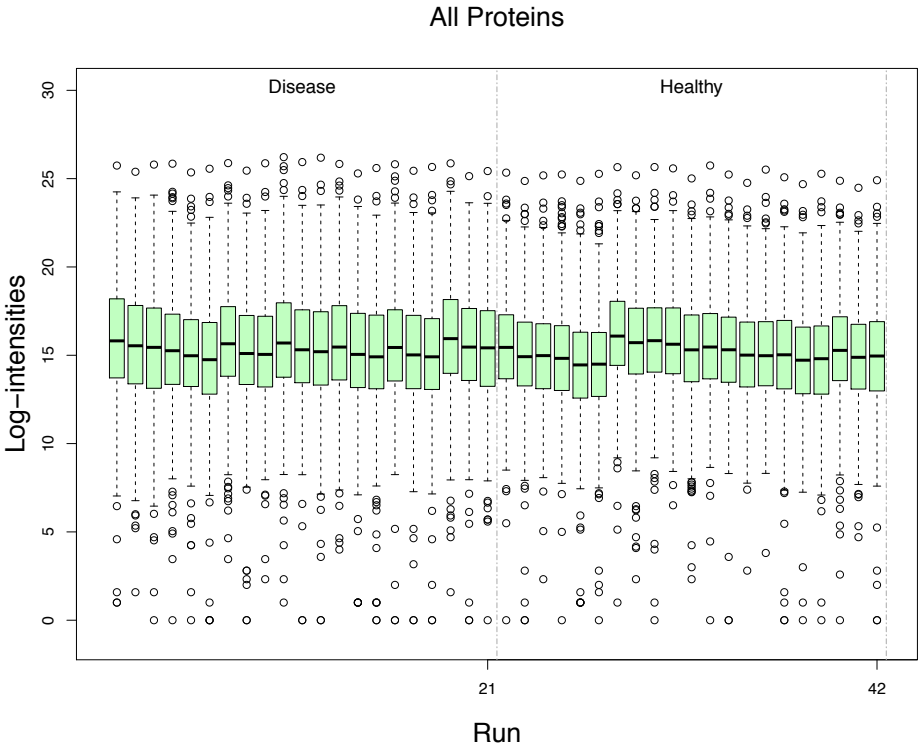
Each Protein	High salt (Disease)			Low salt (Healthy)		
	Sub1	…	Sub7	Sub8	…	Sub14
Tech 1	X	…	X	X	…	X
Tech2	X	…	X	X	…	X
Tech 3	~~X~~	…	X	X	…	X

- Label-free SRM experiment
- Comparison : High Salt – Low Salt (Disease-Healthy)
- Issues
  - Data has 0 (zero) intensities : can't do analysis with zero intensities, Need to change as NA.
  - There are truncated transitions.

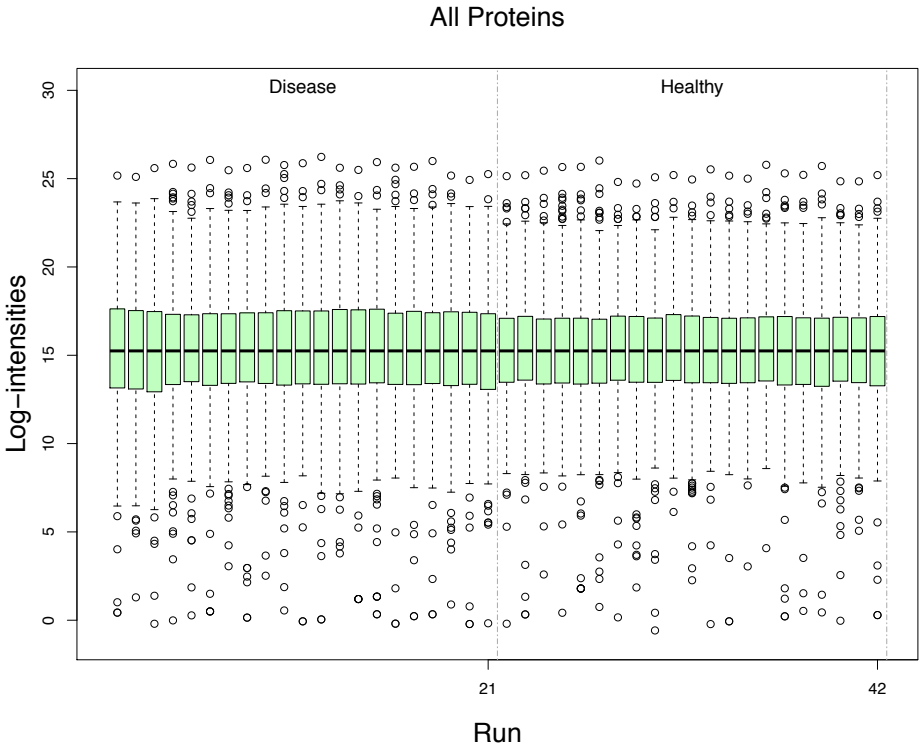
Constant normalization

- Constant Normalization across run for all proteins : default in package

Before Normalization



After Normalization

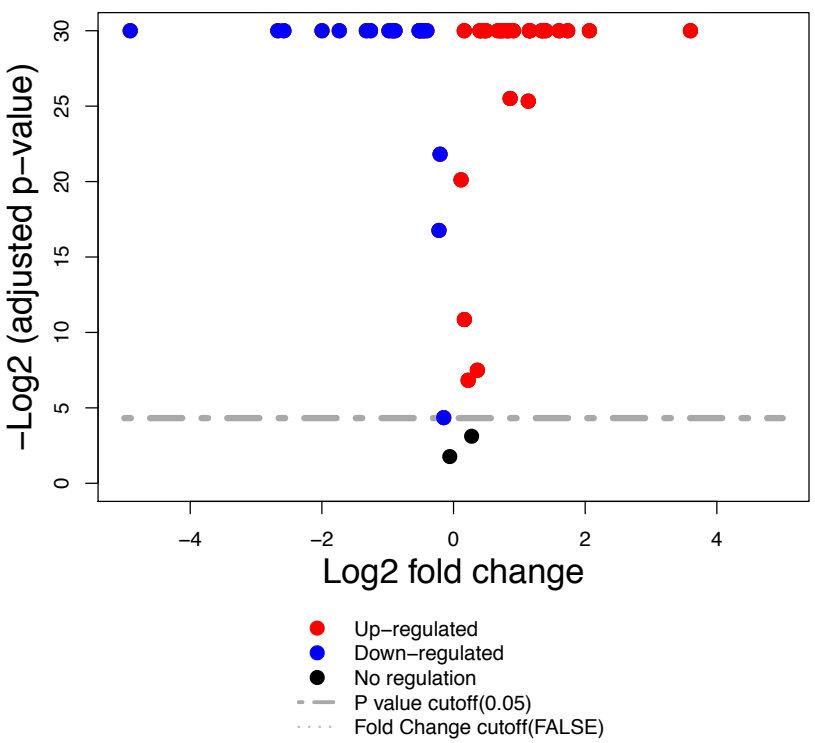


After normalization, the distributions of peaks across MS runs are similar.

