

Definition of valid proteomic biomarkers: solutions to a currently unmet challenge

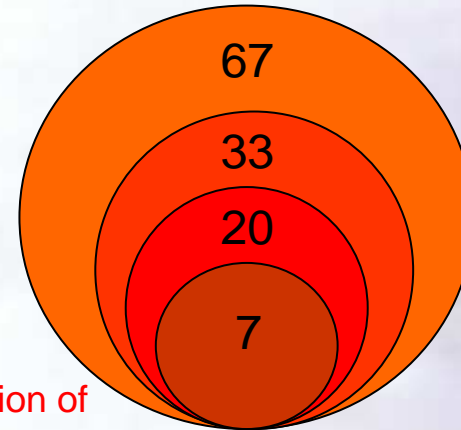
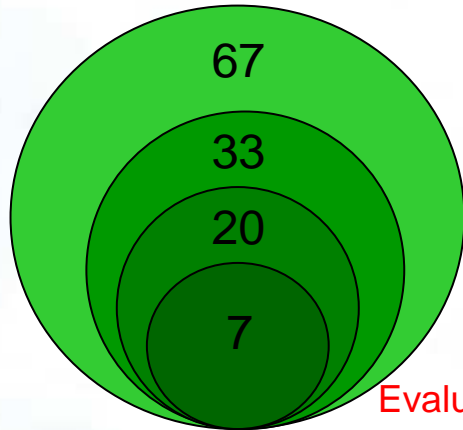
Sebastien Carpentier, Mohammed Dakna, Theodoros Damoulas, Keith Harris, Mark Girolami, Alexandros Kalousis, Walter Kolch, and Harald Mischak

How to assess validity of biomarkers and biomarker models?

- How many samples are required to define useful biomarkers?
- Relevance of Statistics?
- Which classifiers perform best?
- Is assessment in a blinded set necessary/helpful?

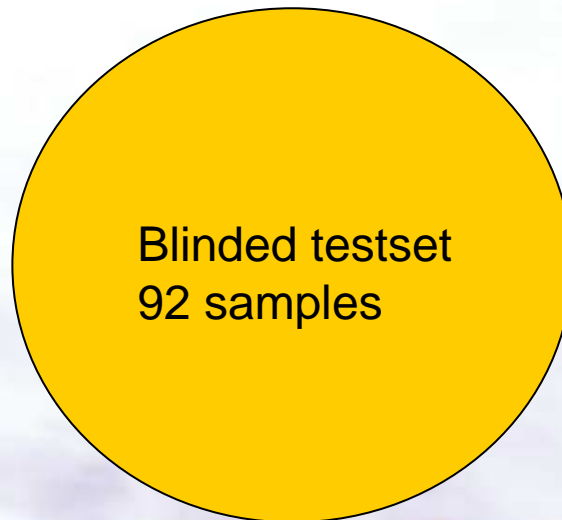
Study Design

healthy female 19-40 years **CE-MS analysis** healthy male 19-40 years

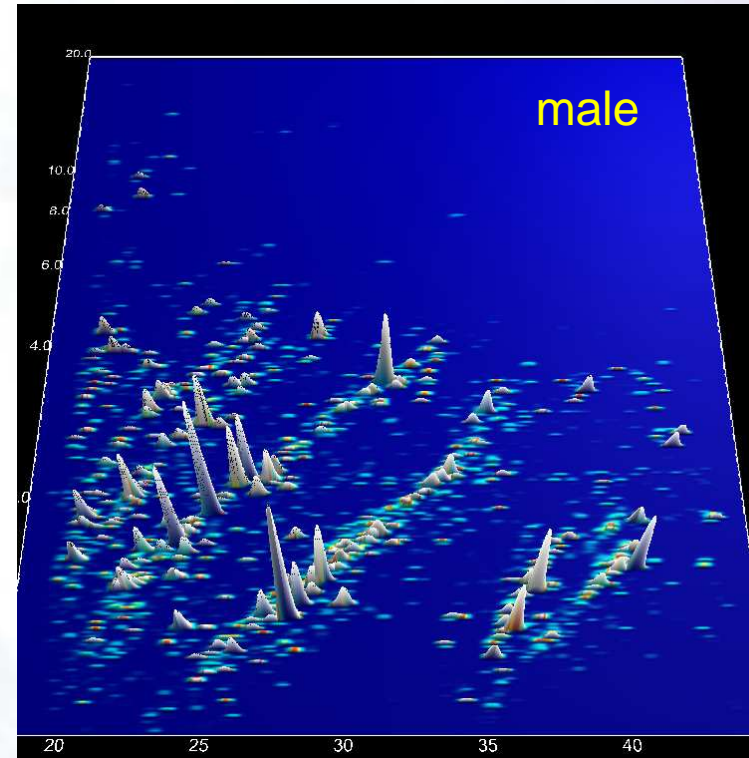
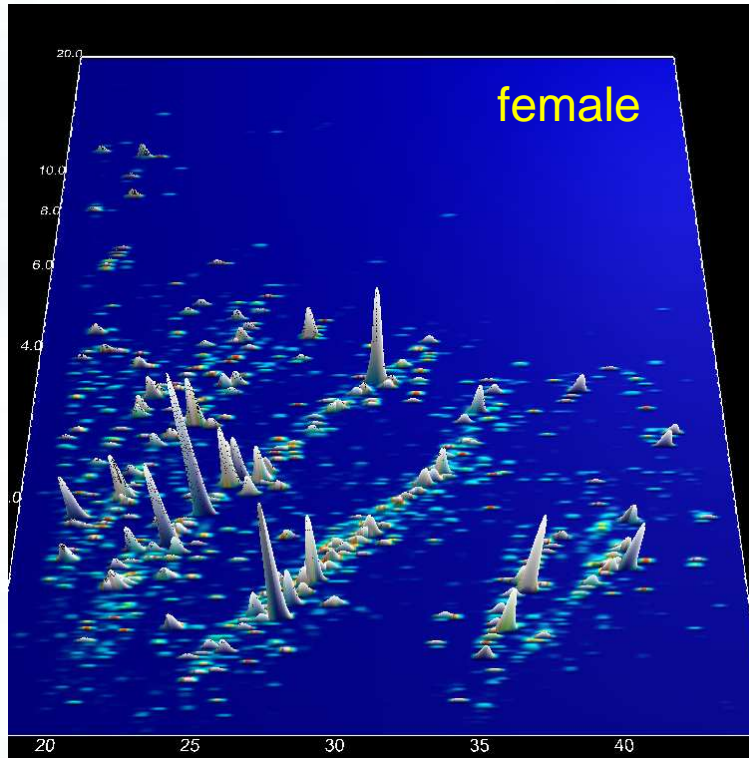


