STATISTICAL METHODS FOR CLASS PREDICTION

Olga Vitek

College of Science

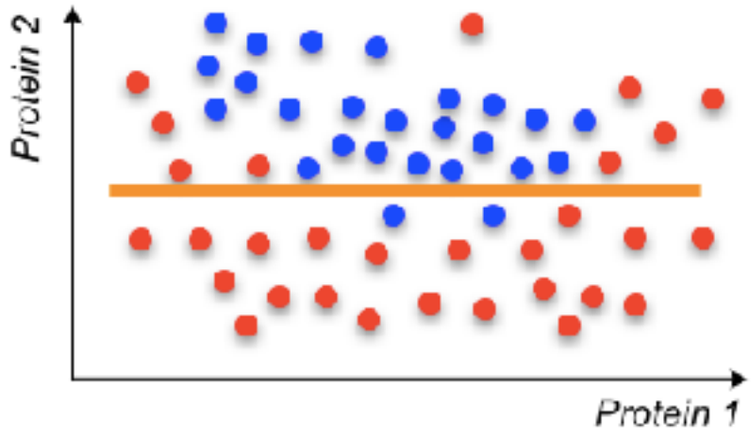
College of Computer and Information Science

Northeastern University

STATISTICAL GOAL 3: CLASS PREDICTION

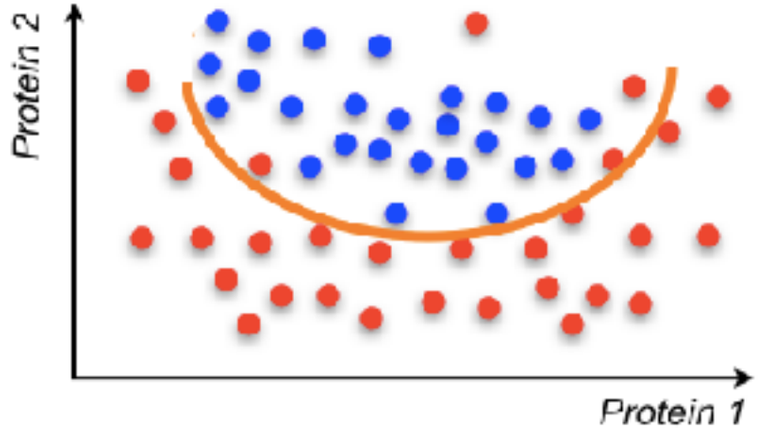
Classify each subject into a known group

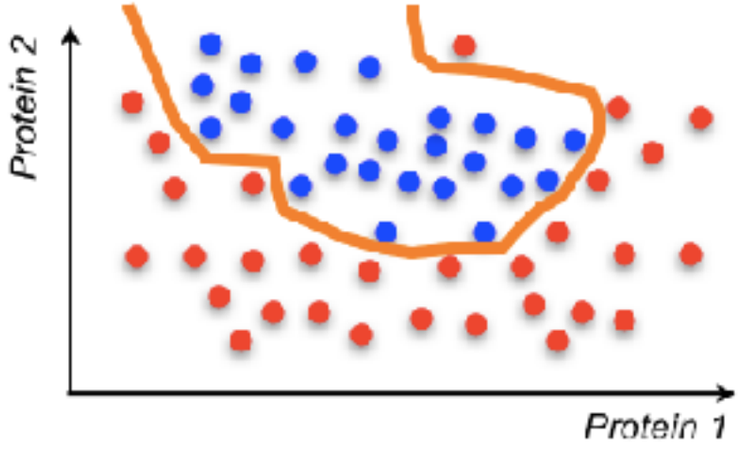
- Known class labels
  - Predict individual subjects
  - Report misclassification error (sensitivity, specificity, predictive value etc)
- Useful when focus on an individual
  - Tier I or Tier II biomarker discovery studies



Linear prediction rule

Quadratic prediction rule



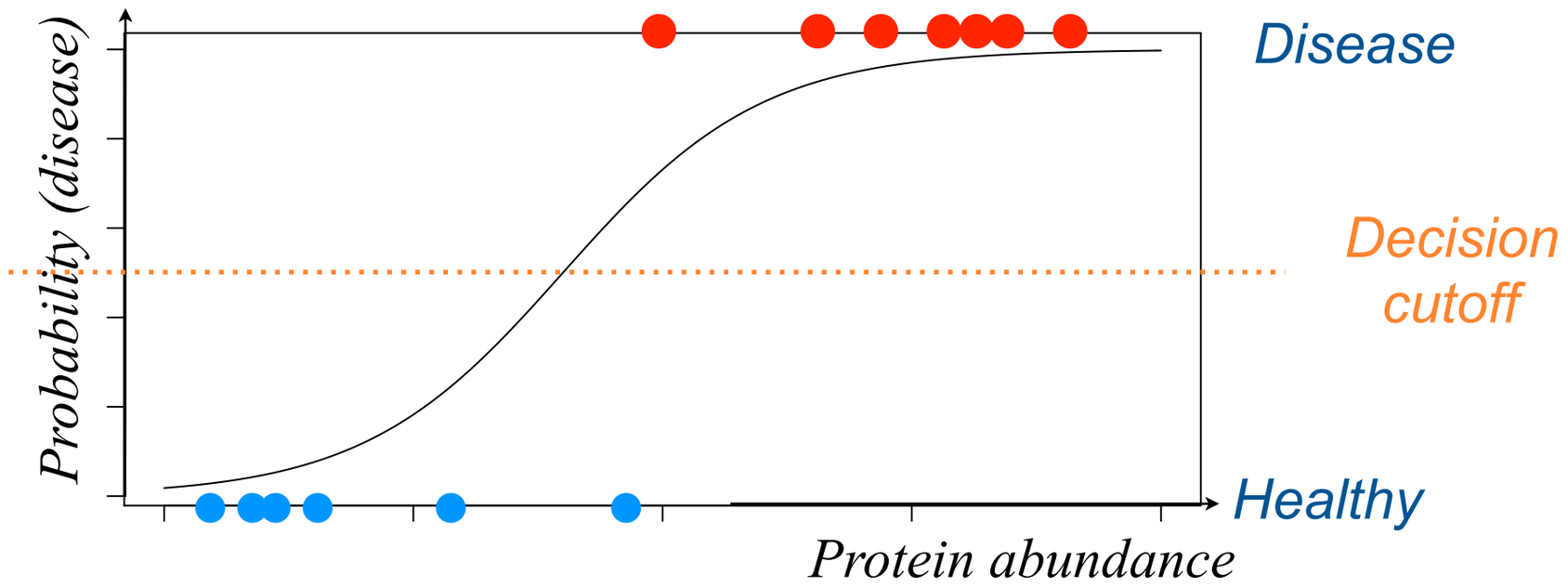


General prediction rule

OUTLINE

- Supervised classification
  - Classifiers, evaluation, issues
- Case studies
  - Alzheimer's diagnosis with plasma signaling proteins
  - Colorectal cancer diagnosis from plasma proteins
  - MACQ II: common practices in assay validation
- Experimental design for prediction
  - Impacts of dimensionality and noise

EXAMPLE: LOGISTIC REGRESSION



The model has a relatively simple form

$$P(disease) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 Protein)}}$$

Add terms to reflect multiple proteins and confounding factors

The parameters have an interpretation that is familiar to clinicians

$$\beta_1 = log \frac{Odds(Protein + 1)}{Odds(Protein)}$$

Selects a subset of proteins as predictors

EVALUATION OF PREDICTIVE ABILITY

Two classes, for a particular probability cutoff

Decision

Truth	Negative	Positive	Totals
Negative	TN	FP	N
Positive	FN	TP	P

Summaries from the decision table

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

One separate table for each probability cutoff

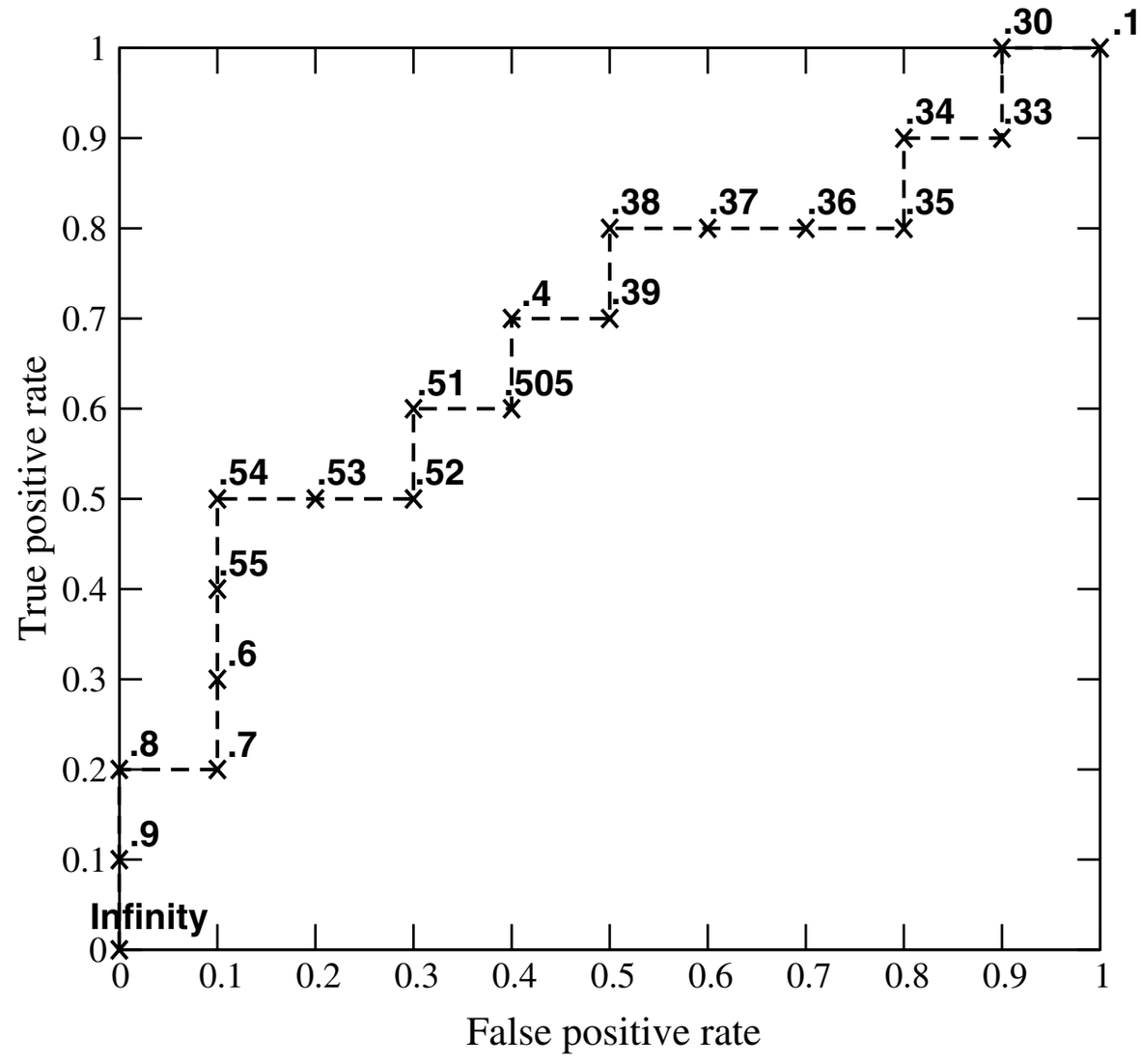
Fawcett, *'An introduction to ROC analysis'*, Pattern Recognition Letters, 2005

SUMMARY OF PREDICTIVE ABILITY

Varying predicted probability cutoff

Inst#	Class	Score	Inst#	Class	Score
1	**p**	.9	11	**p**	.4
2	**p**	.8	12	**n**	.39
3	**n**	.7	13	**p**	.38
4	**p**	.6	14	**n**	.37
5	**p**	.55	15	**n**	.36
6	**p**	.54	16	**n**	.35
7	**n**	.53	17	**p**	.34
8	**n**	.52	18	**n**	.33
9	**p**	.51	19	**p**	.30
10	**n**	.505	20	**n**	.1