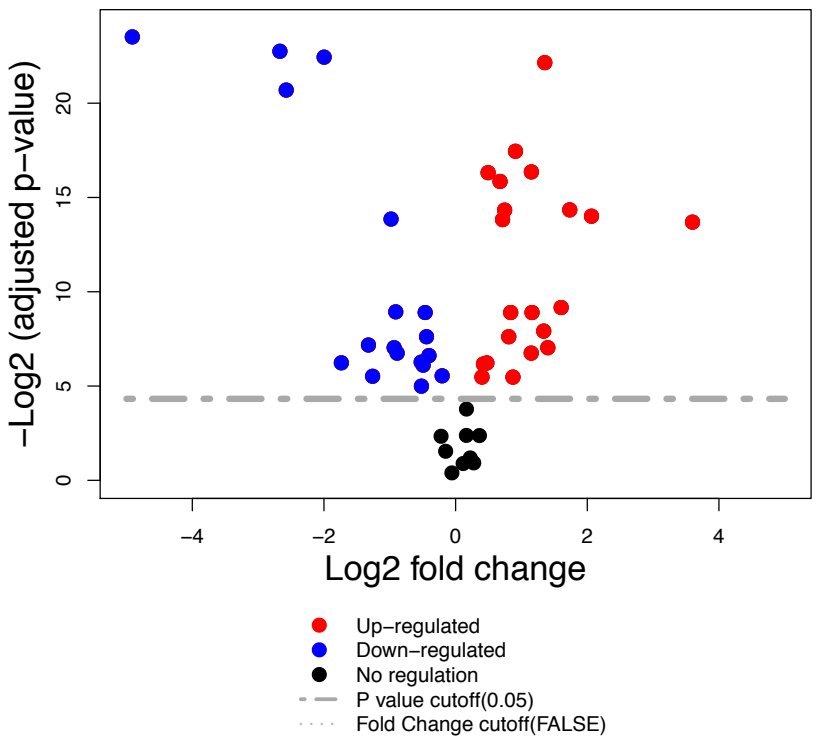
Group Comparison : Volcano plot

- For label-free, no need to specify 'Run' because biological replicates and technical MS runs are confounding.
- Without interference