Evaluation of all individual biomarkers

Evaluation of biomarker models



Compiled datasets of CE-MS data on 67 (each) healthy subjects



Ca. 1200 features per dataset

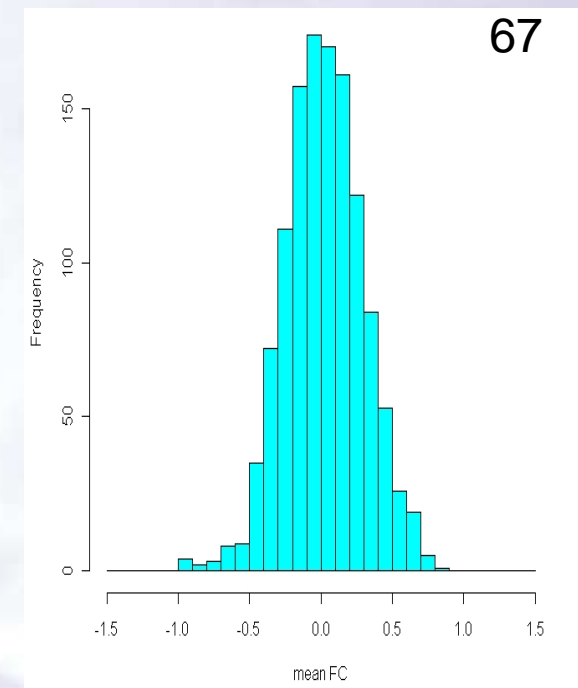
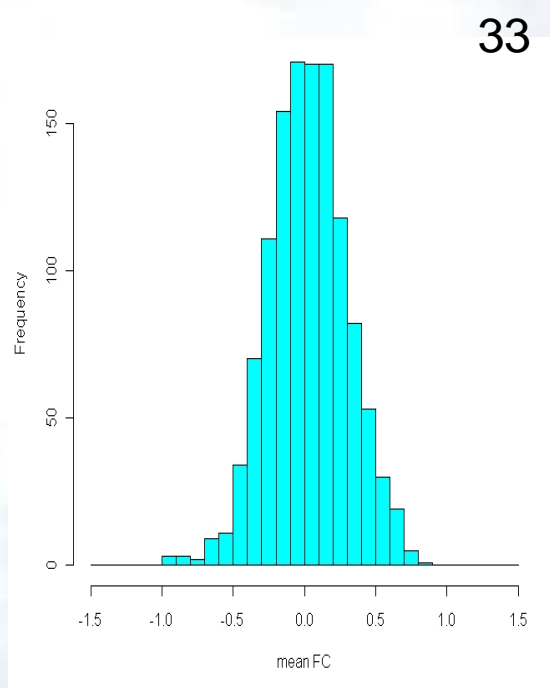
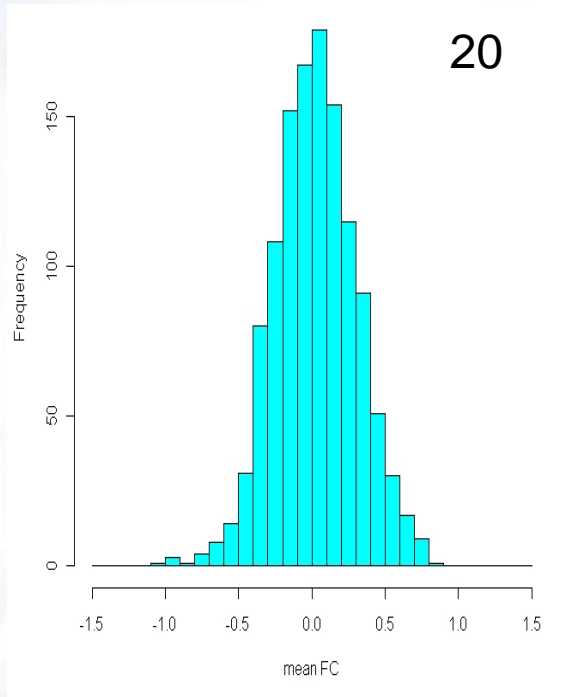
The main Question

What is the sample size required to ensure that the expected probability of classification (PCC) of classifier developed from training data is within a boundary ε (e.g 10%) of the optimal expected correct classification probability.

