***Nature Biotechnology*, August, 2010**

Supplementary Information:

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

MAQC Consortium

Contact: Leming.Shi@fda.hhs.gov or Leming.Shi@gmail.com

Contents:

Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Figure 5. Correlation of candidate model selection between internal cross validation and external validation.

Supplementary Figure 6. Impact of modeling factors on model performance: The empirical BLUPs (Best Linear Unbiased Predictor) of each level for all the factors across 13 endpoints, with clear labeling of interaction terms.

Supplementary Figure 7. A decision-tree model of the relative importance of modeling factors on external validation prediction performance in terms of MCC.

Supplementary Figure 8. Three aspects for assessing the performance of a data analysis protocol (DAP) in decreasing order of priority: prediction performance, robustness, and biological relevance of the gene signatures.

Supplementary Figure 9. Feature landscapes comparing swap features lists to original feature lists.

Supplementary Figure 10. Further analysis on the relationship between stability of feature lists and the level of endpoint predictability.

Supplementary Figure 11. The stability of feature lists is positively correlated with endpoint predictability.

Supplementary Figure 12. Prediction error on a per-sample basis - some samples were consistently misclassified by almost all models.

Supplementary Figure 13. The seven performance metrics measure different aspects of the prediction performance of a model.

Documents in Supplementary Information:

Supplementary Document 1. Validation performance of 318 MAQC-II nominated models

Supplementary Document 2. The performance metrics (MCC, AUC, Accuracy, RMSE)

Supplementary Document 3. Possible explanation of the superior performance of DAT33's model on endpoint A

Supplementary Document 4. The MAQC-II Research Plan (March 22, 2007)

Supplementary Document 5. Standard Operating Procedures (SOPs), Methods and Analysis for MAQC-II

Nature Biotechnology: doi: 10.1038/nbt.1665

Tables in Supplementary Information

Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Table 3. Data Analysis Teams (DATs) in the MAQC-II project

DAT	Org. Code	Organization Name	DAT Leader	Endpoints* (Original)	No. Models (Original)	Endpoints (Swap)	No. Models (Swap)
DAT1	ABT	Abbott Laboratories	Viswanath Devanarayan	8	53		
DAT2	Almac	Almac Diagnostics, UK	Juergen von Frese	1	1		
DAT3	CAS	Chinese Academy of Sciences, China	Tieliu Shi	13	21	13	17
DAT4	CBC	CapitalBio Corporation, China	Liang Zhang	13	26	13	26
DAT5	CDRH	Center for Devices and Radiological Health, FDA	Gene Pennello	3	9		
DAT6	CIPF	Centro de Investigacion Principe Felipe, Spain	Joaquin Dopazo	8	112	8	112
DAT7	Cornell	Weill Medical College of Cornell University	Fabien Campagne	13	614	13	732
DAT8	DKFZ	German Cancer Research Center, Germany	Benedikt Brors	10	34	10	36
DAT9	EPA	U.S. Environmental Protection Agency	Richard Judson	2	1008	2	1011
DAT10	FBK	Fondazione Bruno Kessler, Italy	Cesare Furlanello	13	27	13	13
DAT11	GeneGo	GeneGo Inc.	Weiwei Shi	13	30	13	28
DAT12	GHI	Golden Helix Inc.	Christophe Lambert	13	52	13	26
DAT13	GSK	GlaxoSmithKline	Jie Cheng	13	15	13	13
DAT14	GT	Georgia Institute of Technology – Emory University	May Wang	4	27	6	15
DAT15	JHSPH	Johns Hopkins Bloomberg School of Public Health	Rafael Irizarry	6	12		
DAT16	KU	University of Kansas	Luke Huan	8	19		
DAT17	Ligand	Ligand Pharmaceuticals	Wen Luo	1	1	1	1
DAT18	NCTR	National Center for Toxicological Research, FDA	Weida Tong	13	8580	13	8320
DAT19	NIEHS	National Institute of Environmental Health Sciences	Pierre Bushel	13	311	13	258
DAT20	NWU	Northwestern University	Simon Lin	13	278	13	290
DAT21	Princeton	Princeton University	Jianqing Fan	3	180		
DAT22	Roche	Roche Palo Alto LLC	Hans Bitter	6	5310		
DAT23	SA	SABioscience Corporation	Guozhen Liu	5	112	13	84
DAT24	SAI	Systems Analytics Inc.	John Zhang	13	130	13	130
DAT25	SAS	SAS Institute Inc.	Russ Wolfinger	13	389	13	377
DAT26	SDSU	South Dakota State University	Xijin Ge	8	20		
DAT27	SIB	Swiss Institute of Bioinformatics, Switzerland	Vlad Popovici	10	6	6	18
DAT28	Spheromics	Spheromics, Finland; University of Umeå, Sweden	Andreas Scherer	9	40	2	16
DAT29	Tsinghua	Tsinghua University, China	Xuegong Zhang	13	1660	13	1660
DAT30	UAMS	University of Arkansas for Medical Sciences	Yiming Zhou	2	4		
DAT31	UCLA	Cedars-Sinai Medical Center of UCLA	Xutao Deng	1	12		
DAT32	UIUC	University of Illinois at Urbana-Champaign	Sheng Zhong	13	52	13	52
DAT33	UIUC2	University of Illinois at Urbana-Champaign	Nathan Price	13	19		
DAT34	UML	University of Massachusetts Lowell	Dalila Megherbi	5	5		
DAT35	USM	University of Southern Mississippi	Youping Deng	13	72		
DAT36	ZJU	Zhejiang University, China	Xiaohui Fan	13	52	13	52
Total number of models submitted					19,779		13,287
Total number of nominated models submitted					323		243
Total number of models applied to validation sets					**18,303**		**12,195**
Total number of nominated models applied to validation sets					318		241

*Seventeen (17) of the 36 data analysis teams analyzed all 13 endpoints in the original training-validation experiment, and 16 teams analyzed all 13 endpoints in the swap training-validation experiment.

Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Table 4. Summary information about the options of modeling factors adopted for the 18,060 models in the original analysis (training=>validation). For these 18,060 models, the requested model description information was complete and summarized below.

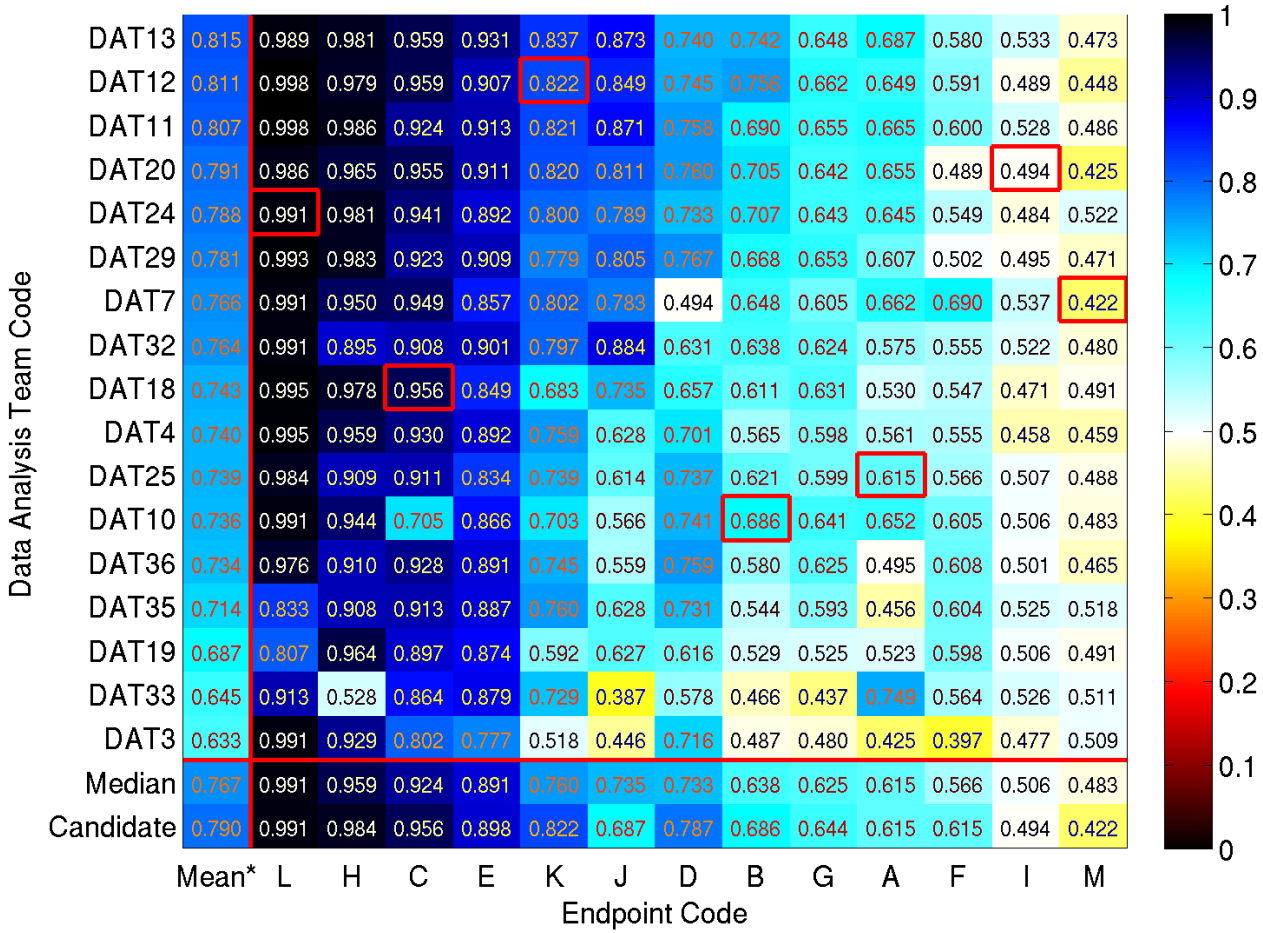
Factor (Levels)	Option	No. of DATs	No. of Models
Feature Selection Method (33 levels)	FC+P	11	5,846
	T-Test	9	609
	SAM	7	4,471
	RFE	6	887
	PAM	3	23
	KS	2	803
	Wilcoxon	2	78
	FC	2	40
	None	2	39
	Genetic Algorithm	2	33
	Fisher	2	5
	Forest	1	1,829
	SA	1	869
	STE	1	861
	ReliefF	1	839
	Genelists	1	313
	SVM-Weights	1	196
	Logistic	1	180
	PCA	1	32
	Permutation	1	19
	Welch	1	16
	Bscatter	1	14
	Vote	1	13
	Barcode	1	12
	PCC	1	10
	Golub	1	7
	BWSS	1	6
	eGOMiner	1	3
	Pathway	1	2
	TAUC	1	2
	CoxModel	1	1
	P	1	1
	eClinical	1	1
Number of Features (4 levels)	10~99	32	10,694
	0~9	25	1,510
	100~999	20	5,301
	>=1000	8	555

Factor (Levels)	Option	No. of DATs	No. of Models
Classification Algorithm (24 levels)	SVM	21	3,600
	KNN	13	9,532
	Tree	9	1,854
	DA	9	1,370
	NB	7	1,099
	PAM	4	51
	Logistic	3	85
	ANN	3	20
	PLS	2	91
	Forest	2	4
	RFE	1	192
	Nearest Centroid	1	26
	ML	1	25
	GLM	1	20
	SMO	1	19
	Barcode	1	12
	SDF	1	11
	PM	1	10
	RF	1	8
	BART	1	7
	BB	1	7
	EN	1	7
	AB	1	6
	K-means	1	4
Batch Effect Removal Method (9 levels)	None	26	9,120
	Mean Shift	5	8,622
	EB	2	23
	ComBat	1	204
	P.Rank	1	30
	OPLS	1	24
	Agilent	1	22
	Barcode	1	12
	Embedded	1	3

Factor (Levels)	Option	No. of DATs	No. of Models
Summary or Normalization Method (17 levels)	MAS5	26	8,568
	Loess	22	6,572
	Median	19	1,100
	RMA	7	92
	refRMA	2	15
	Quantile	2	7
	dChip	1	971
	Mean	1	368
	SVN	1	282
	VSN	1	22
	iset	1	14
	MAS5+Loess	1	12
	PLIER	1	12
	VSN+RMA	1	12
	Genetic Algorithm	1	5
	Bkgd-subonly	1	4
	Raw	1	4
Internal Validation Method (6 levels)	5-CV	30	12,528
	LOOCV	3	206
	10-CV	2	3,387
	7-CV	1	1,808
	12-CV	1	127
	Split Sample	1	4
Number of Internal Validation Iterations (9 levels)	10	27	12,112
	1	6	480
	20	2	116
	100	2	7
	5	1	4,051
	2	1	1,231
	40	1	45
	50	1	12
	30	1	6

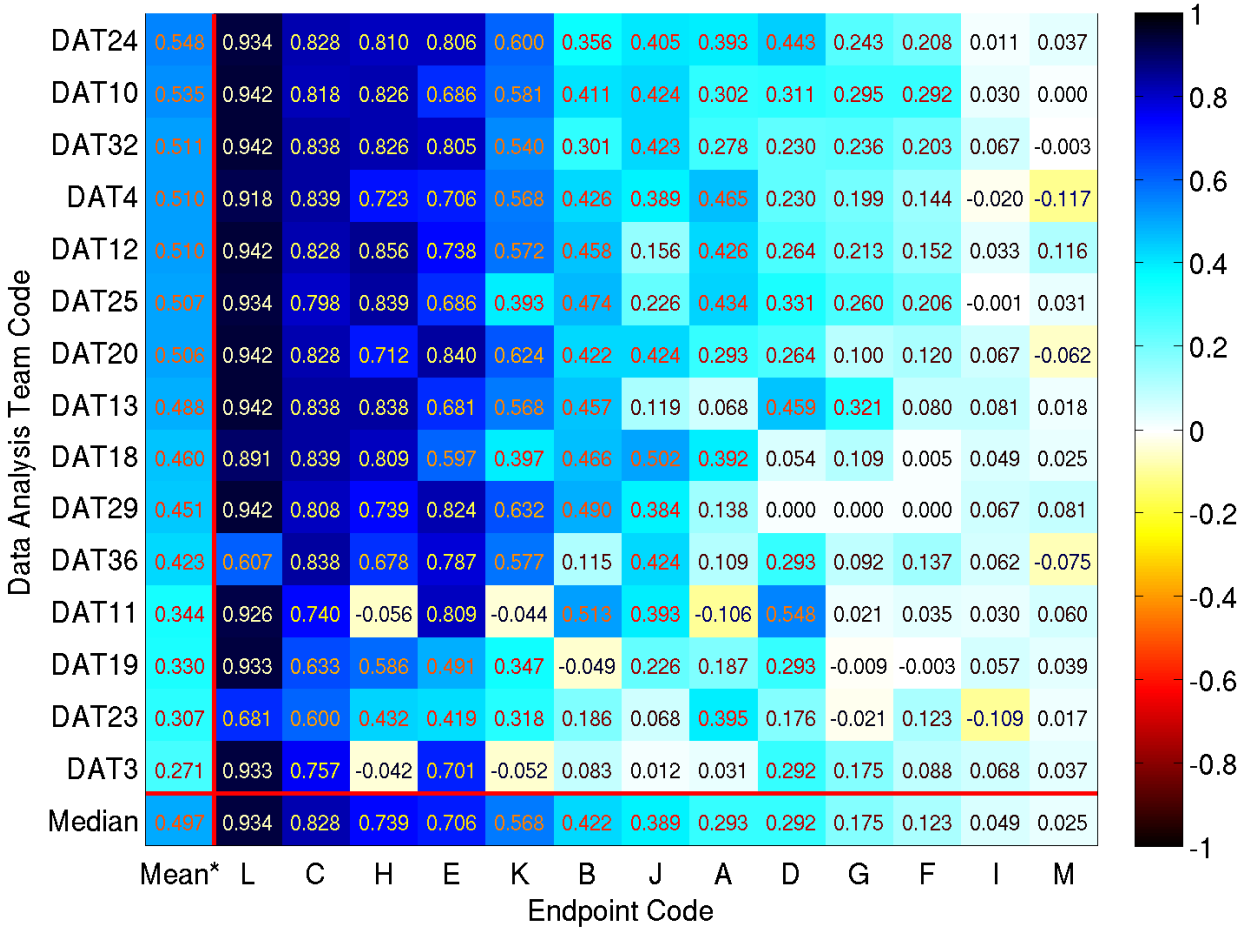
Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Table 5. Original validation AUC of nominated models by 17 data analysis teams (DATs) that analyzed all 13 endpoints in the original training-validation experiment. The median AUC value for an endpoint was calculated based on values from the 17 DATs and the mean AUC value for a DAT was calculated based on values from the 11 non-random endpoints (excluding I and M). The level of predictability of the 13 endpoints is dramatically different, with median AUC varying from 0.991 (L) to 0.483 (M). In addition, the 17 DATs showed large differences in proficiency in developing predictive models, with mean AUC varying from 0.815 (DAT13) to 0.633 (DAT3).



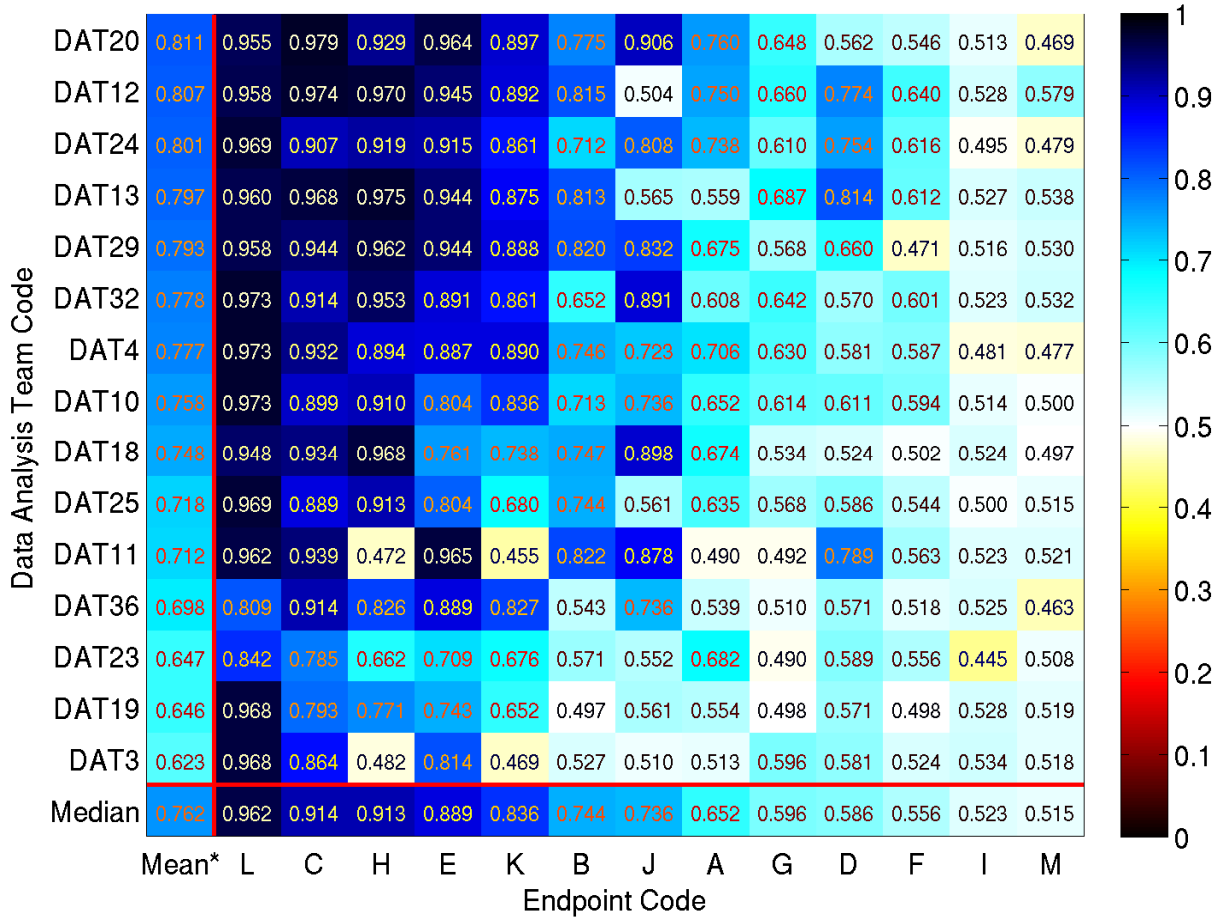
Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Table 6. Swap validation MCC of nominated models by 15 data analysis teams (DATs) that analyzed all 13 endpoints in the swap training-validation experiment. The median MCC value for an endpoint was calculated based on values from the 15 DATs and the mean MCC value for a DAT was calculated based on values from the 11 non-random endpoints (excluding I and M). The level of predictability of the 13 endpoints is dramatically different, with median MCC varying from 0.934 (L) to 0.025 (M). In addition, the 15 DATs showed large differences in proficiency in developing predictive models, with mean MCC varying from 0.548 (DAT24) to 0.271 (DAT3).



Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Table 7. Swap validation AUC of nominated models by 15 data analysis teams (DATs) that analyzed all 13 endpoints in the swap training-validation experiment. The median AUC value for an endpoint was calculated based on values from the 15 DATs and the mean AUC value for a DAT was calculated based on values from the 11 non-random endpoints (excluding I and M). The level of predictability of the 13 endpoints is dramatically different, with median AUC varying from 0.962 (L) to 0.515 (M). In addition, the 15 DATs showed large differences in proficiency in developing predictive models, with mean AUC varying from 0.811 (DAT20) to 0.623 (DAT3).



Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Table 8. Samples consistently predicted wrong by most models for the two positive control endpoints (H and L)

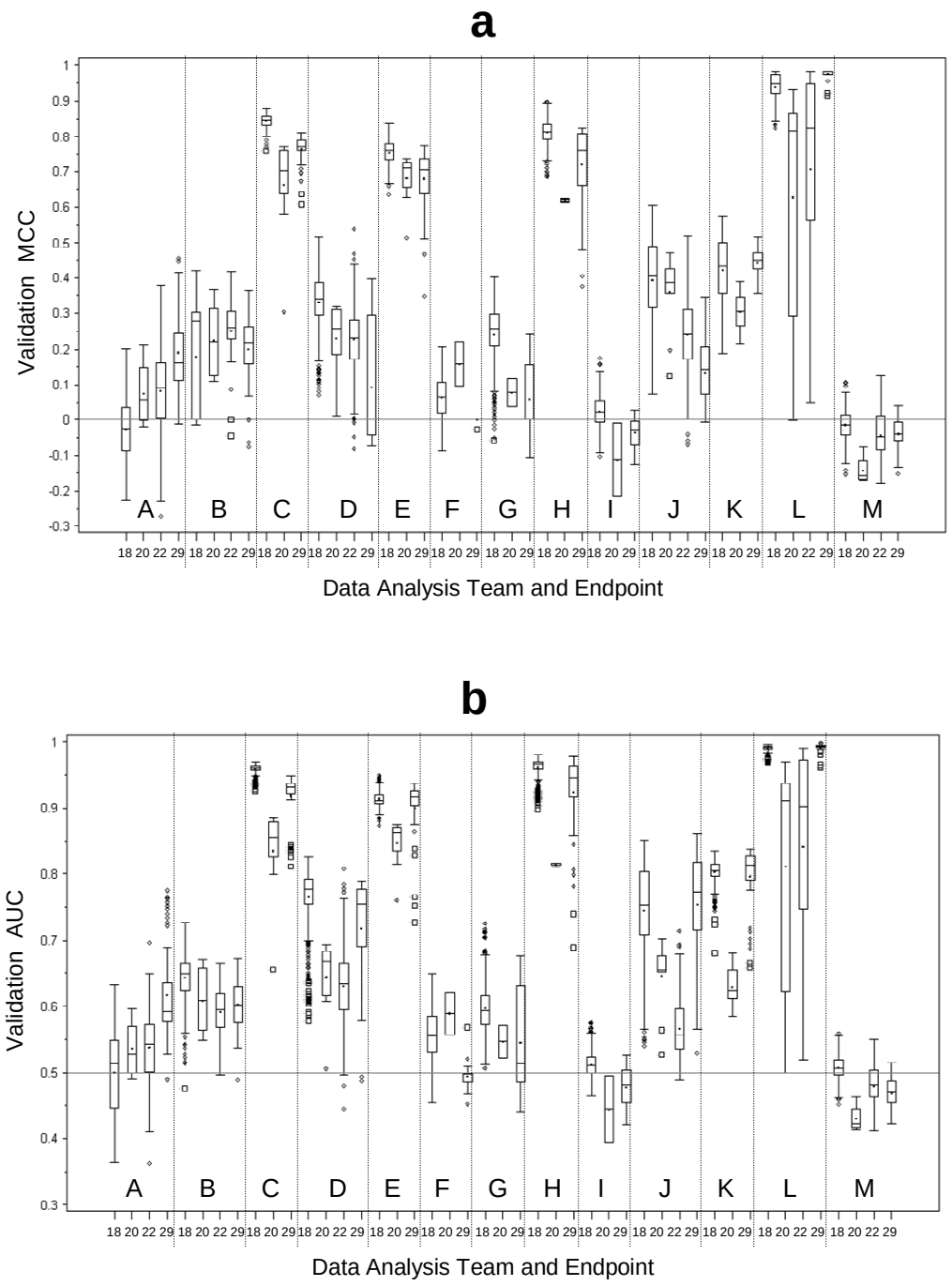
SampleID	Data Set	Endpoint	Total number of models	Number of models with wrong prediction	Fraction of models with wrong prediction	Comment
P1550-01-E679-U133Plus-2	MM	H	889	860	0.967	In original validation set
P1371-01-E336-U133Plus-2	MM	H	889	852	0.958	In original validation set
P1375-01-E113-U133Plus-2	MM	H	888	847	0.954	In original validation set
II-NB407	NB	L	1785	1709	0.957	In original validation set
P0678-01-C118-U133Plus-2	MM	H	827	825	0.998	In original training set
P0067-01-A274-U133Plus-2	MM	H	827	823	0.995	In original training set
P0002-01-FAKE03-U133Plus-2	MM	H	827	821	0.993	In original training set
P0931-01-C648-U133Plus-2	MM	H	827	821	0.993	In original training set
NB528	NB	L	889	889	1	In original training set
NB557	NB	L	889	889	1	In original training set
NB504*	NB	L	889	872	0.981	In original training set
NB560	NB	L	889	871	0.98	In original training set
NB523	NB	L	889	870	0.979	In original training set
NB522*	NB	L	889	870	0.979	In original training set
NB412*	NB	L	889	855	0.962	In original training set

*Clinical information of neuroblastoma patients for whom the positive endpoint L was uniformly misclassified were re-checked, and in three cases (NB412, NB504, and NB522) incorrect sex assignment for endpoint L was revealed. For NB504, the error occurred as the patient's first name in Germany can refer to both a boy and a girl. It had actually been corrected five months before the MAQC-II results became available. For NB412 and NB522, the actual sex was indeed indicated as "W" in the original clinical information table, the German analogue to "F" (female). However, the project leader made a mistake when converting the Sex column into the mock endpoint L (NEP_S): "F" was converted into "0" (Female), and the rest (including the only two cases labeled as “W”) was converted into "1" (Male). It demonstrates that sometimes gene-expression based classification models could help correct human errors. The higher number of consistently mis-predicted samples in the original neuroblastoma training set (endpoint L) might explain why the external validation performance is higher than the internal validation performance (**Figures 2c** of Shi L et al., *Nature Biotechnology,* 2010).