Fixed Subject



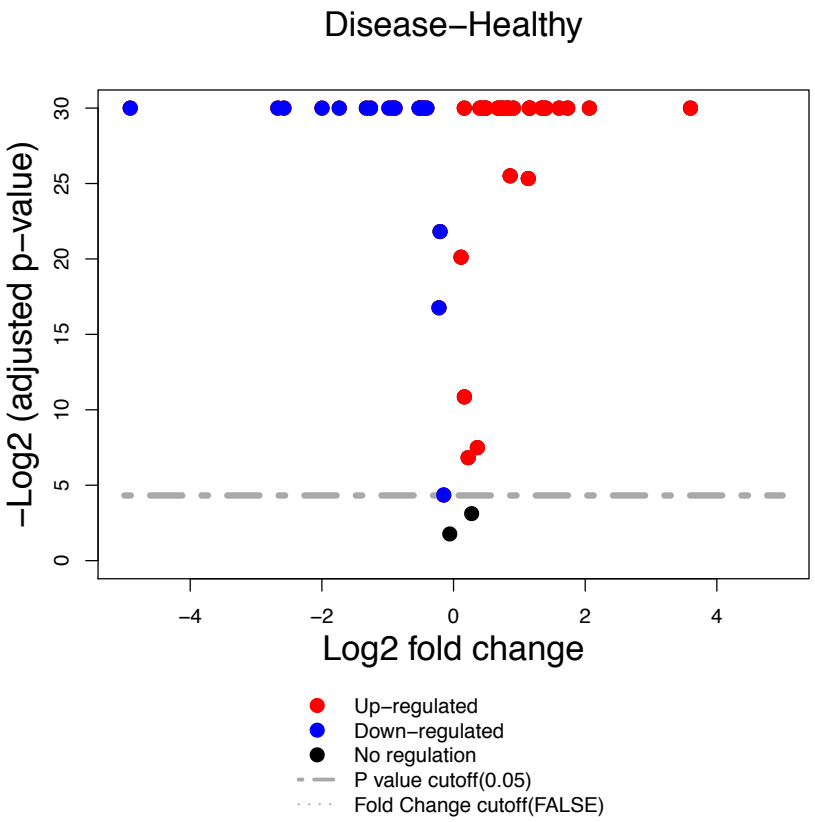
Random Subject



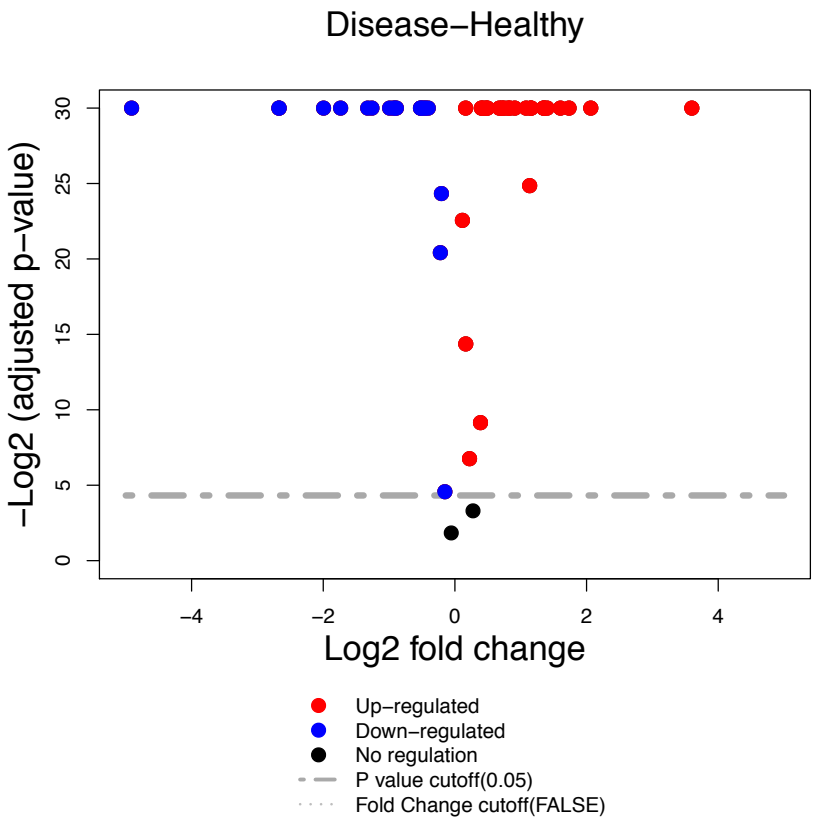
With or without interference

- Interaction may be overfitting the data in label-free
- There is little difference in this dataset. But we expect more in other dataset.