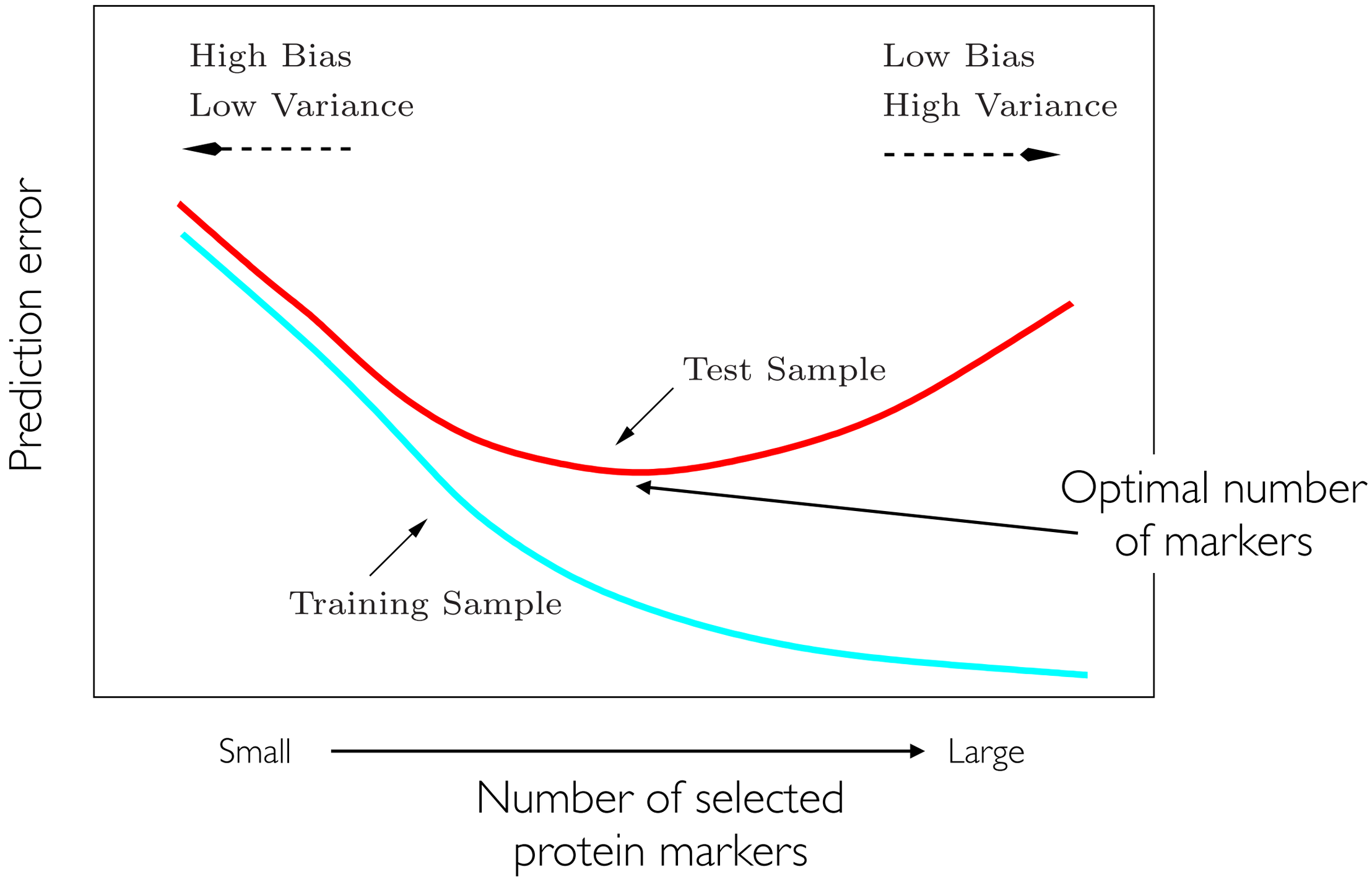
Subjects in the study and their predicted probability of the disease



***ROC curve**, summarizing the predictive ability for various predicted probability cutoffs*

SELECTION OF PREDICTIVE PROTEINS

Bias-variance trade-off



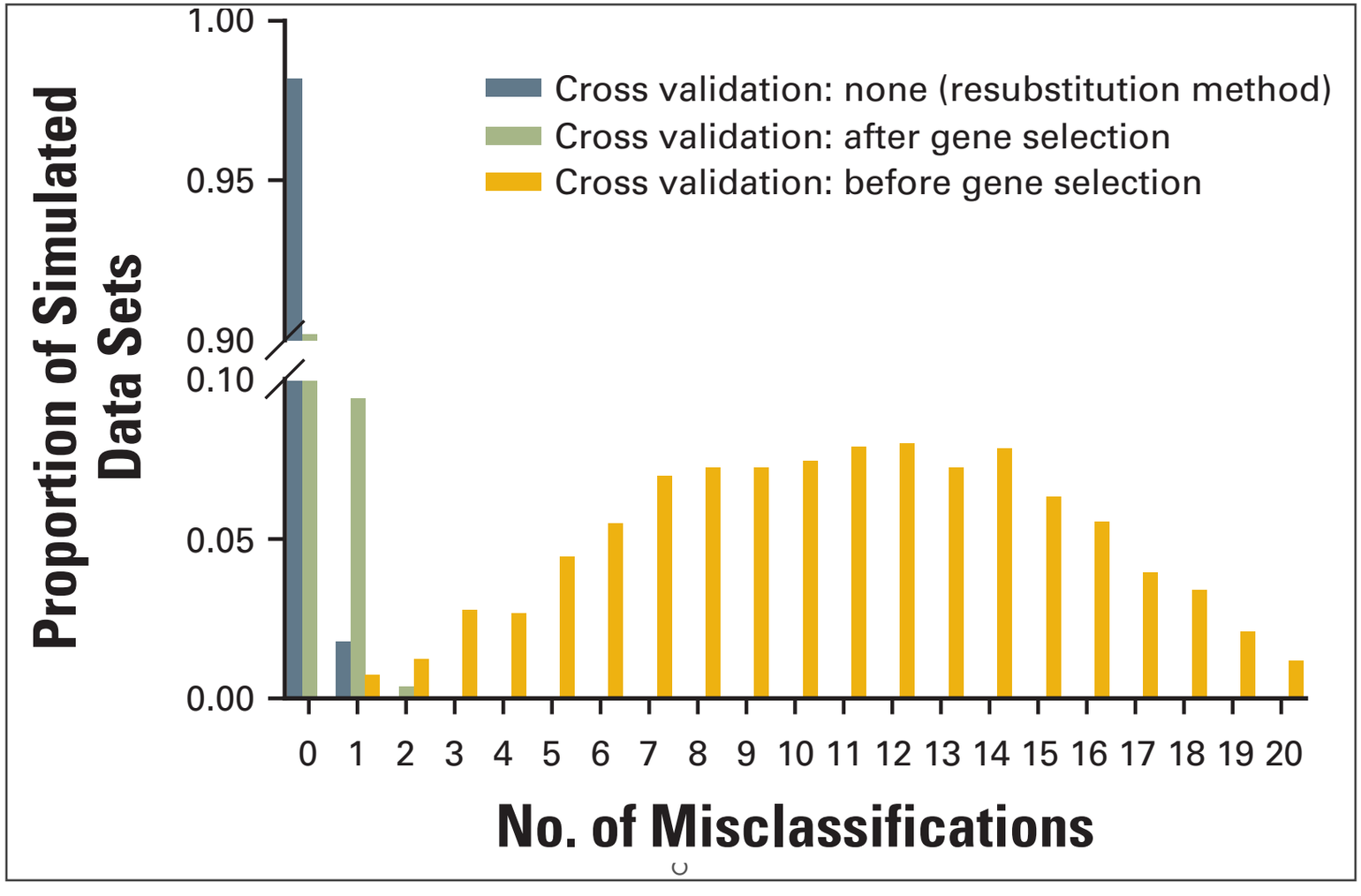
SELECTION OF PREDICTIVE PROTEINS

Should be done within (cross-) validation

- Simulated data with no structure
  - 20 observations with random labels
  - 6,000 possible but unrelated predictors
  - Repeated 200 times

- Estimated predictive accuracy using
  - no cross-validation
  - selecting features on full dataset, then using cross-validation
  - selecting features at each step of cross-validation

Simon et al,. JNCI 2003



OUTLINE

- Supervised classification
  - Classifiers, evaluation, issues
- Case studies
  - Alzheimer's diagnosis with plasma signaling proteins
  - Colorectal cancer diagnosis from plasma proteins
  - MACQ II: common practices in assay validation
- Experimental design for prediction
  - Impacts of dimensionality and noise

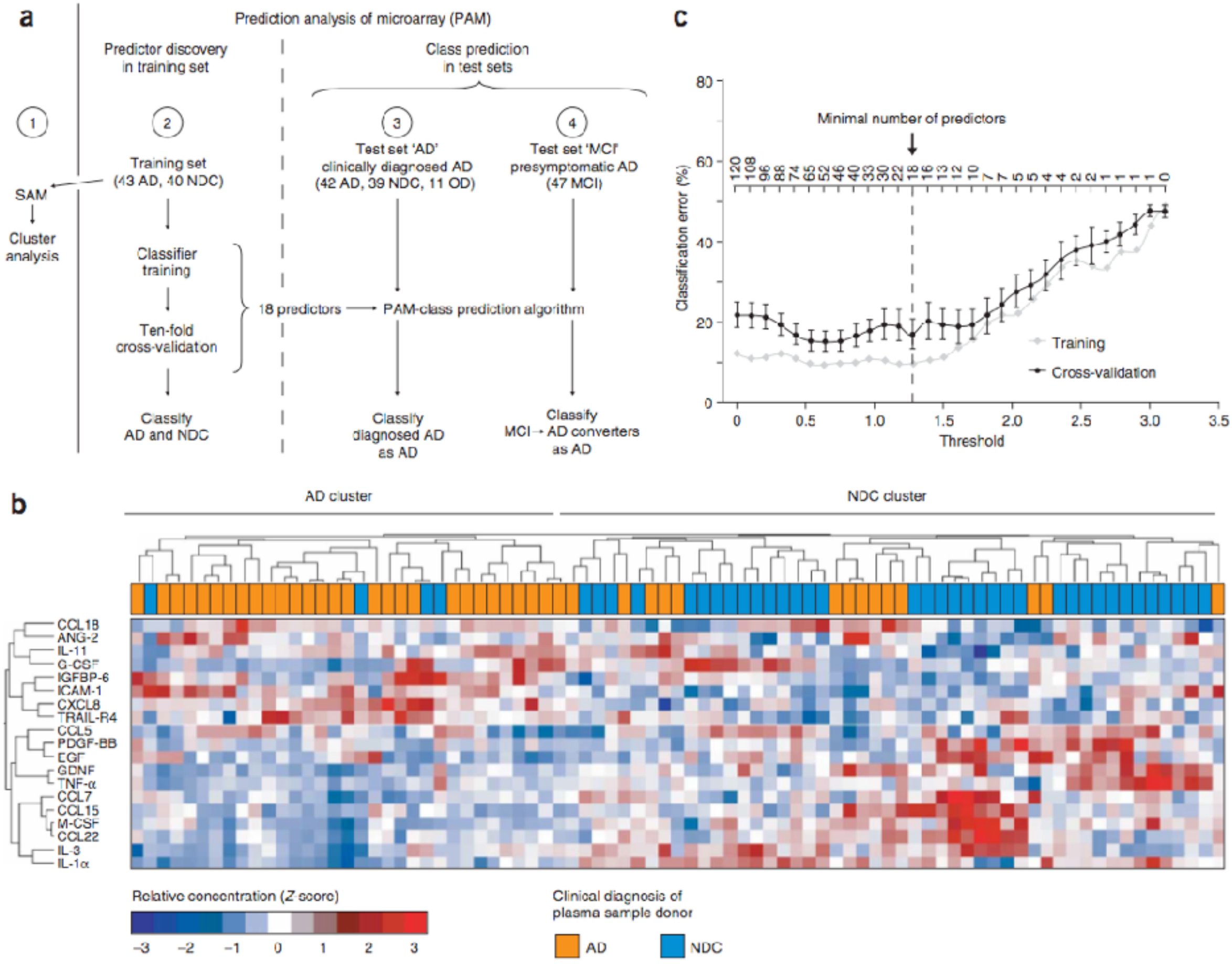
Case study 2: cytokine arrays

nature medicine

Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins

Sandip Ray[1,16], Markus Britschgi[2,16], Charles Herbert[1], Yoshiko Takeda-Uchimura[2], Adam Boxer[3], Kaj Blennow[4], Leah F Friedman[5], Douglas R Galasko[6], Marek Jutel[7], Anna Karydas[3], Jeffrey A Kaye[8], Jerzy Leszek[9], Bruce L Miller[3], Lennart Minthon[10], Joseph F Quinn[8], Gil D Rabinovici[3], William H Robinson[11], Marwan N Sabbagh[12], Yuen T So[2], D Larry Sparks[12], Massimo Tabaton[13], Jared Tinklenberg[5], Jerome A Yesavage[5], Robert Tibshirani[14] & Tony Wyss-Coray[2,15]

Ray et al., Nature Medicine, 2007


a
Prediction analysis of microarray (PAM)
Predictor discovery in training set
Class prediction in test sets
1
2
3
4
Training set (43 AD, 10 NDC)
Test set 'AD' clinically diagnosed AD (42 AD, 39 NDC, 11 OD)
Test set 'MCI' presymptomatic AD (47 MCI)
SAM
Cluster analysis
Classifier training
Ten-fold cross-validation
18 predictors
PAM-class prediction algorithm
Classify AD and NDC
Classify diagnosed AD as AD
Classify MCI→AD converters as AD
c
Minimal number of predictors
120 108 96 88 74 65 52 46 40 33 30 22 18 16 13 12 10 7 7 5 5 4 4 2 2 1 1 1 1 0
Classification error (%)
80
60
40
20
0
0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Threshold
Training
Cross-validation
b
AD cluster
NDC cluster
CCL18
ANG-2
IL-11
G-CSF
IGFBP-6
ICAM-1
CXCL8
TRAIL-R4
CCL5
PDGF-BB
EGF
GDNF
TNF-α
CCL7
CCL15
M-CSF
CCL22
IL-3
IL-1α
Relative concentration (Z score)
−3 −2 −1 0 1 2 3
Clinical diagnosis of plasma sample donor
AD
NDC