Mean foldchange

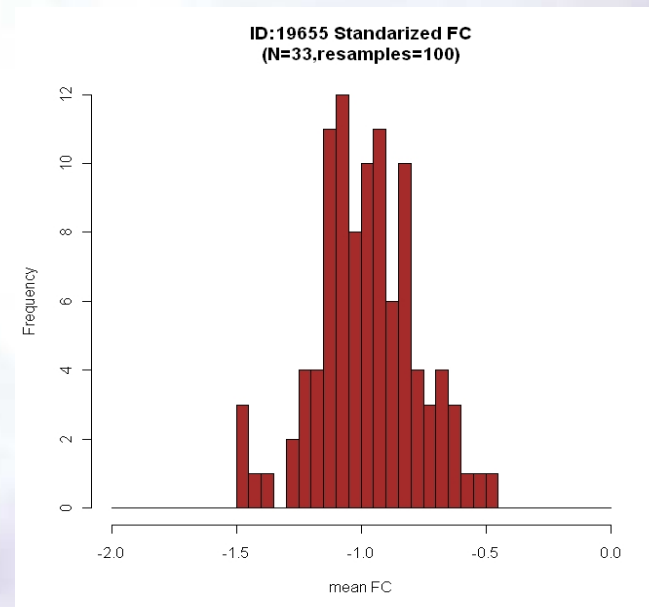
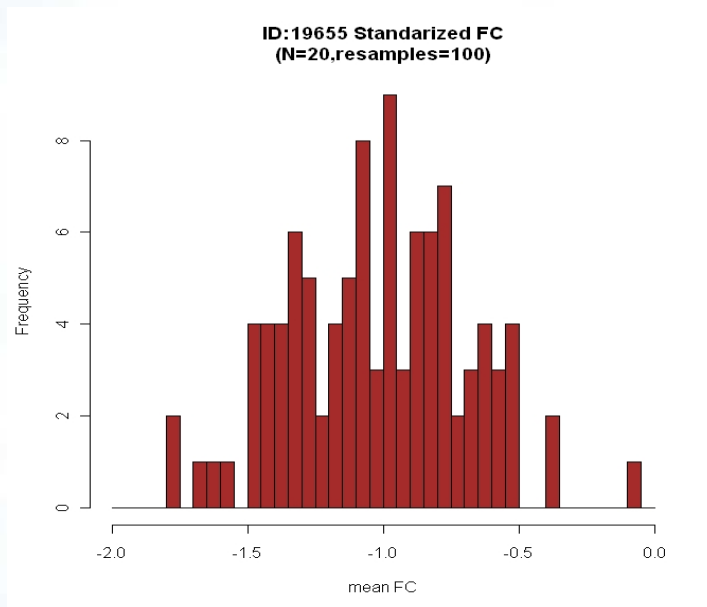
The mean FC of the 1216 markers resulting by setting the frequency at 0.3 and using 100 resamples of sizes 20 and 33. The FC for the full study is also reported

also reported



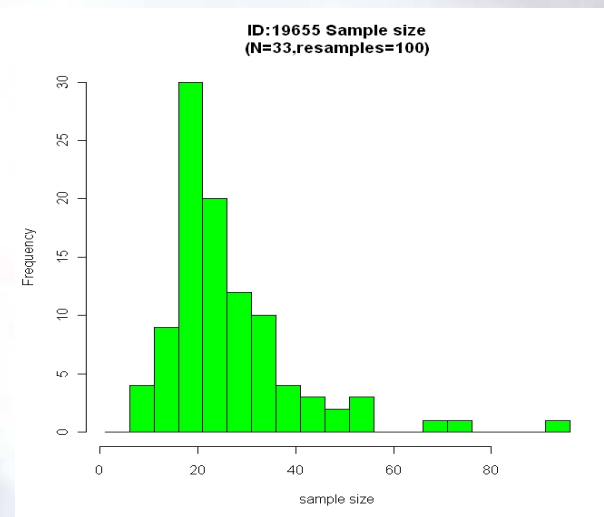
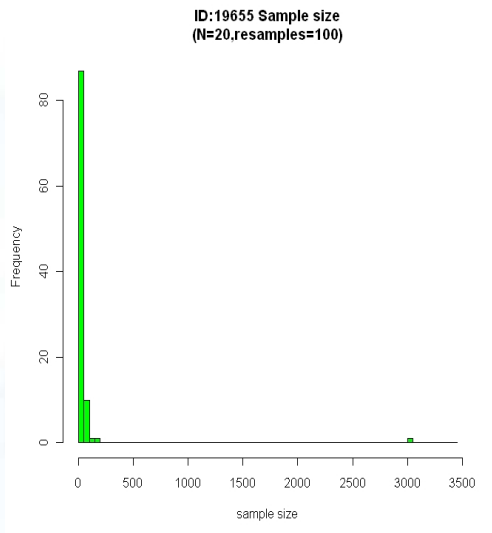
Typical FC of a DE marker

The FC of marker ID:19655 using 100 reamples of sizes 20 (meanFC=-1.021) and 33 (meanFC=-0.974). The FC from the full study (67) for this marker is -0.987.