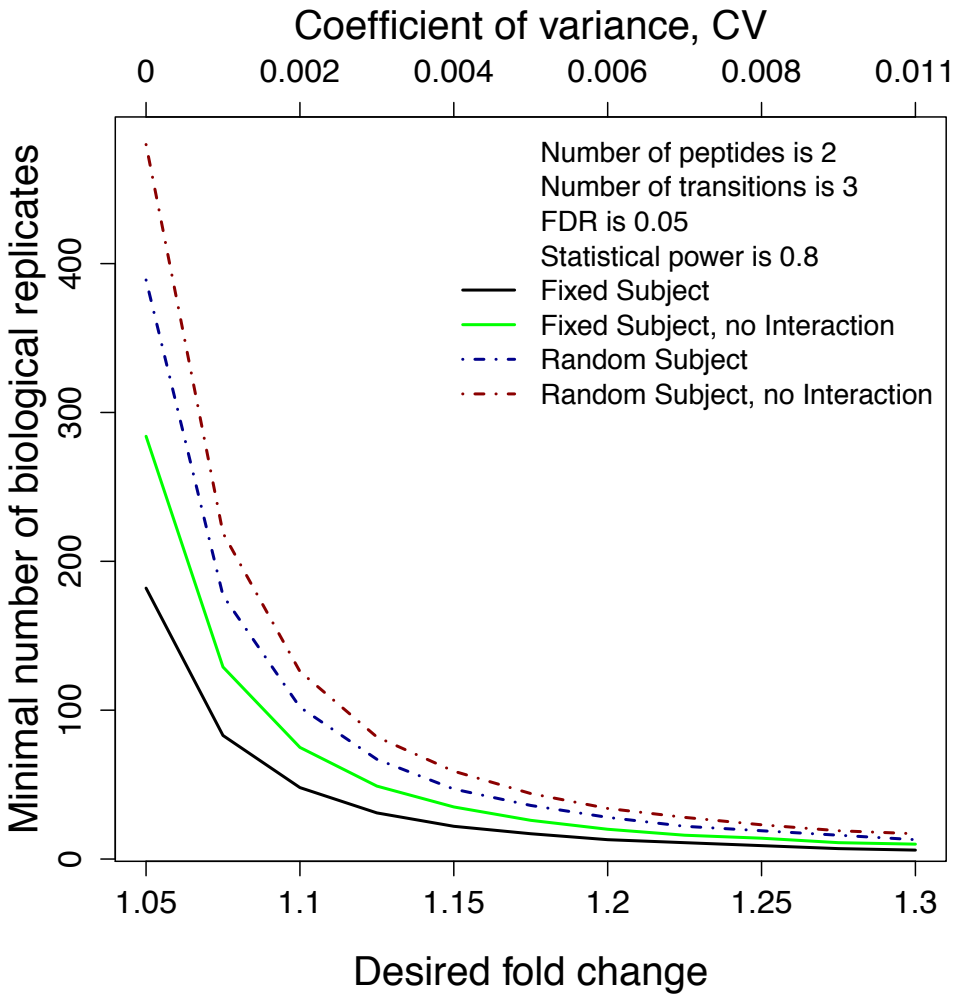
Without interference



With interference

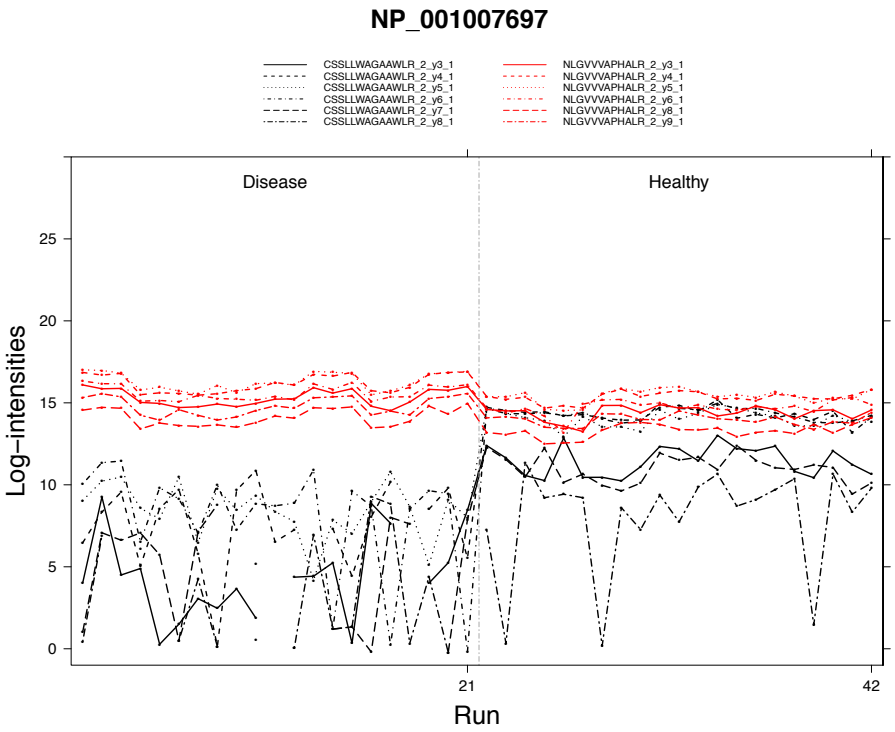


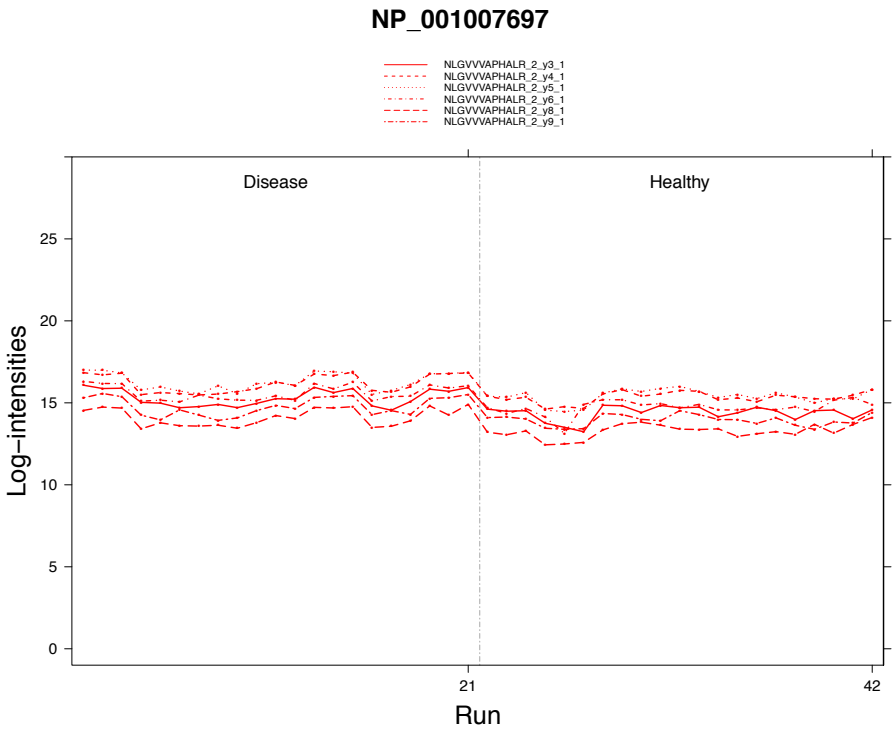
Sample Size Calculation



- Need more sample size because of increasing variation

Examples of poor quality peptides

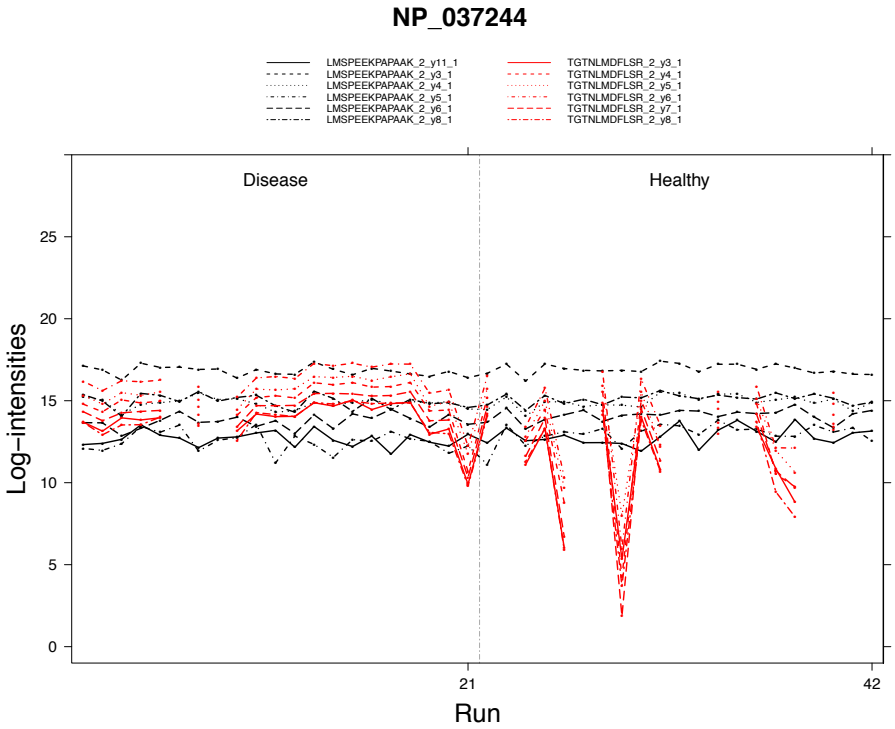


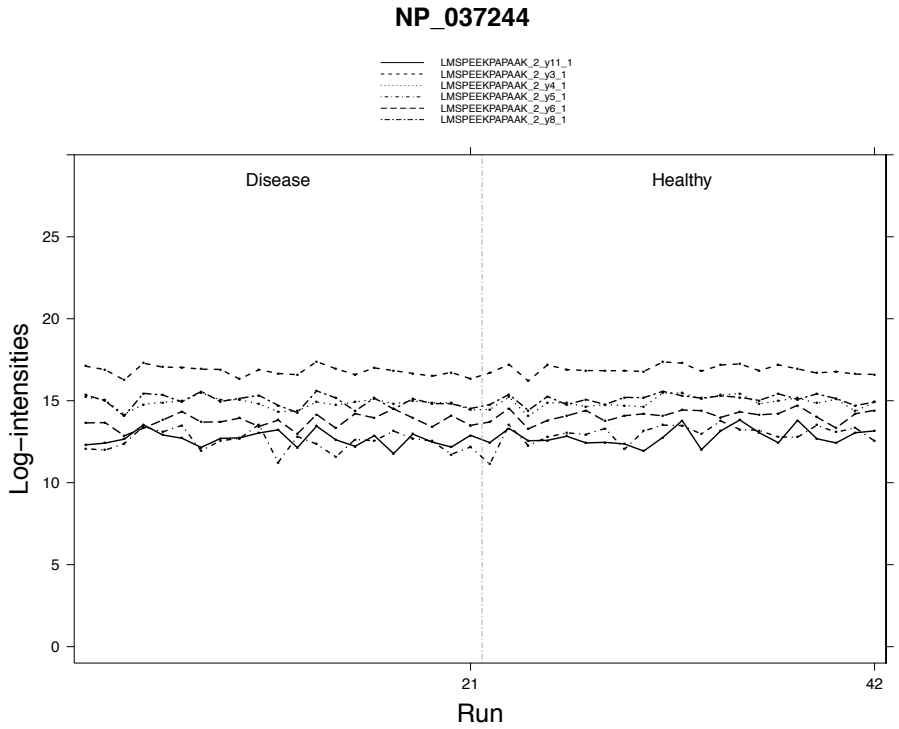


	All features			Remove peptides		
	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value
Fixed Subject	-2.5768	0.2192	<0.0001	0.8899	0.0261	<0.0001
Random Subject	-2.5734	0.2101	<0.0001	0.8899	0.2433	0.0068

Log2 FC and variation are quite different between before and after removing peptides.

Examples of poor quality peptides

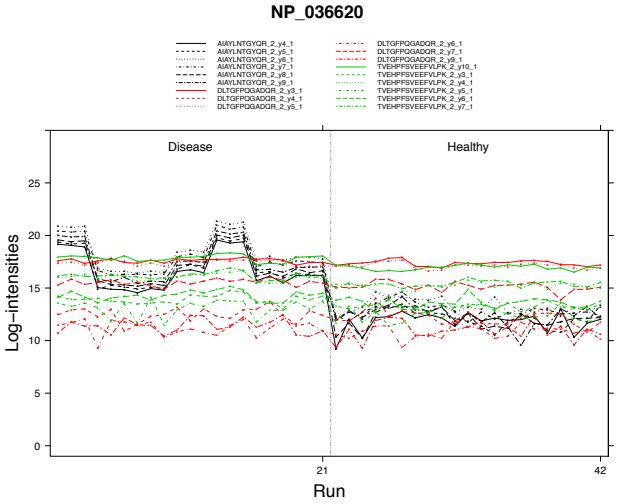


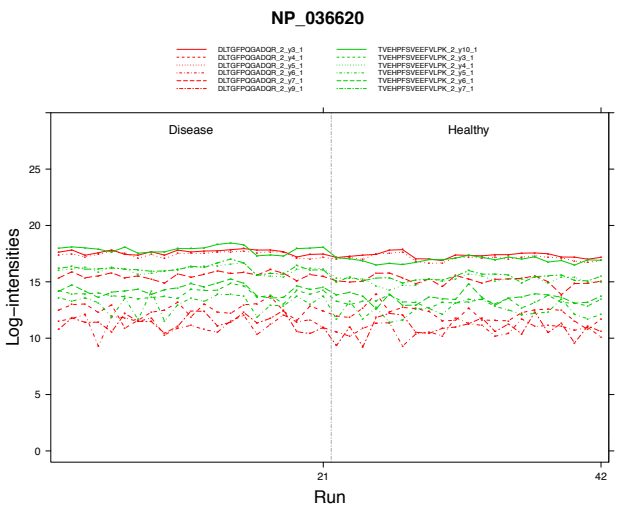


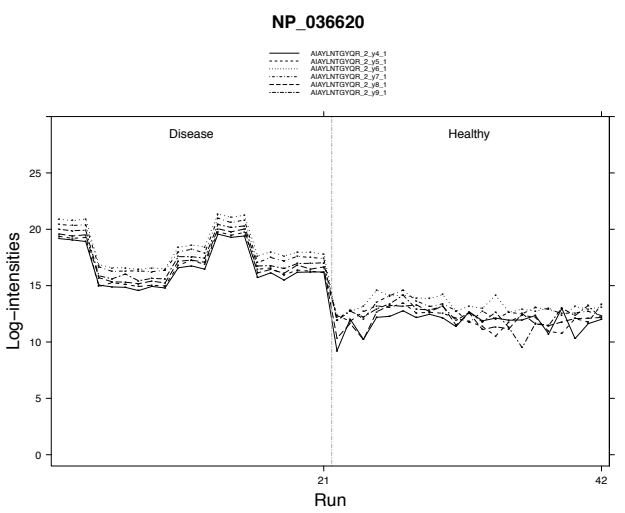
	All features			Remove peptides		
	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value
Fixed Subject	0.8599	0.1494	<0.0001	-0.2133	0.0505	<0.0001
Random Subject	0.8712	0.3175	0.0225	-0.2133	0.109	0.0882

Log2 FC and variation are quite different between before and after removing peptides.

Examples of inconsistent peptides







	All features			Only DLTG and TVEH			Only AIAY		
	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value	log2FC	SE	Adj p-value
Fixed Subject	2.0642	0.0951	<0.0001	0.6167	0.0414	<0.0001	5.0812	0.0591	<0.0001
Random Subject	2.0642	0.2966	<0.0001	0.6167	0.1137	0.0005	5.0812	0.7390	<0.0001

Log2 FC and variation are quite different depending on peptides.

Summary of poor quality peptides

- Less certainty that you look at the correct peptide,
  - suggestion : re-measure in label-based way.
- Need to investigate further a subset of peptides that we find interesting for some reason.
- Can use different models to do extra experimentation.

Overview

1. R packages : MSstats2 and SRMstats
2. Default analysis of a label-based SRM experiment (Human Plasma : Ovarian Cancer)
    1. Whole conceptual analysis
    2. How to analyze in R
    3. How to analyze in Skyline
3. A study of the importance of the quality of peaks
4. Another example of a label-free SRM (Rat plasma)
    1. A study of poor quality or inconsistent peptides
    2. Normalization

Normalization

Normalization for label-free SRM is a bigger problem than in label-based because we don't have references, and that the solution is less obvious.