OUTLINE

- Supervised classification
  - Classifiers, evaluation, issues
- Case studies
  - Alzheimer's diagnosis with plasma signaling proteins
  - Colorectal cancer diagnosis from plasma proteins
  - MACQ II: common practices in assay validation
- Experimental design for prediction
  - Impacts of dimensionality and noise

CASE STUDY

Research Article

TRANSPARENT PROCESS OPEN ACCESS

EMBO Molecular Medicine

Prediction of colorectal cancer diagnosis based on circulating plasma proteins

Silvia Surinova[1,†], Meena Choi[2], Sha Tao[3,‡], Peter J Schüffler[4], Ching-Yun Chang[2], Timothy Clough[2], Kamil Vysloužil[5], Marta Khoylou[6], Josef Srovnal[6], Yansheng Liu[1], Mariette Matondo[1], Ruth Hüttenhain[1], Hendrik Weisser[1], Joachim M Buhmann[4], Marián Hajdúch[6], Hermann Brenner[3,7], Olga Vitek[2,8,9,*] & Ruedi Aebersold[1,10,**]

Research Article

TRANSPARENT PROCESS OPEN ACCESS

EMBO Molecular Medicine

Non-invasive prognostic protein biomarker signatures associated with colorectal cancer

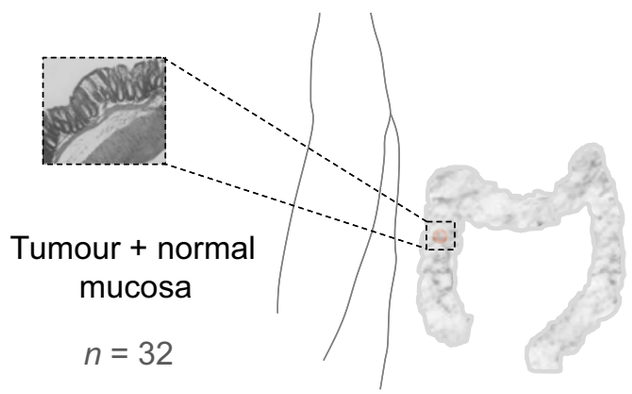
Silvia Surinova[1,†], Lenka Radová[2,‡], Meena Choi[3], Josef Srovnal[2], Hermann Brenner[4,5], Olga Vitek[3,6,7,*], Marián Hajdúch[2] & Ruedi Aebersold[1,8,**]

OVERVIEW

A

Phase 1

Candidate discovery



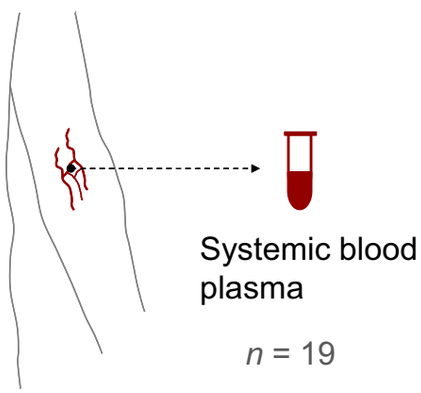
Tumour + normal mucosa

n = 32

CRC glycoproteome profiling by discovery-driven MS
2301 peptides and 673 proteins identified
334 protein candidates selected

Phase 2

Candidate screening



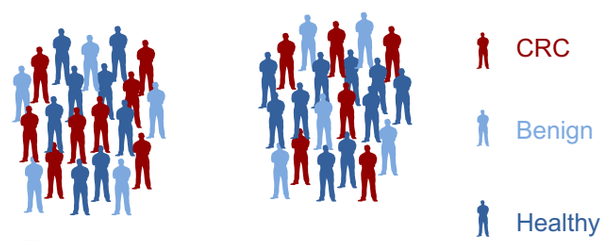
Systemic blood plasma

n = 19

CRC candidate screening in the patient circulation by targeted MS
88 protein candidates consistently detected and quantified

Phase 3

Biomarker signature development



Training cohort

n = 200

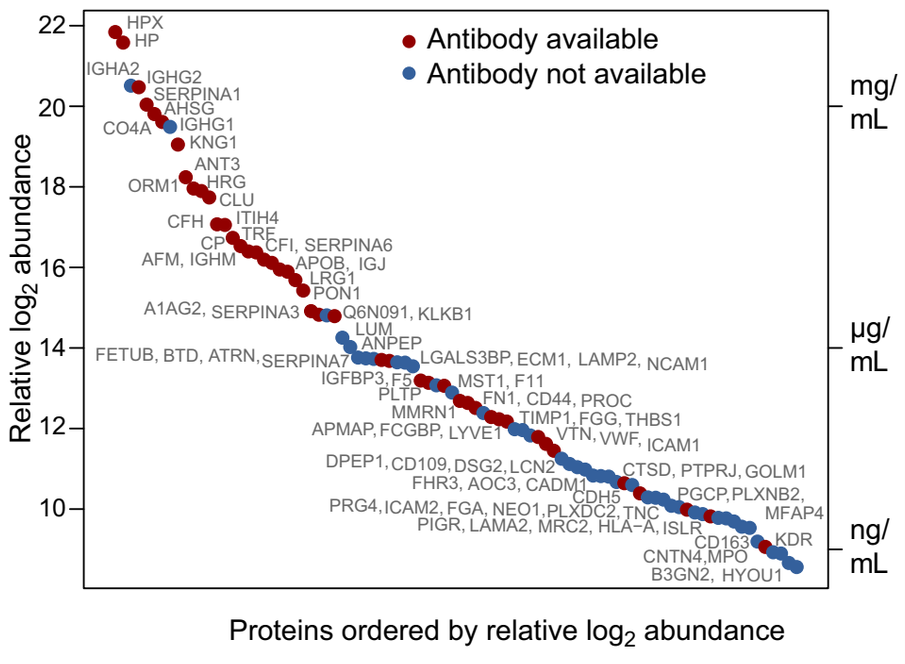
Validation cohort

n = 269

Prediction of CRC and control cases in two clinical cohorts
90-plex quantification method
5-protein biomarker signature

B

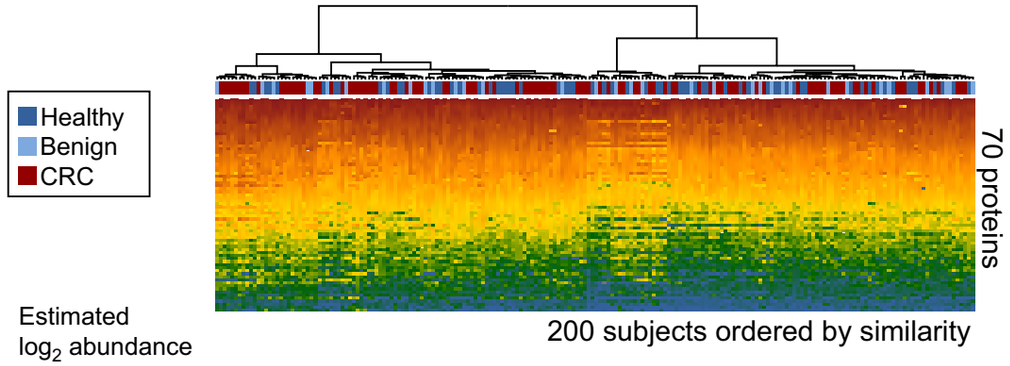
Abundance range of 88 candidate proteins in phase 2



Proteins ordered by relative $\log_2$ abundance

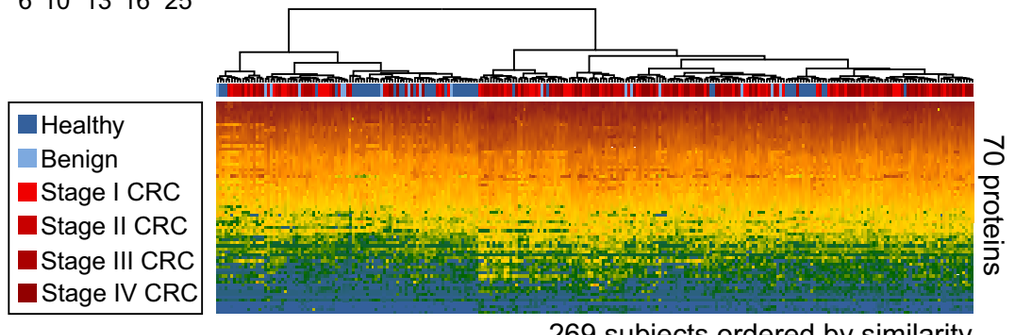
C

Training cohort in phase 3



200 subjects ordered by similarity

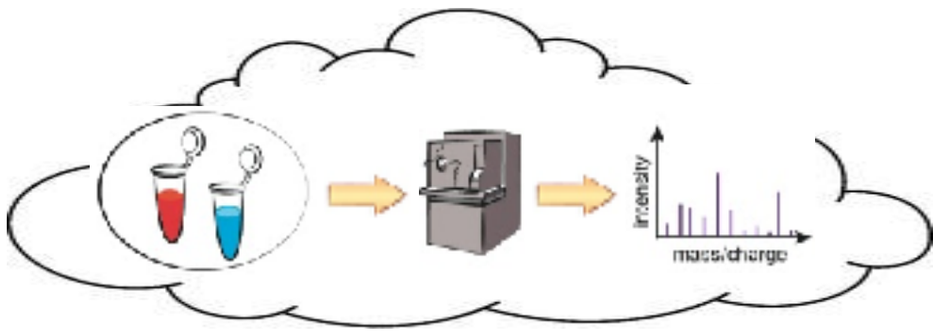
Validation cohort in phase 3



269 subjects ordered by similarity

ANALYSIS WORKFLOW WITH MSSTATS

Experimental design



QC and normalization

Statistical modeling

Sample quantification

Classifier development

Performance evaluation

MSstats

Statistical Tool For Quantitative Mass Spectrometry-Based Proteomics

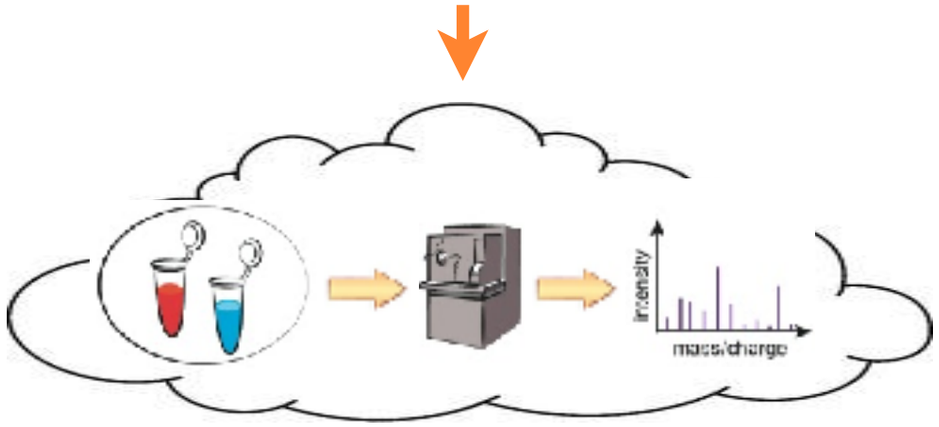
M. Choi et al. *BMC Bioinformatics., 2014*

Skyline external tool

Bioconductor

ANALYSIS WORKFLOW WITH MSSTATS

Experimental design



QC and normalization

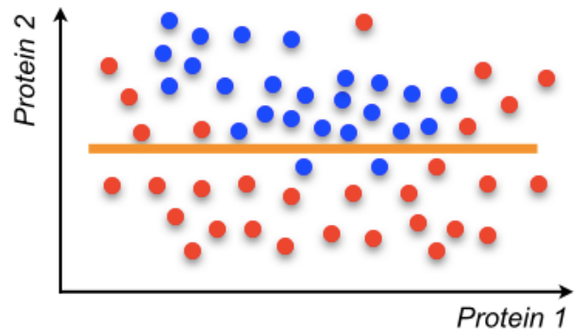
Statistical modeling

Sample quantification

Classifier development

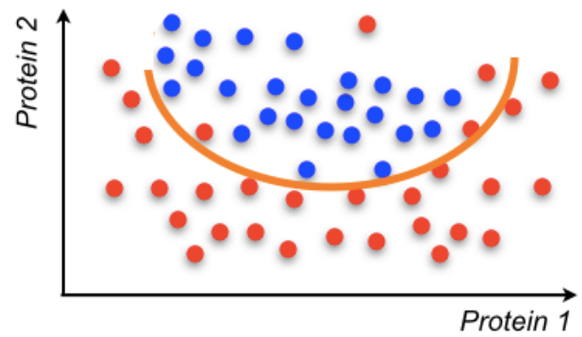
Performance evaluation

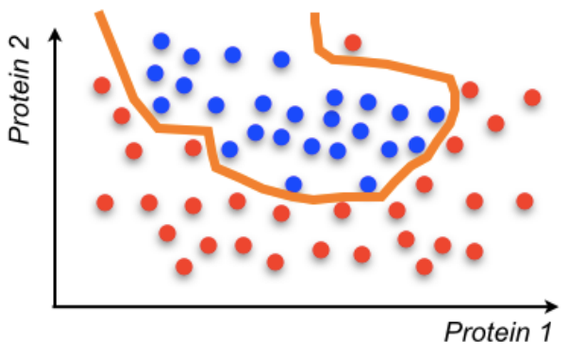
- Objective: Train a model that can classify each subject into correct class.