Sample size for a marker

The sample size to detect the marker ID:19655 ($\alpha=0.05, \beta=0.9$) using 100 resamples of sizes 20 (meanSS=59) and 33 (meanSS=26). The sample size from the full study (67) for this marker is 23.



Sample size of the training set

Using the method described in Kevin Dobbin and Richard Simon, *Biostatistics* 8:101-17, 2007 we found that:

By requiring confidence >0.9

48 male and 47 female are required to obtain a classifier with a performance of $\varepsilon=5\%$ of the optimal classifier.

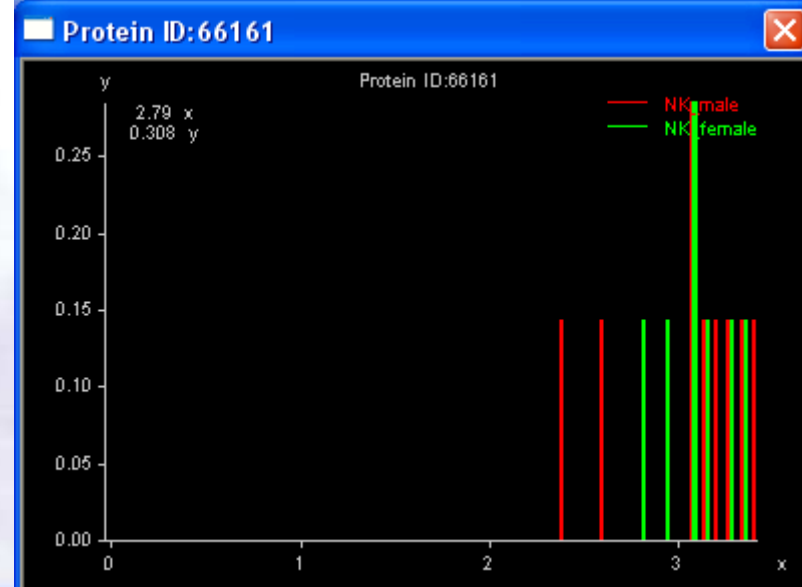
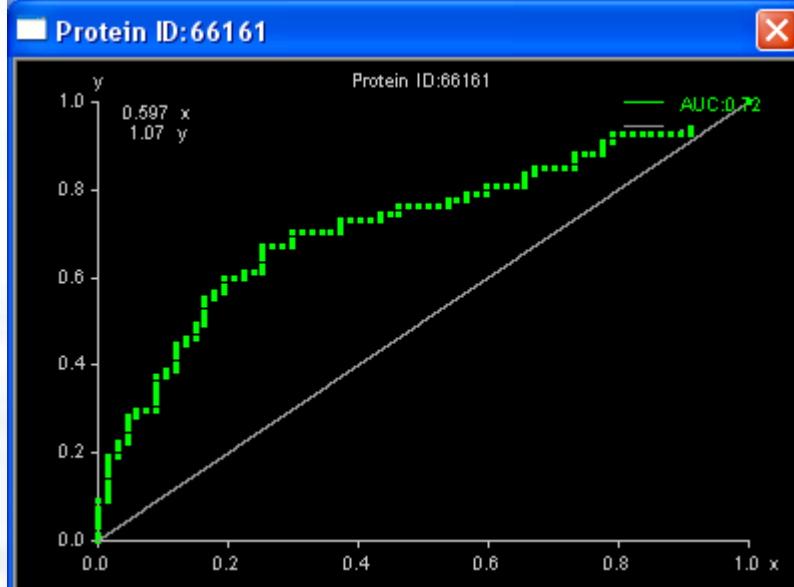
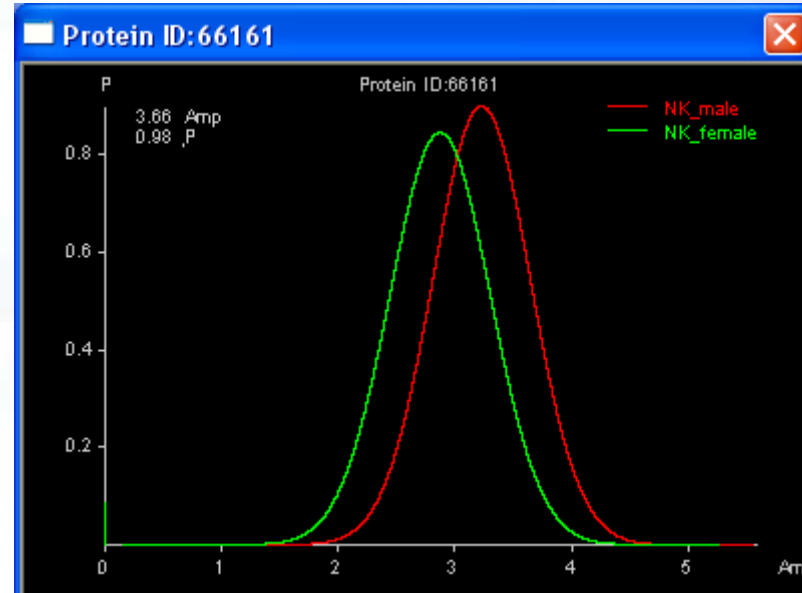
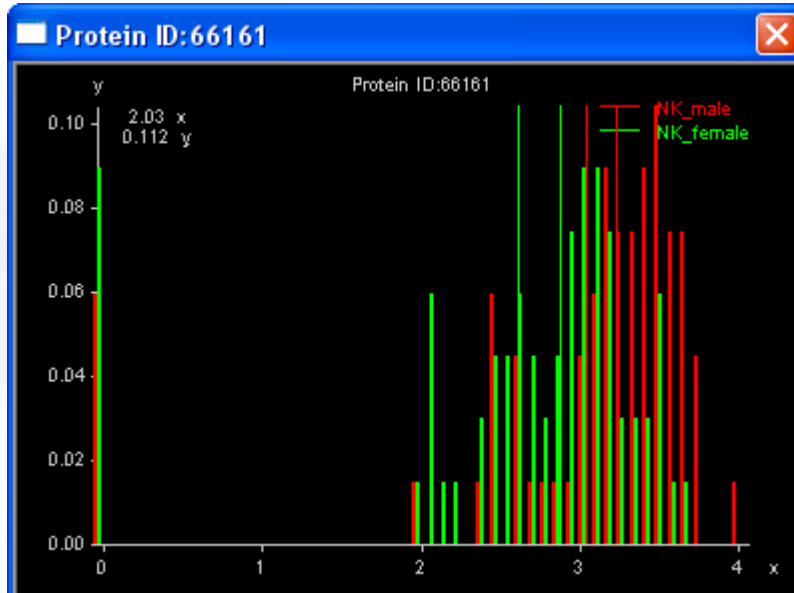
whereas