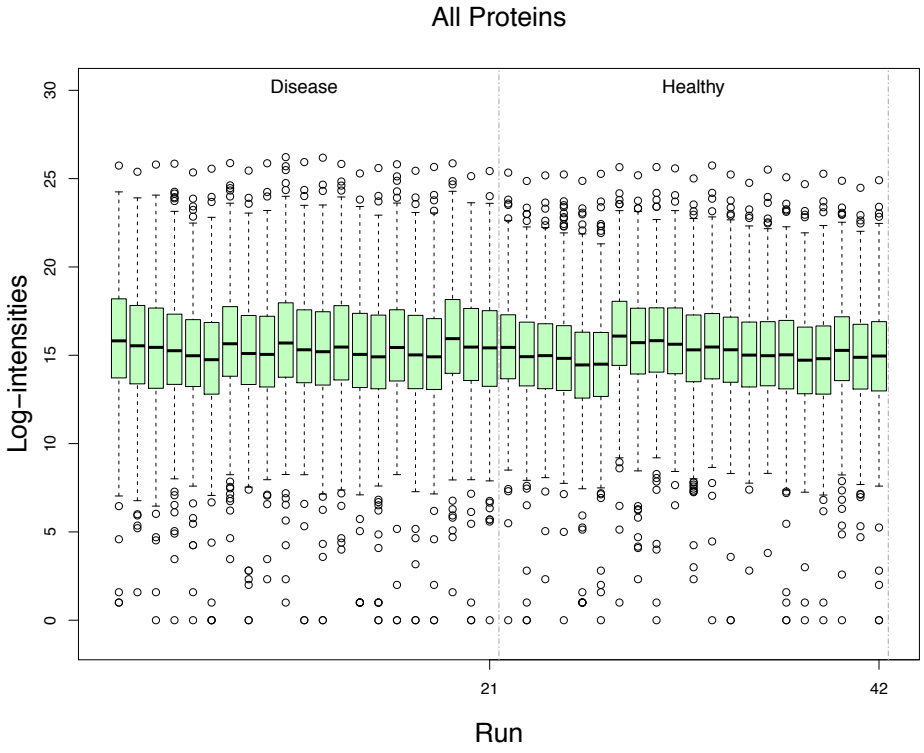
There are three normalizations we can do:

1. Constant normalization across run with all proteins : default in package
2. Quantile normalization across run with all proteins
3. Constant normalization with reference peptides

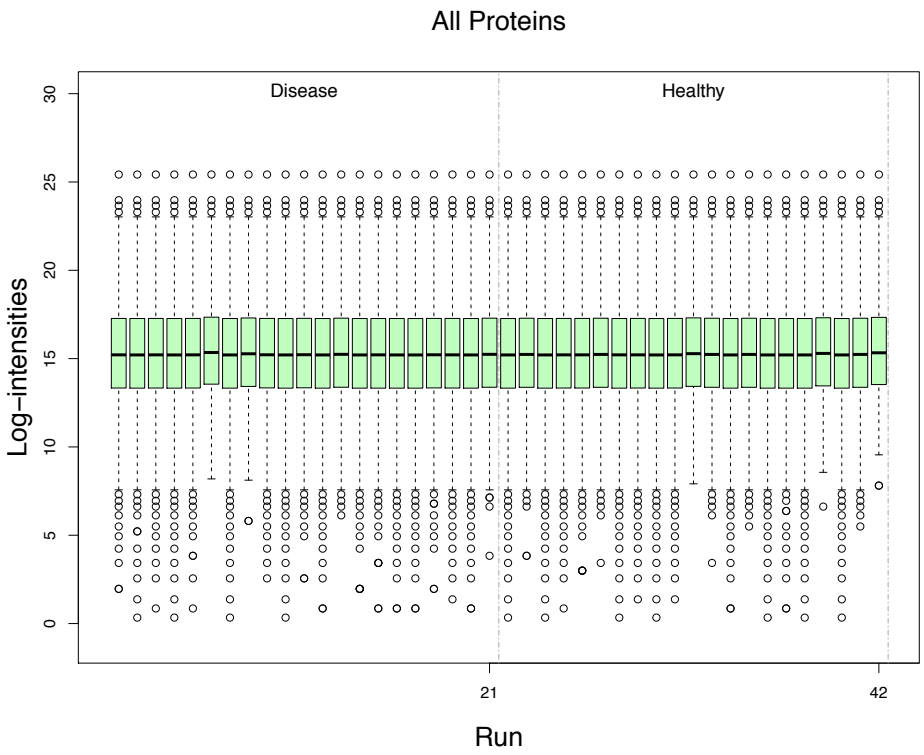
Quantile normalization

- Quantile Normalization across run for all proteins

Before Normalization



After Normalization

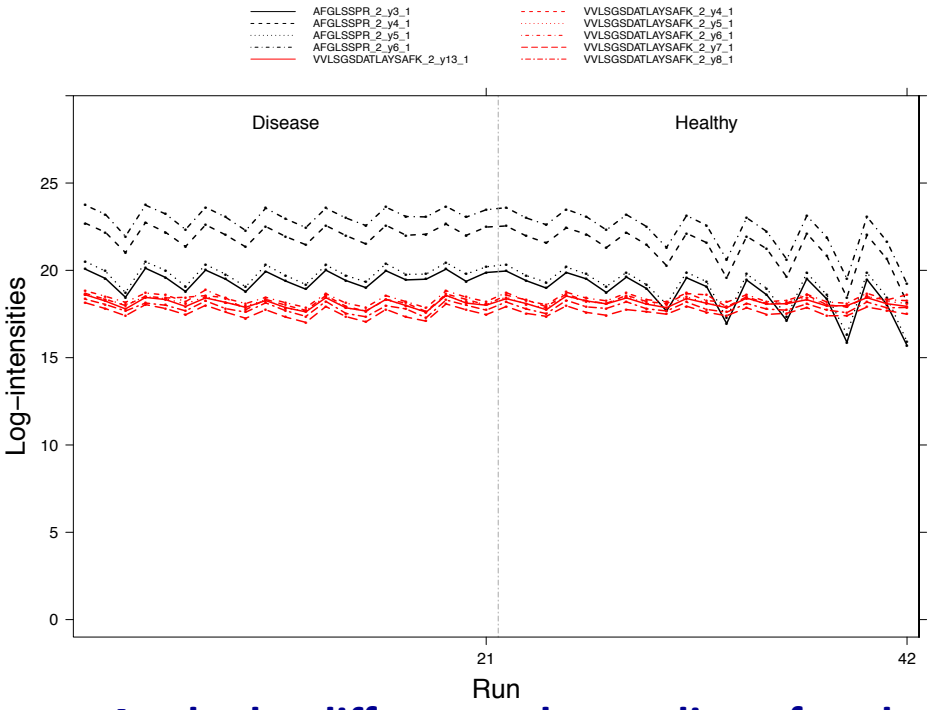


After normalization, the distributions of peaks across MS runs are the same.