Linear prediction rule

Quadratic prediction rule





General prediction rule

SELECTION OF BIOMARKER PANEL

Phase 3, training cohort: 100 CRC and 100 control subjects

Step 1 **Discovery of a multivariate signature within 10-fold cross-validation**

- Systematically rotate samples between 10 folds
  - Select differentially abundant proteins on 9/10 of samples ► about 20 proteins per fold
  - Fit logistic regression models with stepwise protein selection ► 4-5 proteins per model
  - Predict the disease status of 1/10 'left-out' samples ►AUROC (mean=0.62, SE=0.04)

  10x
- Generate a consensus logistic regression model with proteins selected in ≥ 5 folds:

  $\ln\{\text{Prob}(\text{CRC})/(1-\text{Prob}(\text{CRC}))\} = -15.15 + 0.73\times\text{CP} - 1.06\times\text{PON1} + 0.69\times\text{SERPINA3} + 0.37\times\text{LRG1} + 0.41\times\text{TIMP1}$

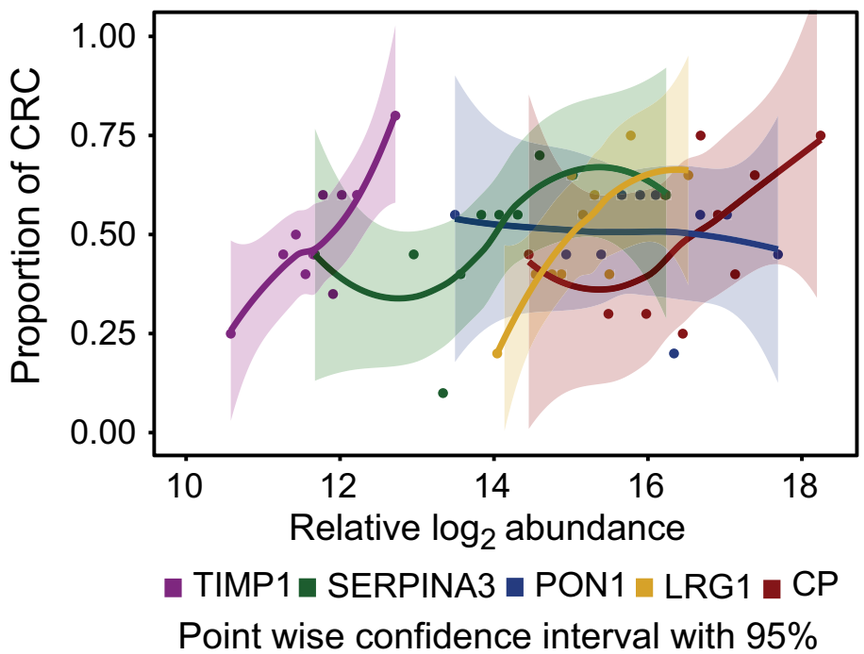
Step 2 **Characterization of selected candidate biomarkers on the full training cohort**

- Quantify protein fold changes and standard errors

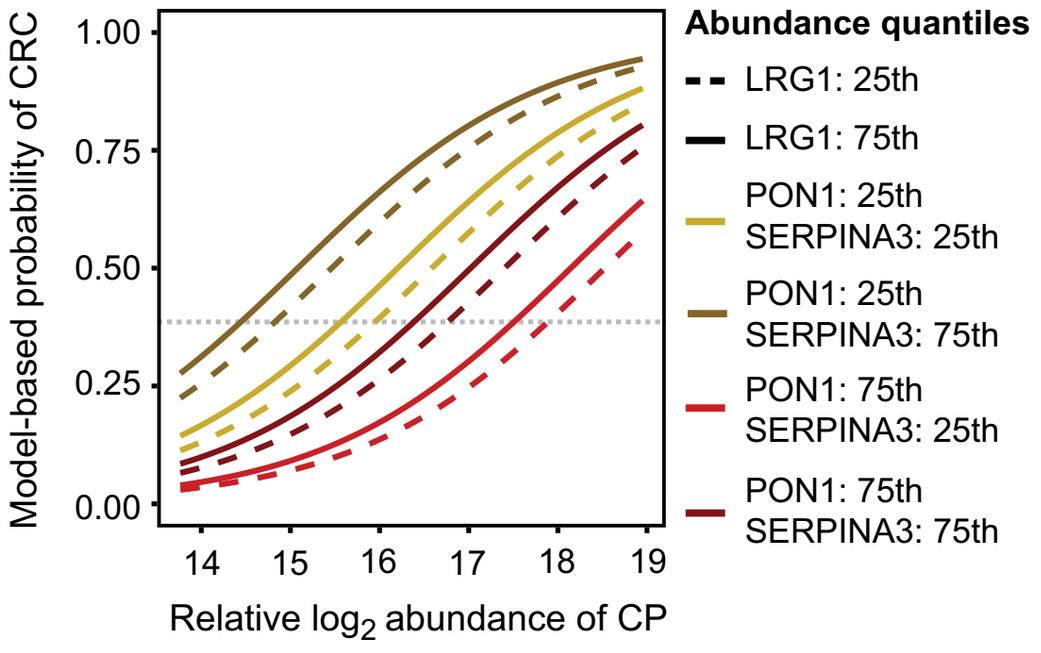
	CP	PON1	SERPINA3	LRG1	TIMP1
Fold change (SE)	1.3 (0.05)	0.9 (0.04)	1.3 (0.06)	1.2 (0.03)	1.2 (0.03)

- Illustrate the range of protein intensities and their univariate contribution to disease (plot 1)
- Illustrate model-based estimated probability of disease (plot 2)

Plot 1



Plot 2



EVALUATION OF BIOMARKER PANEL

Phase 3, validation cohort: 202 CRC and 67 control subjects

Step 3 **Assessment of the predictive ability of the biomarker signature**

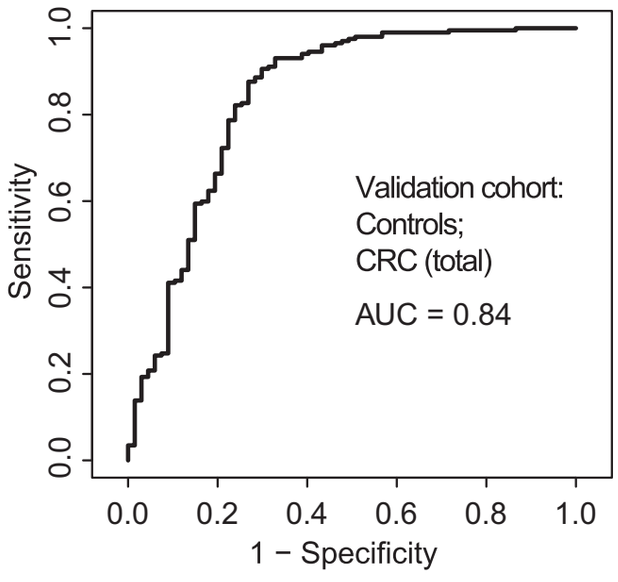
- Predict CRC and controls on the full validation dataset ►AUROC=0.84

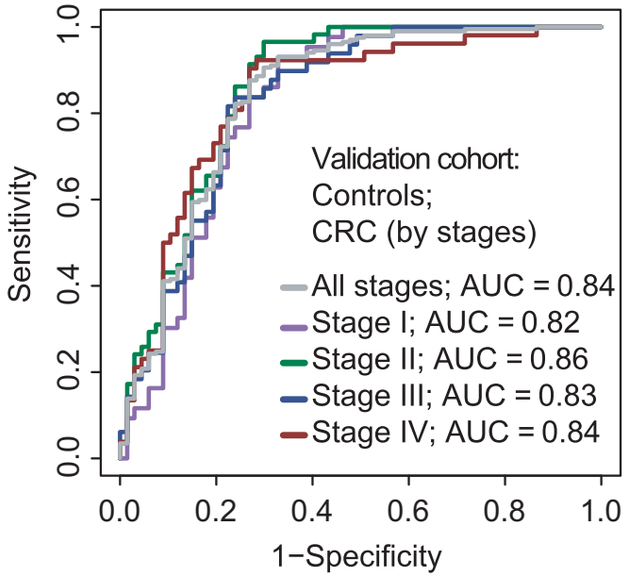
Characterization of selected candidate biomarkers on the validation cohort

Step 4

- Evaluate the reproducibility of fold changes and standard errors

	CP	PON1	SERPINA3	LRG1	TIMP1
Fold change (SE)	1.6 (0.03)	0.8 (0.03)	1.2 (0.04)	1.5 (0.04)	1.4 (0.04)





Summary statistics

		99% CI
AUC	0.84	0.75 – 0.92
Specificity	0.79	0.64 – 0.91
Sensitivity	0.70	0.38 – 0.95
Accuracy	0.72	0.65 – 0.79

AUC comparisons

	p-value				99% CI
Stage	II	III	IV	All	
I	0.49	0.87	0.75	0.74	0.72–0.92
II		0.59	0.71	0.68	0.76–0.94
III			0.88	0.87	0.71–0.92
IV				1.00	0.73–0.93

OUTLINE

- Supervised classification
  - Classifiers, evaluation, issues
- Case studies
  - Alzheimer's diagnosis with plasma signaling proteins
  - Colorectal cancer diagnosis from plasma proteins
  - MACQ II: common practices in assay validation
- Experimental design for prediction
  - Impacts of dimensionality and noise

Case study 1: MACQCII

Shi et al., 2010

nature biotechnology

ARTICLES

The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models

MAQC Consortium*

Gene expression data from microarrays are being applied to predict preclinical and clinical endpoints, but the reliability of these predictions has not been established. In the MAQC-II project, 36 independent teams analyzed six microarray data sets to generate predictive models for classifying a sample with respect to one of 13 endpoints indicative of lung or liver toxicity in rodents, or of breast cancer, multiple myeloma or neuroblastoma in humans. In total, >30,000 models were built using many combinations of analytical methods. The teams generated predictive models without knowing the biological meaning of some of the endpoints and, to mimic clinical reality, tested the models on data that had not been used for training. We found that model performance depended largely on the endpoint and team proficiency and that different approaches generated models of similar performance. The conclusions and recommendations from MAQC-II should be useful for regulatory agencies, study committees and independent investigators that evaluate methods for global gene expression analysis.