34 male and 33 female are required to obtain a classifier with a performance of $\varepsilon=10\%$ of the optimal classifier

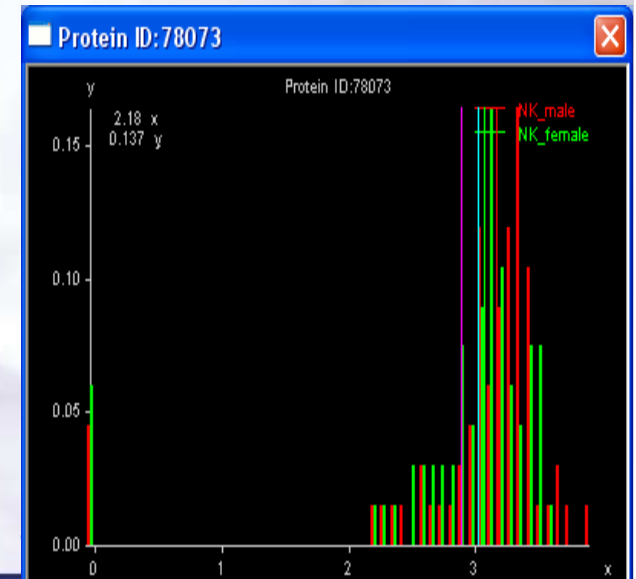
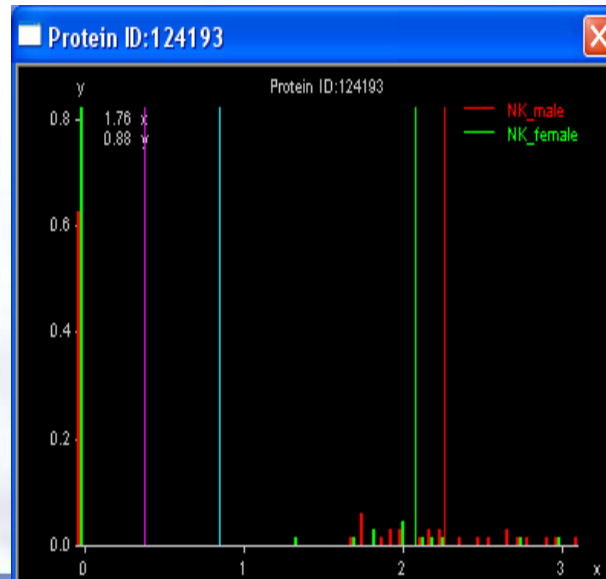