Nature Biotechnology: doi: 10.1038/nbt.1665

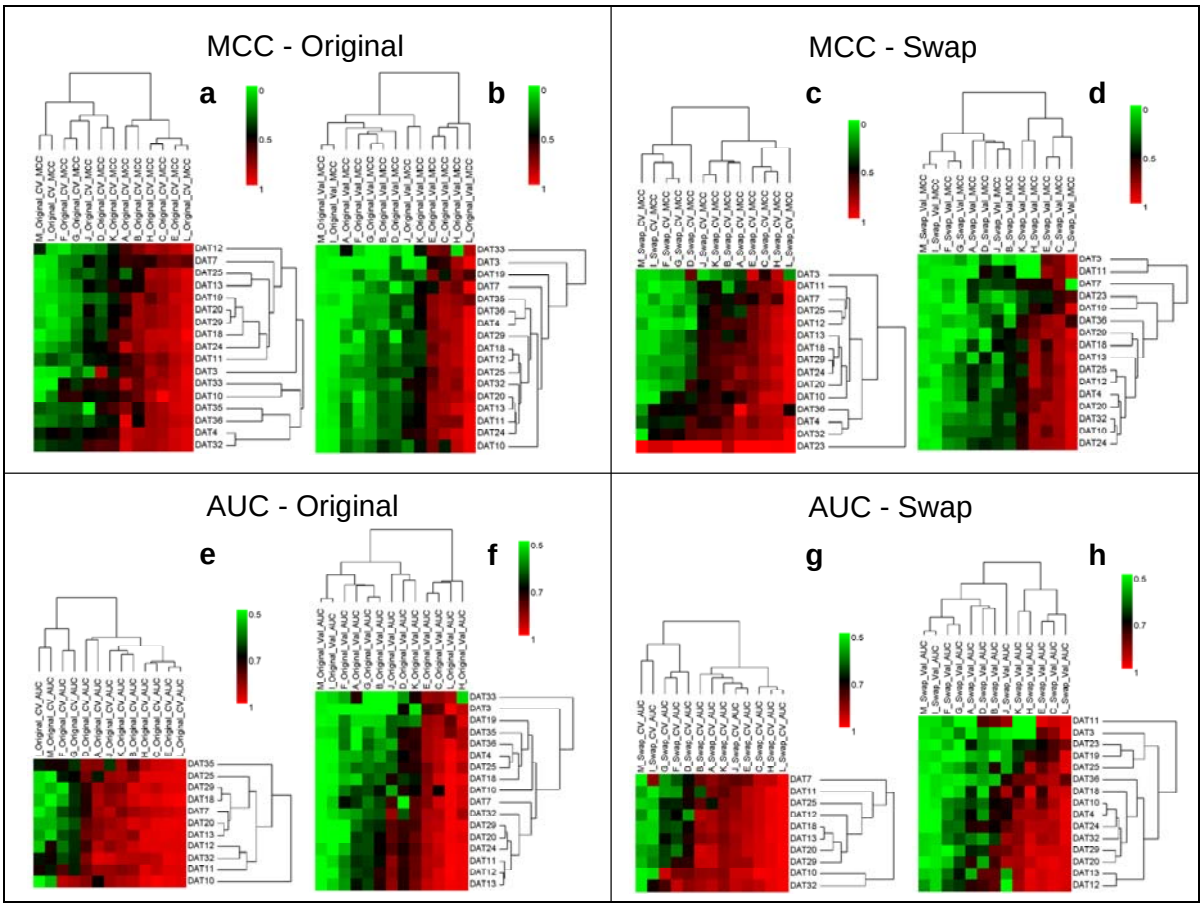
Figures in Supplementary Information

Nature Biotechnology: doi: 10.1038/nbt.1665



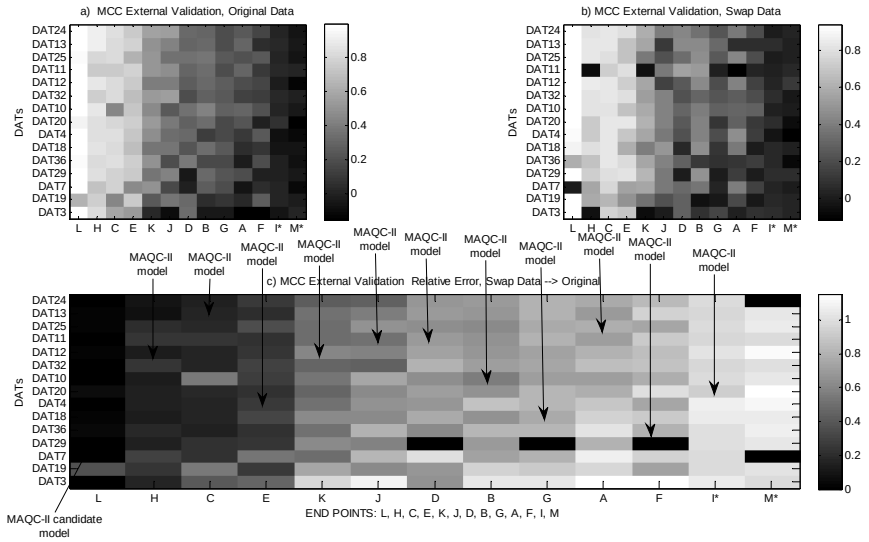
Supplementary Figure 1. KNN models developed by different data analysis teams showed significant differences in external validation performance. **a**. MCC; **b**. AUC. Further investigation into the data analysis protocols revealed differences in tunable modeling factors options not captured in the model summary information tables (**Supplementary Tables 1 and 2**). See Parry RM et al., *The Pharmacogenomics Journal*, 2010 for more details.

Nature Biotechnology: doi: 10.1038/nbt.1665

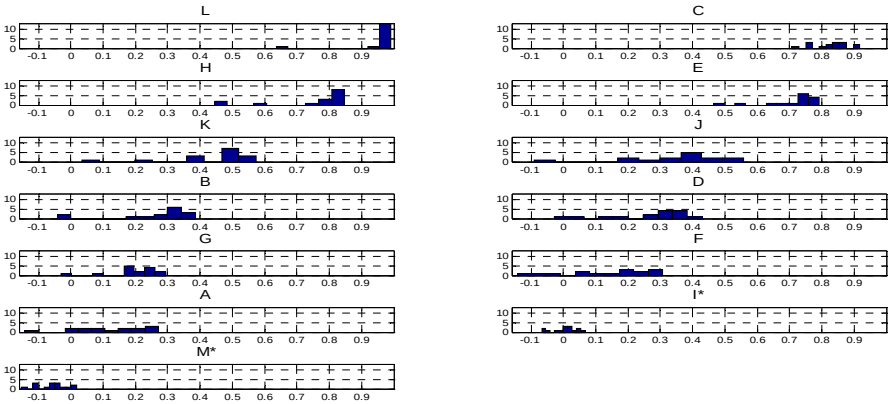


Supplementary Figure 2. The pattern of performance estimates of the nominated models across 13 endpoints. **a**-**d**: MCC as performance metric; **e**-**h**: AUC as performance metric; **a**, **c**, **e**, **g**: internal cross-validation; **b**, **d**, **f**, **h**: external validation. The patterns across the 13 endpoints are similar between the original training-validation results and the swap training-validation results. The deterioration of external validation performance from internal validation performance estimate is obvious for some models. Some DATs clearly reported over-optimistic cross-validation performance for at least some endpoints; however, their external prediction performance appeared to be consistent with the majority of the other teams. Models nominated by DAT7 appeared to have under-estimated both the cross-validation performance and external validation performance. Variability in external validation performance from other teams could be due to various causes including clerical error.

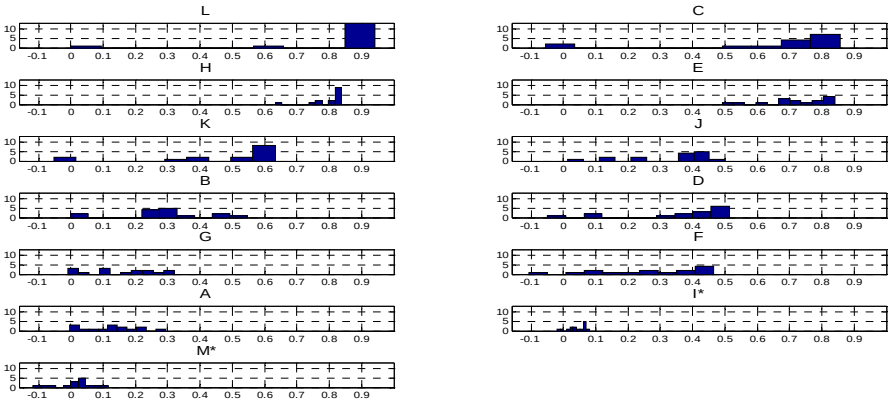
Nature Biotechnology: doi: 10.1038/nbt.1665



d)



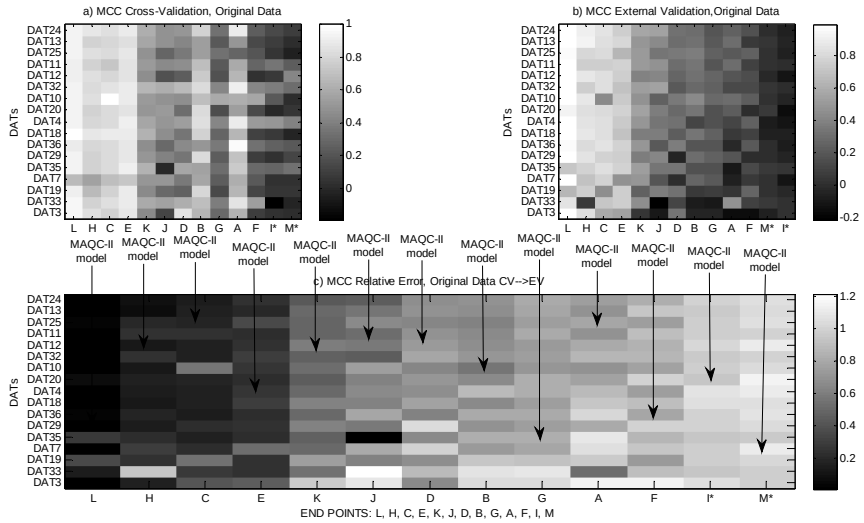
e)



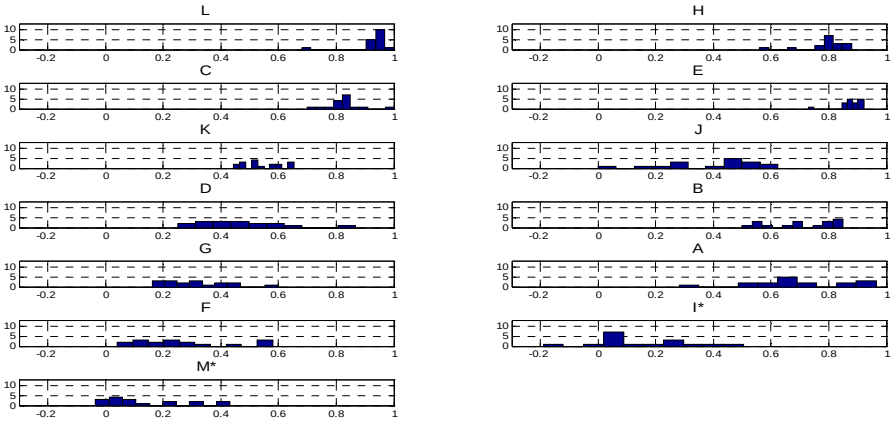
Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Figure 3. Analysis and visualization of Original and Swap validation models' prediction performance MCC of nominated models by 15 data analysis teams (DATs) that analyzed all 13 endpoints in the original training-validation experiment. Gray-level "images" (not heat-maps) of respectively: the MCC scores for the external independent validation original data **(a**) and swap data **(b)**, for endpoints (in that order) L, H, C, E, K, J, D, B, G, A, F, I*, M*, and Organizations, (in that order) DAT24, DAT13, DAT25, DAT11, DAT12, DAT32, DAT10, DAT20, DAT4, DAT!8, DAT36, DAT29, DAT7, DAT19, DAT3 . As one can see the darker the image pixel intensity is the lower the MCC score is. One can, by simple observation of the images, see that, in general, there are brighter (whiter) pixel intensities in image (b) than there are in image (a); as the swap MCC values seem to be generally a bit higher than the original data MCC values. Additionally, one can easily also observe/confirm, again, by looking globally at the image that the first 4 endpoints, L,H, C, E in both the swap and original data seem to be easily predictable compared to end points G, A, F, M*, I*,. One can also observe that it appears that although both endpoints I* and M* are virtually non-predictable or hard to predict, endpoint M* seems to be slightly more predictable than endpoint I* in the case of the swap data. **(c**) shows the gray-level image of the relative errors between the original and swap data External Validation MCC values for same end points and organizations described in (a) and (b). Brighter (whiter) pixels show the larger differences (and/or inconsistencies) *in percentile* (%) in the prediction of the related endpoints and organizations. Note that since for some end points the MAQC-II models were not derived by these DAT teams, in the 2-dimensional "gray image" of Figure S11c the exact and/or approximate locations of the MAQC-II candidate models are pointed to by arrows, for illustration. **(d**) & **(e**) illustrates the number of organizations that made about the same prediction (MCC value), in the External Validation of respectively the Original and Swap data, for each of the endpoints, L, H, C, E, K, J, D, B, G, A, F, I*, M*. One can observe by inspection of (c), (d) and (e) how model prediction consistency between Original and Swap Data generally varies and decreases as a function of the 13 End points and the 15 DAT models.

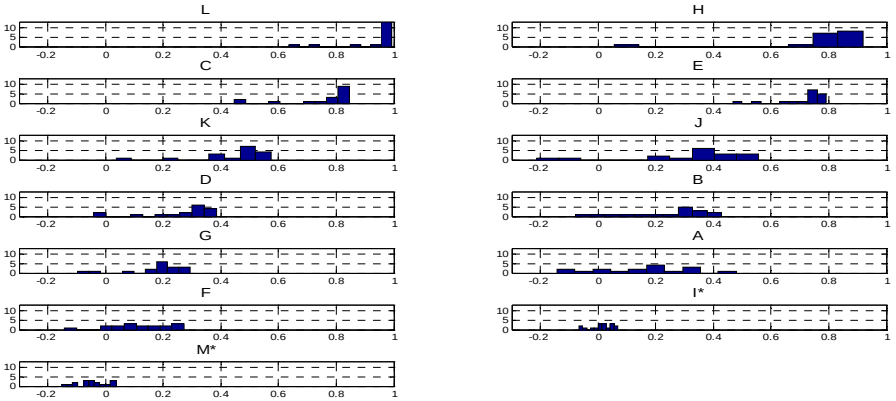
Nature Biotechnology: doi: 10.1038/nbt.1665



d)



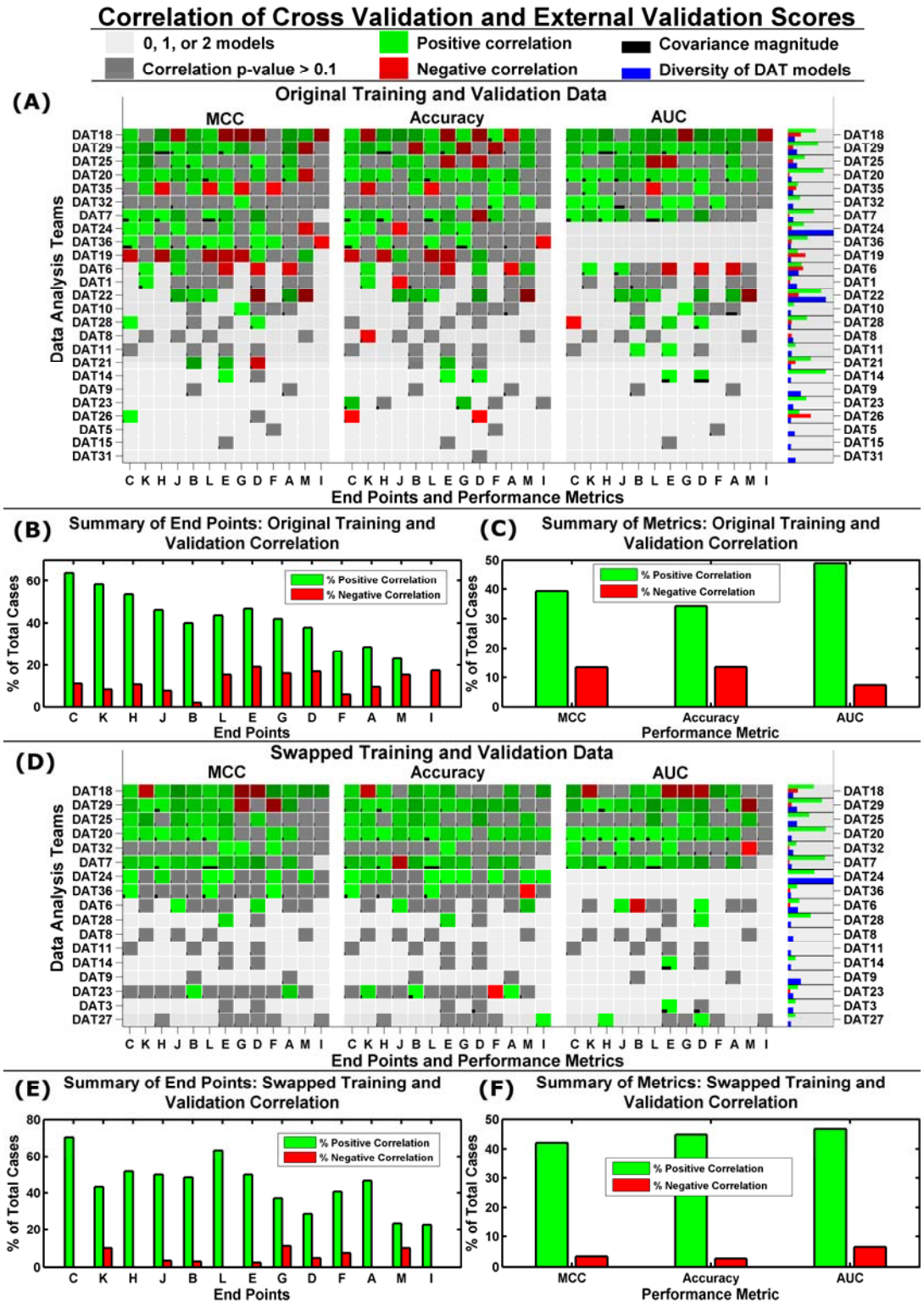
e)



Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Figure 4. Analysis and visualization of Original data External and Cross validation models' prediction performance MCC of nominated models by 17 data analysis teams (DATs) that analyzed all 13 endpoints in the original training-validation experiment. Gray-level "images" (not heat-maps) of respectively: the MCC scores for **(a**) the External Validation (EV) of Original Data and **(b)**, Cross Validation (CV) of Original Data for the 13 End Points and 17 DATs (DAT24, 13, 25, 11, 12, 32, 10, 20, 4, 8, 36, 29, 35, 7, 33 and, 3) in this order. As one can see the darker the image pixel intensity is the lower the MCC score is. One can, by simple observation of the images, see that, in general, there are brighter (whiter) pixel intensities in image (a) than there are in image (b); as the CV MCC values are generally higher than EV MCC values. In particular, one can easily observe from (a) that endpoints A and B were over-estimated in general by most DATs. Additionally, one can also bserve/confirm, again, by looking globally at the image that the first 4 endpoints, L, H, C, and E in both the CV and EV seem to be easily predictable compared to end points G, A, F, M*, and I*. **(c**) shows the gray-level image of the relative errors between the EV and CV MCC values for same end points and organizations described in (a) and (b). Brighter (whiter) pixels show the larger differences (and/or inconsistencies) *in percentile* (%) in the prediction of the related endpoints and organizations. One can observe that the differences are more significant than in the case of Figure S11. Note that since for some endpoints the MAQC-II candidate models were not derived by these DAT teams, in the 2-dimensional "gray image" of Figure S11c the exact and/or approximate locations of the MAQC-II candidate models are pointed to by arrows, for illustration. **(d**) & **(e**) illustrates respectively the number of organizations that made about the same prediction (MCC value), in the CV and EV of the Original data for each of the endpoints, L, H, C, E, K, J, D, B, G, A, F, I*, M*. One can observe by inspection of (c), (d) and (e) how model prediction overestimation (CV→EV) varies and increases as a function of the 13 End points and 17 DATs models.

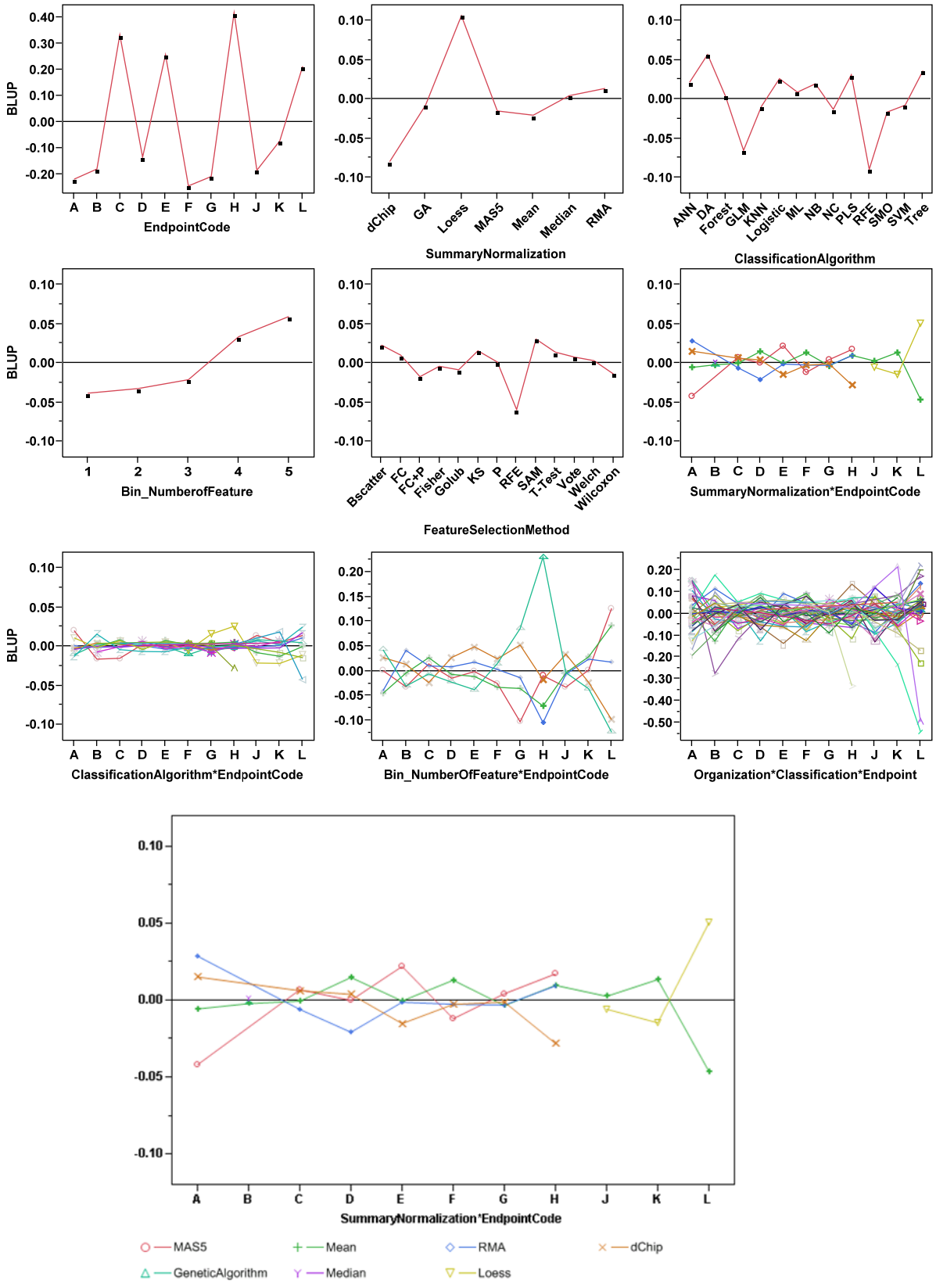
Nature Biotechnology: doi: 10.1038/nbt.1665



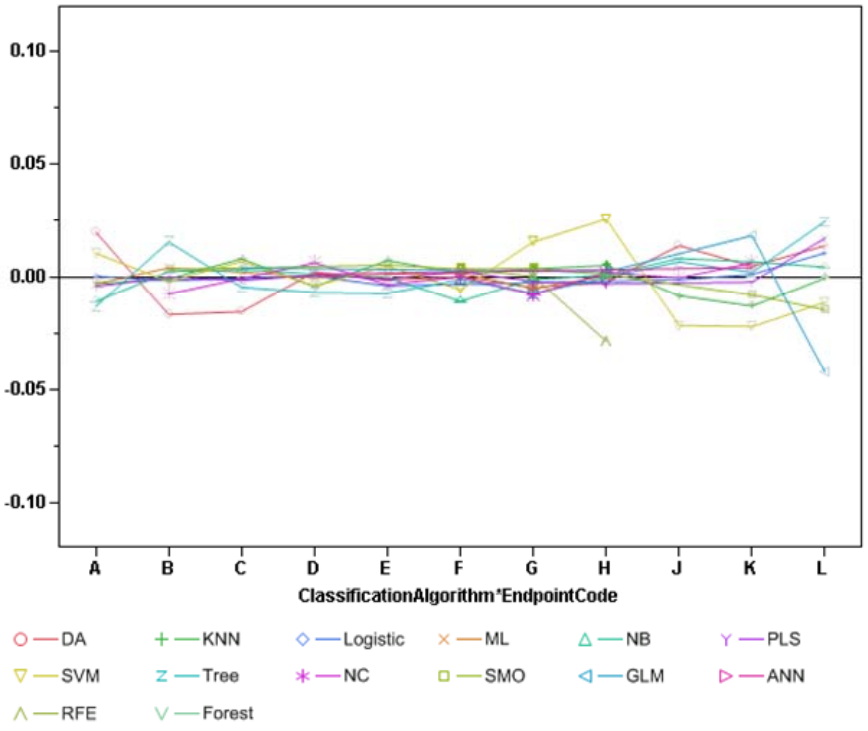
Nature Biotechnology: doi: 10.1038/nbt.1665

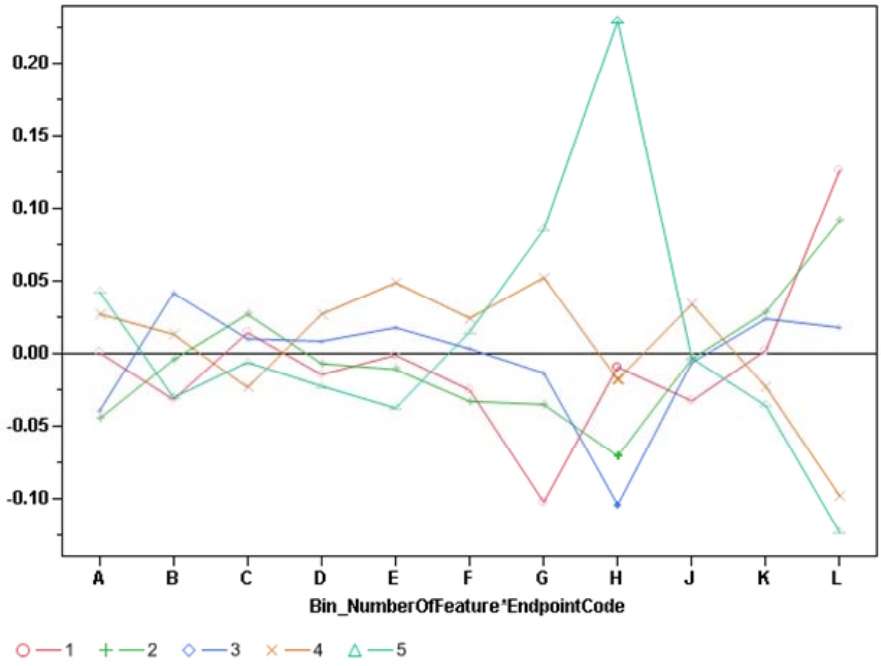
Supplementary Figure 5. Correlation of candidate model selection between internal cross validation and external validation. (A) DATs computed classification model performance using three performance metrics (MCC, accuracy, and AUC) and averaging 10 iterations of 5-fold cross validation. At least three models from both internal cross validation and external validation are required to compute correlation for each DAT and endpoint pair. Light gray squares indicate that only zero, one, or two models are available. Because the calculation of correlation requires three data points, the data analysis teams who were not listed are those who have not provide enough data to compute correlation for at least one endpoint. For example, GHI would have been included in the original training-validation panel if it provided internal performance estimation for the 26 additional models added in validation. Currently, because only 26 of the 52 models listed in Table S1 had both internal and external performance, and 2 models per endpoint, no correlation could be calculated. As another example, FBK only provided one model per endpoint during swap; therefore no correlation can be calculated on a per endpoint basis. Green squares indicate a positive correlation between internal cross validation scores and external validation scores. Red squares indicate negative correlation. The brightness of red and green squares indicates the degree of correlation, i.e., a larger absolute Pearson's correlation coefficient results in a lighter square. Dark gray squares indicate that the p-value of correlation is larger than 0.1. The black bar within each box represents the absolute covariance. Data analysis teams are sorted from top to bottom by decreasing number of endpoints analyzed, then by decreasing total number of models. Endpoints are sorted from left to right by increasing percentage of positive correlations minus negative correlations. The image bar on the right summarizes each DAT with the percentage of positive correlations (green), negative correlations (red), and relative diversity of the DAT (blue). Diversity is a measure of the number of unique feature selection/classification methods used. (**D**) Similar results for the swapped data. (**B**, **E**) Summary of the positive and negative correlations for each end point. (**C**, **F**) Summary of the positive and negative correlations for each performance metric.