Negative controls:
randomly assigned endpoints I & M

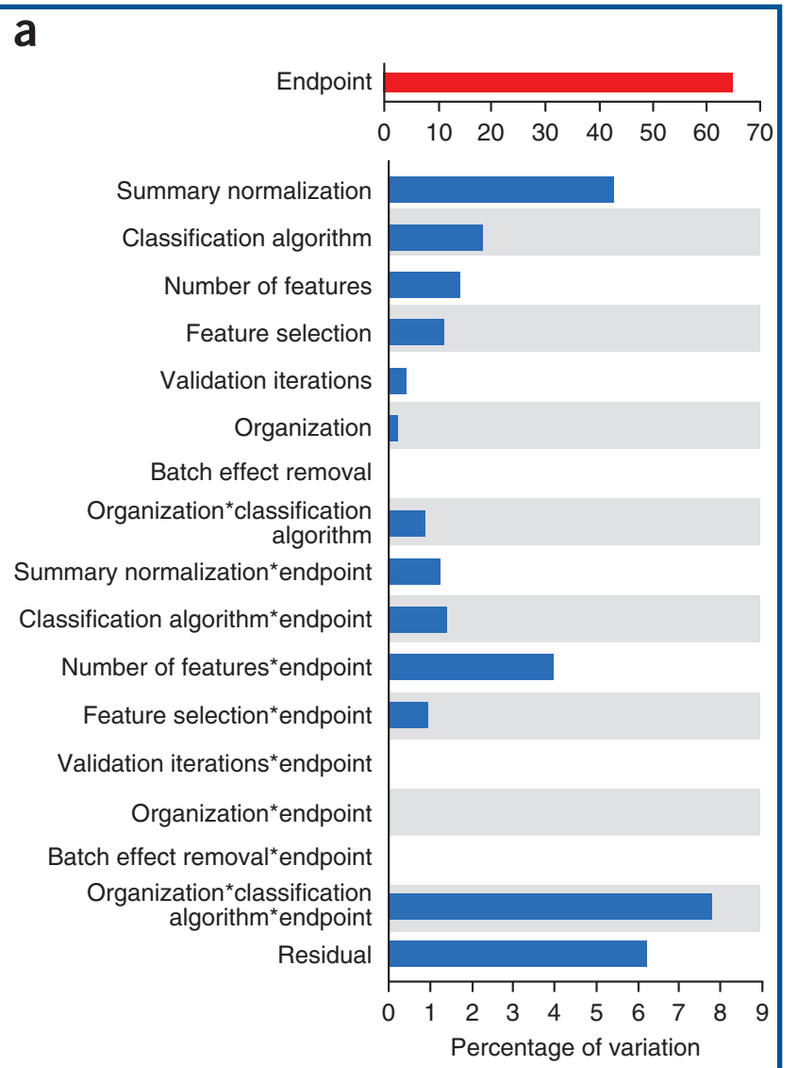
Positive controls:
gender H & L

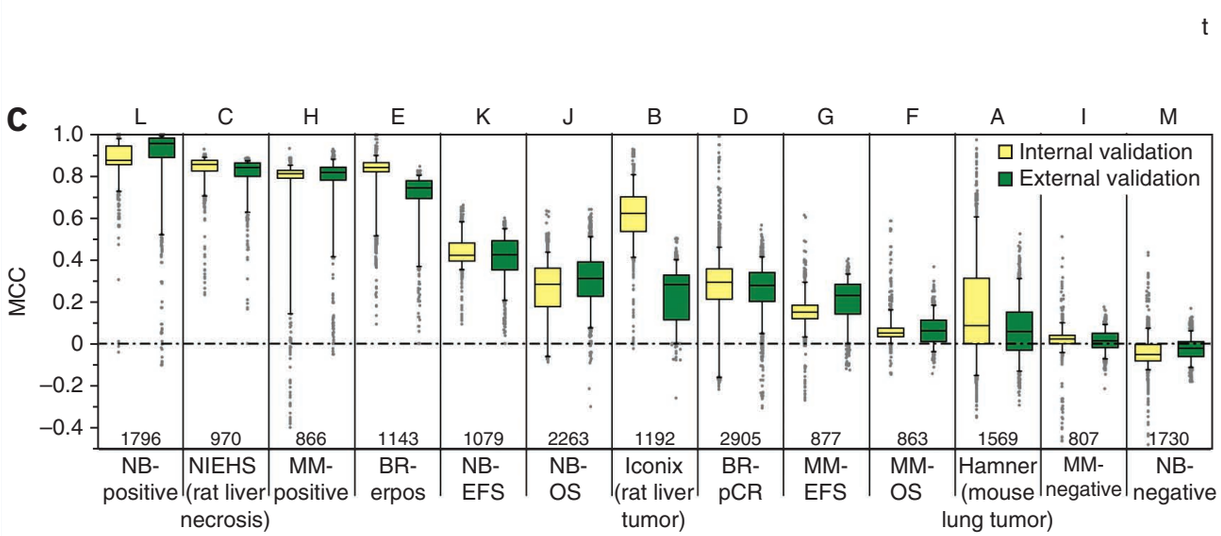
Summary of the results

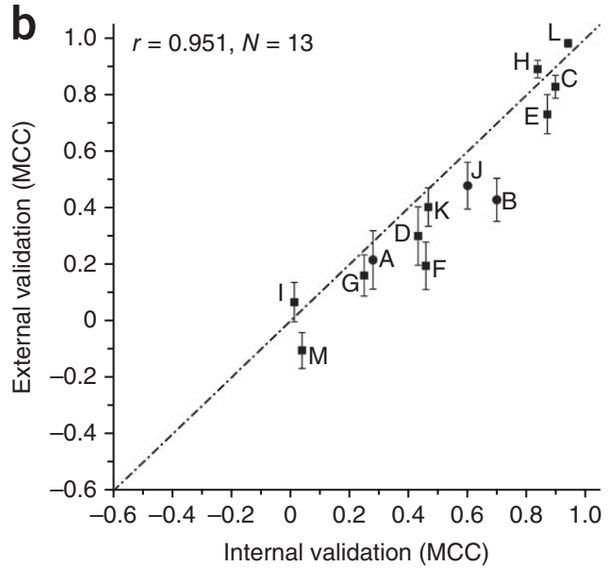
Table 2 Modeling factor options frequently adopted by MAQC-II data analysis teams

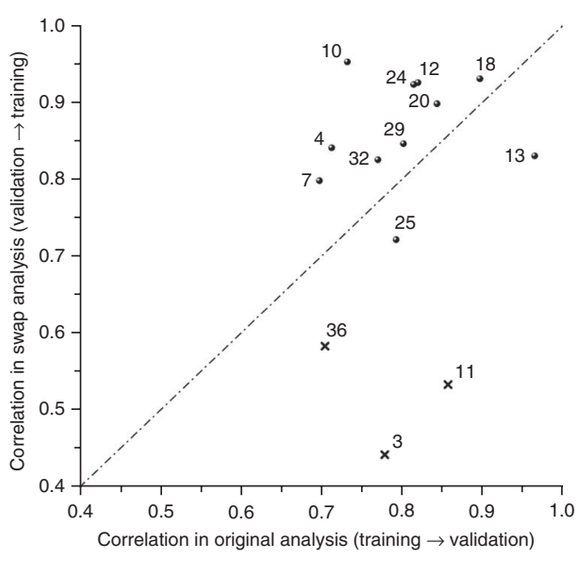
		Original analysis (training => validation)		
Modeling factor	Option	Number of teams	Number of endpoints	Number of models
Summary and normalization	Loess	12	3	2,563
	RMA	3	7	46
	MAS5	11	7	4,947
Batch-effect removal	None	10	11	2,281
	Mean shift	3	11	7,279
Feature selection	SAM	4	11	3,771
	FC+P	8	11	4,711
	T-Test	5	11	400
	RFE	2	11	647
Number of features	0~9	10	11	393
	10~99	13	11	4,445
	≥1,000	3	11	474
	100~999	10	11	4,298
Classification algorithm	DA	4	11	103
	Tree	5	11	358
	NB	4	11	924
	KNN	8	11	6,904
	SVM	9	11	986

Analytic options used by two or more of the 14 teams that submitted models for all endpoints in both the original and swap experiments. RMA, robust multichip analysis; SAM, significance analysis of microarrays; FC, fold change; RFE, recursive feature elimination; DA, discriminant analysis; Tree, decision tree; NB, naive Bayes; KNN, K-nearest neighbors; SVM, support vector machine.







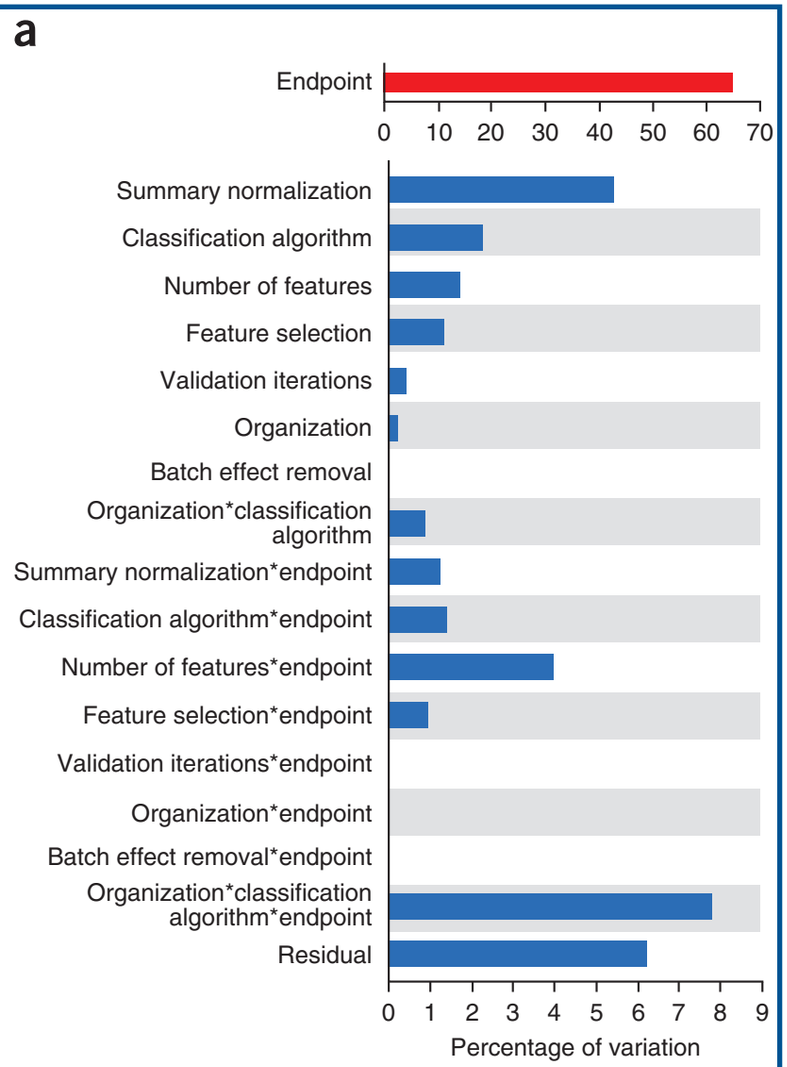


Summary of the results

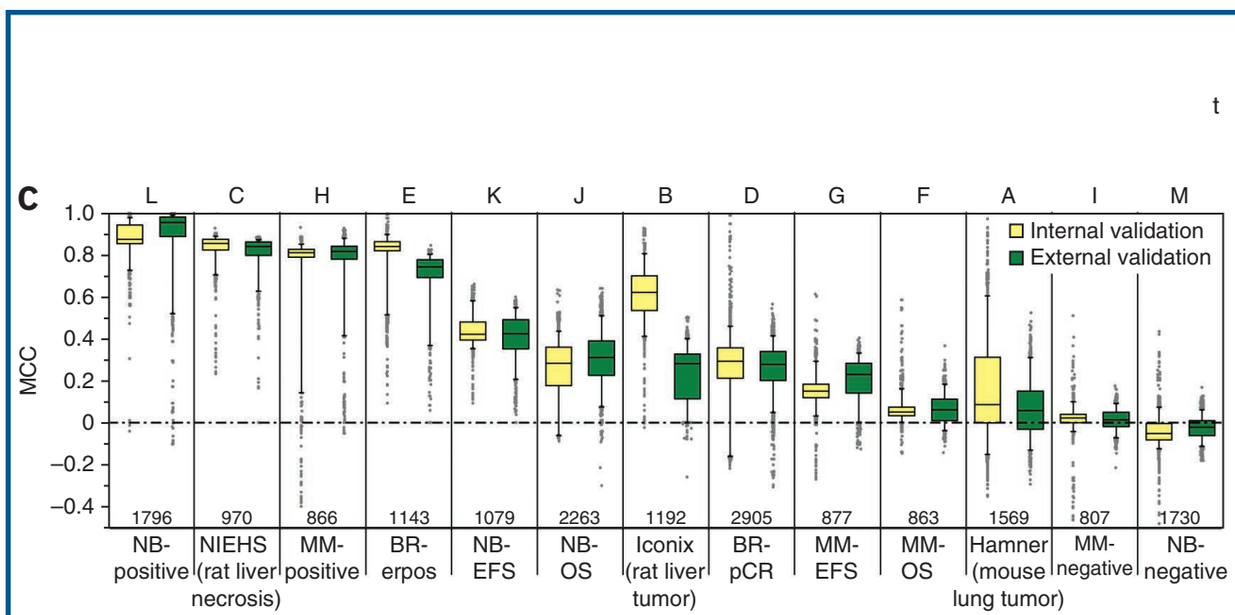
Table 2 Modeling factor options frequently adopted by MAQC-II data analysis teams

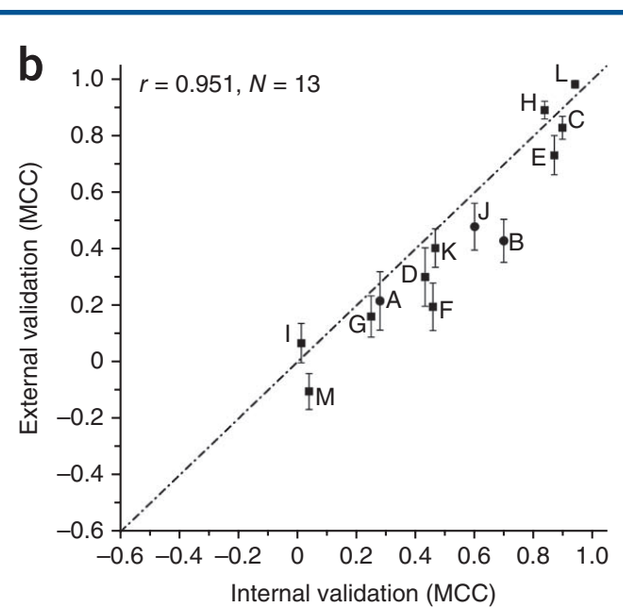
		Original analysis (training => validation)		
Modeling factor	Option	Number of teams	Number of endpoints	Number of models
Summary and normalization	Loess	12	3	2,563
	RMA	3	7	46
	MAS5	11	7	4,947
Batch-effect removal	None	10	11	2,281
	Mean shift	3	11	7,279
Feature selection	SAM	4	11	3,771
	FC+P	8	11	4,711
	T-Test	5	11	400
	RFE	2	11	647
Number of features	0~9	10	11	393
	10~99	13	11	4,445
	≥1,000	3	11	474
	100~999	10	11	4,298
Classification algorithm	DA	4	11	103
	Tree	5	11	358
	NB	4	11	924
	KNN	8	11	6,904
	SVM	9	11	986