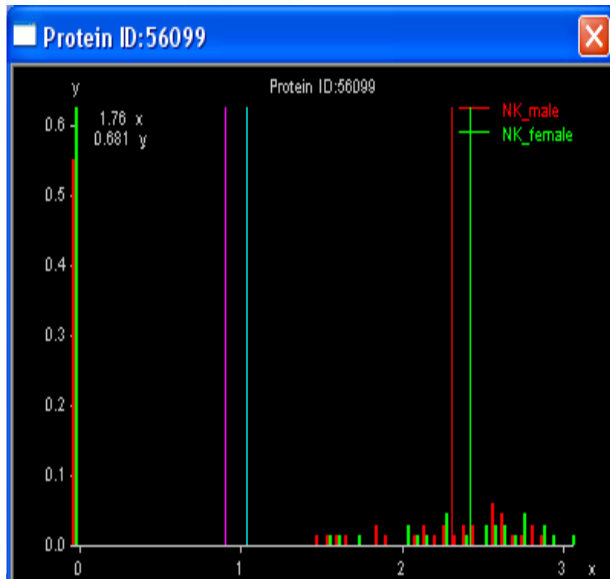
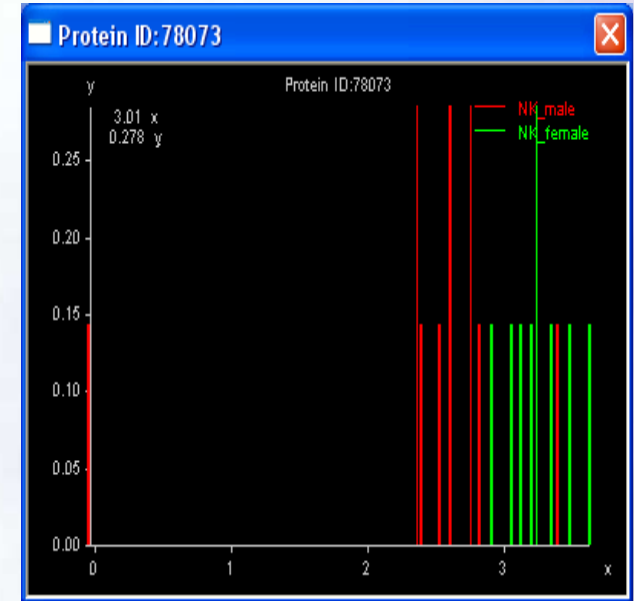
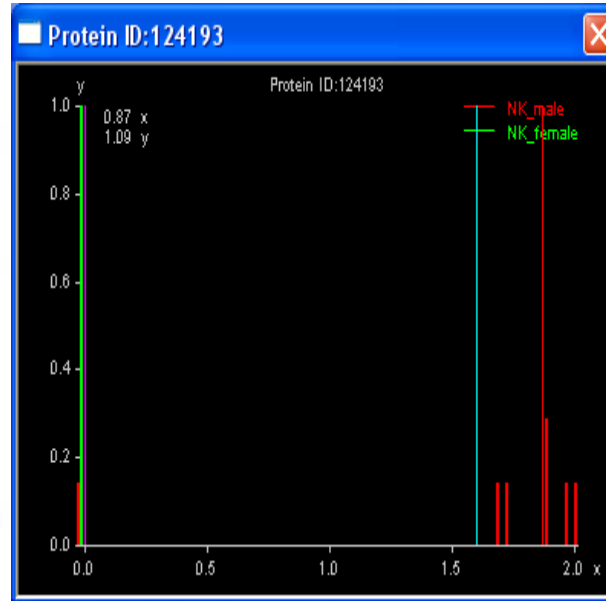
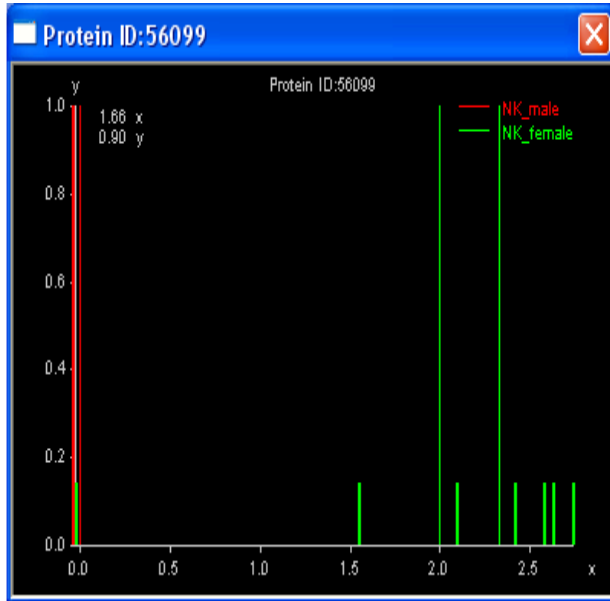
Results of statistical assessment