Nature Biotechnology: doi: 10.1038/nbt.1665

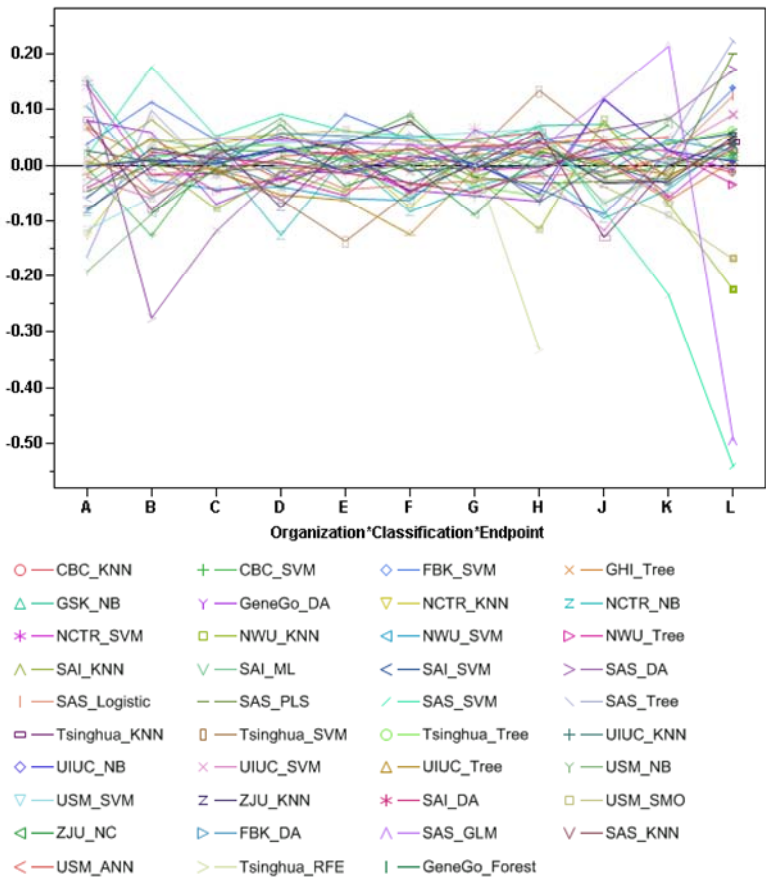


Nature Biotechnology: doi: 10.1038/nbt.1665



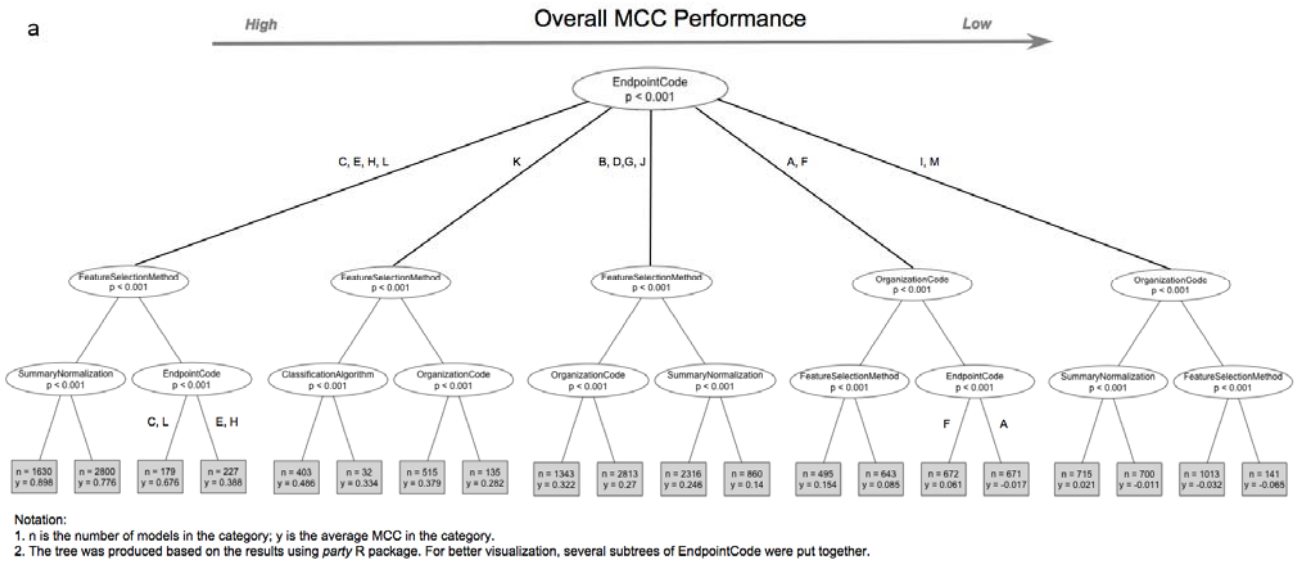


Nature Biotechnology: doi: 10.1038/nbt.1665



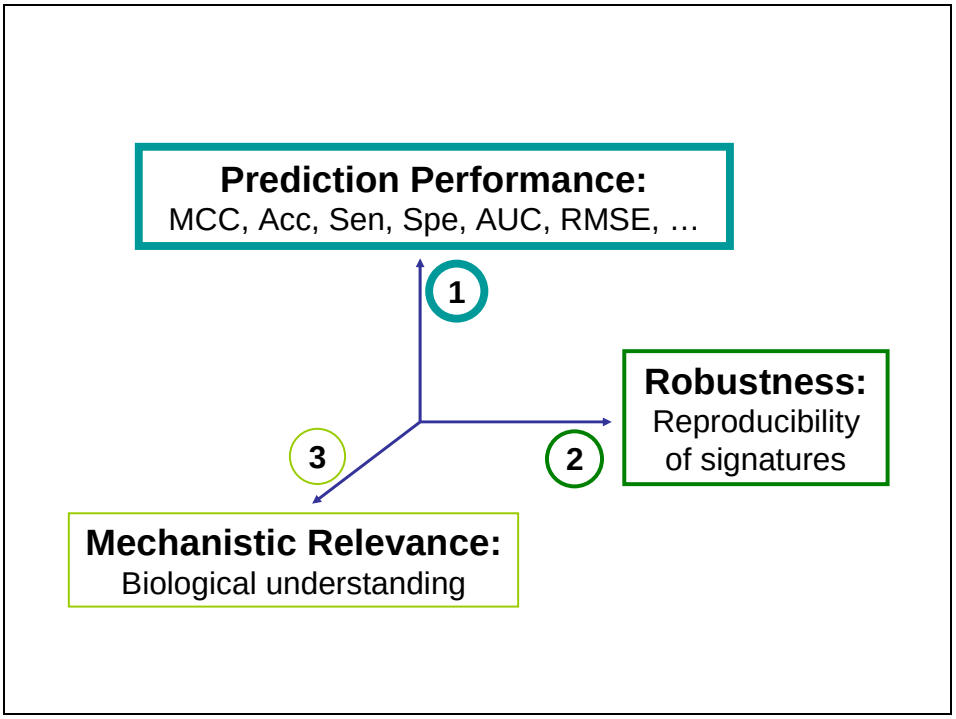
Supplementary Figure 6. Impact of modeling factors on model performance: The empirical BLUPs (Best Linear Unbiased Predictor) of each level for all the factors across 13 endpoints, with clear labeling of interaction terms. See **Figure 4** for more information.

Nature Biotechnology: doi: 10.1038/nbt.1665



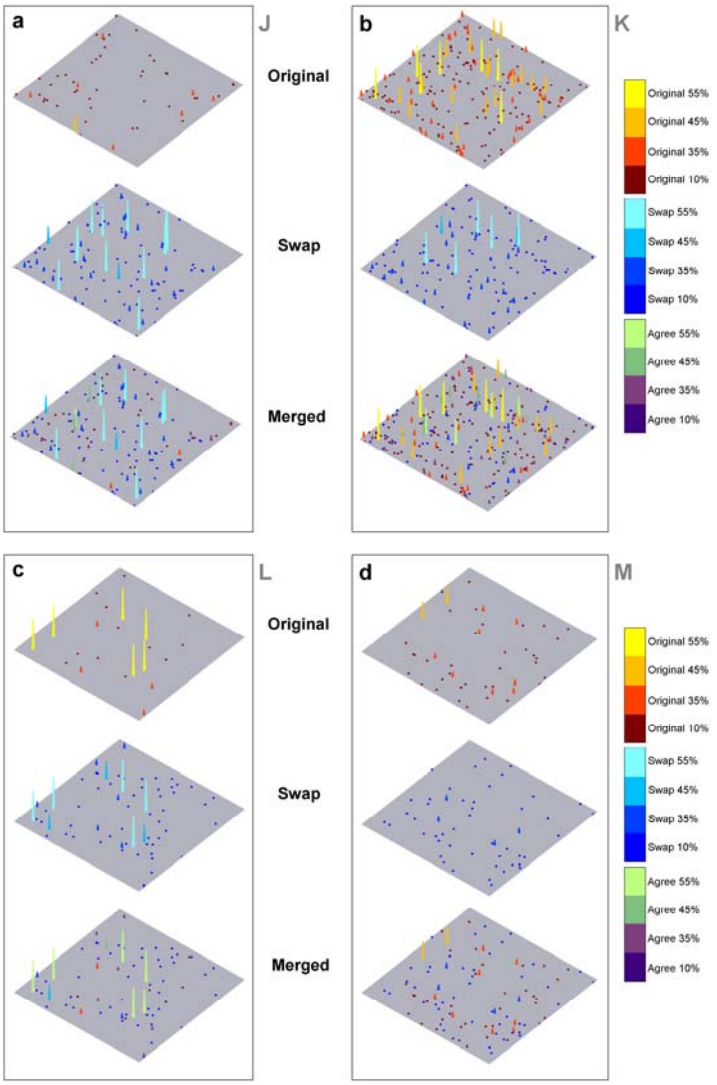
Supplementary Figure 7. A decision-tree model of the relative importance of modeling factors on external validation prediction performance in terms of MCC.

Nature Biotechnology: doi: 10.1038/nbt.1665



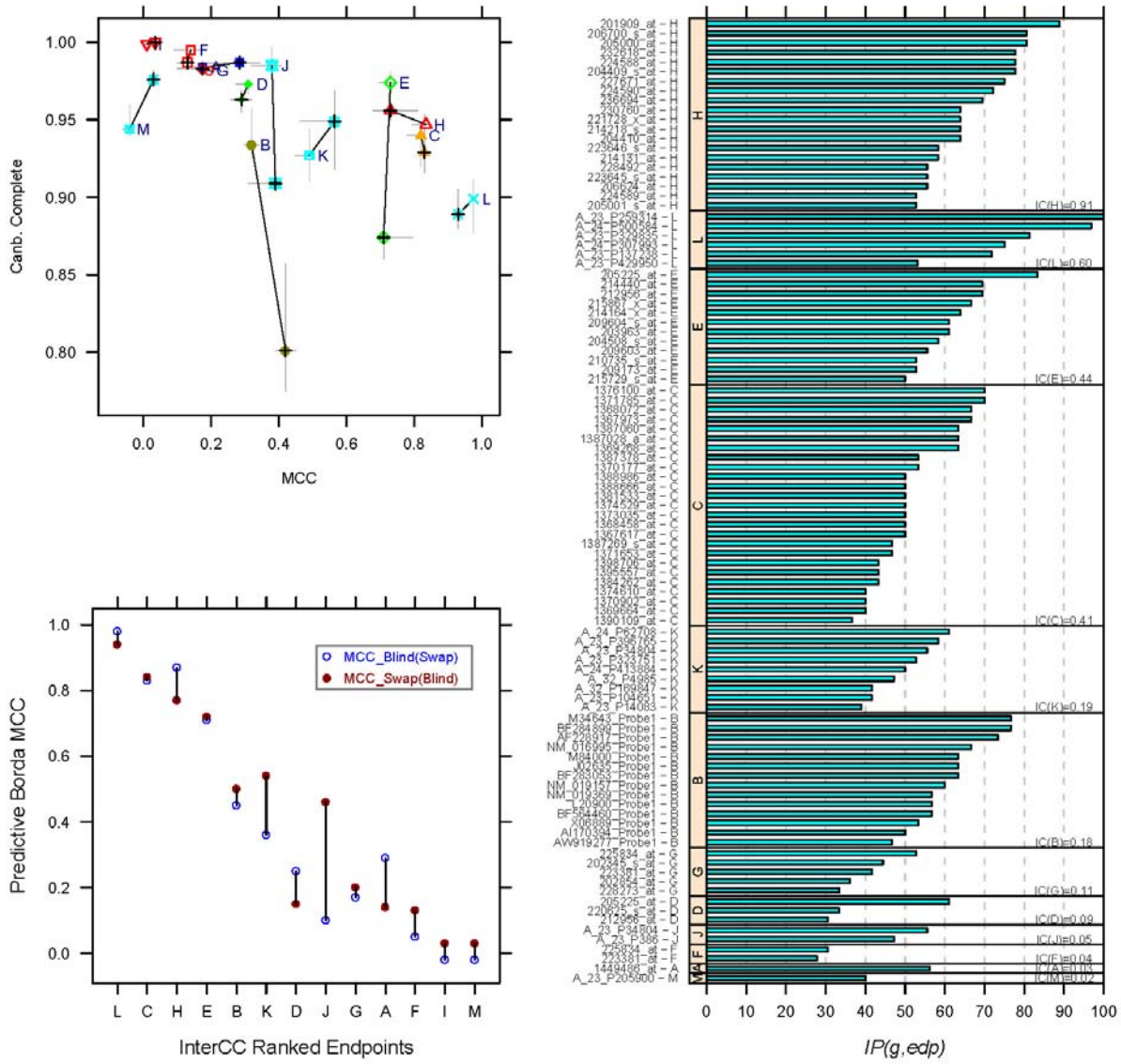
Supplementary Figure 8. Three aspects for assessing the performance of a data analysis protocol (DAP) in decreasing order of priority: prediction performance, robustness, and biological relevance of the gene signatures. Robustness, defined here as the reproducibility of gene signatures across experiments was an important criterion for comparing MAQC-II model performance. Lack of robustness is more likely when features are selected based on ranking significant genes by p-value from a t-test, especially when a small number of genes are selected (Shi, L. et al. *Nat Biotechnol* 24, 1151-1161 (2006) and *BMC Bioinformatics* 9 Suppl 9, S10 (2008)). While the approach can achieve a high sensitivity in selecting true positive genes, the high variance of individual genes with few replicates can result in discordant significant genes lists even in high quality experiments. In contrast, features selected based on the magnitude of expression changes, or fold change, combined with a non-stringent p-value cutoff, tend to be more concordant, providing more robust models. A compromise entails calculating gene variance as a weighted average of individual gene measured variance and the average gene variance for the entire array. Criticism of microarray data for lack of reproducibility has been largely discounted since discordant gene lists are largely the result of selecting genes by small p-value when, in fact, it is the consequence of poor estimates of true variance used to determine the t statistic. MAQC models were also compared on the basis of ability to embody mechanistic relevance of the endpoints. With increasing data on gene and protein functions and the biological and metabolic pathways, as well as growing empirical data on the molecular mechanisms of disease or toxicity, a significant gene list of a model can be examined for consistency with known biology. Consistency between the model's features and the biology can be viewed as enhancing model validity, whereas inconsistency even with accurate prediction could mean that improvements in the model are possible. Biological considerations can also be used to guide feature selection such that a model developed that includes such genes would inherently include the corresponding mechanistic relevance.

Nature Biotechnology: doi: 10.1038/nbt.1665



Supplementary Figure 9. Feature landscapes comparing swap features lists to original feature lists. (a) endpoint J (overall survival milestone outcome of neuroblastoma patients); (b) endpoint K (event-free survival milestone outcome of neuroblastoma patients); (c) endpoint L (gender of neuroblastoma patients); (d) endpoint M (randomly assigned class label). These landscapes show that the endpoint determines the reproducibility of features identified by different teams and the model performance after data swap. The largest peaks indicate that more than 55% of contributing teams selected that feature in their nominated models. The medium peaks indicate that 45%-55% teams selected that feature. The smaller peaks represent features chosen in 35%-45% of the nominated models. Green and purple peaks in the merged landscape indicate peaks chosen at the same rate in the original and the swap analyses. Peaks of different sizes may also be overlapped in the merged image, producing multicolored peaks. Endpoints with few, small peaks indicate difficult biological problems where teams arrived at many diverse solutions with little agreement in feature lists (a). On the other hand, large peaks in the emerged image indicate an easy endpoint to predict and different teams arrived at models that used many common features.

Nature Biotechnology: doi: 10.1038/nbt.1665

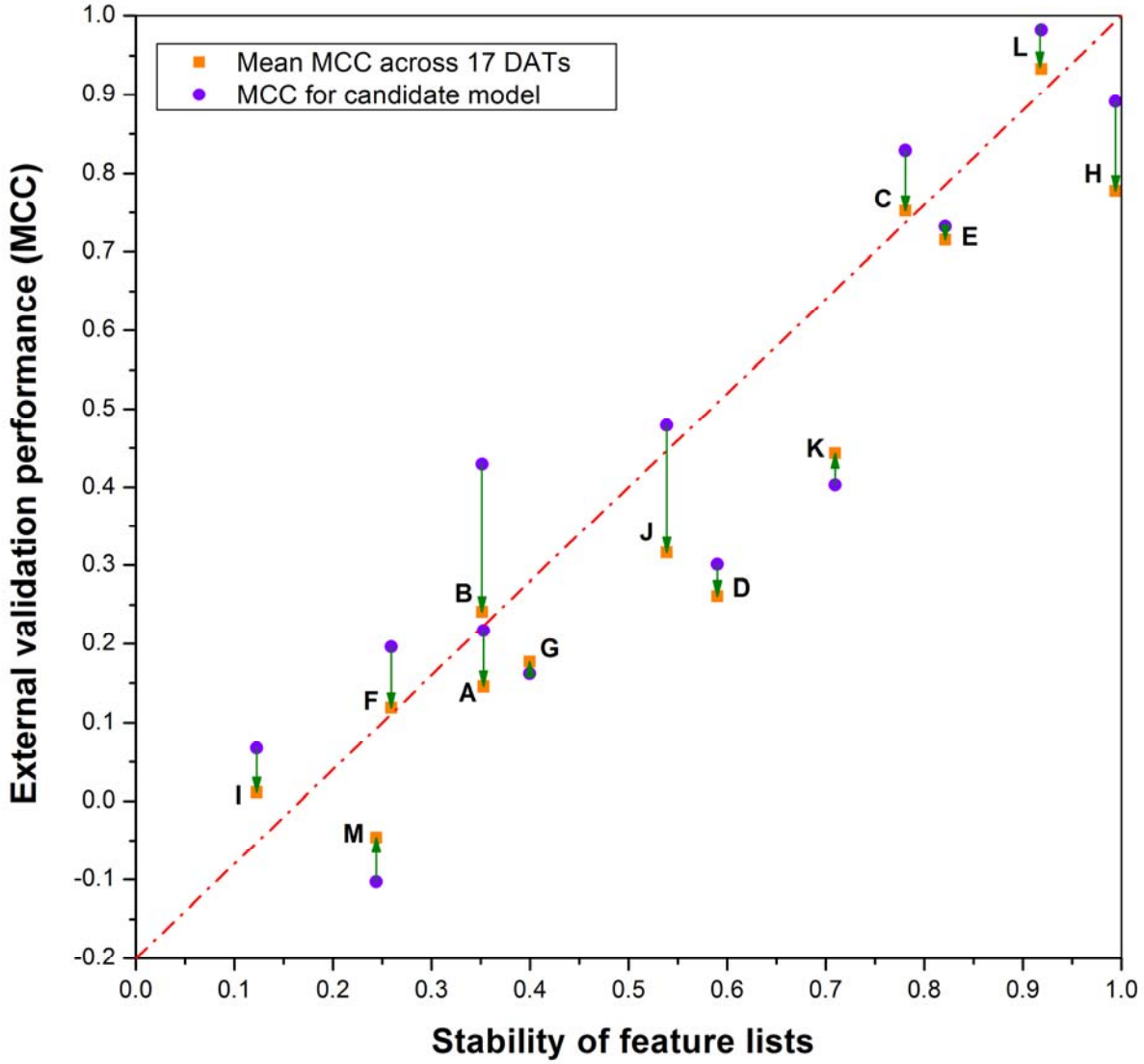


Supplementary Figure 10. Further analysis on the relationship between stability of feature lists and the level of endpoint predictability. The plot in panel (***a)*** , top-left, displays the performance-stability analysis for the candidate models at endpoints(both Blind and Swap), as submitted by the FF organizations. The ordinate is the IntraCC stability to study how lists *change* varying data from Blind to Swap, while differences *between* the lists in the two experiments are analyzed in Panel (***c***), right. The MCC coordinate is the median of the MCCs for endpoint and same experiment. For each point, the rectangles are defined by CI (95% bootstrap Student's *t*). The area of the plot can be split into 4 main zones: LowerRight (good classification performances and good stability), UpperRight (good classification performances and bad stability), UpperLeft (bad classification performances and bad stability) and LowerLeft (bad classification performances and bad stability). Letters nearby points indicates the endpoint, while a plus sign marks the points corresponding to the Swap experiments. Blind and Swap experiments are

Nature Biotechnology: doi: 10.1038/nbt.1665

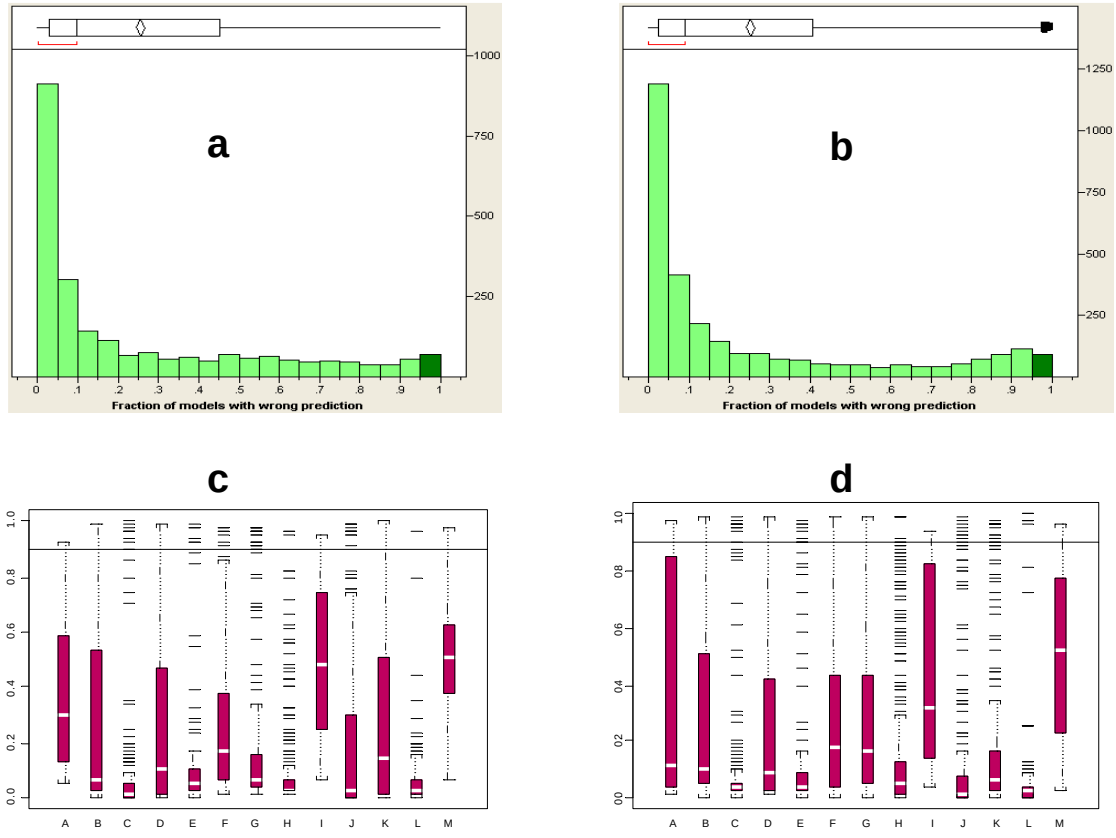
connected by segments. Panel (**b**), bottom-left, analyzes most extracted common features for each endpoint. We first compute the median m(*edp*, *exp*) of the length of the FF submitted lists for each endpoint *edp* and for each experiment *exp* (*m* can be different in the Blind and the Swap experiment for the same endpoint). Then we compute the corresponding Borda lists and consider its first m(*edp*,*exp*) elements obtaining the top-m(*edp*, *exp*) partial Borda lists B(*edp*, *exp*). For each endpoint *edp* we then consider the intersection between the Borda partial lists for the Blind and the Swap experiments, i.e. the number N(*edp*) of genes occurring in both the lists for each endpoint. The indicator IC(*edp*)=2*N(*edp*)/(med(*edp*,Blind)+med(*edp*,Swap)) is used in Panel **b** to decreasingly rank the endpoints. In each subplot, a silhouette graph describes the common genes for the endpoint: decreasing bar lengths indicate for each gene *g* the mean number of its extractions normalized by the number L(*edp*) of features for the endpoint: IP(g,edp)=(E(g,edp,Blind)+E(g,edp,Swap))/(2*L(edp)). The randomly labeled endpoints are lowest ranked, while L, E, and H are ranking high. Panel (**c**): Analysis of inter-experiment list stability compared it with external MCC performance. For each pair *(edp, exp)* the partial top-med(*edp, exp)* Borda list B*(edp, exp)* is used in alternate experiment. Swap experiment is run with features belonging to B(*exp, Blind*) by means of a 10x5-CV with linear Support Vector Machine models, retrieving a MCC_{Blind}(Swap) value. Analogously, we compute the MCC_{Swap}(Blind) for the same endpoint. InterCC on the x axis ranks endpoints for increasing list variability between experiments.

Nature Biotechnology: doi: 10.1038/nbt.1665



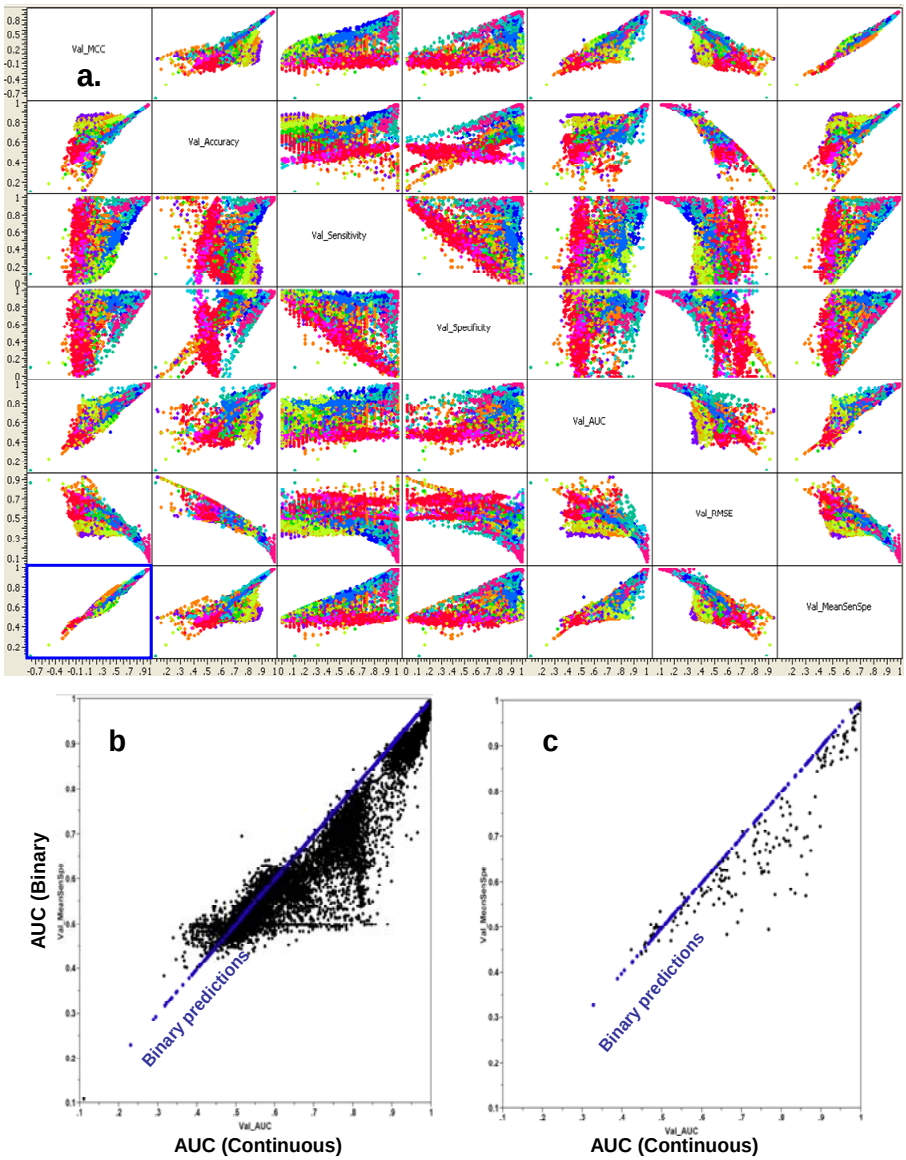
Supplementary Figure 11. The stability of feature lists is positively correlated with endpoint predictability. The feature lists used in the nominated models from the original training-validation experiment were compared to those from the swap experiment. Contingency tables and the Fisher's exact test statistic were used to summarize a probability that the degree of overlap between the lists is the result of random chance. The *P* values were averaged over all pairs of feature lists. The resulting measure of the stability (1 – *P*) of feature lists positively correlates with external validation performance.

Nature Biotechnology: doi: 10.1038/nbt.1665



Supplementary Figure 12. Prediction error on a per-sample basis - some samples were consistently misclassified by almost all models. (**a**) prediction of the original validation set; (**b**) prediction of the swap validation set (i.e., the original training set); (**c**) sample prediction error rates by endpoint for the original validation; and (**d**) sample prediction error rates by endpoint for the swap validation. One reason for hard-to-predict samples is that the "true" label is wrong. For example, for endpoints H and L (sex), a female patient may be recorded as male (or vice versa). Another reason is that the hard-to-predict samples represent a distinct subset of samples which just cannot be reliably predicted given the information in the datasets. These samples may represent subsets that would be identifiable given other data or just subsets whose class membership follows a different set of rules (e.g., outcomes that follow a different causal pathway). Whatever the reason that hard-to-predict samples are hard to predict, it seems that most if not all nontrivial datasets have them. They even occur in datasets in which class labels have been assigned at random.

Nature Biotechnology: doi: 10.1038/nbt.1665



Supplementary Figure 13. The seven performance metrics measure different aspects of the prediction performance of a model. Shown in (**a**) is the pair-wise correlation between seven prediction performance metrics for 18,303 models. AUC values favor modes for which continuous prediction values are provided (**b** and **c**). Notice that for most models with continuous prediction values, AUC (Continuous) is greater than AUC (Binary) in which the continuous prediction values were dichotomized based on the prediction decision value pre-defined at the training stage. Different performance metrics (MCC and AUC) can lead to different model selection, highlighting the importance of the choice of performance metric. Some models showed reasonably good performance in terms of AUC; however, their performance in terms of MCC appeared to be low. The main reason for this discrepancy was that the decision threshold determined during the model development stage turned out to be non-optimal for the validation data (see **Supplementary Documents 1** and **2** in **Supplementary Data**). Decision threshold is an important part of a predictive model; it must be determined during the training stage and should be frozen. This illustrates the need for guidelines for determining decision thresholds.