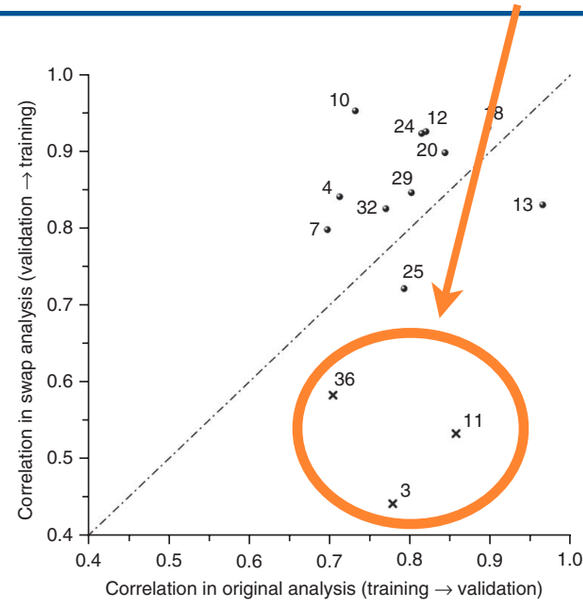
Analytic options used by two or more of the 14 teams that submitted models for all endpoints in both the original and swap experiments. RMA, robust multichip analysis; SAM, significance analysis of microarrays; FC, fold change; RFE, recursive feature elimination; DA, discriminant analysis; Tree, decision tree; NB, naive Bayes; KNN, K-nearest neighbors; SVM, support vector machine.



operator errors



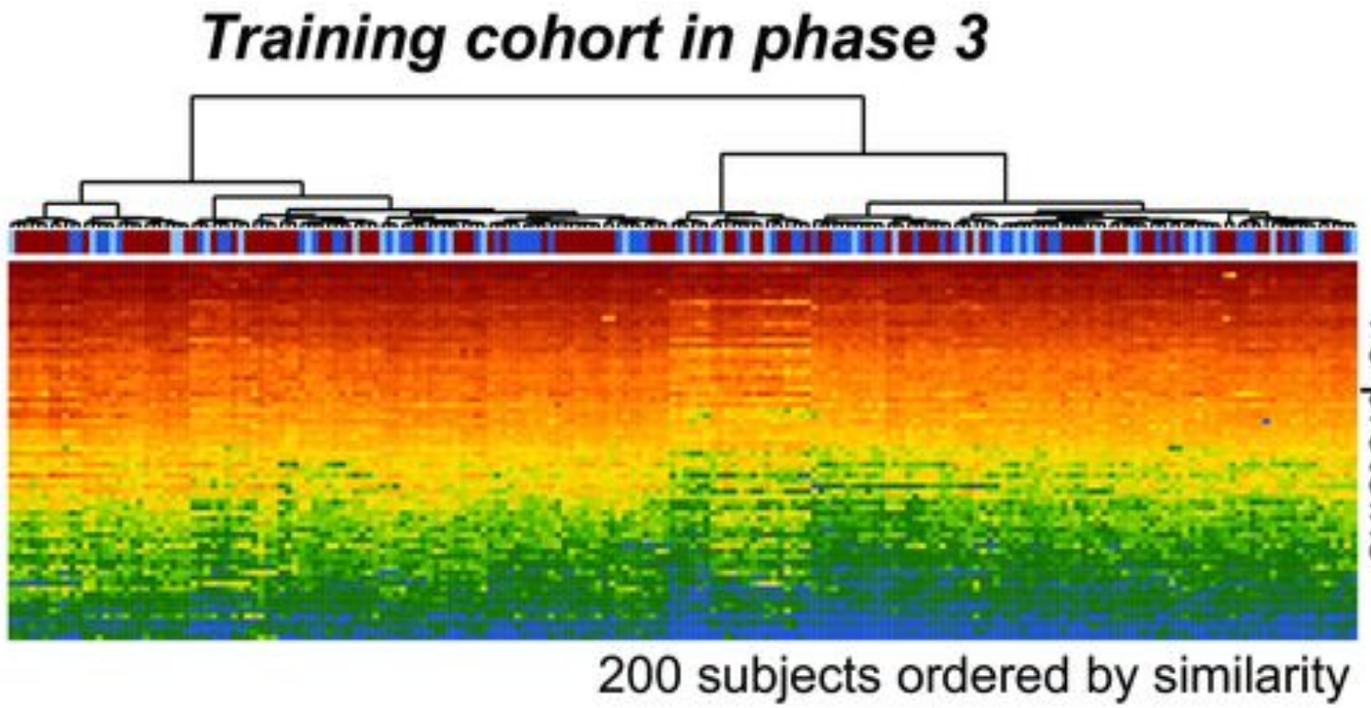




OUTLINE

- Supervised classification
  - Classifiers, evaluation, issues
- Case studies
  - Alzheimer's diagnosis with plasma signaling proteins
  - Colorectal cancer diagnosis from plasma proteins
  - MACQ II: common practices in assay validation
- Experimental design for prediction
  - Impacts of dimensionality and noise

DOES THE CLASSIFICATION ACCURACY DEPEND ON THE PROTEIN/SAMPLE SPACE?



More subjects are better for biomarker discovery?

	DDA	SRM/PRM	DIA
Proteomic coverage	High (1000s~10000)	Low (10s~100s)	Medium (1000s)
Sample throughput	Low (10~100s, even one sample)	High (100s~1000s)	Medium (100s)
Sensitivity	Low	High	Medium

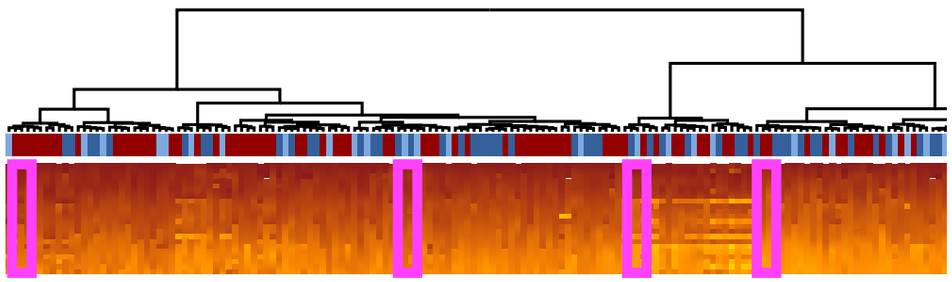
More proteins are better for biomarker discovery?

Aebersold and Mann, *Nature,* 2016

REPRODUCIBILITY EVALUATION

Resample subjects in training set, while fixing proteins number

Training cohort in phase 3



Protein rank: random forest classifier

10 predictive proteins

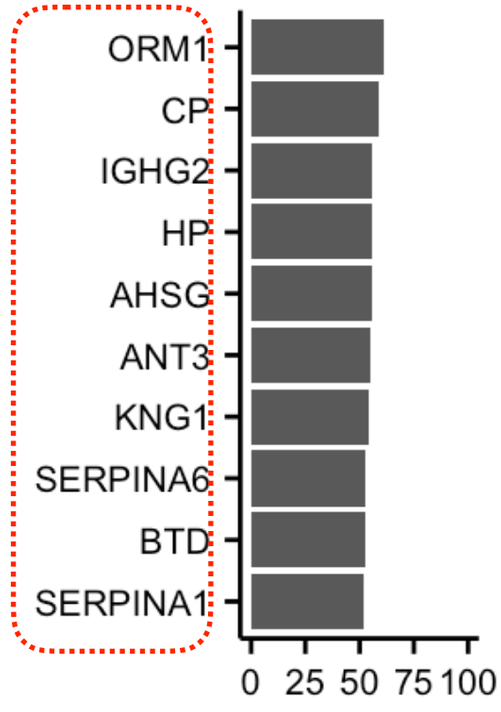
Protein	Mean accuracy decrease
ACTG1	4.29792551
H2AFX	4.17272432
ACTA2	3.64730166
HIST2H2AC	3.26940544
HIST1H2AB	3.10230189
ACTA1	3.07322088
TUBA1C	2.76765106
ACTG2	2.28619635
HSP90AA1	2.02174233
TUBB4A	1.99468791
HIST1H4A	1.86925866
SPTAN1	1.84453388
TUBB2B	1.75602473
COL6A3	1.72005229
TUBB4B	1.70743012
HIST1H2BK	1.68147503
HBB	1.61645207
HIST2H2AB	1.50013107

Repeat 100 times

Decision

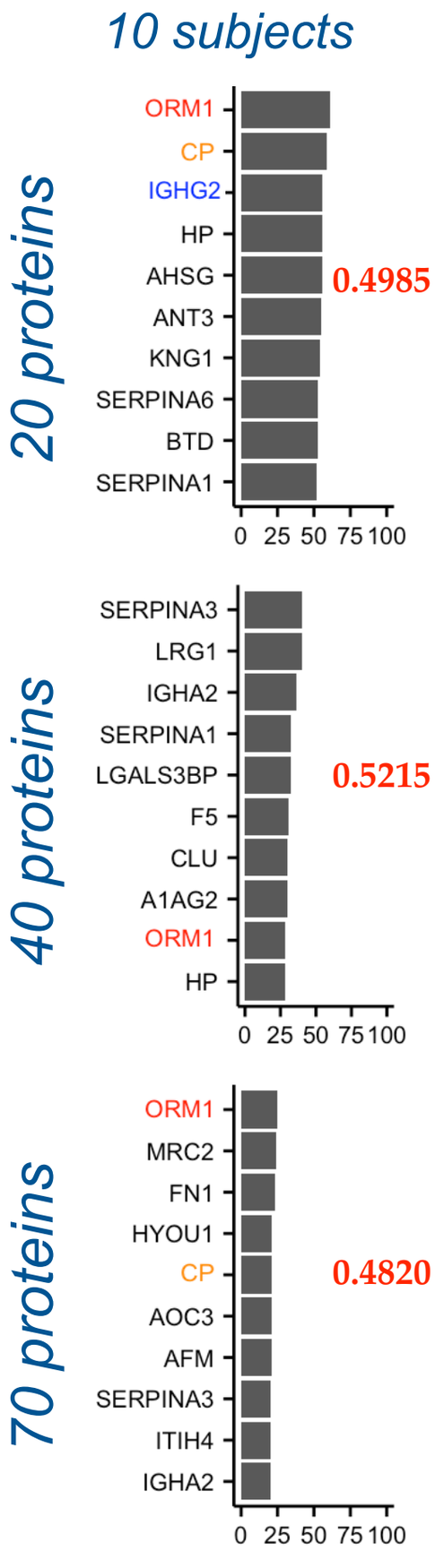
Truth	CRC	Control
CRC	TP	FN
Control	FP	TN