No of samples	No of biomarkers @ p-value < 0.05	No of biomarkers @ p-value < 0.05 after FDR adjustment	No of biomarkers @ p-value < 0.05 after Bonferroni	Biomarkers found valid (p-value < 0.05) in testset)
14	41	0	0	1 / 0 / 0
40	151	0	0	25 / 0 / 0
66	232	48	1	67 / 18 / 1
134	323	136	23	124 / 71 / 19

Biological variability requires large datasets for definition of valid biomarkers



Inappropriately low numbers of independent samples result in „erroneous biomarkers“

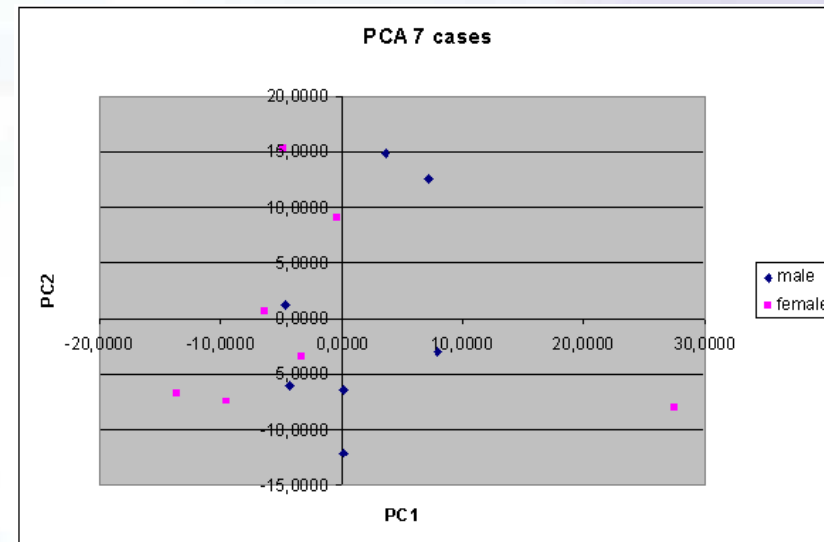
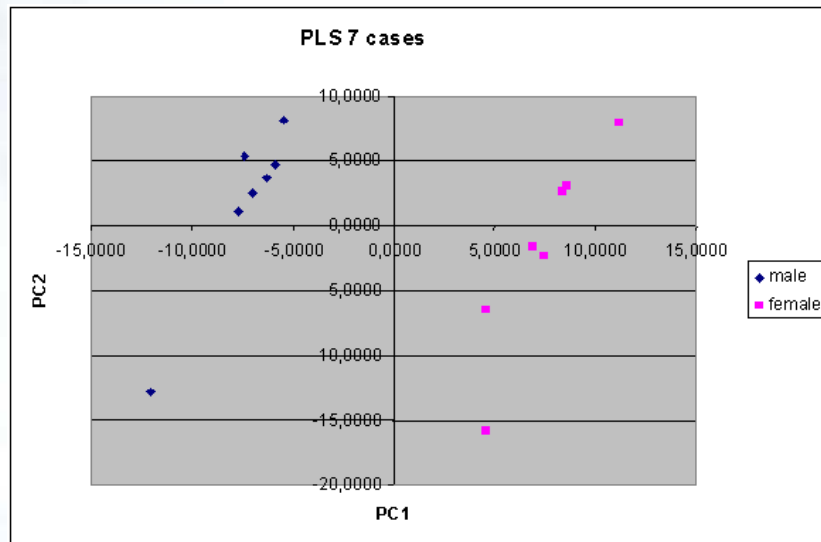


Conclusion

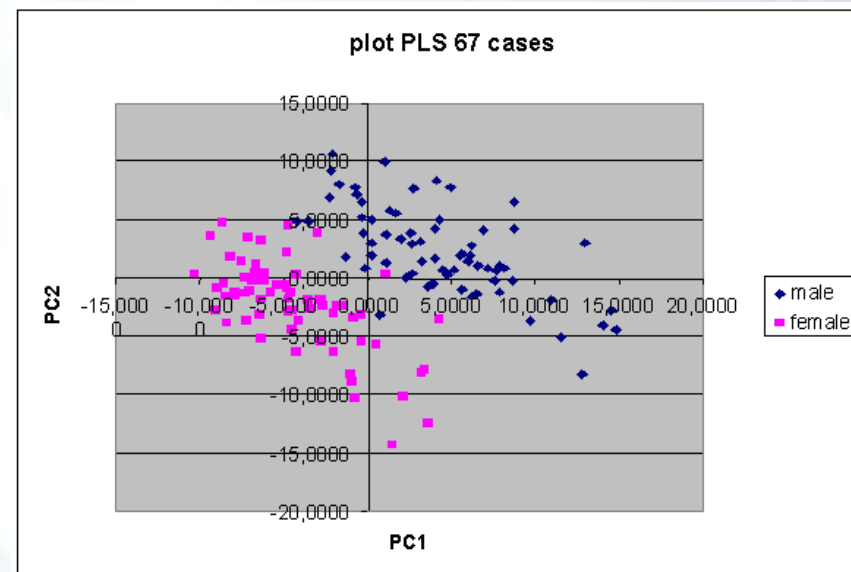
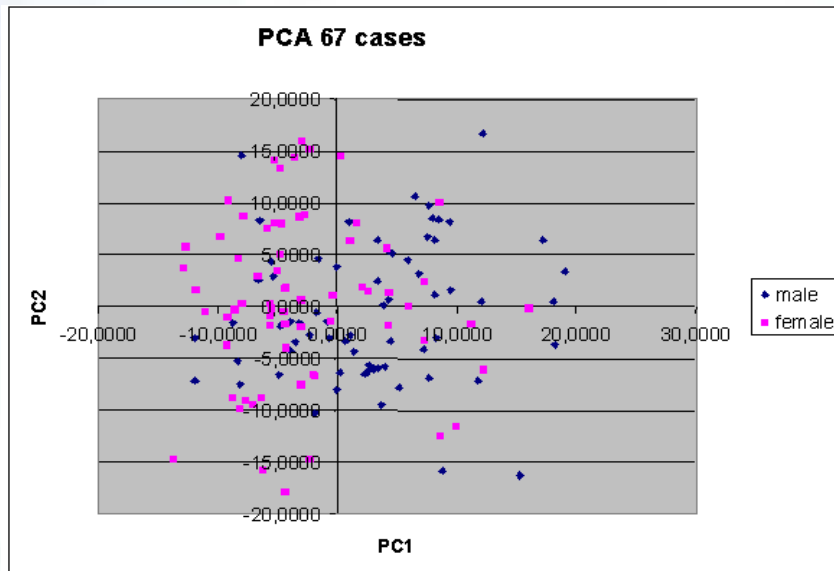
- Small number of datasets will result in erroneous determination of true distribution of features, results of such studies appear meaningless.
- Unadjusted p-values are essentially meaningless in a typical (multiparametric) proteomics experiment.
- Adjusted p-values still result in an overestimation of validity, likely due to generally unknown underlying bias.