Nature Biotechnology: doi: 10.1038/nbt.1665

Documents in Supplementary Information

Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Document 1:

Validation performance of 318 MAQC-II nominated models

In the following pages of this document, one page contains the results for one particular endpoint (A, B, C, … , M).

On each page, there are six plots:

1. [Top Left] ROC curves: for models giving continuous or quasi-continuous outputs, the empirical ROC curves are plotted (green solid line). For models giving binary outputs, a dot is plotted at (1-Specificity, Sensitivity) (black circle). The RBWG candidate model is plotted in red color.

2. [Top Right] AUC: the estimated AUC and its approximate 95% confidence interval. Green: models giving continuous or quasi-continuous outputs; Black: models giving binary outputs; Red: the RBWG candidate model.

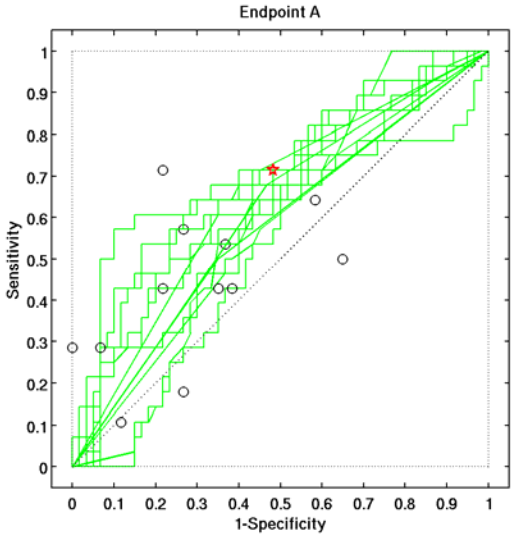
3. [Middle Left] Plot of the binary performance in the ROC space: for models giving continuous or quasi-continuous outputs, a pre-specified dichotomizing threshold of 0.5 is applied (green).

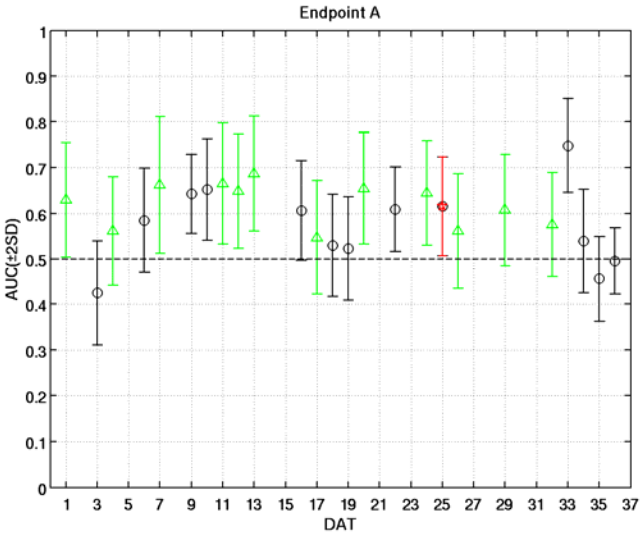
4. [Middle Right] MCC: for models giving continuous or quasi-continuous outputs, two MCC results are given. Green: dichotomizing the outputs using the pre-specified threshold of 0.5; Blue: dichotomizing the outputs using the threshold value that maximizes the MCC on the validation dataset.

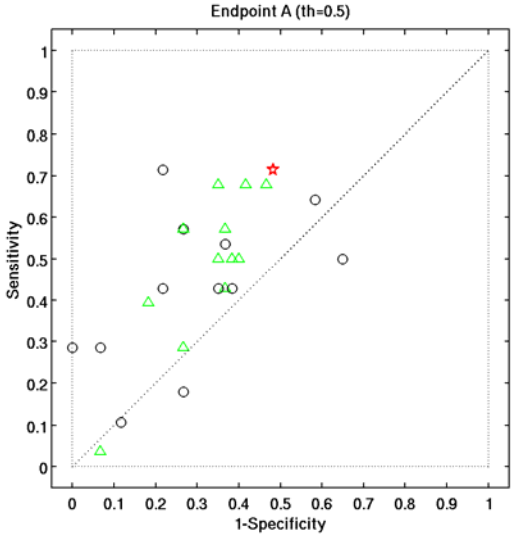
5. [Bottom Left] Plot of the binary performance in the ROC space: for models giving continuous or quasi-continuous outputs, the threshold value that maximizes the MCC on the validation dataset is applied (blue).

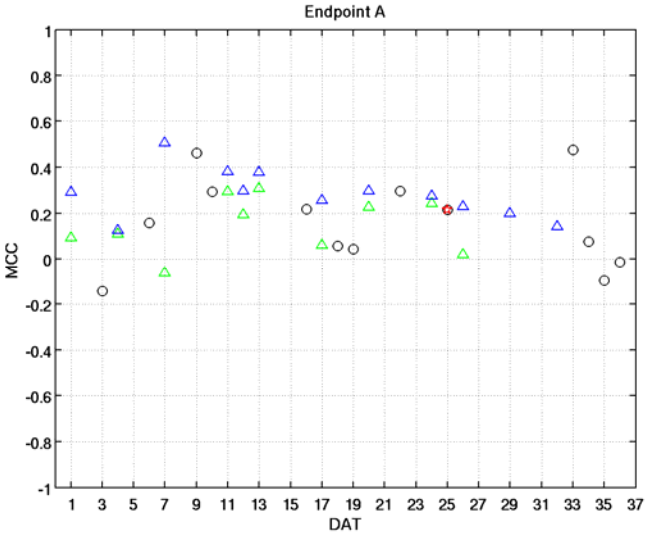
6. [Bottom Right] Plot of the constant MCC contours in the ROC space for the prevalence of the endpoint.

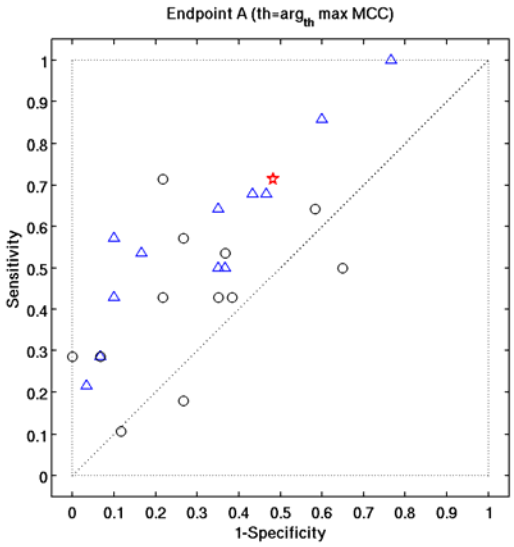
Nature Biotechnology: doi: 10.1038/nbt.1665

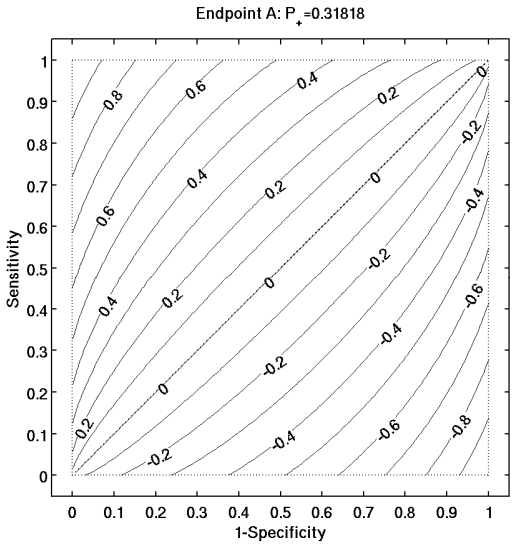




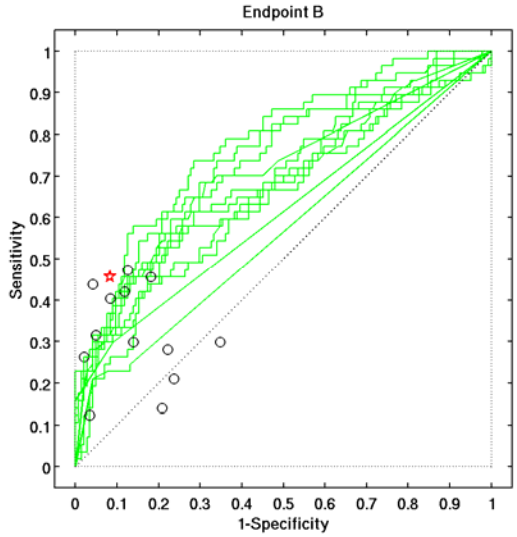


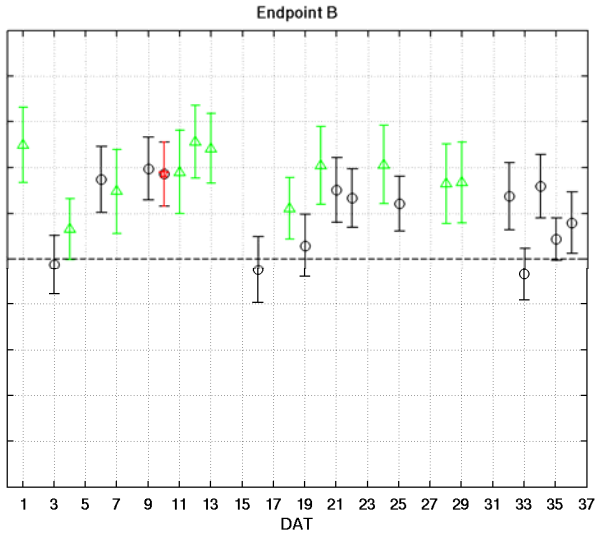


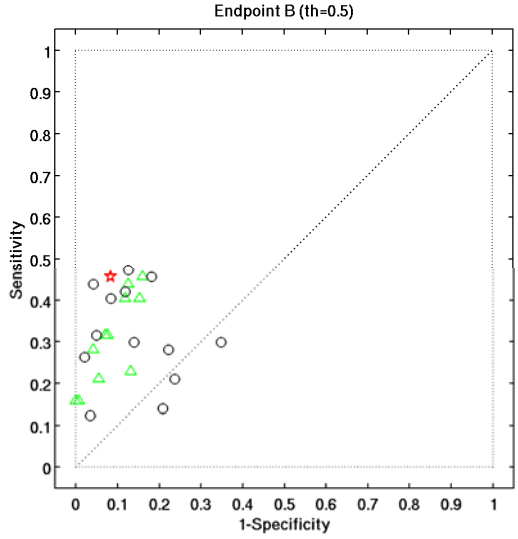


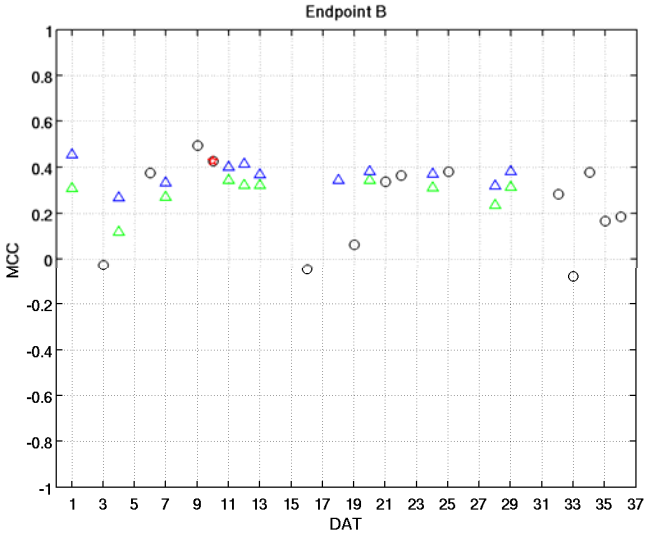


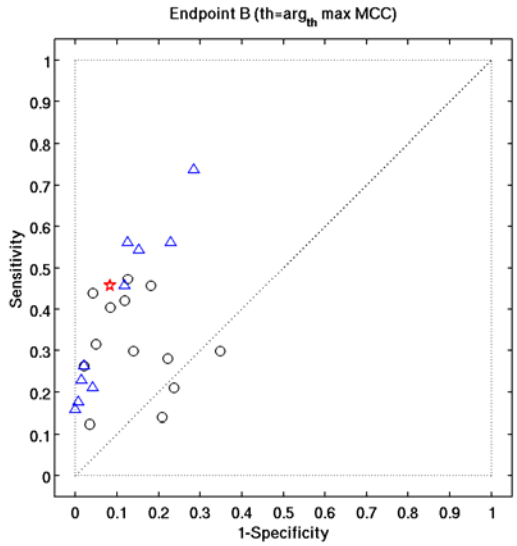
Nature Biotechnology: doi: 10.1038/nbt.1665

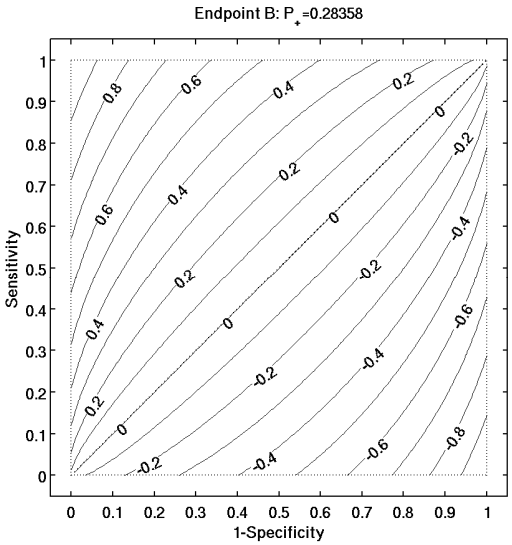




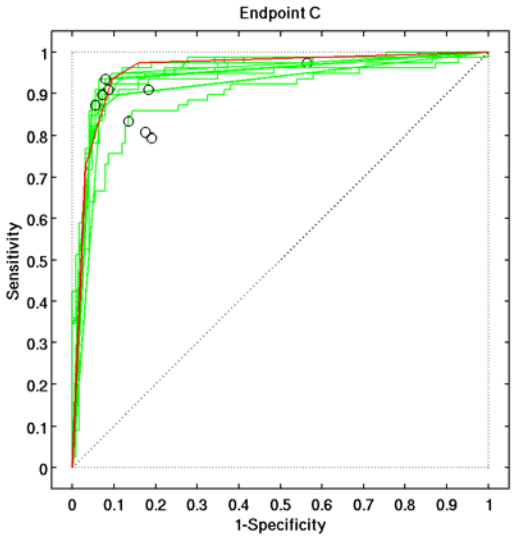


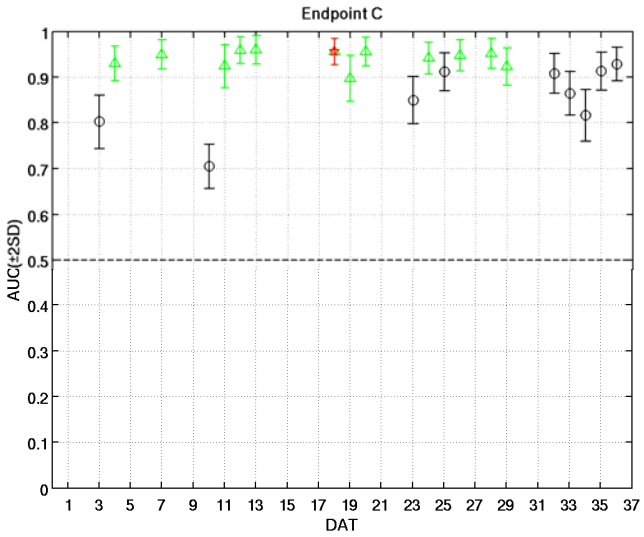


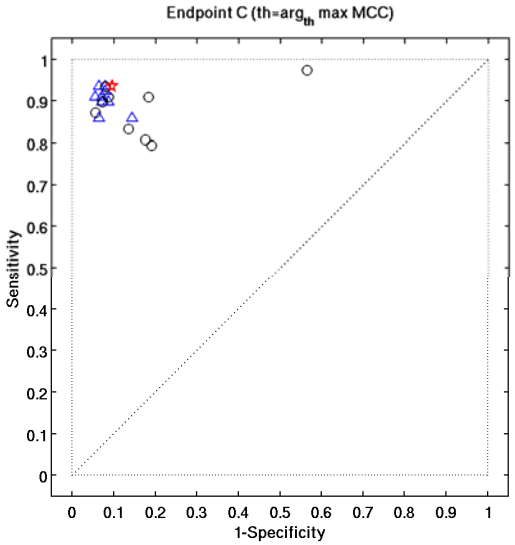


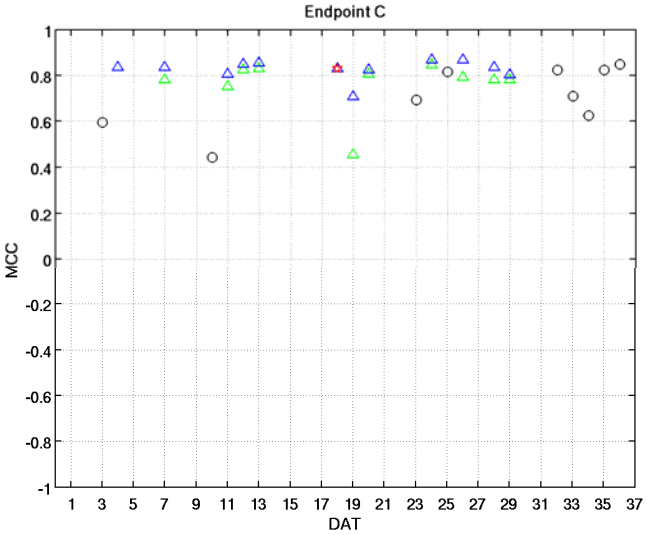


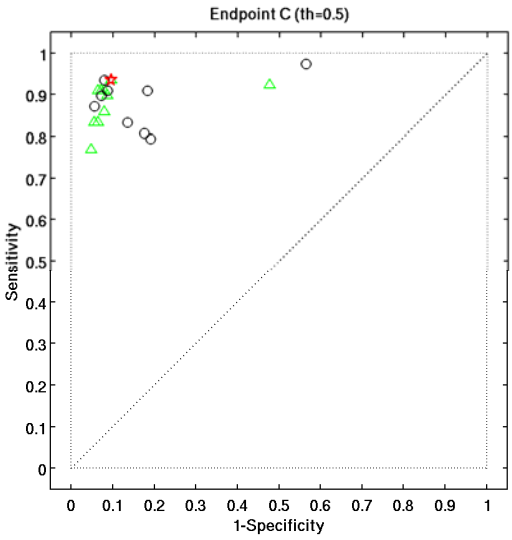
Nature Biotechnology: doi: 10.1038/nbt.1665

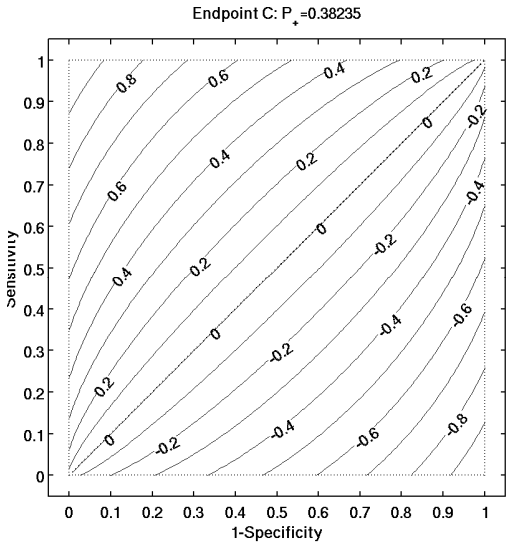




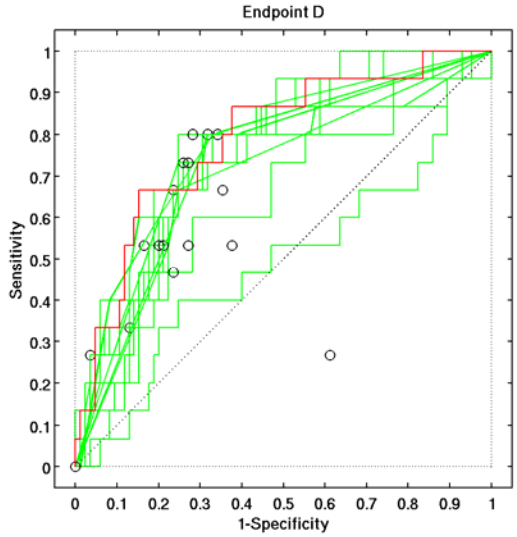


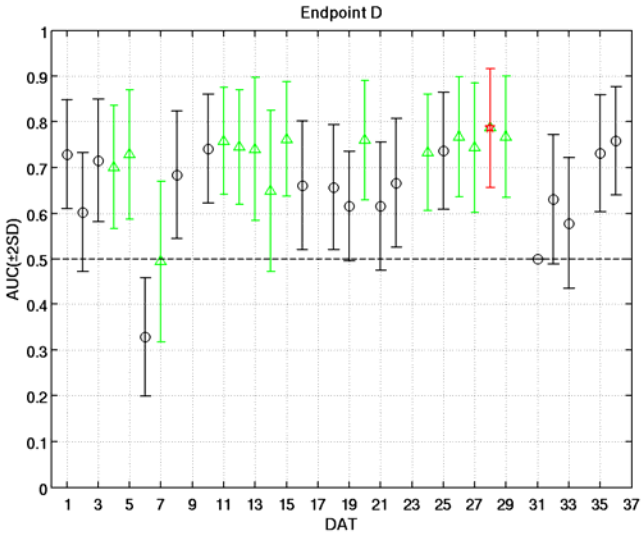


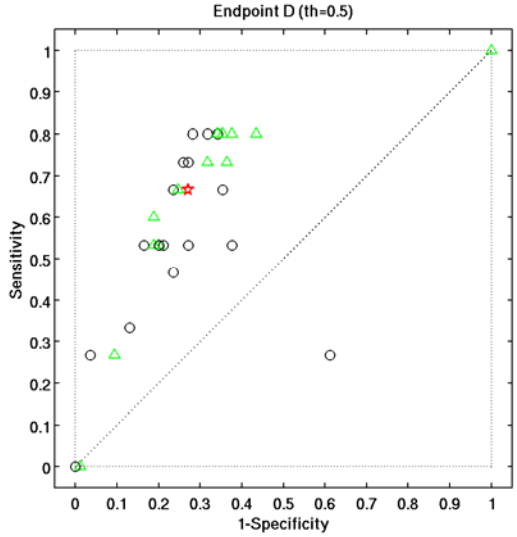


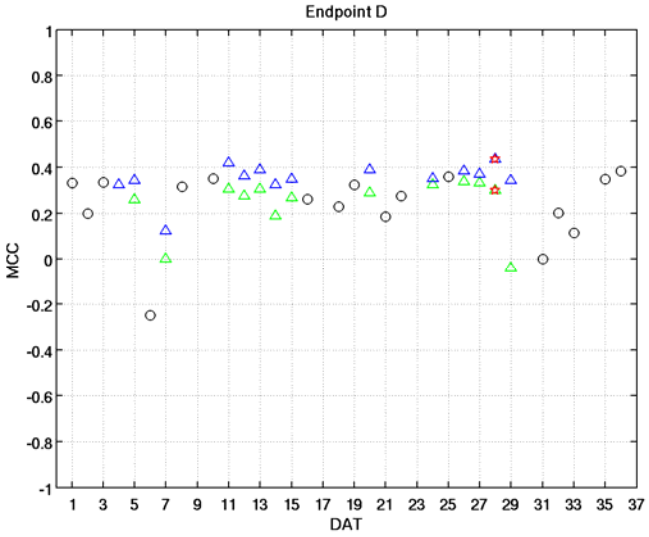


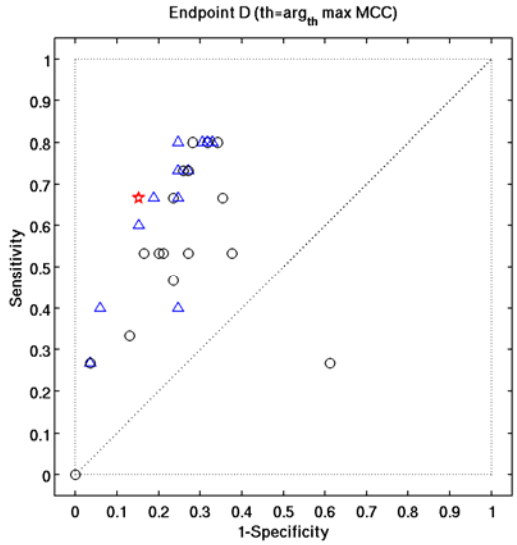
Nature Biotechnology: doi: 10.1038/nbt.1665

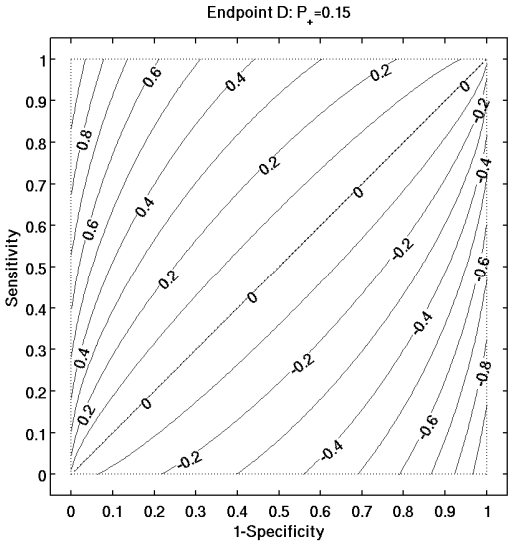




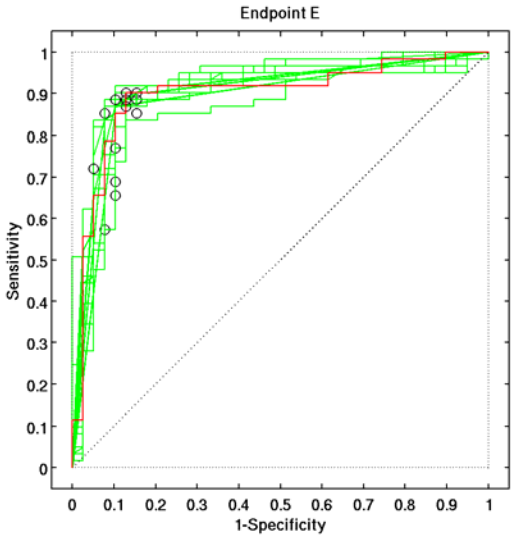


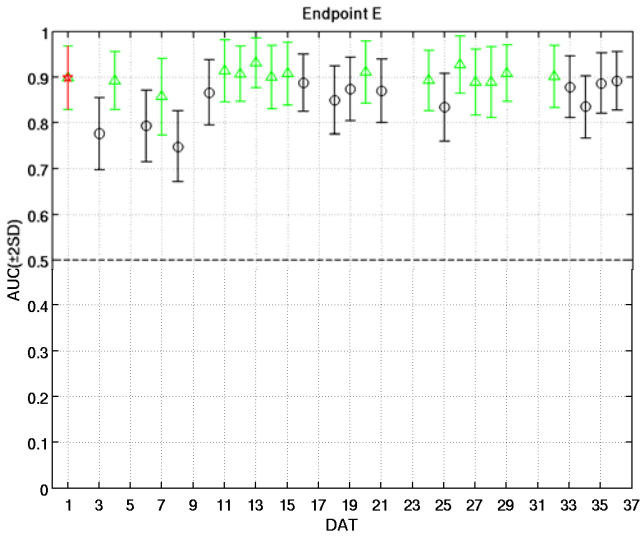


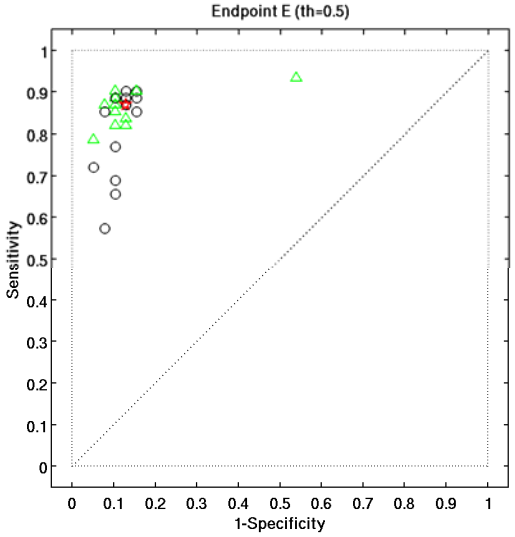


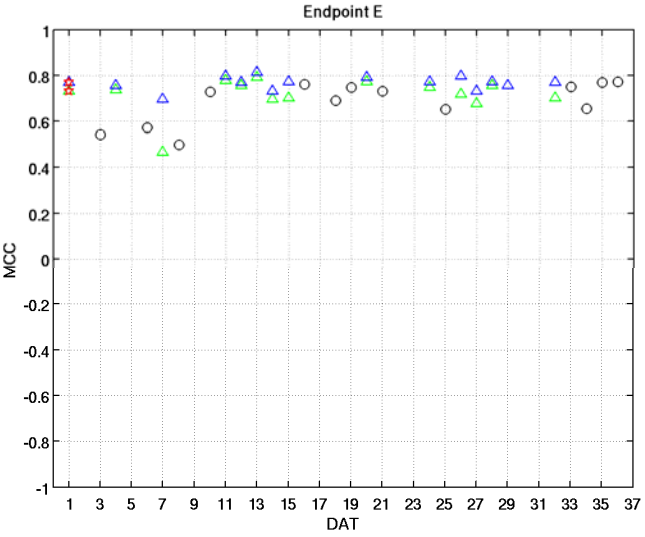


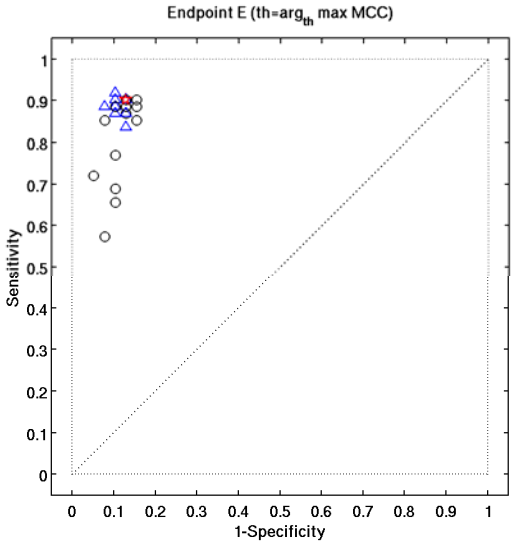
Nature Biotechnology: doi: 10.1038/nbt.1665