Top 10 most frequently selected proteins

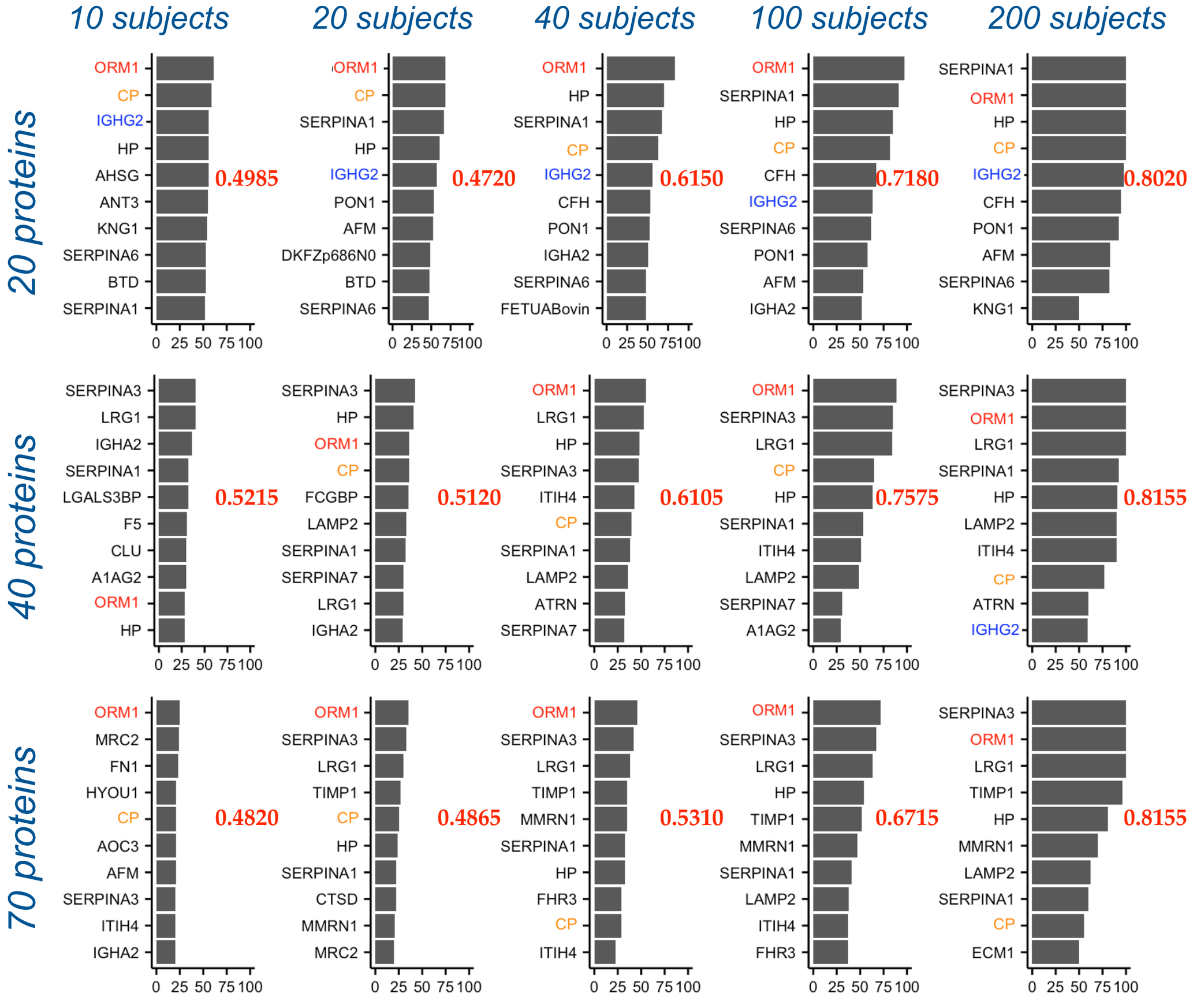


Number of iterations that a protein is identified predictive

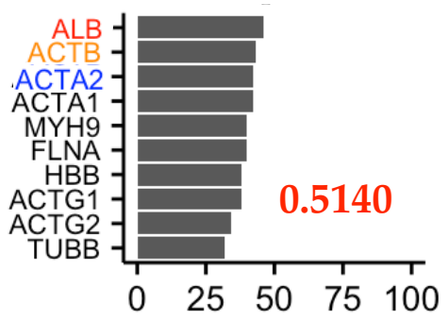
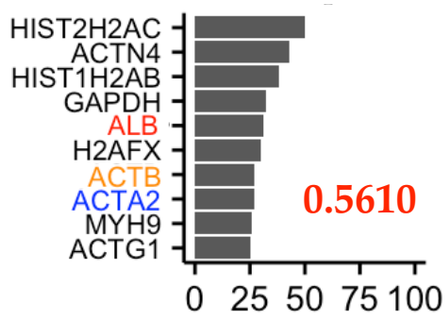
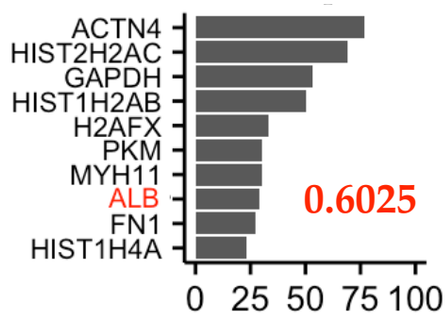
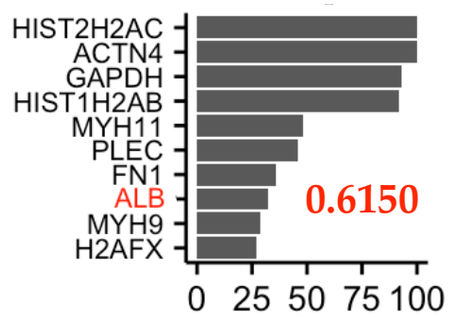
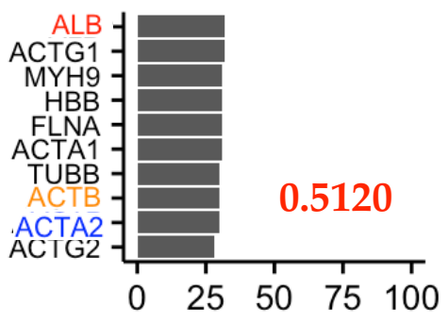
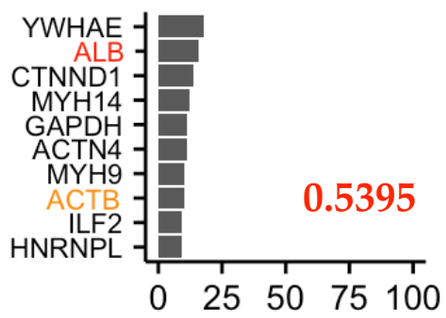
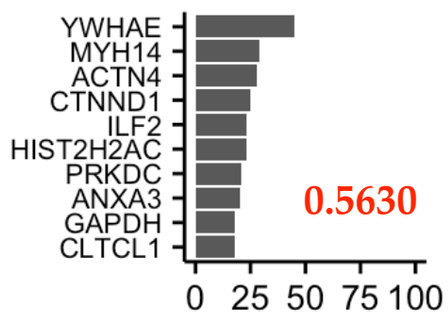
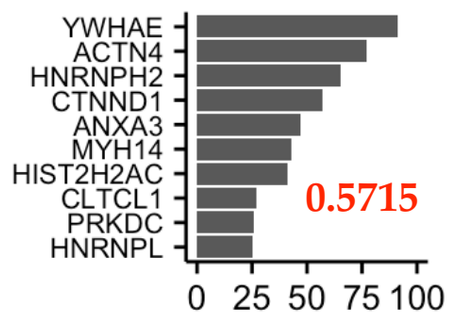
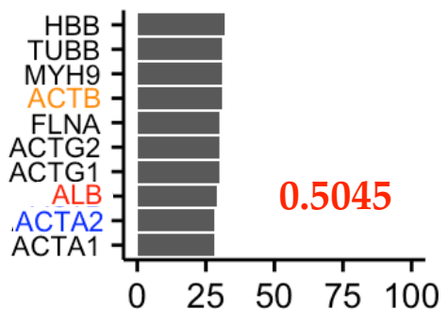
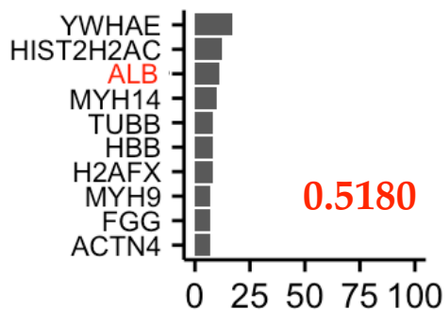
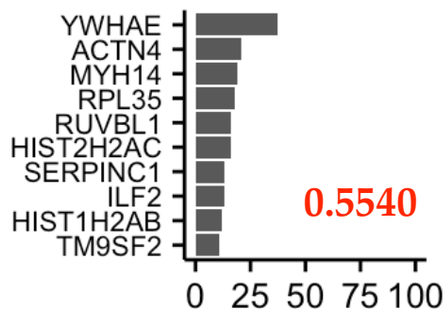
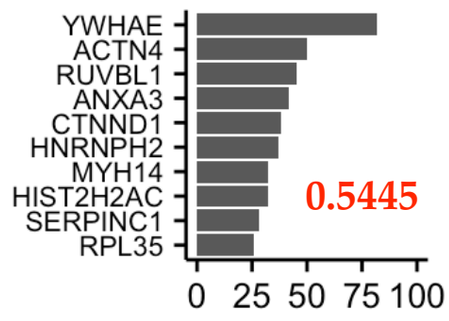
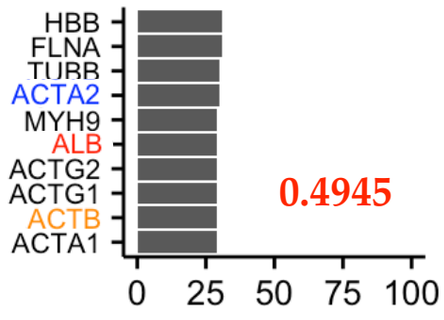
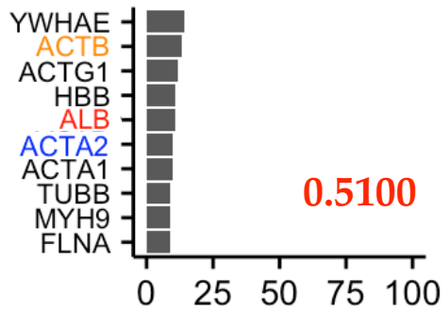
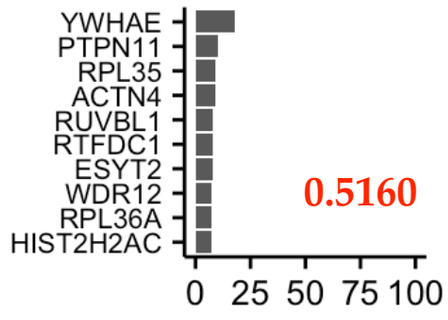
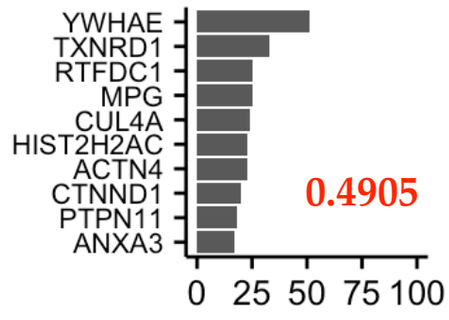
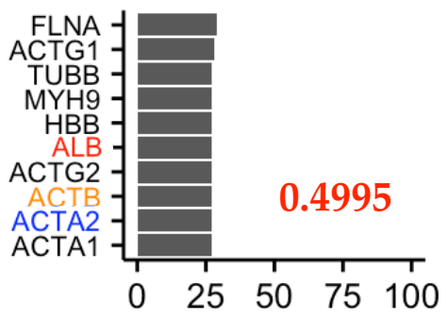
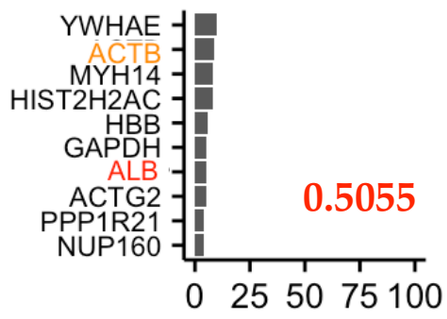
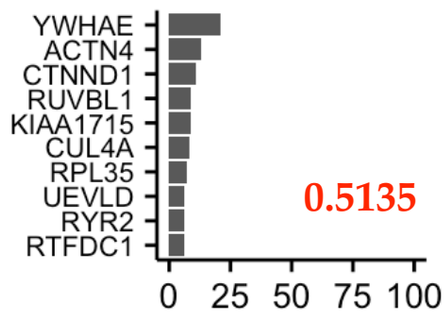
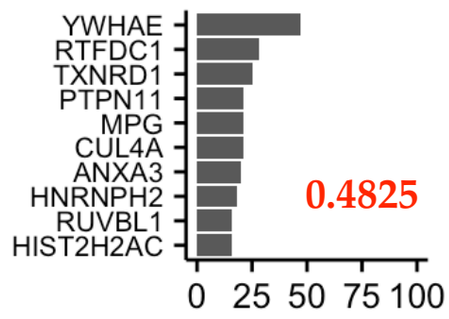
200 COLORECTAL CANCER SUBJECTS: SRM