Application of Principal Component Analysis PCA and Partial Least Squares (PLS) onto male/female dataset

Apparently good results of PCA and PLS on 7 cases and controls



Apparently good results of PLS on 67 cases and controls, failure of PCA



However, classification of the blinded test set was not successful with PCA when using the features selected with PLS

Conclusions

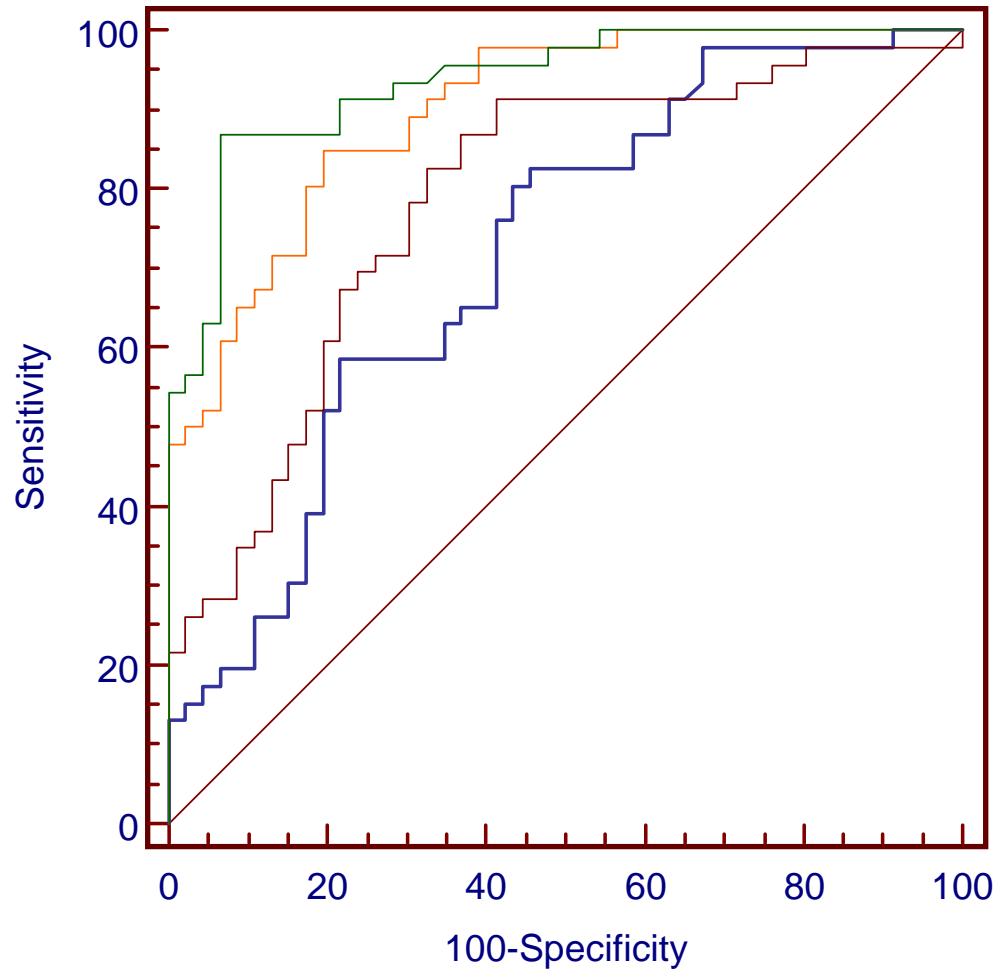
- PCA: not suitable, too much variability not related to the two classes
- PLS: suitable to explain the variability of a small sample population but not powerful enough to classify a real big population faultless
- → issue of missing values plays an important role

Results of classifier assessment in testset, based on 134 cases and controls

Classifier	error in testset
SVM	0.119
Random Forest	0.141
Ada boost	0.173
Linear	0.282
Decision Tree	0.315

Machinelearning algorithms outperform linear combination

Classification performance using different size training sets



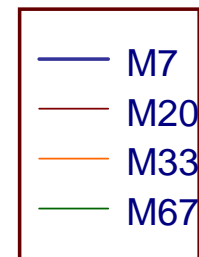
Results of crossvalidation in Trainingset:

M7: 100 % accuracy

M20 : 95% accuracy

M33: 84% accuracy

M67: 94% accuracy



M7 Area under the ROC curve = 0,715

M20 Area under the ROC curve = 0,786

M33 Area under the ROC curve = 0,900

M67 Area under the ROC curve = 0,936

M7 vs M20 Significance level P = 0,222

M20 vs M33 Significance level P = 0,005

M33 vs M67 Significance level P = 0,192

Summary

- Simple Wilcoxon or Students T-test are inappropriate for identification of useful biomarkers in a typical proteomics experiment, adjustment for multiple testing/false discovery rate is mandatory.
- Combination of biomarkers to a biomarker model can successfully be accomplished using any of the multidimensional algorithms tested (SVM, Random Forest, Adaboost, Baesian algorithm). These tools generally outperform linear combination and decision trees based on single biomarker.
- Crossvalidation of the trainingset is inappropriate to assess performance of machine learning tools, assessment in independent testset is mandatory.
- Machine learning tools do not require statistical assessment of data; larger feature sets generally result in increased performance. However: the minimum number of datasets required for the definition of true statistically valid biomarkers coincides with the number of datasets that is required by machine learning tools to achieve good performance, that cannot be significantly improved by inclusion of additional datasets.