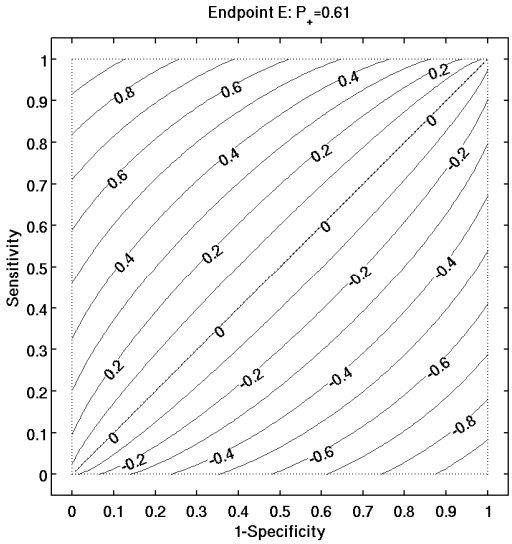




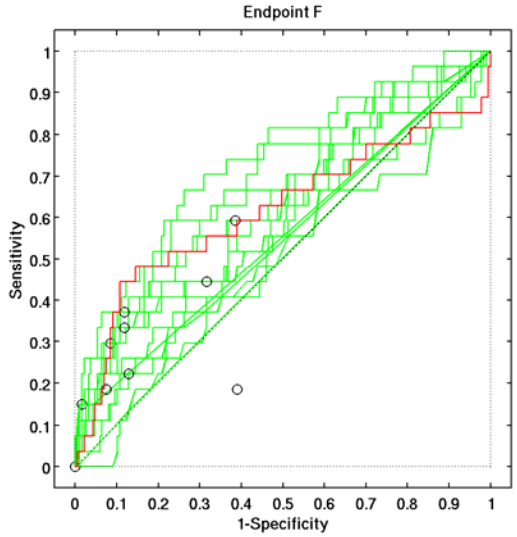


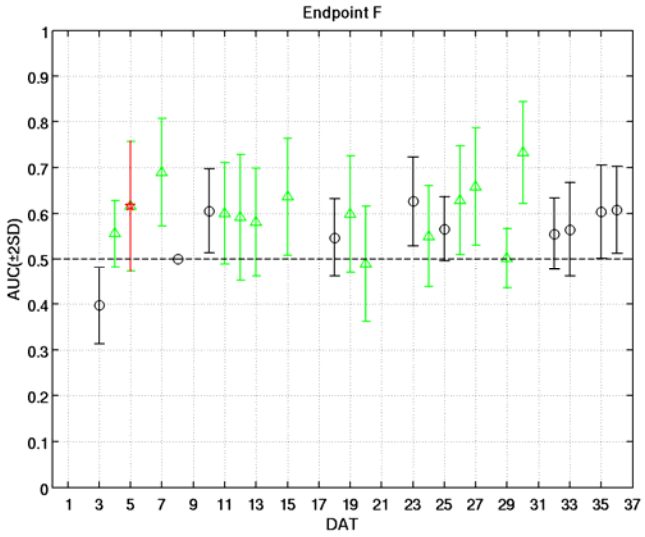


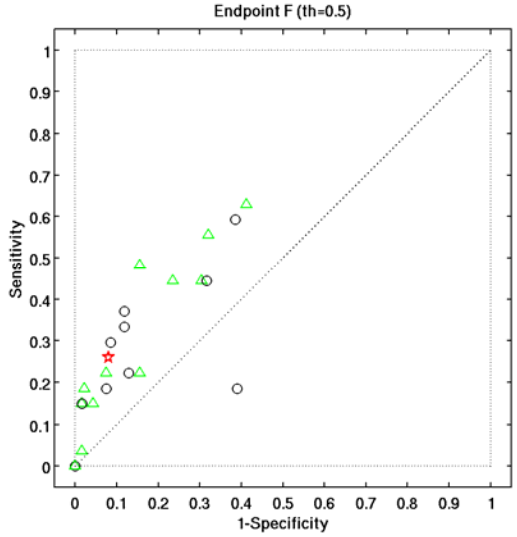


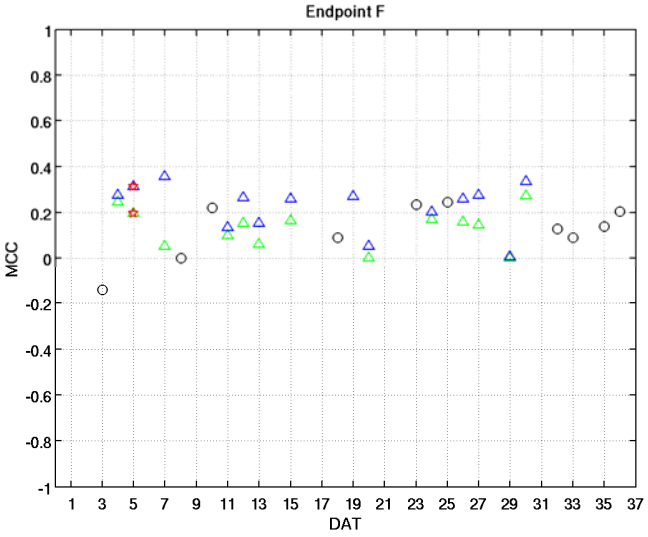


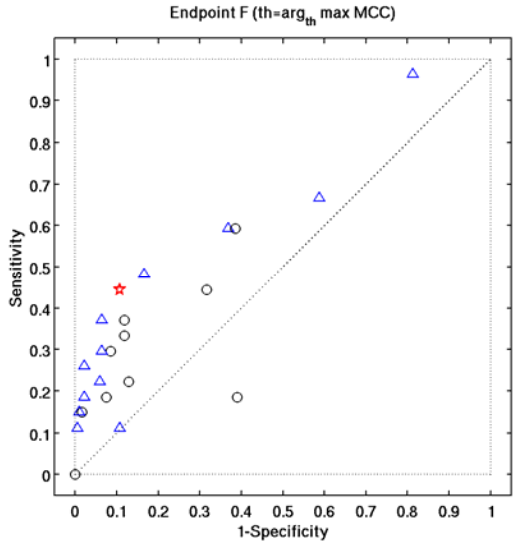
Nature Biotechnology: doi: 10.1038/nbt.1665

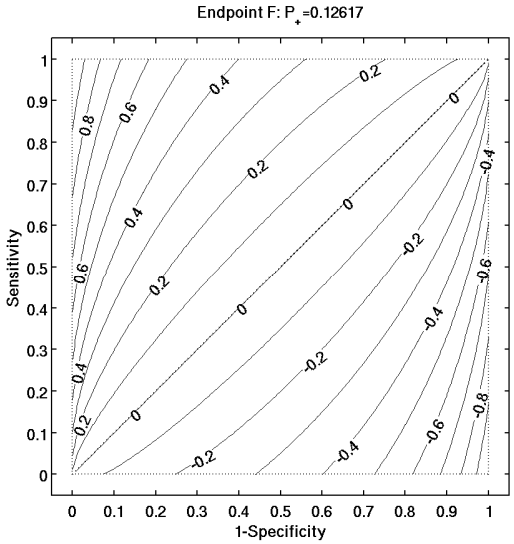




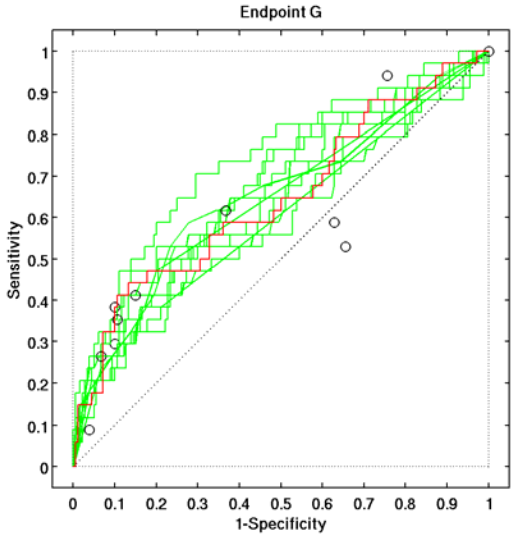


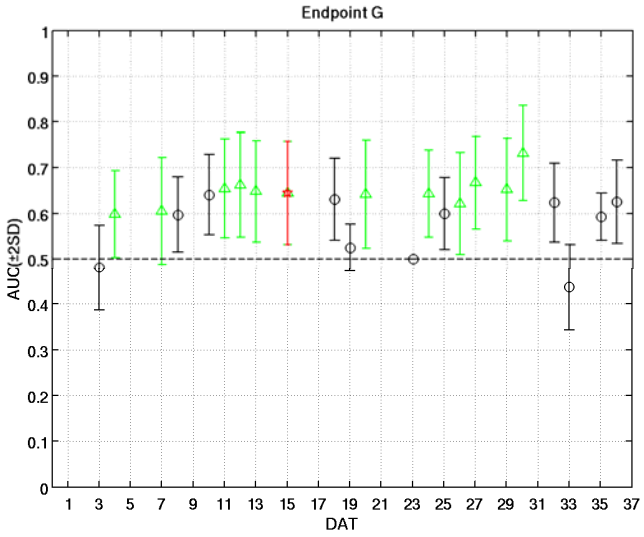


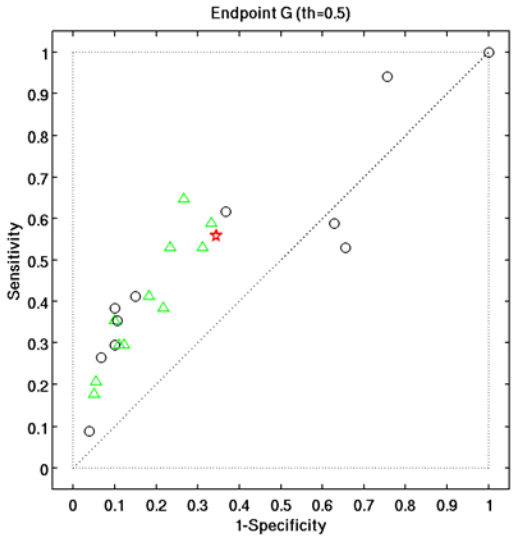


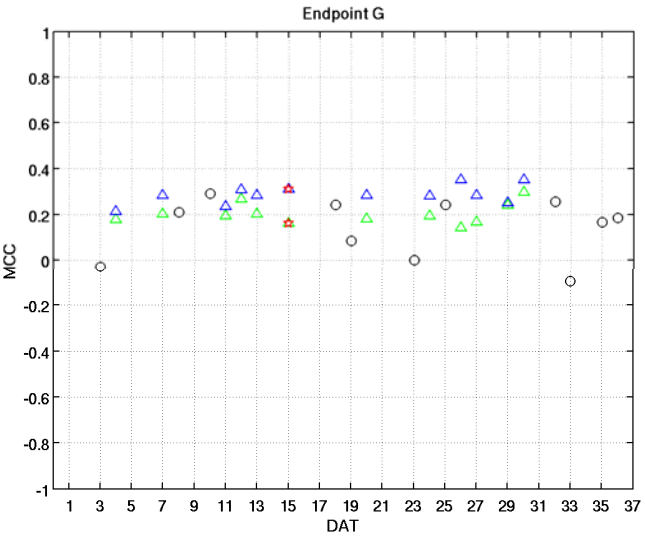


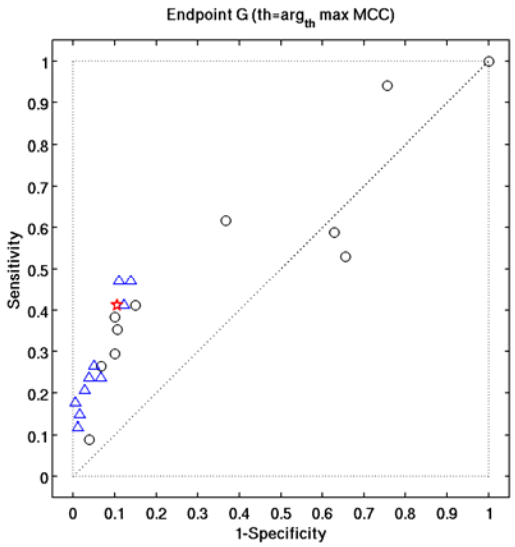
Nature Biotechnology: doi: 10.1038/nbt.1665

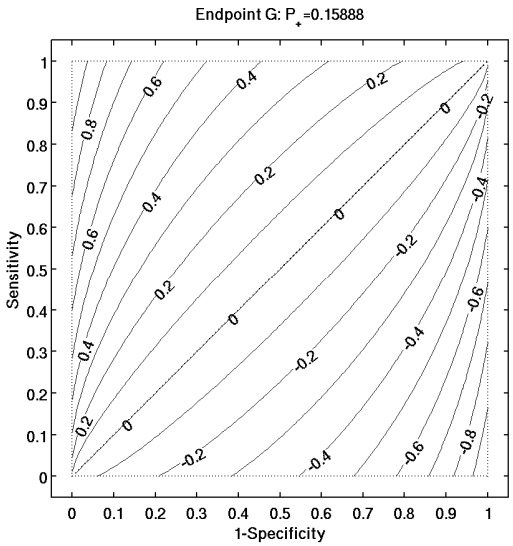




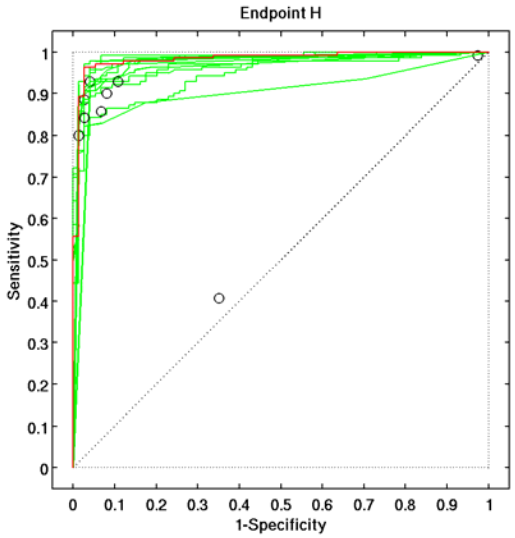


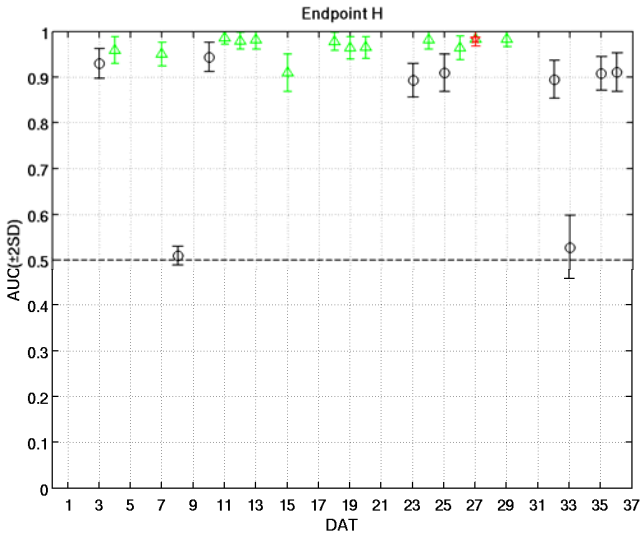


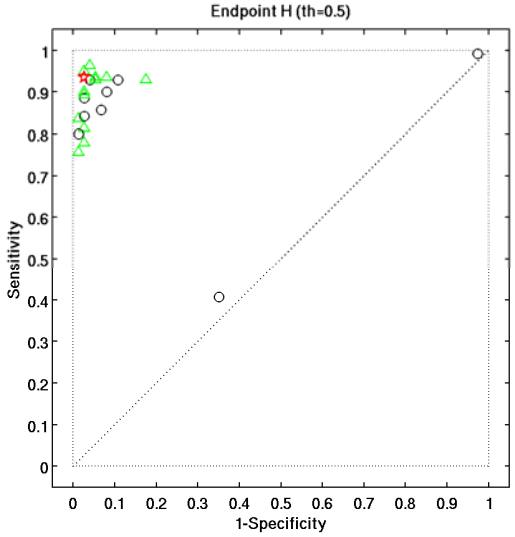


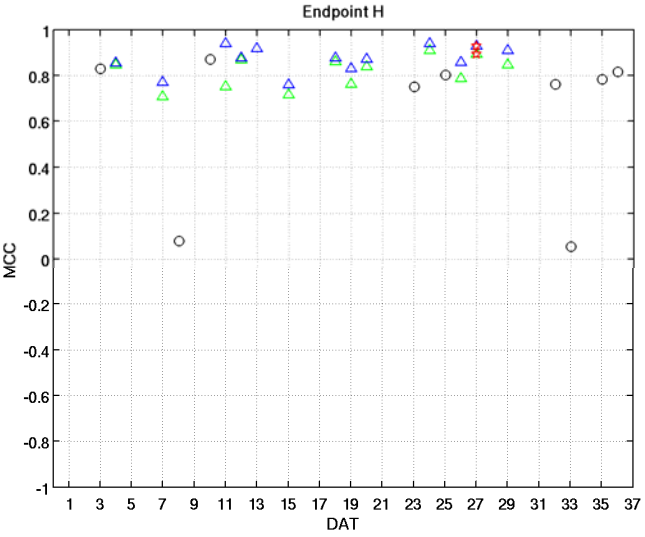


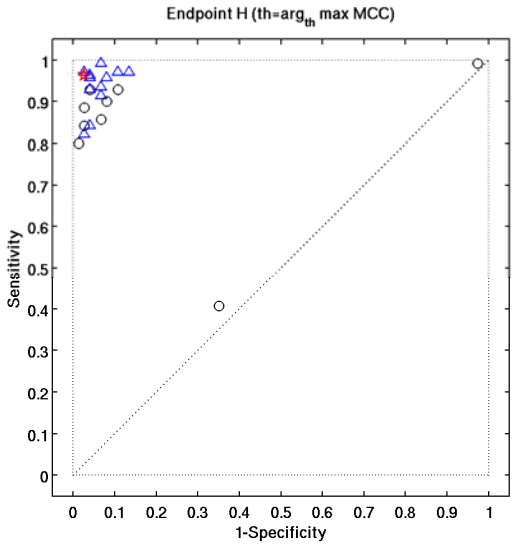
Nature Biotechnology: doi: 10.1038/nbt.1665

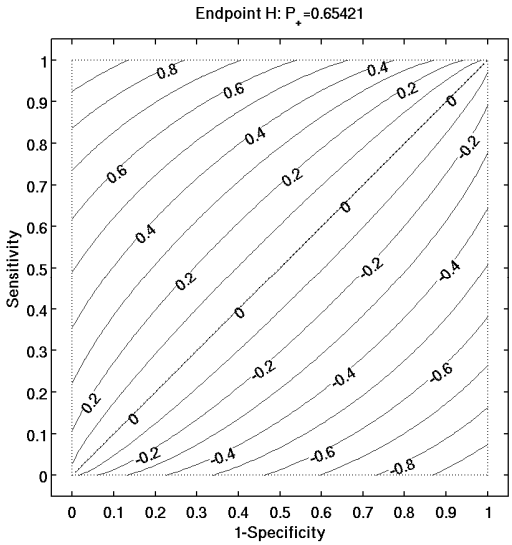




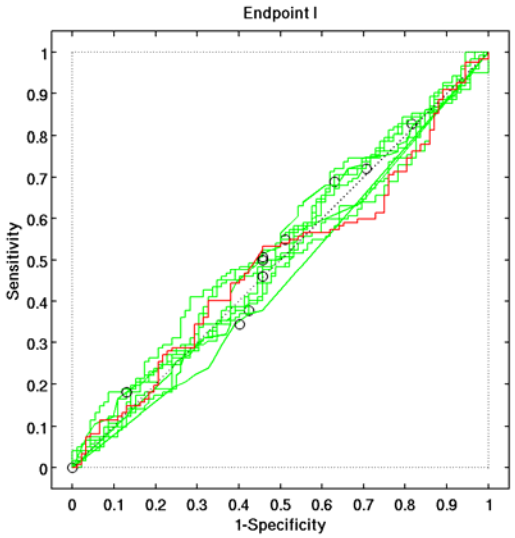


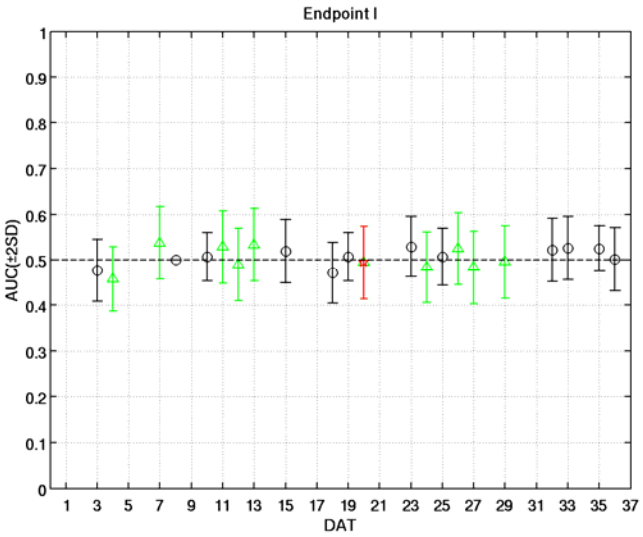


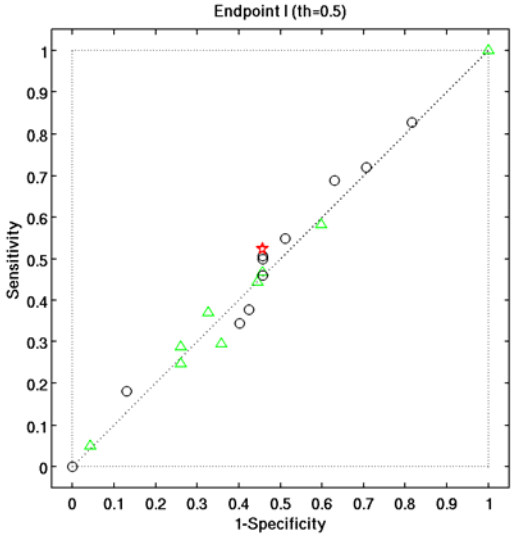


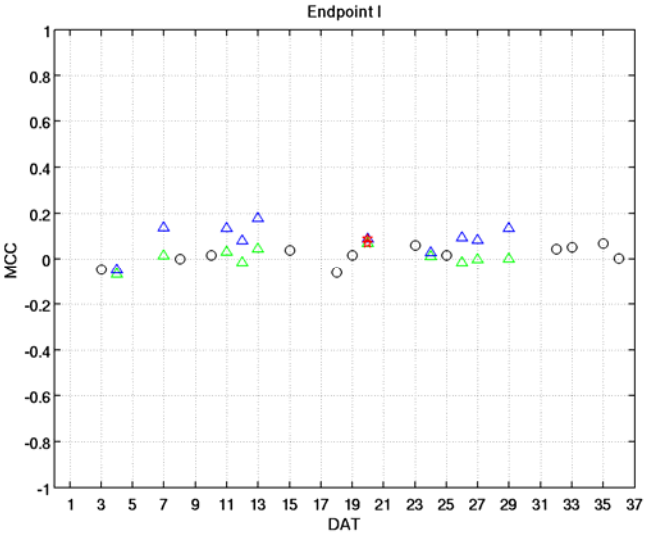


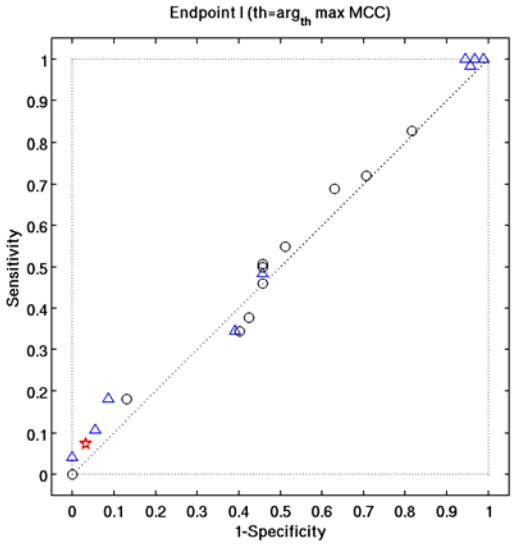
Nature Biotechnology: doi: 10.1038/nbt.1665

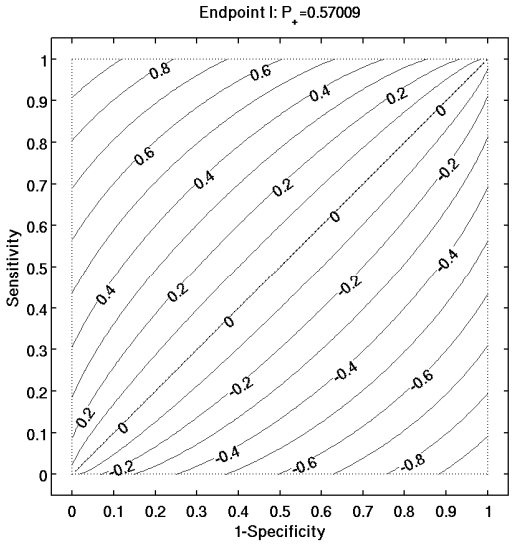




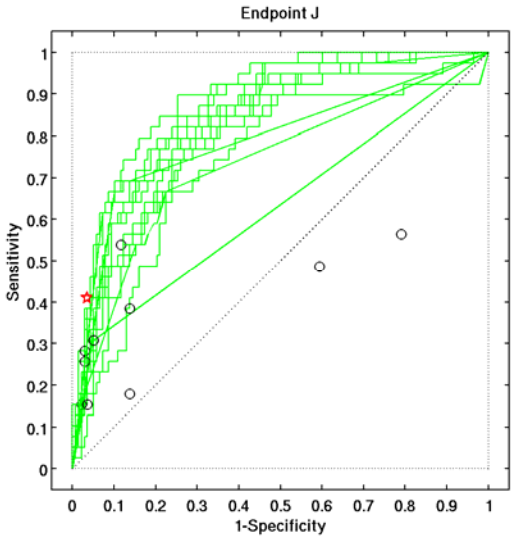


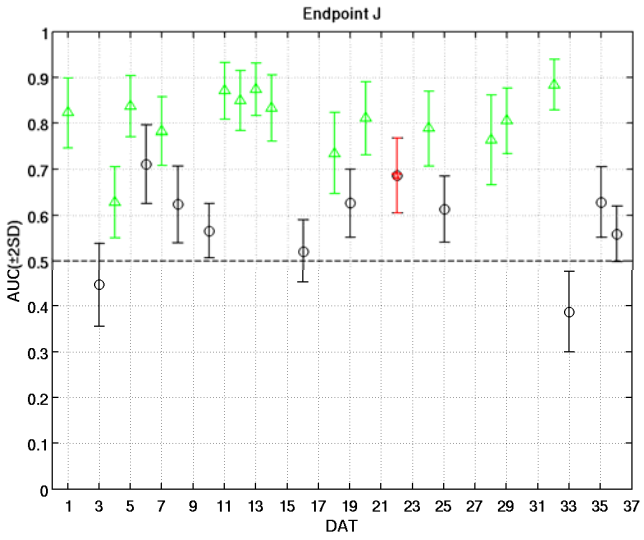


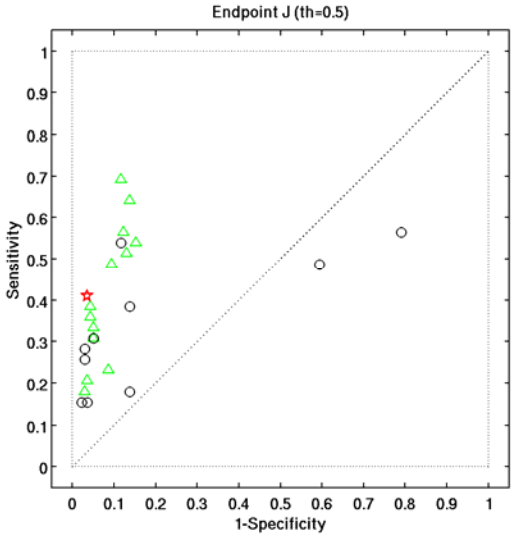


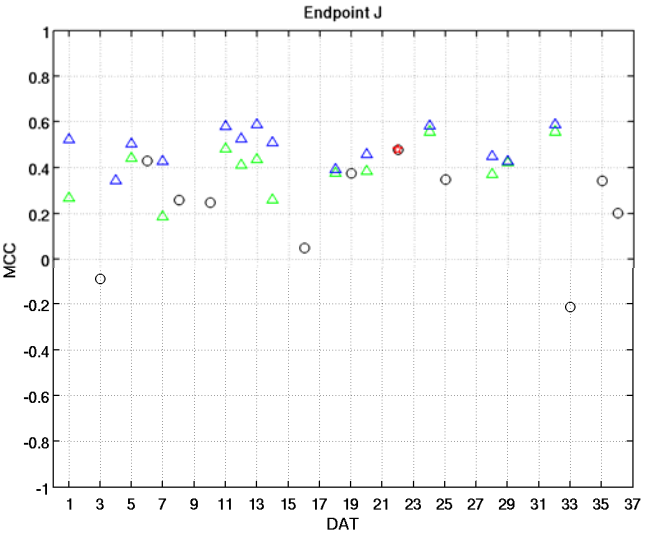


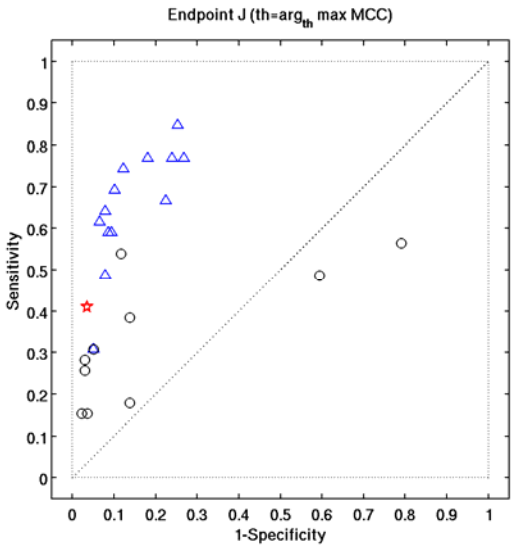
Nature Biotechnology: doi: 10.1038/nbt.1665