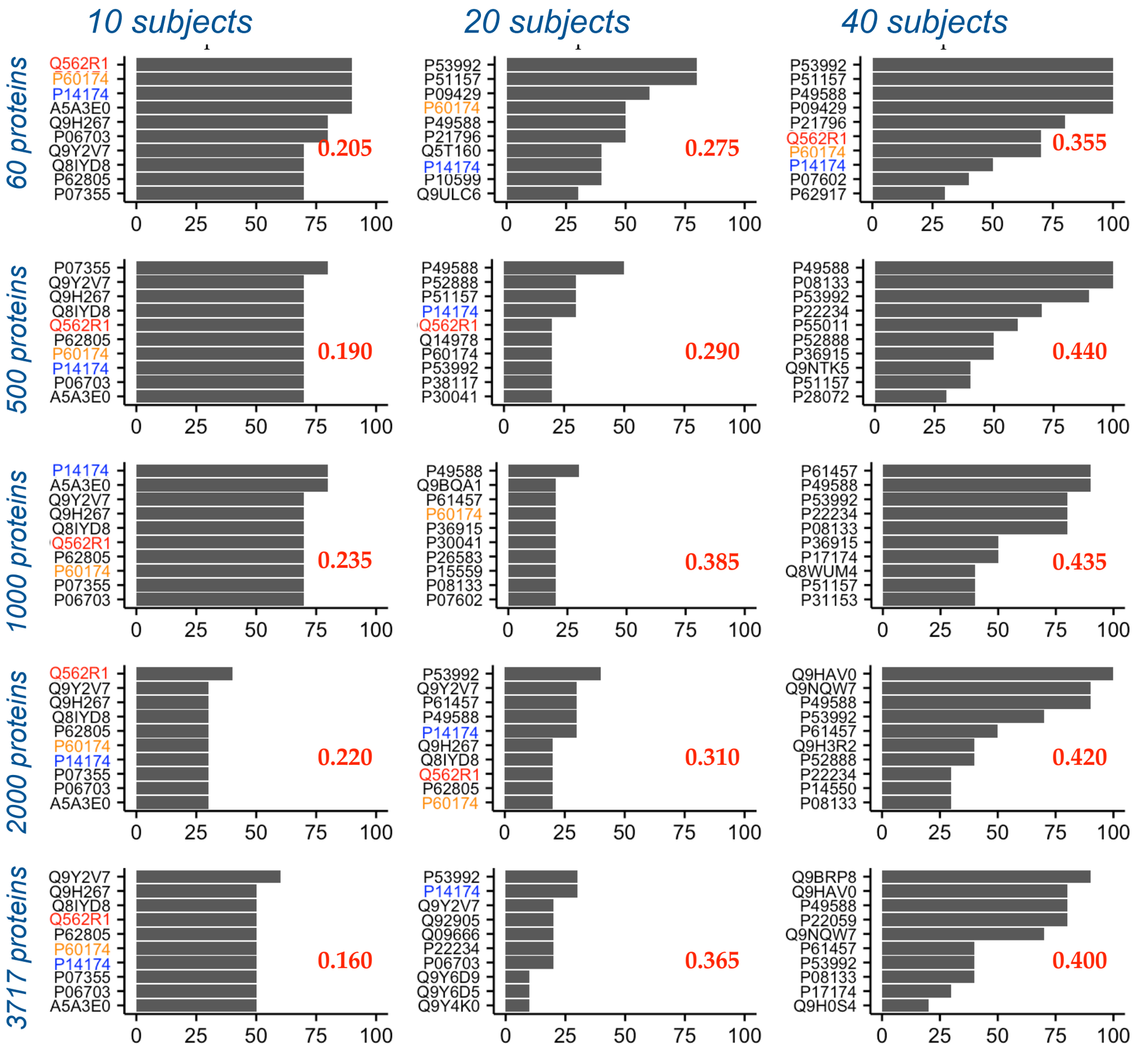
200 COLORECTAL CANCER SUBJECTS: SRM



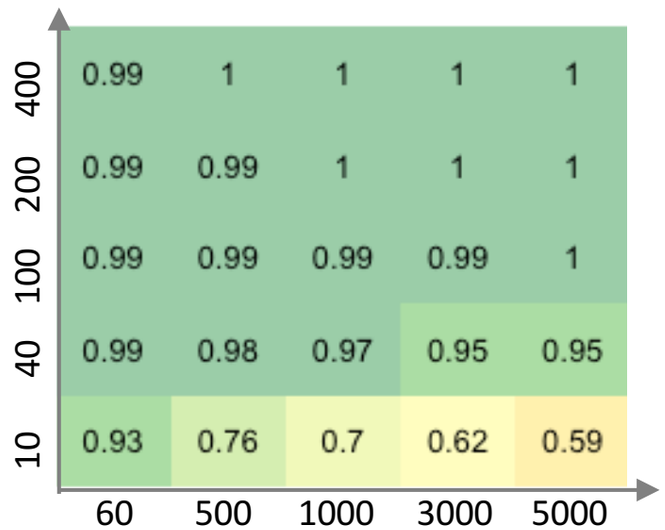
75 COLORECTAL CANCER SUBJECTS: DDA

	10 subjects	20 subjects	40 subjects	74 subjects
60 proteins				
500 proteins				
1000 proteins				
4000 proteins				
7211 proteins				

60 CELL LINES : DIA



INCREASING PROTEIN NUMBER DECREASES ACCURACY



INCREASING PROTEIN NUMBER DECREASES ACCURACY

High S/N | *Medium S/N* | *Low S/N*

CRC DDA

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.99	1	1	1	1
200	0.99	0.99	1	1	1
100	0.99	0.99	0.99	0.99	1
40	0.99	0.98	0.97	0.95	0.95
10	0.93	0.76	0.7	0.62	0.59

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.97	0.98	0.98	0.99	1
200	0.97	0.98	0.98	0.98	1
100	0.97	0.97	0.97	0.97	1
40	0.95	0.92	0.91	0.89	0.98
10	0.84	0.67	0.65	0.58	0.61

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.88	0.9	0.92	0.97	1
200	0.87	0.89	0.9	0.96	1
100	0.87	0.87	0.87	0.92	1
40	0.82	0.77	0.74	0.78	0.99
10	0.62	0.58	0.57	0.57	0.63

CRC SRM

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.84	0.87	0.9	0.9	0.89
200	0.83	0.85	0.88	0.86	0.83
100	0.84	0.83	0.8	0.77	0.72
40	0.7	0.68	0.7	0.62	0.6
10	0.66	0.57	0.58	0.52	0.56

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.78	0.82	0.86	0.88	0.89
200	0.76	0.79	0.82	0.82	0.83
100	0.74	0.75	0.76	0.74	0.73
40	0.63	0.68	0.7	0.59	0.57
10	0.52	0.54	0.49	0.56	0.52

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.66	0.73	0.81	0.85	0.89
200	0.66	0.69	0.77	0.78	0.85
100	0.59	0.62	0.68	0.7	0.72
40	0.58	0.57	0.62	0.62	0.6
10	0.52	0.51	0.56	0.49	0.47

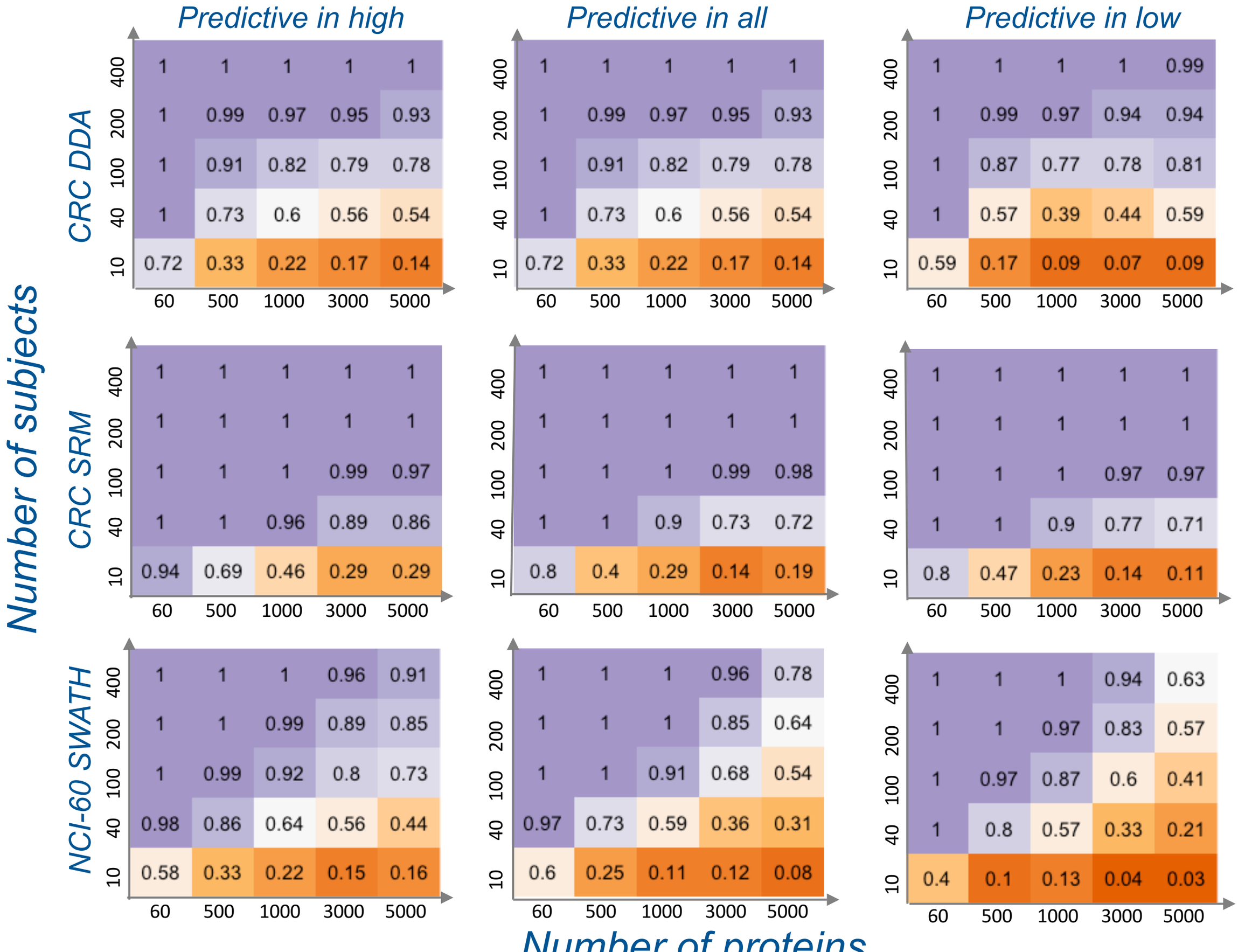
NCI-60 SWATH

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.97	0.99	0.99	0.99	0.98
200	0.97	0.99	0.99	0.99	0.98
100	0.97	0.98	0.98	0.97	0.95
40	0.93	0.94	0.92	0.84	0.81
10	0.88	0.74	0.65	0.57	0.52

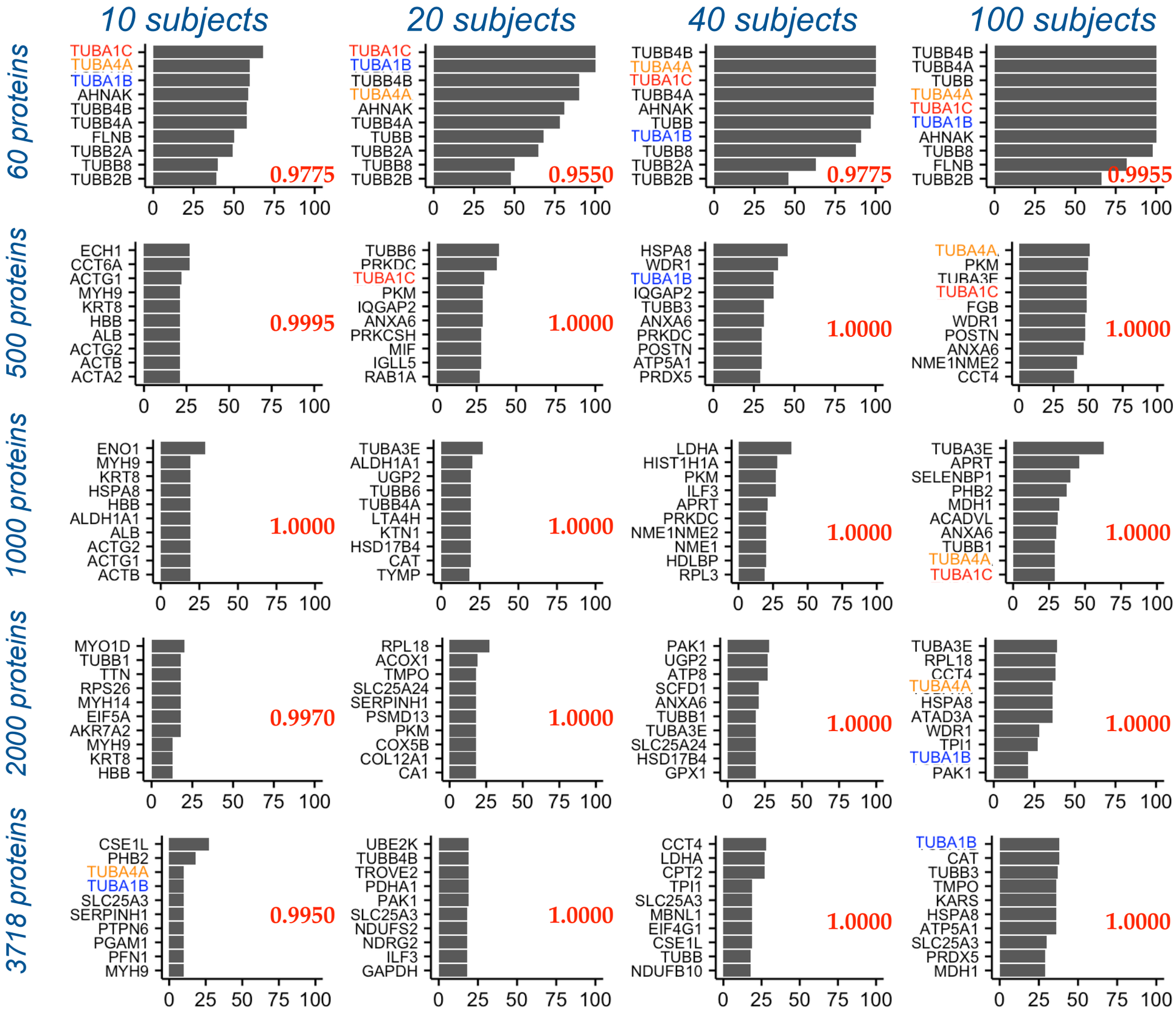
Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.92	0.96	0.97	0.97	0.96
200	0.92	0.96	0.97	0.96	0.94
100	0.91	0.96	0.96	0.94	0.89
40	0.87	0.84	0.86	0.79	0.7
10	0.7	0.67	0.55	0.55	0.61

Number of subjects \ Number of proteins	60	500	1000	3000	5000
400	0.81	0.89	0.91	0.91	0.9
200	0.79	0.87	0.9	0.89	0.85
100	0.78	0.85	0.86	0.81	0.79
40	0.74	0.78	0.75	0.72	0.71
10	0.56	0.55	0.57	0.54	0.51

MORE REPLICATES SELECT MORE PREDICTIVE PROTEINS



PERFECT CLASSIFICATION: BIASED DESIGN



ACKNOWLEDGEMENTS

Northeastern University

Kylie Bemis
Meena Choi
Eralp Dogu
Dan Guo
April Harry
Ting Huang
Cyril Galitzine
Robert Ness
Sara Taheri
Tsung-Heng Tsai

ETH Zurich

Ruedi Aebersold
Tiannan Guo
Ruth Huttenhain
Paola Picotti
Silvia Surinova
Bernd Wollscheid

University of Washington

Michael MacCoss
Brendan MacLean
Jarrett Egertson

Biognosis

Lukas Reiter

iPRG 2015

Zeynep Eren-Dogu
Chris Colangelo
John Cottrell
Michael Hopman
Eugene Kapp
Santa Kim
Henry Lam
Tom Neubert
Magnus Palmblad
Brett Phinney
Sue Weintraub,