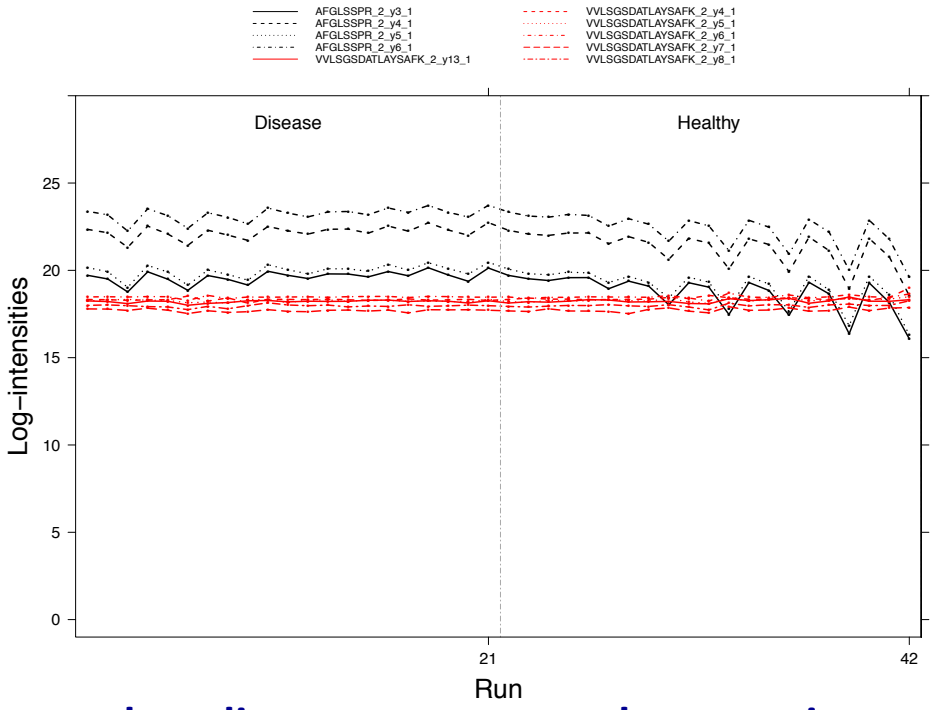
Constant normalization with reference peptides