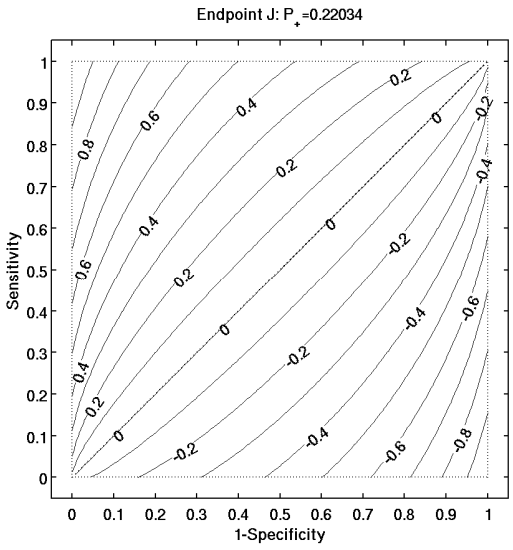




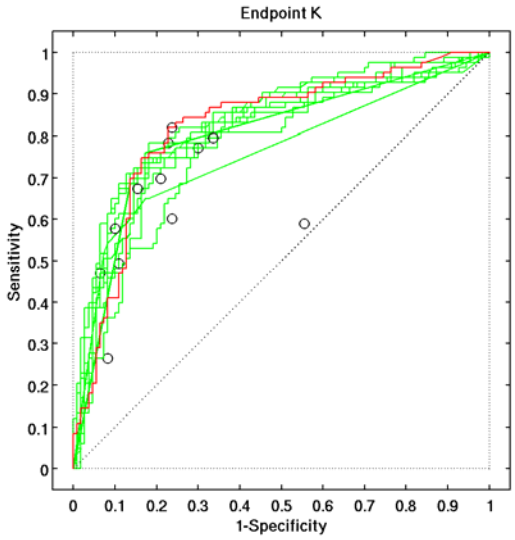


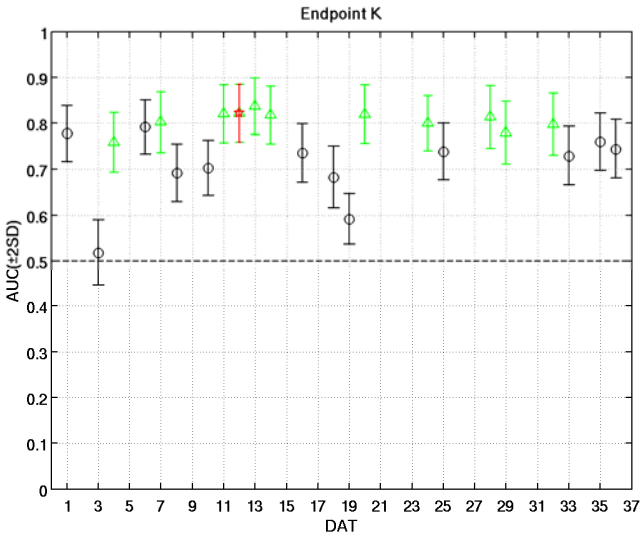


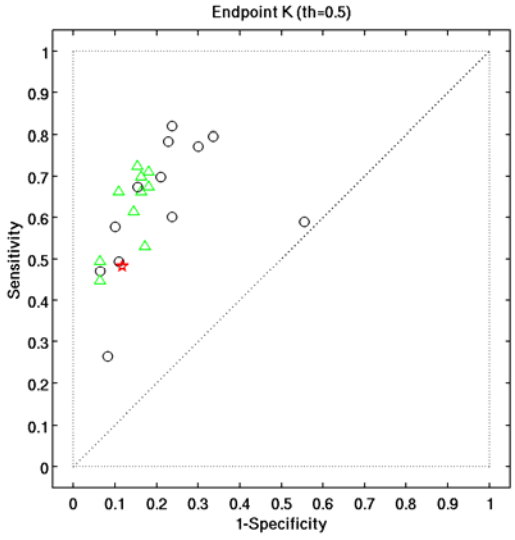


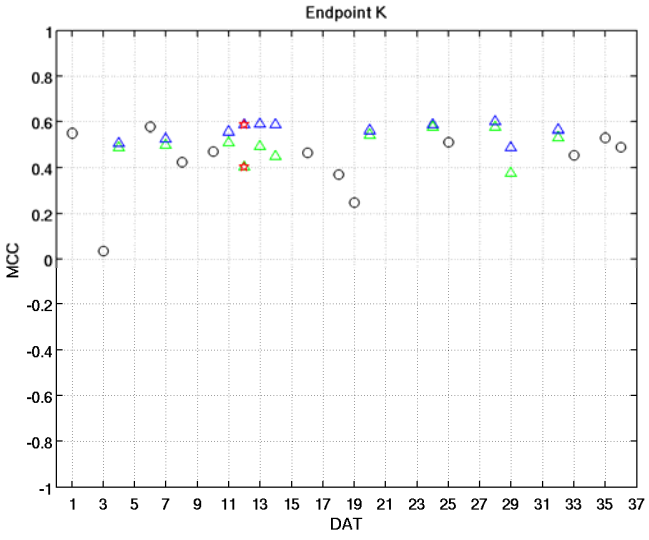


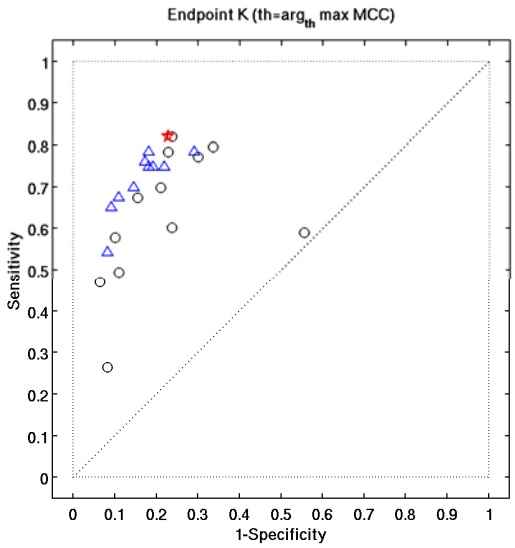
Nature Biotechnology: doi: 10.1038/nbt.1665

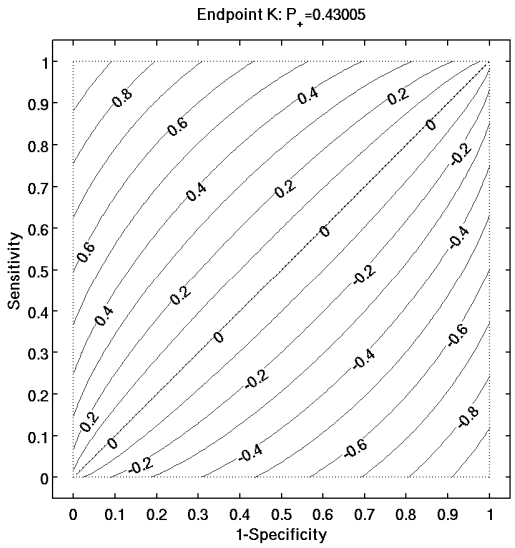




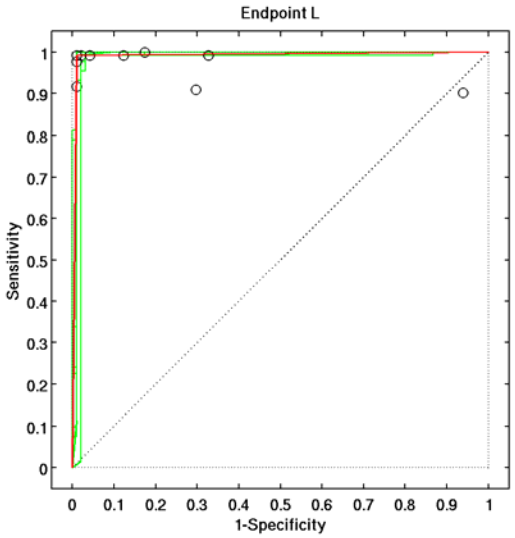


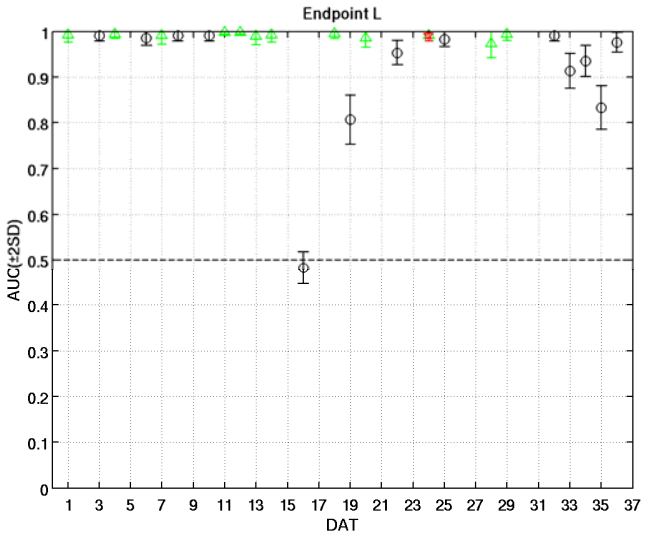


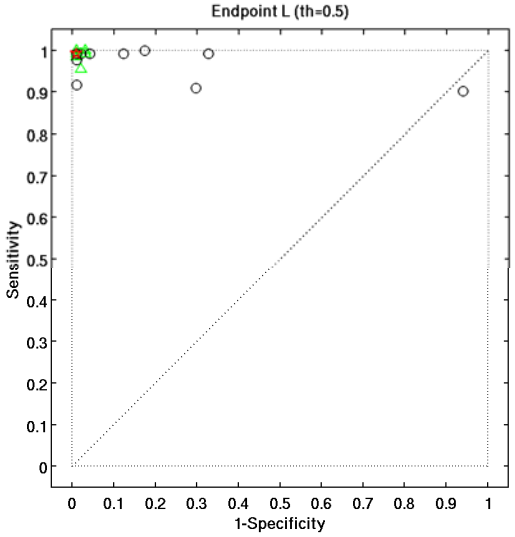


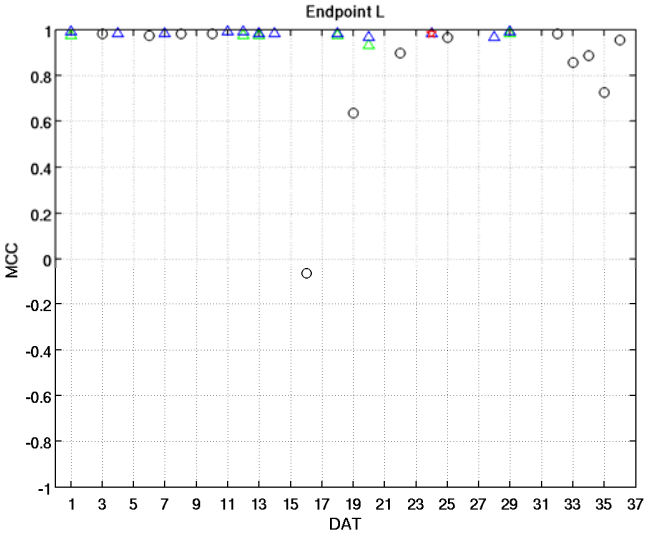


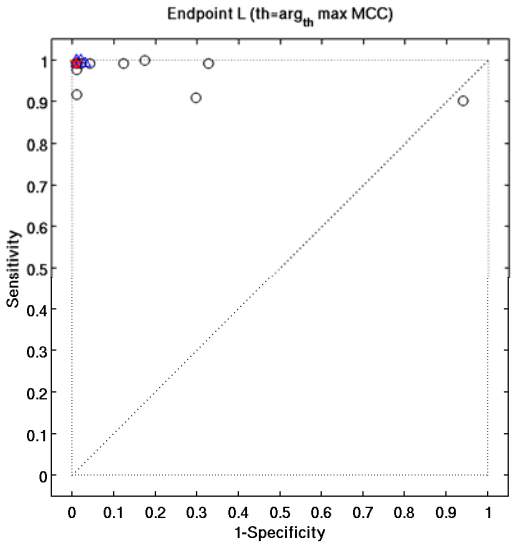
Nature Biotechnology: doi: 10.1038/nbt.1665

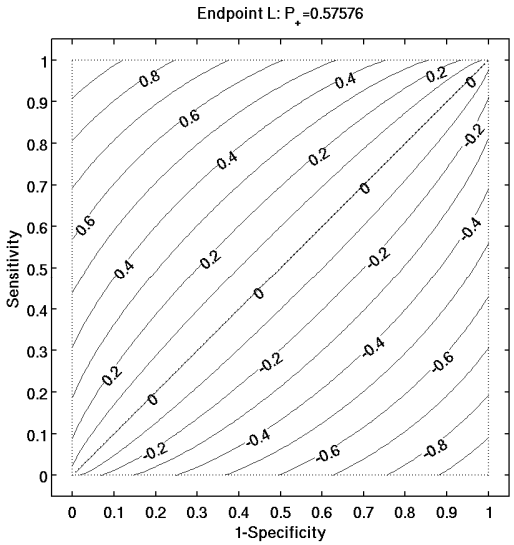




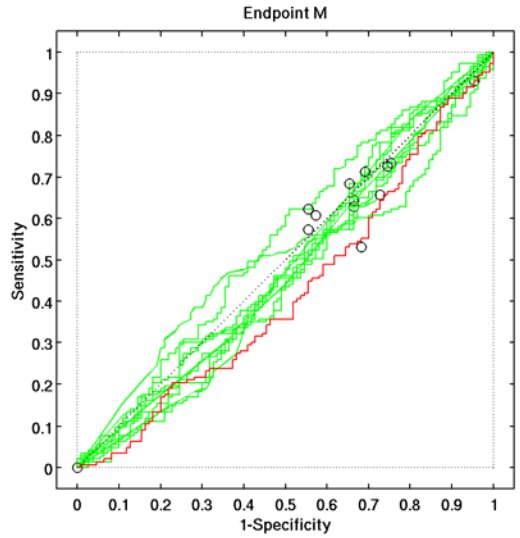


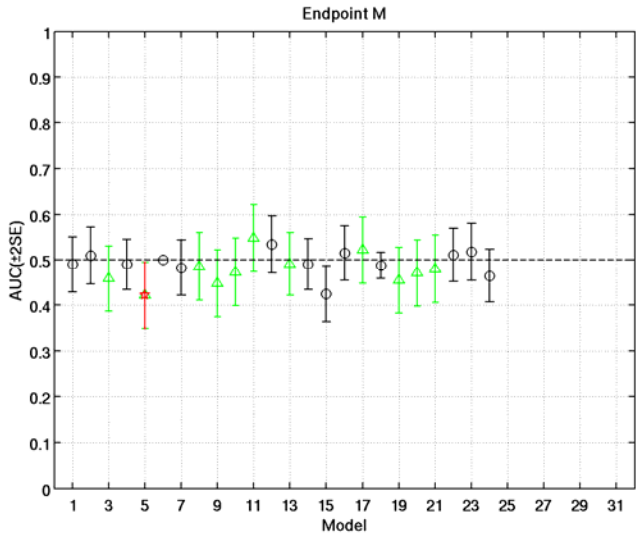


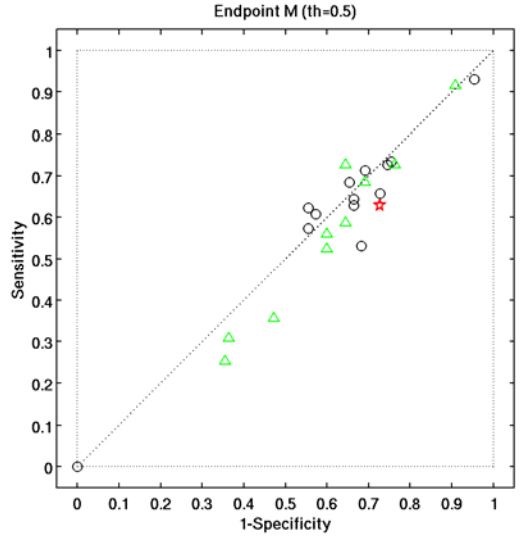


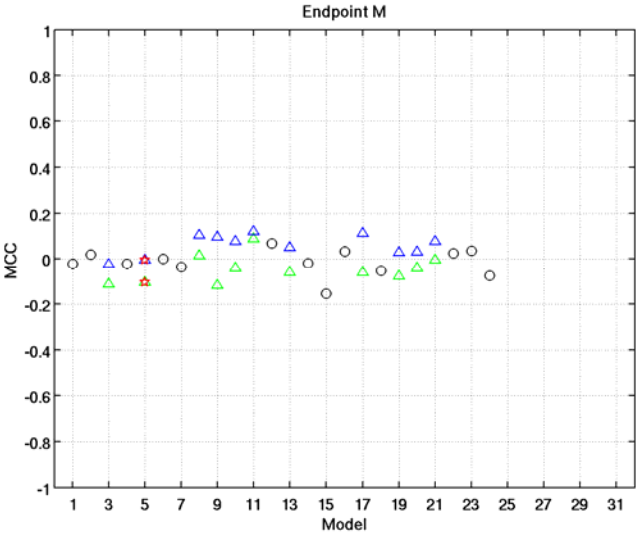


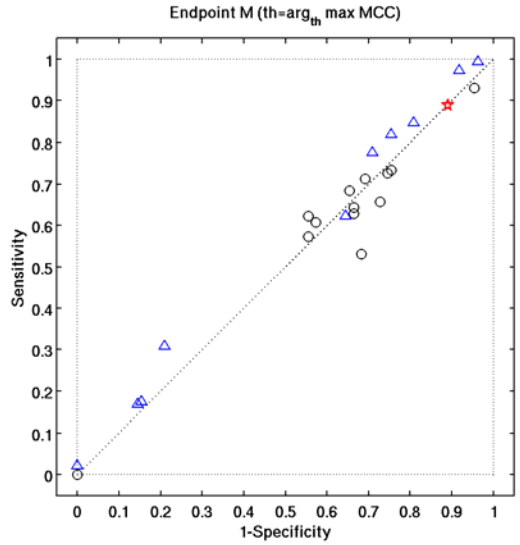
Nature Biotechnology: doi: 10.1038/nbt.1665

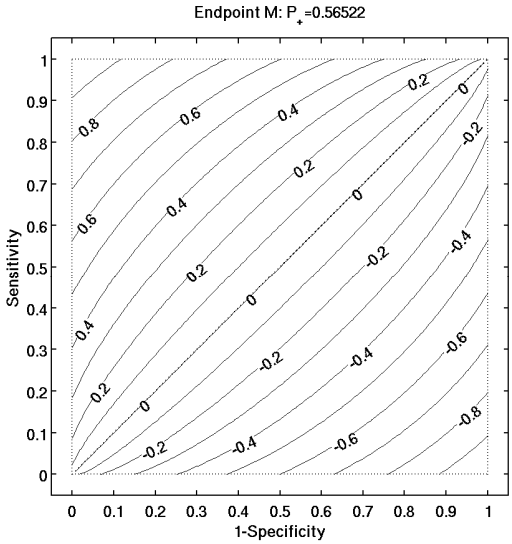












Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Document 2:

The performance metrics (MCC, AUC, Accuracy, RMSE)

Performance Metrics in Binary Medical Classification Problems
Brandon D. Gallas and Weijie Chen
US FDA, Center for Devices and Radiological Health

This discussion is based on the following classical references in Signal Detection Theory and Medical Decision Making: Green and Swets (1966) and Metz (1978). A nice overview of the development of this field from its original application, diagnostic medical imaging, to the modern microarray technologies can be found in Wagner (2007).

Fundamental Performance Metrics

For the binary classification task, many prediction models first yield a score that is compared to a threshold to make the binary decision, the diagnosis. We can use a 2×2 truth/decision table to present the frequency of each of the four possible outcomes in an experiment (See Table 1, which gives the truth/decision validation results for RBWG candidate model, Endpoint A). There are two correct decisions (true positives (TP) and true negatives (TN)) and two incorrect decisions (false positives (FP) and false negatives (FN)). We often summarize this table by a pair of metrics: sensitivity (Se) and specificity (Sp). Sensitivity is the rate at which you correctly call diseased patients diseased; it is also referred to as the TP fraction (TPF) . Specificity is the rate at which you correctly call normal patients normal; it is also referred to as the TN fraction (TNF). The rates of incorrect decisions are the FP fraction $(\mathrm{FPF}=1-\mathrm{TNF})$ and the FN fraction $(\mathrm{FNF}=1-\mathrm{TPF})$.

There is an alternate "transposed" performance perspective that also summarizes Table 1. Instead of normalizing by the truth (row-sums), as we did for $(\mathrm{Se},\mathrm{Sp})$, we normalize by the decisions (col-sums). This pair of metrics includes the positive and negative predictive values $(\mathrm{PPV},\mathrm{NPV})$ and focuses on the patient; $(\mathrm{Se},\mathrm{Sp})$ focuses on the population. The PPV is the probability that if a patient is diagnosed as (disease) positive, the patient actually has the disease. The NPV is the probability that if a patient is diagnosed as (disease) negative, the patient is actually normal.

Finally, there is a less common, but especially useful, way to summarize Table 1 using positive and negative likelihood ratios $(\mathrm{PLR},\mathrm{NLR})$. Like $(\mathrm{PPV},\mathrm{NPV})$, likelihood ratios focus on the patient. The PLR tells how likely subjects with disease are to have a positive result compared to subjects without disease. The NLR tells how likely subjects with disease are to have a negative result compared to subjects without disease.

Nature Biotechnology: doi: 10.1038/nbt.1665

Table 1: Truth/Decision table of outcomes plus definitions of performance.

	Decision: Negative	**Decision**: Positive
Truth: Normal N_0 = 60 cases	TN = # of true neg = 31 TNF = Sp = $\frac{TN}{TN+FP}$ = 0.52	FP = # of false pos = 29 FPF = 1-Sp = $\frac{FP}{TN+FP}$ = 0.48
Truth: Disease N_1 = 28 cases	FN = # of true neg = 8 FNF = 1-Se = $\frac{FN}{TP+FN}$ = 0.29	TP = # of true pos = 20 TPF = Se = $\frac{TP}{TP+FN}$ = 0.71
p = prevalence $= \frac{TP+FN}{TP+FN+TN+FP}$	NPV = $\frac{TN}{TN+FN}$ = 0.79 = $\frac{Sp\times(1-p)}{Sp\times(1-p)+(1-Se)\times p}$	PPV = $\frac{TP}{TP+FP}$ = 0.68 = $\frac{Se\times p}{Se\times p+(1-Sp)\times(1-p)}$
	NLR = $\frac{1-Se}{Sp}$ = 0.55 = $\frac{FN}{TP+FN}\frac{TN+FP}{TN}$	PLR = $\frac{Se}{1-Sp}$ = 1.48 = $\frac{TP}{TP+FN}\frac{TN+FP}{FP}$

During model development, $(\mathrm{Se},\mathrm{Sp})$ and $(\mathrm{PLR},\mathrm{NLR})$ are more appropriate than $(\mathrm{PPV},\mathrm{NPV})$ for one simple reason. The pairs $(\mathrm{Se},\mathrm{Sp})$ and $(\mathrm{PLR},\mathrm{NLR})$ do not depend on prevalence (p), whereas the pair $(\mathrm{PPV},\mathrm{NPV})$ does depend on prevalence. This is because the cases used in model development are often obtained by convenience rather than by prospectively or randomly sampling from an intended-use population. Thus the prevalence used in model development may not reflect the prevalence in the intended use population. Let's consider an example.

Ignoring measurement error for a moment, let's assume that the fractions given in Table 1 are the true rates for a diagnostic device. If this model is validated in a low-prevalence screening population, where $p=0.01$, then $(\mathrm{Se},\mathrm{Sp})$ and $(\mathrm{PLR},\mathrm{NLR})$ do not change, but $(\mathrm{PPV},\mathrm{NPV})=(0.016,0.993)$ is quite different from that measured on the training dataset $(\mathrm{PPV},\mathrm{NPV})=(0.68,0.79)$ depicted in the table. If instead, this model were tested in a high-prevalence high-risk population, where $p=0.4$, then again, $(\mathrm{Se},\mathrm{Sp})$ and $(\mathrm{PLR},\mathrm{NLR})$ do not change, but $(\mathrm{PPV},\mathrm{NPV})=(0.525,0.672)$ is again different from that measured in the training dataset.

Nature Biotechnology: doi: 10.1038/nbt.1665

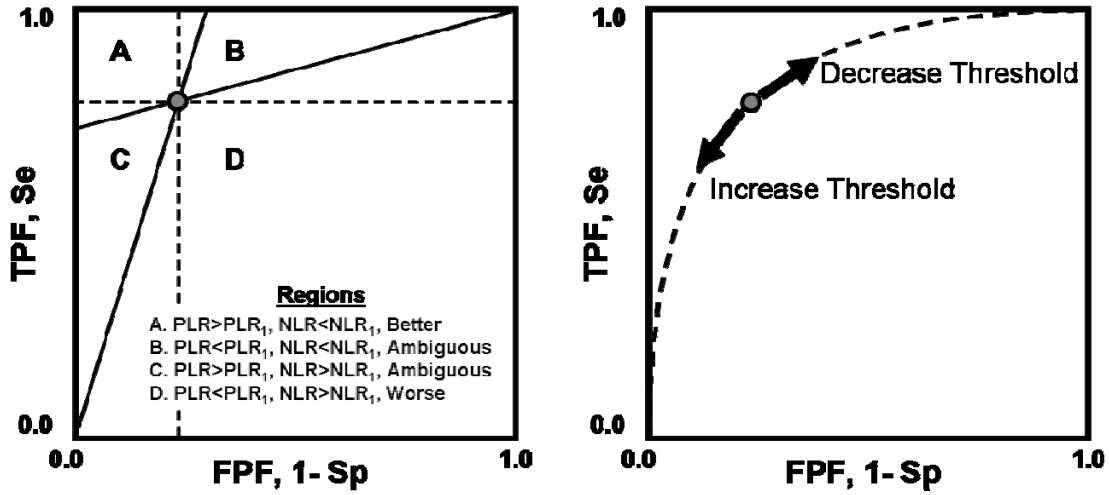


Figure 1: Left: A single (Se,Sp) operating point in ROC space defining regions where otherpoints are better (A), worse (D), and ambiguous (B,C). Right: An operating point on a hypothetical ROC curve and its dependence as the decision threshold is varied.

For $(\mathrm{PPV}, \mathrm{NPV})$ to be appropriate, the proportion of diseased cases in the validation dataset should approximate the prevalence of disease in an intended-use population. This may not be the case for the MAQC II datasets. There was no discusion about whether the datasets were collected in a prospective way from a well-defined intended-use population. Consequently, we do not want to base our performance analysis on the prevalence-dependent metrics like $(\mathrm{PPV}, \mathrm{NPV})$.

Unfortunately (Se, Sp) and (PLR, NLR) are themselves troublesome. First, there is possible ambiguity when comparing prediction models. One model may have better Se but worse Sp than another. Likewise, one model may have better PLR but worse NLR. Note that these conditions are not equivalent. Biggerstaff (2000) has shown that the comparison could be done with (PLR, NLR), not (Se, Sp). The (PLR, NLR) pair of metrics has smaller regions of ambiguity than (Se, Sp), shown in the left plot of Fig. 1 as Regions B,C bounded by the solid lines with slopes PLR and NLR passing through the model's performance operating point (Se, Sp). The regions of ambiguity for (Se, Sp) are Regions B,C bounded by the dashed lines.