- Constant Normalization across run with reference peptides

Before Normalization

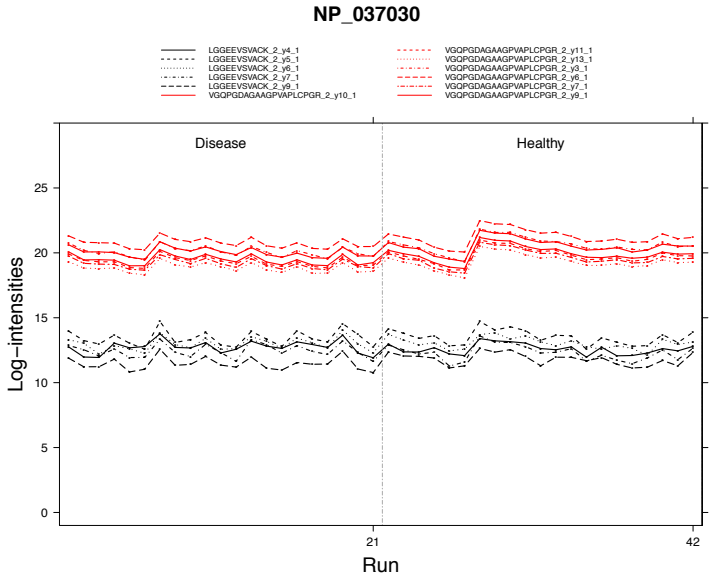


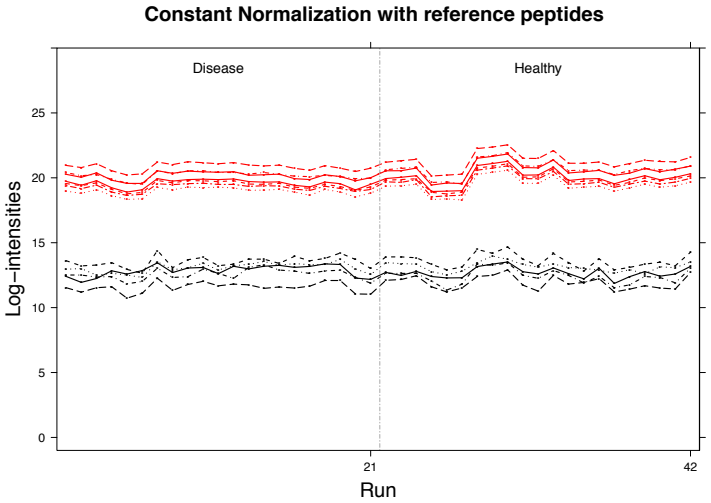
After Normalization

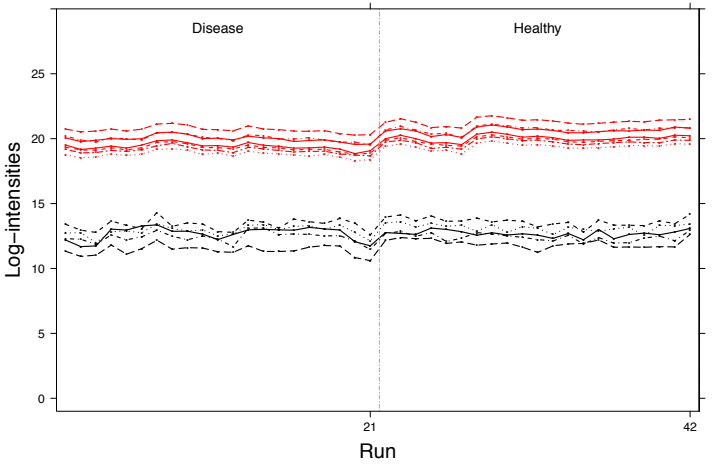


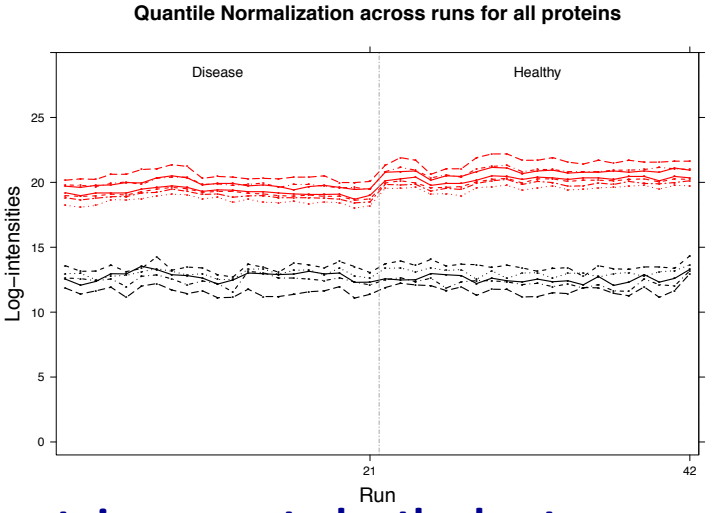
Apply the difference the median of each run and median across run to other proteins for each run.

Profile plots with normalizations



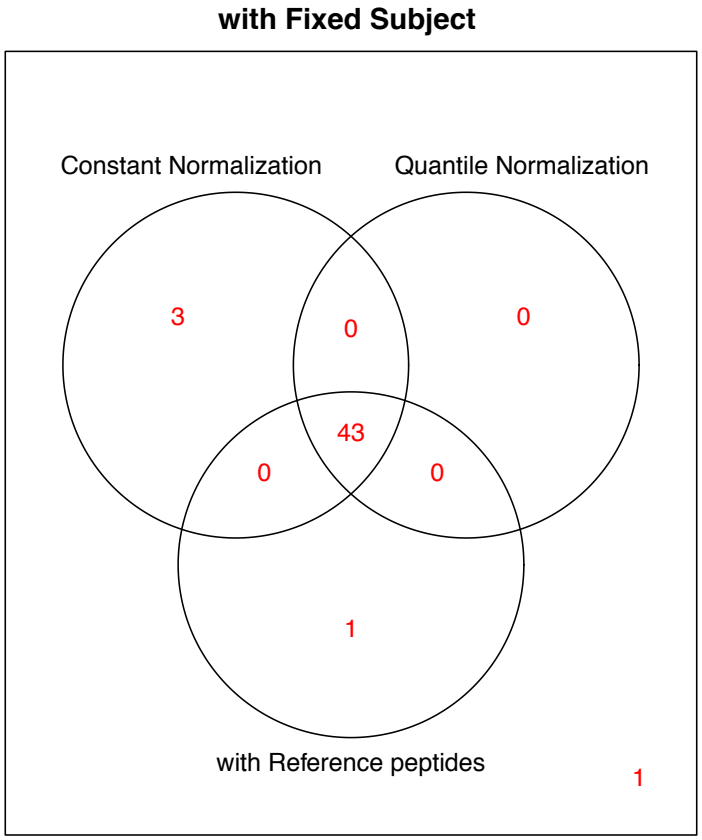


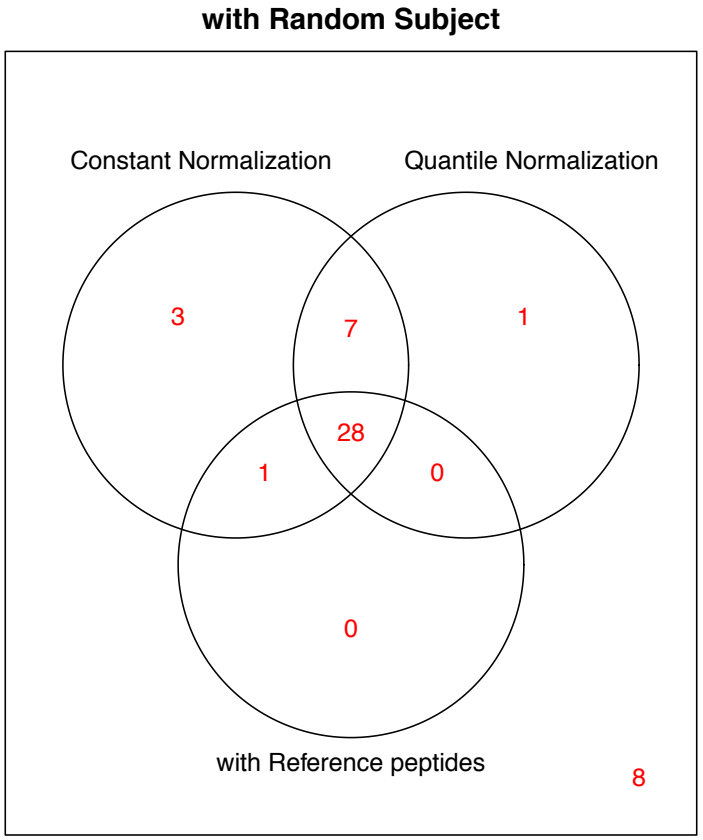




Constant Normalization across runs with all proteins seem to be the best.

How different the significant proteins among normalizations





Constant Normalization across runs with all proteins has highest sensitivity.

Future plan

- MSstats2
  - Develop the tools for sparse data, clustering, Biomarker study, network analysis
  - Other technical workflow : SWATH

- Integrated with Skyline
  - Tools for experimental design : randomization
  - Add various options
  - User friendly interface

Conclusion

- Label-based vs Label-free : compromise between accuracy of quantification, confidence in identification, and expense.
- These are good tools for figuring out and following up experimentally with more synthetic peptides or other low-throughput assays.

Contact

Meena Choi
Statistics, Purdue University
choi67@purdue.edu

Brendan MacLean
MacCoss Lab, Genome Sciences, U.Washington
brendanx@proteinms.net

Olga Vitek
Statistics and Compouter Science, Purdue University
ovitek@purdue.edu