The other troublesome characteristic of $(\mathrm{Se}, \mathrm{Sp})$ and $(\mathrm{PLR}, \mathrm{NLR})$ stems from the ambiguity in selecting the threshold. Optimally selecting a threshold can be done by maximizing the expected benefit/utility (or minimizing the expected risk/cost) [Green and Swets 1966, Metz 1978, Patton 1978, Wagner et al 2004]. The selection depends on the prevalence of the disease or condition of interest, and the utility parameters--the positive utilities of the two kinds of correct decisions (U_{TP} and U_{TN}) and the negative utilities of the two kinds of incorrect decisions (U_{FN} and U_{FP}). The expected utility is then

$$\overline{U} = (U_{TP} \times \mathrm{Se} + U_{FN} \times (1 - \mathrm{Se})) \times p + (U_{FP} \times (1 - \mathrm{Sp}) + U_{TN} \times \mathrm{Sp}) \times (1 - p). \qquad \text{Eq. 1}$$

We have already indicated that the proportion of diseased cases in the MAQC II validation datasets may not approximate the prevalence of disease in an intended-use population. Therefore, we should not use this dataset to estimate the prevalence. The utility parameters, however, are free parameters that are highly dependent on the what can and will be done to clinically manage and treat the patient. While they can be clearly stated in the abstract,

specifying them for an application is very context dependent, making for a very complex (and often very subjective) task. Since there has been no discussion of how to utilize the classifiers for clinical management, there is no way to specify utilities.

We can avoid selecting a threshold by considering all thresholds. The receiver operating characteristic (ROC) curve maps out the trajectory of the $(\mathrm{Se},\mathrm{Sp})$ pair as that threshold is varied over its entire range (See the right plot of Fig. 1). The ROC curve is the fundamental prevalence-independent picture of model performance in the binary task, as well as a pre-requisite for optimizing the expected benefit (where the prevalence and utilities then enter).

It may appear that we have complicated our ability to compare two prediction models by not selecting a threshold, but in fact, by considering all thresholds, we can make definitive comparisons more often. By mapping out the ROC curve of each model, we may find that one is higher (better) everywhere than the other. In other words, one model has a higher sensitivity than the other at every specificity. Alternatively, we may find that the two models yield the same ROC curve. This means that the two prediction models have the same performance characteristics. Finally, we may find that the two ROC curves cross. This case is troublesome; picking the better prediction model requires an expected utility analysis.

To reduce the information in the ROC curve to a single summary metric of performance without specifying the prevalence and utilities, we can use the area under the ROC curve (AUC). If the two curves do not cross, the one with the higher AUC has a higher sensitivity at every specificity. Furthermore, the AUC has several meaningful interpretations. First, the AUC has a nice elementary interpretation: it is the sensitivity averaged over all specificities,

$$\mathrm{AUC} = \int \mathrm{Se}\, d(\mathrm{Sp}). \qquad \text{Eq. 2}$$

The next interpretation of AUC comes from its nonparametric estimate. This estimate is a rescaled version of an elementary nonparametric statistic called the Wilcoxon-Mann-Whitney (WMW) statistic, a simple measure of the separation of two distributions. This statistic is given by

$$\mathrm{AUC} = \frac{1}{N_0 N_1} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} s_{ij}, \qquad \text{Eq. 3}$$

where

$$s_{ij} = s(t_{1j} - t_{0i}) = \begin{cases} 1 & t_{1j} - t_{0i} > 0 \\ 1/2 & t_{1j} - t_{0i} = 0 \\ 0 & t_{1j} - t_{0i} < 0 \end{cases} \qquad \text{Eq. 4}$$

are success indicators of whether the prediction model successfully scores the j^{th} diseased case higher than i^{th} normal case, and t_{0i}, t_{1j} are the prediction model scores for the $i = 1, 2, ..., N_0$ normal and $j = 1, 2, ..., N_1$ diseased cases.

The last and most common interpretation of AUC is probabilistic. This interpretation considers the probability that a randomly selected diseased case will be scored higher than a randomly selected normal case, plus 0.5 times the probability that the cases are tied:

$$E(\mathrm{AUC}) = E(s_{ij}) = P(t_{1j} > t_{0i}) + 0.5P(t_{1j} = t_{0i}). \qquad \text{Eq. 5}$$

This probability is related to the triage task, where a doctor has two patients and must decide who to treat first.

Combining Sensitivity and Specificity

Nature Biotechnology: doi: 10.1038/nbt.1665

There are a few types of prediction models that can only make the binary decision, which yields a single operating point (se, Sp) in ROC space. To summarize the performance of such models with a single summary performance metric, we must combine Se and Sp. At the extremes, we may consider Se and ignore Sp, or vice versa. The optimal combination, as stated above, depends on prevalence and the utility parameters. In the following, we shall describe a few combinations proposed by the MAQC II consortium, and relate them to AUC and the expected utility when possible. These measures are not recommended for the MAQC II model analyses as they either implicitly assume certain prevalences and utilities, or are explicitly prevalent dependent.

Accuracy

Accuracy and its complement, the probability of misclassification (PMC), or error rate, are the most straightforward summaries of performance:

$$\mathrm{Accuracy} = 1 - \mathrm{PMC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FN} + \mathrm{TN} + \mathrm{FP}}. \qquad \text{Eq. 6}$$

This metric pools cases across truth status and returns the proportion of correct decisions. With a little algebra, we can show that Accuracy is a prevalence weighted average of sensitivity and specificity:

$$\mathrm{Accuracy} = p \times \mathrm{Se} + (1 - p) \times \mathrm{Sp}. \qquad \text{Eq. 7}$$

In terms of expected utility, Accuracy implicitly assumes that the utilities of correct decisions equal 1.0, and utilities of incorrect decisions equal 0.0. More generally, Accuracy is linearly related to the expected utility whenever the utilities of correct decisions are equal and the utilities of incorrect decisions are also equal.

AUC(Binary Scores)

Another summary of sensitivity and specificity is their average with equal weights. Interestingly, this average equals the empirical AUC defined in Eq. 3 when the scores are limited to zeros and ones. This summary has the same implicit assumptions for the utilities as Accuracy, and additionally assumes that the population prevalence equals 1/2.

Matthew's Correlation Coefficient

The Matthews Correlation Coefficient (MCC) is not exactly an average of sensitivity and specificity. It is normally written as a function of the truth/decision table cell counts (Table 1); namely,

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}. \qquad \text{Eq. 8}$$

Consequently, its connection to expected utility is a bit more complicated.

The relationship between MCC and Se and Sp is

$$\mathrm{MCC} = \sqrt{(\mathrm{Se} + \mathrm{Sp} - 1)(\mathrm{PPV} + \mathrm{NPV} - 1)}, \qquad \text{Eq. 9}$$

the geometric mean of a prevalence independent term and a prevalence dependent term. To show the equivalence of Eqs 8&9, replace the rate quantities in Eq. 9 with their equivalent expressions using count quantities $(\mathrm{TP}, \mathrm{TN}, \mathrm{FP}, \mathrm{FN})$. For more insight into this equation, notice that the first term is a rescaling of the average of Se and Sp, which as mentioned above, is the empirical AUC when the scores are binary. The second term is a like quantity from the transposed perspective.

Nature Biotechnology: doi: 10.1038/nbt.1665

The result above contradicts a current misbelief that MCC is independent of prevalence. This misbelief comes from a perceived characteristic that MCC is less dependent on prevalence than Accuracy and more stable when the number of case mix is unbalanced. Equation 9 shows that this is the result of the balancing of the prevalent dependent part $(\mathrm{PPV}+\mathrm{NPV}-1)$ by the rescaled AUC $(\mathrm{Se}+\mathrm{Sp}-1)$. As such MCC is less stable than AUC when the case mix is unbalanced; AUC does not depend on case mix.

Not only does MCC depend on prevalence, but the dependence is nonlinear. A little algebra on Eq. 8 or Eq. 9 can express MCC as an explicit function of *Se*, *Sp* and prevalence *p* as follows,

$$MCC = \frac{Se + Sp - 1}{\sqrt{(1 - Sp + \frac{p}{1-p} Se)(1 - Se + \frac{1-p}{p} Sp)}} \qquad \text{Eq. 9a}$$

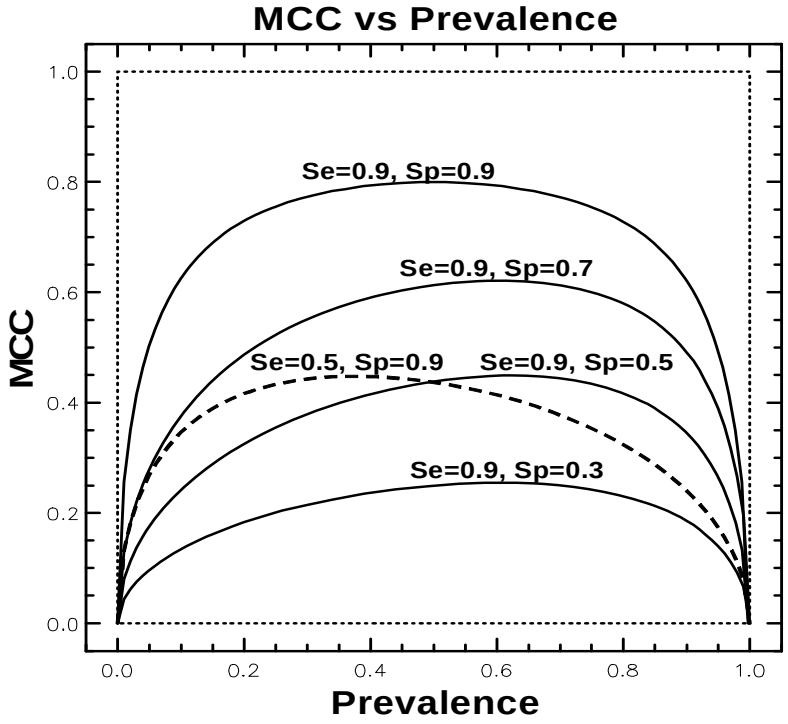


Figure 2: A plot of the prevalence dependence of Matthews correlation coefficient (MCC).

We demonstrate the prevalence dependence of MCC in Fig. 2 for several different performance operating points. Note the skew when $\mathrm{Se} \neq \mathrm{Sp}$, and the symmetry when Se, Sp are swapped. For comparison, a plot of Accuracy for $(\mathrm{Se}, \mathrm{Sp}) = (0.9, 0.5)$ would be a straight line connecting $(0.0, 0.5)$ and $(1.0, 0.9)$ and its symmetric twin when Se, Sp are swapped; a plot of AUC from binary scores with $(\mathrm{Se}, \mathrm{Sp}) = (0.9, 0.5)$ would be flat at 0.7.

MSE

The mean squared error (MSE) is

Nature Biotechnology: doi: 10.1038/nbt.1665

$$\mathrm{MSE} = \frac{1}{N_0 + N_1}\left[\sum_{k=1}^{N}(t_k - y_k)^2\right], \qquad \text{Eq. 10}$$

where t_k are the prediction model scores for all $k = 1,2,...,N = N_0 + N_1$ cases, regardless of truth status, and y_k are the truth labels. Please notice that, when the scores and truth are binary,

$$\mathrm{MSE} = \frac{1}{N_0 + N_1}\left[\sum_{i=1}^{N_0}(t_{0i})^2 + \sum_{j=1}^{N_1}(t_{1j} - 1)^2\right]. \qquad \text{Eq. 11}$$

The terms in the sums are ones and zeros, which are unchanged when squared. Consequently,

$$\mathrm{MSE} = \mathrm{PMC} = 1 - \mathrm{Accuracy}. \qquad \text{Eq. 12}$$

We have already shown that Accuracy is prevalence dependent. As such, MSE is prevalence dependent as well.

Unlike the previous metrics in this section, MSE is not limited to situations where the scores are binary. However, in this case MSE has the undesirable characteristic that it is not invariant to monotonic transformations of the data. Monotonic transformations do not change the rank ordering of the scores, the ROC curve, or the AUC. Monotonic transformations also do not change the maximum expected utility for a given prevalence and set of utility parameters.

Variance and Covariance

The performance metrics given in the previous section are all random quantities when estimated with finite datasets. As such they are meaningless without an assessment of their uncertainty. In this section, we describe a bootstrapping method for comparing $(\mathrm{Se}, \mathrm{Sp})$ points in ROC space and give the U-statistic variance estimates of Se, Sp, and AUC that we use in the analyses of the prediction models.

The bootstrapping method we use to compare $(\mathrm{Se}, \mathrm{Sp})$ points in ROC space is as follows. In each of 1000 iterations, we sample with replacement N_0 normal cases and N_1 disease cases from the available dataset. Then we generate the $(\mathrm{Se}, \mathrm{Sp})$ pair or the difference in $(\mathrm{Se}, \mathrm{Sp})$ pairs from two models. The 1000 pairs represent the distribution of $(\mathrm{Se}, \mathrm{Sp})$ from independent sampling, assuming the empirical distribution reflects the true distribution. This assumption is supported by the fact that the empirical distribution is the *maximum likelihood estimate* (MLE) of the true distribution.

The U-statistic variance estimates seem natural since the WMW AUC given in Eq 3 is the U-statistic estimate of the population AUC given in Eq. 5 [Randles and Wolfe 1979]. U-statistics are also nice because they are unbiased, nonparametric, and usually have the smallest variance among all unbiased estimators.

For Se and Sp, the U-statistic variance and covariance estimates come from the sample variance and covariance of the corresponding Bernoulli trials. In other words, let t_{0ir}, t_{1jr} be the scores for the normal and diseased cases for the r^{th} prediction model. Then the Bernoulli successes are $1 - t_{0ir} = 1$ when model r scores normal-case i correctly and $t_{1jr} = 1$ when model r scores disease-case j correctly. The variances and covariances are

Nature Biotechnology: doi: 10.1038/nbt.1665

$$\hat{\mathrm{var}}(\hat{Sp}_r) = \frac{\hat{\sigma}_0^2}{N_0} = \sum_{i=1}^{N_0} \frac{\left(1 - t_{0ir} - \hat{Sp}_r\right)^2}{N_0(N_0 - 1)} = \frac{\hat{Sp}_r(1 - \hat{Sp}_r)}{N_0 - 1} \quad \text{Eq. 13}$$

$$\hat{\mathrm{var}}(\hat{Se}_r) = \frac{\hat{\sigma}_1^2}{N_1} = \sum_{j=1}^{N_1} \frac{\left(1 - t_{1jr} - \hat{Se}_r\right)^2}{N_1(N_1 - 1)} = \frac{\hat{Se}_r(1 - \hat{Se}_r)}{N_1 - 1} \quad \text{Eq. 14}$$

$$\mathrm{cov}\left(\hat{\mathrm{Sp}}_r, \hat{\mathrm{Sp}}_{r'}\right) = \sum_{i=1}^{N_0} \frac{\left(1 - t_{0ir} - \hat{\mathrm{Sp}}_r\right)\left(1 - t_{0ir'} - \hat{\mathrm{Sp}}_{r'}\right)}{N_0(N_0 - 1)}, \quad \text{Eq. 15}$$

$$\mathrm{cov}\left(\hat{\mathrm{Se}}_r, \hat{\mathrm{Se}}_{r'}\right) = \sum_{i=1}^{N_1} \frac{\left(t_{1ir} - \hat{\mathrm{Se}}_r\right)\left(t_{1ir'} - \hat{\mathrm{Se}}_{r'}\right)}{N_1(N_1 - 1)}. \quad \text{Eq. 16}$$

The (co)variance calculation for AUC is a bit more complicated. While we use U-statistic estimates [Bamber 1975, Campbell et al. 1988, Gallas 2006], there are other nonparametric options[DeLong et al. 1988], as well as parametric options[Dorfman and Alf 1969, Metz et al. 1998, Hanley 1989].

The expressions for the U-statistic estimates we show are based on the work of Gallas 2006. The estimates make use of the success indicators (Eq. 4) with the added model index r. The U-statistic estimate of the variance of AUC and the covariance between the AUC of two (fixed) models is

$$\hat{\mathrm{var}}(AUC_r) = c_1\hat{M}_{1rr} + c_2\hat{M}_{2rr} + c_3\hat{M}_{3rr} + c_4\hat{M}_{4rr} - \hat{M}_{4rr} \quad \text{Eq. 17}$$

$$\hat{\mathrm{cov}}(AUC_r, AUC_{r'} \mid r, r') = c_1\hat{M}_{1rr'} + c_2\hat{M}_{2rr'} + c_3\hat{M}_{3rr'} + c_4\hat{M}_{4rr'} - \hat{M}_{4rr'} \quad \text{Eq. 18}$$

where the coefficients c_k and moment estimates $\hat{M}_{krr'}$ for $k = 1,2,3,4$ are given in Table 2.

Coefficients	**U-statistic Moment Estimates**
$c_1 = \frac{1}{N_0 N_1}$	$M_{1rr'} = \sum_{i,j} \frac{s_{ijr} s_{ijr'}}{N_0 N_1}$
$c_2 = \frac{N_0 - 1}{N_0 N_1}$	$M_{2rr'} = \sum_{i,j,i' \neq i} \frac{s_{ijr} s_{i'jr'}}{N_0 N_1 (N_0 - 1)}$
$c_3 = \frac{N_1 - 1}{N_0 N_1}$	$M_{3rr'} = \sum_{i,j,j' \neq j} \frac{s_{ijr} s_{ij'r}}{N_0 N_1 (N_1 - 1)}$
$c_4 = \frac{(N_0 - 1)(N_1 - 1)}{N_0 N_1}$	$M_{4rr'} = \sum_{i,j,i' \neq i,j' \neq j} \frac{s_{ijr} s_{i'j'r'}}{N_0 N_1 (N_0 - 1)(N_1 - 1)}$

Results

In the following, we demonstrate some performance analyses. We limit ourselves here to MAQC II validation data endpoint A and the $N_2 = 25$ data analysis teams (DAT) candidate models submitted. From these 25 candidates, the RBWG selected one model as the endpoint A candidate. Much of these analyses are repeated for the other MAQC II endpoints and the results can be found in the supplemental material.

Nature Biotechnology: doi: 10.1038/nbt.1665

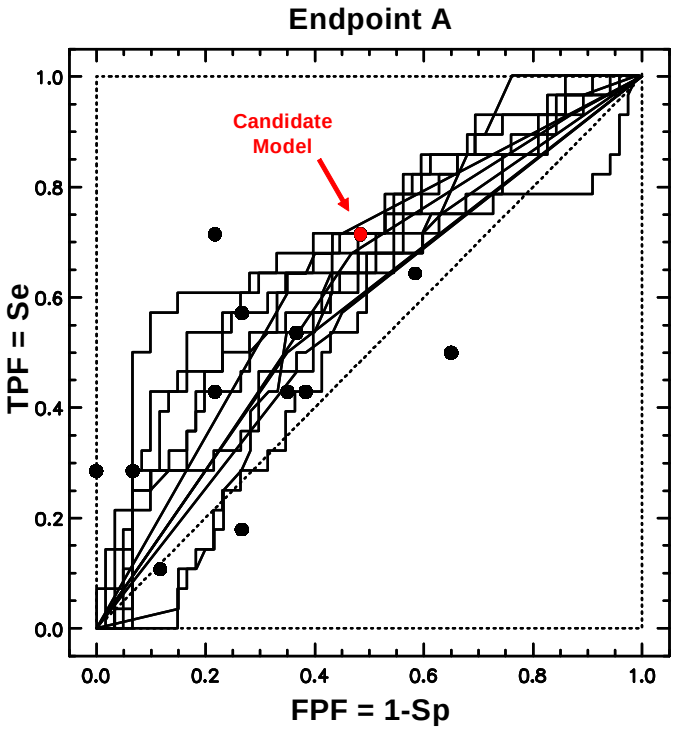


Figure 3: Empirical ROC curves for prediction models with continuous scores, (Se,Sp) operating points for models with binary scores, and the binary-score RBWG candidate model for Endpoint A (see arrow).

Comparing an ROC curve to an operating point

The MAQC II standard operating procedures for the DATs recommended that continuous scores be given and required that binary prediction outcomes be provided. As such, some DATs provided continuous scores and all DATS provided prediction outcomes. The only way to analyze this data without assuming a prevalence (explicitly or implicitly) is to present this data graphically. In Fig. 3 we show the ROC curves for the Endpoint A prediction models. The dots show the $(\mathrm{Se},\mathrm{Sp})$ of the prediction models that only gave binary scores; they only have one operating point. The lines connect $(\mathrm{Se},\mathrm{Sp})$ operating points for different continuous-score models. Most of the dots and lines lie above the diagonal chance line, but a few do not.

The RBWG candidate model is binary, and its single operating point is close to the outer edge of the ROC curves. With Sp near 0.5, the $\mathrm{Se} = 0.7$ of this model is pretty good. However, there is another binary score model with similar Se, but better $\mathrm{Sp} = 0.78$.

Nature Biotechnology: doi: 10.1038/nbt.1665

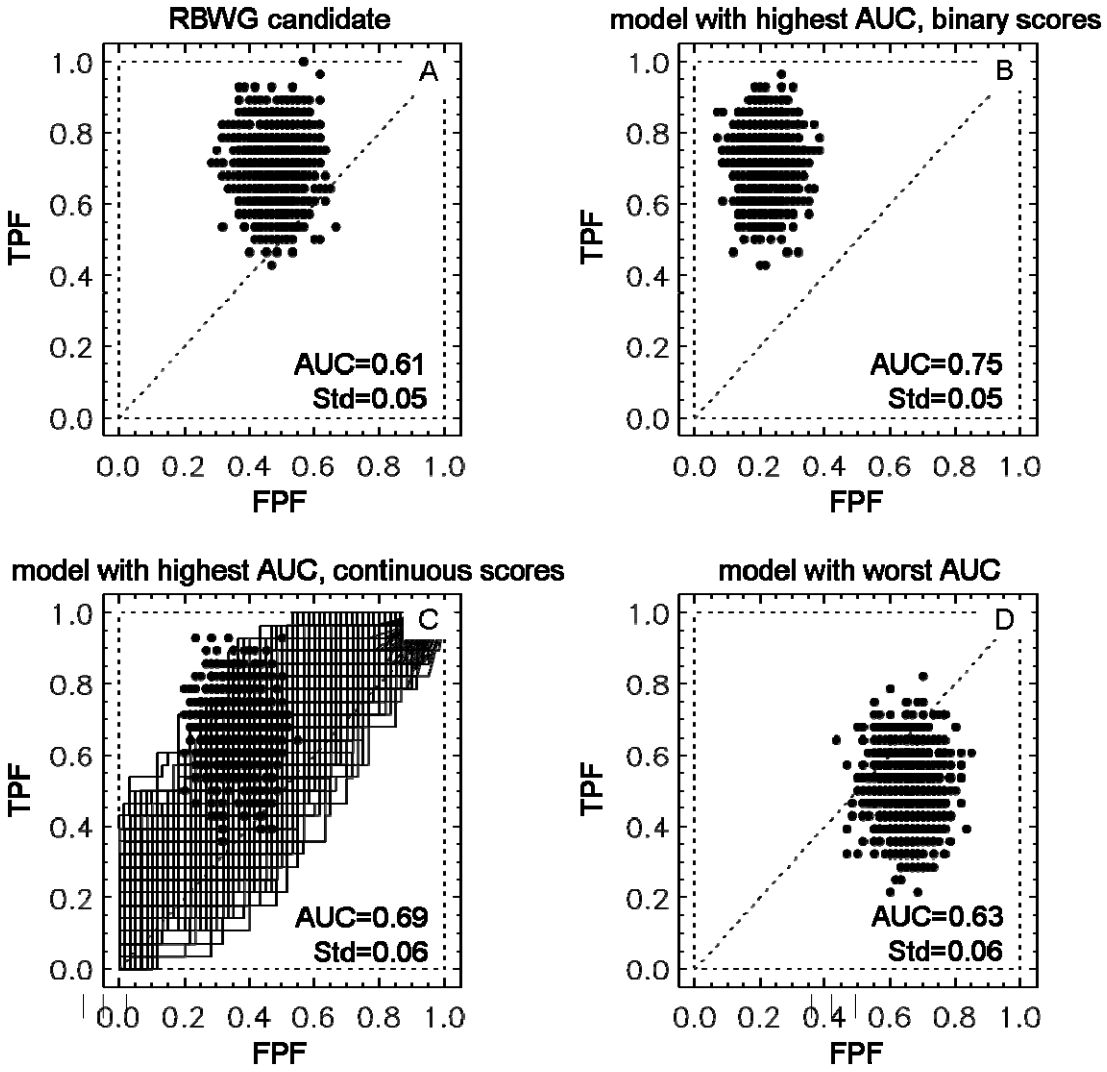


Figure 4: Bootstrap ROC curves and (Se,Sp) operating points for the binary-score RBWG candidate prediction model (A), the highest AUC binary-score model (B), the highest AUC continuous-score model (C), and the lowest AUC model (D). In the bottom-left plot, we also show the bootstrap (Se,Sp) operating points for the continuous-score model after applying a threshold of 0.5.

Given the size of the datasets, the uncertainty in our estimates of $(\mathrm{Se},\mathrm{Sp})$ is fairly substantial. In Fig. 4 we show the bootstrapping results for the RBWG candidate model (top left), the highest AUC binary-score model (top right), the highest AUC continuous-score model (bottom left), and the lowest AUC model (bottom right). For the continuous-score model, we can plot the ROC curves from bootstrapping, as well as the $(\mathrm{Se},\mathrm{Sp})$ points. The difference between the lowest and highest Se in each plot is about 0.5, while the difference between the lowest and highest Sp in each plot is about 0.3.

Nature Biotechnology: doi: 10.1038/nbt.1665

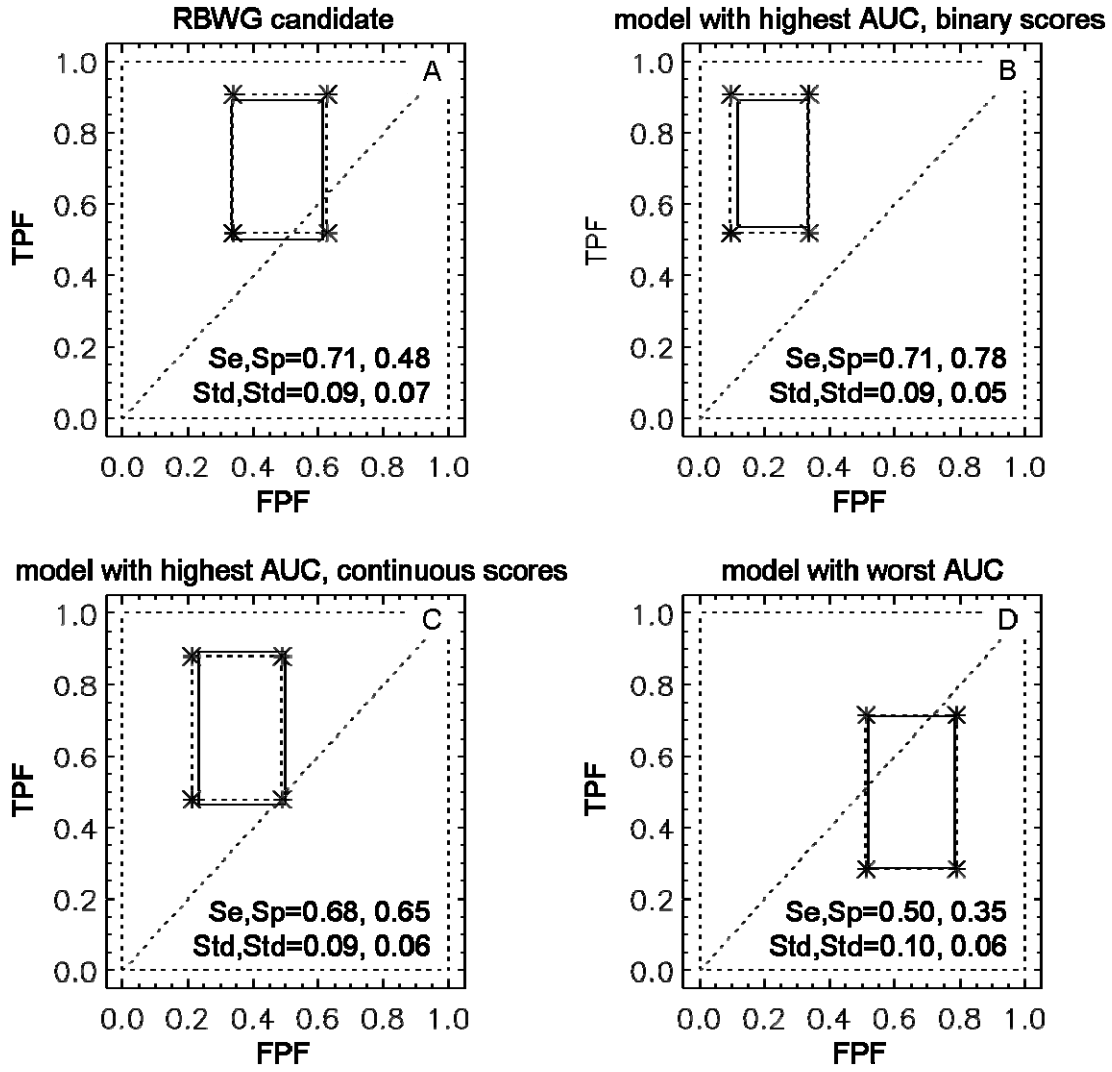


Figure 5: Confidence boxes derived by the product of 97.5% confidence intervals in the separate Se and Sp directions. The solid lines show the bootstrap results and the dotted lines with asterisk corners show the analytical results.

We can use the bootstrap distributions of $(\mathrm{Se},\mathrm{Sp})$ to define confidence boxes in the 2D ROC space. These are the solid line boxes shown in Fig. 5. Likewise, we can use the variance estimates for Se and Sp to define the boxes. These are the dashed boxes with the asterisks in the corners. Note that the 2D boxes are the product of 97.5% 1D confidence intervals in the separate $(\mathrm{Se},\mathrm{Sp})$ dimensions. The consequence of this product is that the 2D box is not a 97.5% confidence box; it contains about 97.5% x 97.5%=95% of the samples. If we were to plot these boxes on top of one another, the RBWG candidate model would overlap considerably with the highest AUC continuous-score model, be adjacent to the highest AUC binary-score model, and would have a bit of an overlap with the lowest AUC model. However, we would not be accounting for

Nature Biotechnology: doi: 10.1038/nbt.1665

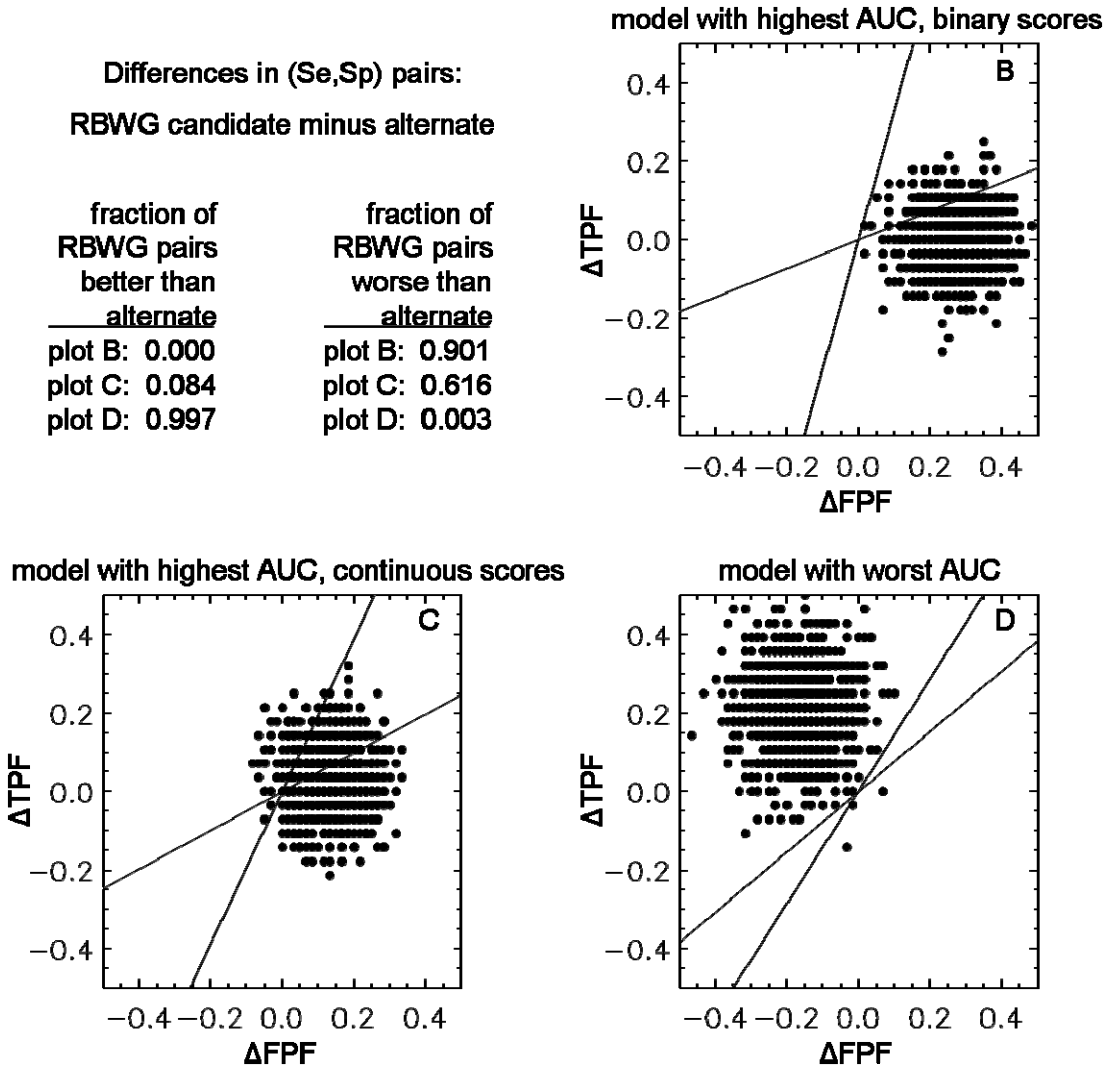


Figure 6: The difference in (Se,Sp) operating points between the RBWG candidate models and the alternate models: highest AUC binary-score model (B), the highest AUC continuous-score model (C), and the lowest AUC model (D).

the correlations in the data that exist because each model is tested on the same data.

In Fig. 6 we show the difference between the $(\mathrm{Se},\mathrm{Sp})$ points for the RBWG candidate model and the three other models. The lines in these plots show the lines of constant PLR and NLR of the three comparison models. As such, when points fall in the upper left region, they indicate that the RBWG candidate performed better on that bootstrap sample. When points fall in the lower right region, they indicate that the RBWG candidate performed worse on that bootstrap sample. Unlike comparing the confidence boxes, this analysis accounts for the correlations generated from the models being tested on the same data. Please refer to the small table listing the probabilities that the RBWG candidate model is better and worse than the three comparison models according to the

Nature Biotechnology: doi: 10.1038/nbt.1665

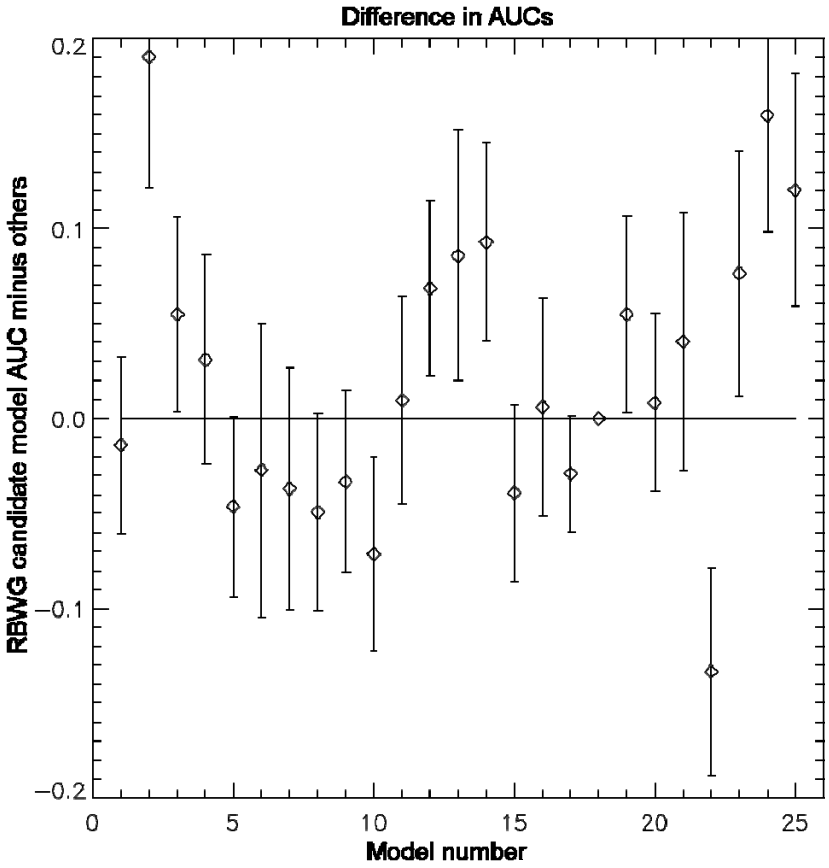


Figure 7: The differences between the RBWG candidate model and the 25 other DAT candidates with error bars indicating +/- two standard deviations.

performance pair $(\mathrm{Se}, \mathrm{Sp})$. Finally, we can compare the RBWG model against all the other prediction models with the singular statistic AUC. In Fig. 7 we show the difference between the RBWG model AUC and the AUC of the other models. The error bars on the differences are two standard deviations of the AUC estimate that accounts for the correlations generated by testing the models on the same data:

$$STERR(\mathrm{AUC}_{RBWG} - \mathrm{AUC}_r) = \sqrt{\mathrm{var}(\mathrm{AUC}_{RBWG}) + \mathrm{var}(\mathrm{AUC}_r) - 2\,\mathrm{cov}(\mathrm{AUC}_{RBWG}, \mathrm{AUC}_r \,|\, RBWG, r)}. \quad \text{Eq. 19}$$

Ignoring multiplicity issues, the RBWG candidate is statistically larger than several other models, statistically smaller than two, and undistinguishable from about half.

Nature Biotechnology: doi: 10.1038/nbt.1665

Bibliography

D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *J Math Psych* 12 (1975) 387—415.

B. J. Biggerstaff, Comparing diagnostic tests: A simple graphic using likelihoodratios, *Stat Med* 19 (5) (2000) 649—663.

G. Campbell, M. A. Douglas, J. J. Bailey, Nonparametric comparison of two tests of cardiac function on the same patient population using the entire ROC curve, in: *Proceedings of the IEEE Computers in Cardiology Conference,1988, Computer Society of the IEEE*, 1988.

E. R. DeLong, D. M. Delong, D. L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach, *Biometrics* 44 (1988) 837—845.

D. D. Dorfman, E. Alf, Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data, *J Math Psychol 6* (1969) 487—496.

B. D. Gallas, One-shot estimate of MRMC variance: AUC, *Acad Radiol* 13 (3) (2006) 353—362.

D. M. Green, J. A. Swets, Signal Detection Theory and Psychophysics, Wiley, New York, 1966, [reprint (Krieger, New York, 1974)].

J. A. Hanley, B. J. McNeil, The meaning and use of the area under a receiver
Operating characteristic (ROC) curve, *Radiology* 143 (1) (1982) 29—36.

C. E. Metz, Basic principles of ROC analysis, *Semin Nucl Med* 8 (4) (1978) 283—298.

C. E. Metz, B. A. Herman, J. Shen, Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data, *Statistics in Medicine* 17 (9) (1998) 1033—1053.

D. D. Patton, Introduction to clinical decision making, *Semin Nucl Med* 8 (4) (1978) 273—282.

R. H. Randles, D. A. Wolfe, Introduction to the Theory of Nonparametric Statistics, John Wiley and Sons, New York, 1979.

R. F. Wagner, From medical images to multiple-biomarker microarrays, *Med Phys* 34 (12) (2007) 4944—4951.

R. F. Wagner, C. A. Beam, S. V. Beiden, Reader variability in mammography and its implications for expected utility over the population of readers and cases, *Med*
Decis Making 24 (6) (2004) 561—572.

Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Document 3:

Possible explanation of the superior performance of DAT33's model on endpoint A

One interesting observation was the higher accuracy measured for the UIUC2 classifier on endpoint A relative to other approaches. Endpoint A was an interesting case because of its high batch variability. The UIUC2 group was testing the performance of the k-Top Scoring Pairs approach (Tan et al. 2005). This approach focuses on identifying pairs of markers that yield simple classifier variables of the form: *If A > B, then class 1, else class 2*, where A and B represent, for example, the expression of two genes. These rules result in a ratio-based data transformation to binary variables. A primary advantage of this transformation is that the accuracy of the classifier becomes independent of monotonic data normalization (e.g. quantile or mean normalization of expression). The method also maximizes the size of the relative expression reversal amongst the most accurate pairs. Thus, the method was developed in order to identify classifiers that are robust to variance due to batch effects and differences across platforms. In previous studies, this method was found to be accurate even when tested using a different measurement platform than that on which it was trained (Xu et al. 2005; Price et al. 2007). A potential weakness of the TSP approach, reflected in its average score in Table 2, is that the parsimony of the classifiers considered means that on complex phenotypes there can sometimes be no TSPs that have good accuracy even on the training set, in which case it can be identified as not being a correct candidate model choice for these situations. For Endpoint A, the question arises: is the higher performance observed on Endpoint A an accurate assessment or is it an outlier observed by chance? To really settle this question, additional results need to be repeated on multiple data sets with similar batch variability properties. However, the results observed herein is an intriguing clue that ratio-based data transformation is beneficial for identifying classifiers that are robust to high batch variance.

Price, N. D., J. Trent, A. K. El-Naggar, D. Cogdell, E. Taylor, K. K. Hunt, R. E. Pollock, L. Hood, I. Shmulevich and W. Zhang (2007). "Highly accurate two-gene classifier for differentiating gastrointestinal stromal tumors and leiomyosarcomas." Proc Natl Acad Sci U S A **104**(9): 3414-9.

Tan, A. C., D. Q. Naiman, L. Xu, R. L. Winslow and D. Geman (2005). "Simple decision rules for classifying human cancers from gene expression profiles." Bioinformatics **21**(20): 3896-904.

Xu, L., A. C. Tan, D. Q. Naiman, D. Geman and R. L. Winslow (2005). "Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data." Bioinformatics **21**(20): 3905-11.

Supplementary Document 4:

MAQC-II Research Plan

Development and Validation of Predictive Models Based on Microarray Gene Expression Profiles

Leming Shi
National Center for Toxicological Research
US Food and Drug Administration
3900 NCTR Road, Jefferson, Arkansas 72079, USA
Tel: +1-870-543-7387, Fax: +1-870-543-7854
Leming.Shi@fda.hhs.gov

March 22, 2007

The MicroArray Quality Control Consortium (MAQC)
http://edkb.fda.gov/MAQC/

Toward Consensus on "Best Practices" for the Generation, Analysis, and Application of Microarray Data in the Discovery, Development, and Review of FDA-regulated Products

Nature Biotechnology: doi: 10.1038/nbt.1665

Purpose of This Research Plan

The MicroArray Quality Control (MAQC) project is already one of the most ambitious and comprehensive studies to date on microarray quality control, addressing such issues as cross-laboratory and cross-platform comparisons and performance evaluation of data analysis methods for (1) the identification of differentially expressed genes (MAQC-I, *i.e.*, Phase I) and (2) the development and validation of predictive models or classifiers (MAQC-II, *i.e.*, Phase II). A consortium of many government, academic, and commercial participants has contributed and will continue to contribute substantial resources, time, and expertise to make this project a success. To achieve the intended, long-term benefits of the MAQC project, proper management and control is needed before the distribution of data sets or experimental processing of samples begins.

The validity of the MAQC project can be seriously compromised if all participants of the project do not operate under the same set of rules and guidelines. We recognize that the number of MAQC participants has been growing rapidly, and recent teleconferences and face-to-face meetings have demonstrated that the level of shared understanding between participants is variable. This Research Plan has been prepared to ensure a common basis of understanding for participants of the MAQC project. It summarizes the background of the MAQC project and MAQC-I results and outlines the scope of MAQC-II toward establishing consensus on the appropriate approaches to development and validation of predictive models based on microarray gene expression profiles.

This Research Plan is an integral part of the MAQC Confidential Information Disclosure and Transfer Agreement (CIDTA). MAQC participants are expected to carefully read this document and the attached SOP on Data Analysis by the MAQC Regulatory Biostatistics Working Group, and closely follow the guidelines outlined in the Research Plan and the SOP.

Disclaimer

The US Food and Drug Administration (FDA) has solicited DNA microarray gene expression data sets as well as proposals to analyze these data sets in order to evaluate the impact of different analysis protocols on the selection of genes and their associated predictive models for biomarker pattern development (*Federal Register*, 71(77), 20707-8, April 21, 2006; available at http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/docs/FederalRegister_MAQC_FollowUp.pdf). The MAQC project is being coordinated by the FDA, but there are no regulatory rights conveyed to anyone by the participation of FDA personnel in this project. Although FDA personnel are involved in the MAQC project, the views expressed here in this MAQC-II Research Plan are not FDA guidance and do not necessarily represent FDA policy.

Participation in the MAQC project is completely voluntary. No fund whatsoever is available from the MAQC to any participant. Participants agree to cover all their own costs as a result of voluntary involvement in the MAQC project.

Nature Biotechnology: doi: 10.1038/nbt.1665

Abbreviations

MAQC: MicroArray Quality Control project
MAQC-I: Phase I of MAQC project (identification of differentially expressed genes)
MAQC-II: Phase II of MAQC project (development and validation of predictive models/classifiers)
WG: Working Group
CWG: Clinical Working Group of MAQC-II
RBWG: Regulatory Biostatistics Working Group of MAQC-II
TGxWG: Toxicogenomics Working Group of MAQC-II
TitrationWG: Titration Working Group of MAQC-II
A: RNA sample A ("Apples"), Startagene's Universal Human Reference RNA
B: RNA sample B ("Bananas"), Ambion's Human Brain Reference RNA
C: 3A:1B mixture (titration)
D: 1A:3B mixture (titration)
DEG: Differentially expressed gene
CIDTA: Confidential information disclosure and transfer agreement
IRB: Institutional Review Board
SOP: Standard operating procedure
FDA: US Food and Drug Administration
CBER: Center for Biologics Evaluation and Research, FDA
CDER: Center for Drug Evaluation and Research, FDA
CDRH: Center for Devices and Radiological Health, FDA
CFSAN: Center for Food Safety and Applied Nutrition, FDA
CVM: Center for Veterinary Medicine, FDA
NCTR: National Center for Toxicological Research, FDA

List of Tables

List of Figures

Nature Biotechnology: doi: 10.1038/nbt.1665

Table of Contents

1. MicroArray Quality Control (MAQC) Project

1.1 Microarrays and FDA's Critical Path Initiative

On March 16, 2004, the US Food and Drug Administration (FDA) released a report on "*Innovation/Stagnation: Challenge and Opportunity on the Critical Path to New Medical Products*", addressing the recent slowdown in innovative medical products submitted to the FDA for approval. The report described the urgent need to modernize the medical product development process – the Critical Path from bench to bed side – so that the product development process will be more predictable and efficient. On March 16, 2006, HHS Secretary Mike Leavitt and FDA Commissioner Andrew von Eschenbach released the Critical Path Opportunities List and Report that provided concrete focus for public and private efforts and investments in new tools that could revolutionize medical product development. Among the 76 opportunities in fields such as genomics and proteomics, imaging, and bioinformatics, "*Biomarker qualification*" and "*Standards for microarray and proteomics-based identification of biomarkers*" were cited as the top two opportunities.

Microarray technology was identified by the FDA's Critical Path Initiative (http://www.fda.gov/oc/initiatives/criticalpath/) as a key tool that holds "vast potential" for advancing medical product development and personalized medicine through the identification of biomarkers. However, a gap exists between technological levels in use today and those required for application during product development and regulatory decision making. For example, recent publications have raised concerns about the reliability of microarray technology because of the apparent lack of reproducibility between lists of genes (*i.e.*, potential biomarkers) identified as differentially expressed from similar or identical study designs with different platforms or laboratories[1, 2]. In addition, the reliability and utility of classification models for the prediction of patient outcomes has been questioned in recent literature[3-5]. Moreover, a recent survey of publications of on the prediction of cancer outcomes based on microarrays revealed serious flaws in the statistical analysis of microarray data[6].

1.2 MAQC Project in Response to FDA's Critical Path Initiative

On February 11, 2005, in response to the FDA Critical Path Initiative, scientists at the FDA's National Center for Toxicological Research (NCTR), Jefferson, Arkansas formally launched the MicroArray Quality Control (MAQC) project (http://edkb.fda.gov/MAQC/; FDA/NCTR research protocol number: E0720701; PI: Leming Shi) in order to address reliability concerns as well as other performance, standards, quality, and data analysis issues[7]. Phase I of the MAQC project (MAQC-I, from February 11, 2005 to September 8, 2006) focused on assessing technical reliability of microarray technology for the identification of differentially expressed genes between a pair of well-established reference RNA samples. MAQC-I involved 137 scientists from 51 organizations including the six FDA centers (CBER, CDER, CDRH, CFSAN, CVM, and NCTR), government agencies (the US Environmental Protection Agency, the National Institutes of Health, and the National Institute of Standards and Technology), manufacturers of microarray platforms and RNA samples, microarray service providers, academic laboratories, and other stakeholders. All MAQC participants freely donated their time and reagents for the completion of MAQC-I. Phase II of the MAQC project (MAQC-II) was officially launched on September 21, 2006 at the NCTR and another meeting was held in CDER on November 28-29, 2006. MAQC-II focuses on the development and validation of predictive models or classifiers in clinical and preclinical (toxicogenomic) applications. The MAQC project has been listed as one

of the Critical Path Opportunities initiated by FDA during 2006 (http://www.fda.gov/oc/initiatives/criticalpath/opportunities06.html).

1.3 Two Phases of MAQC Project: MAQC-I (Gene Lists) and MAQC-II (Predictive Models)

Microarray gene expression profiling is being used for a variety of applications, two of which are (1) understanding general expression differences in various biological populations, classes, states, or conditions, which typically leads to the identification of lists of differentially expressed genes (DEGs) that distinguish populations and classes, and (2) the development of predictive models or classifiers that accurately predict outcomes of an *individual* based on a gene expression profile. These two types of applications have important ramifications and distinctions. In the first, information about a population or differences between populations is inferred. In the second, something about an individual member of a population is inferred or predicted. Although signatures can be used to classify individuals (e.g., assign or associate the individual with a subtype of a particular disease), MAQC-II is primarily focused on prediction of health outcomes based on microarray measurement of biological samples. These can putatively be used to predict response to treatment regimens, patient prognosis, recurrence of disease, survival, etc. The two types of applications are being addressed in Phase I and Phase II of the MAQC project, *i.e.*, MAQC-I and MAQC-II, respectively.

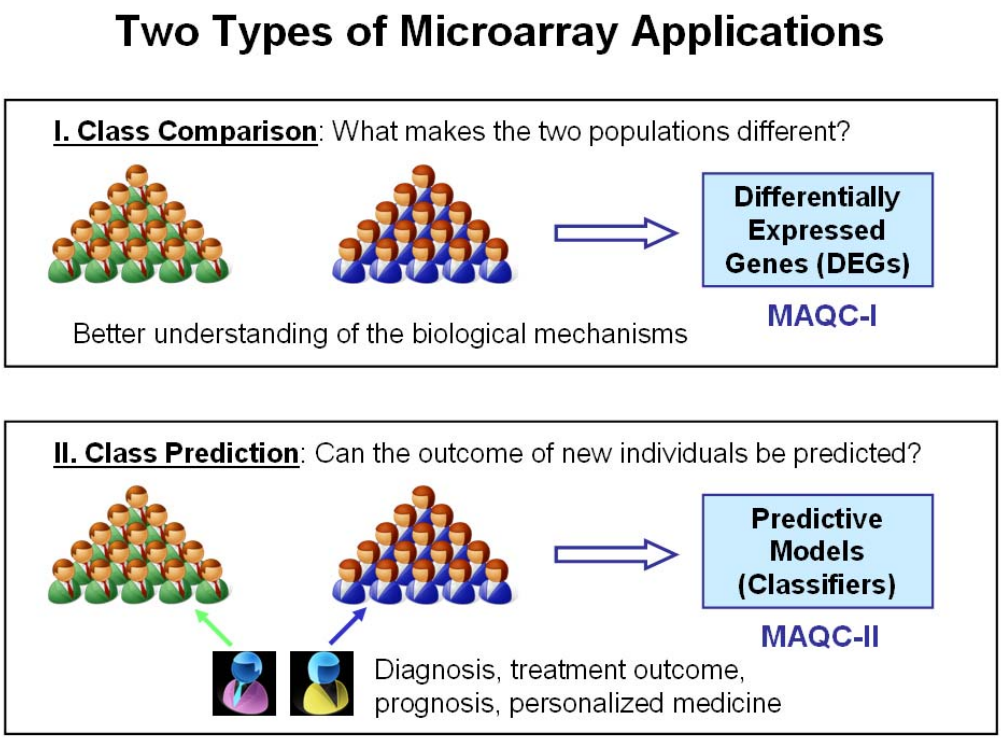


Figure 1. The two major types of applications of microarray technology are being addressed in Phase I and Phase II of the MAQC project, *i.e.*, MAQC-I and MAQC-II, respectively.

1.4 MAQC-I Results: Microarrays Are Reproducible and Reliable

Gene expression data on four titration pools from two distinct, commercially available reference RNA samples (samples A and B, see Abbreviations on page 3) were generated at multiple test sites using a variety of microarray-based and alternative technology platforms. The resulting rich reference data set consists of over 1,300 microarray hybridizations, and additional measurements for over 1,000 genes with alternative technologies such as qPCR. The MAQC project observed high intraplatform reproducibility across test sites, as well as interplatform concordance in terms of genes identified as differentially expressed. Platforms with divergent approaches to the assay generated comparable results in terms of differential gene expression. In other words, the differential gene expression patterns reflected the same biology despite differences in platform technology. Similar results were observed from a realistic rat toxicogenomics experiment[8], in support of the major findings from data generated from the reference RNA samples.

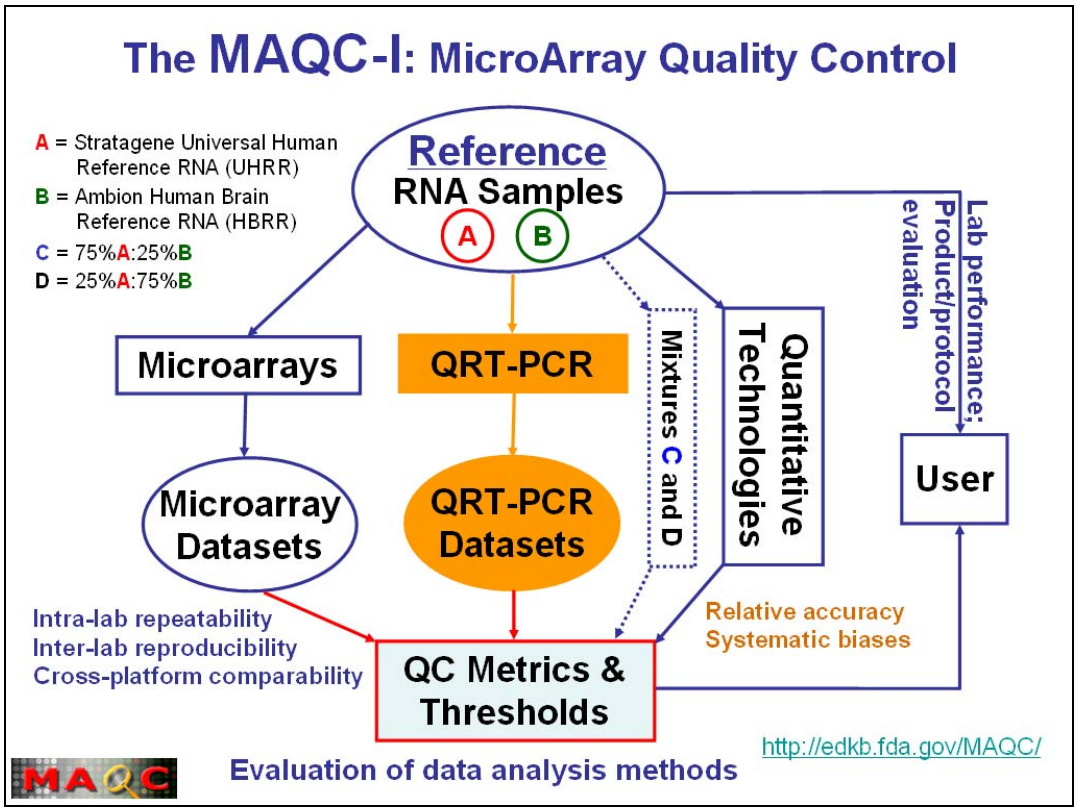


Figure 2. The design of Phase I of the MAQC project for evaluating the technical performance of microarray platforms and data analysis methods in identifying differentially expressed genes.

One important goal of the MAQC Phase I was to assess the best performance achievable with microarray technology under consistent experimental conditions so that future end users will have a benchmark to judge whether the quality of their microarray data is comparable. In doing so, procedural failures of a laboratory or operator may be identified and corrected before precious study samples are profiled. The commercial availability of the two reference RNA samples coupled with the large reference data sets would also allow for the objective evaluation of new array products, reagents, or protocols.

Nature Biotechnology: doi: 10.1038/nbt.1665

Several unique features set the MAQC project apart from previous cross-platform comparison studies: (1) the enthusiastic participation of the microarray community in an extraordinary team effort; (2) the scale of the MAQC data set with over 1,300 microarrays from more than 40 test sites and 20 microarray platforms; (3) the large number of additional gene expression measurements with alternative technology platforms; (4) the commercial availability to the community of the same batches of the two reference RNA samples used in the MAQC study for subsequent quality control, performance evaluations, and proficiency testing; (5) the extensive sequence-based mapping of probes across platforms; and (6) last but not least, the identification of statistical explanations for some misconceptions on the comparability of microarray results.

Major findings of the first phase of the MAQC project were published in six research papers on the September 8, 2006 issue of *Nature Biotechnology*[7-12]. Also published in the same issue was an Editorial[13] by *Nature Biotechnology*, a Foreword by Dr. Daniel Casciano (former FDA/NCTR Director) and Dr. Janet Woodcock (FDA Deputy Commissioner), "*Empowering microarrays in the regulatory setting*"[14], three Commentaries from the FDA[15], the EPA[16], and Stanford University[17], and a Glossary[18]. All the MAQC papers are freely available at *Nature Biotechnology*'s website (http://www.nature.com/nbt/focus/maqc/index.html). In addition, all the MAQC papers were published as a supplement to the Nature Publishing Group in October 2006 and distributed to a wide readership. Data are available through GEO (series accession number: GSE5350), ArrayExpress (accession number: E-TABM-132), ArrayTrack (http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/), and the MAQC website (http://edkb.fda.gov/MAQC/MainStudy/upload/). The MAQC project has attracted international attention as can be seen from the positive reporting by *Cell*[19], *Nature*[20], *Science*[21], *Nature Methods*[22], *Analytical Chemistry*[23], and other scientific publications.

1.5 MAQC-I Debate on Microarray Data Analysis Continues

A major challenge to the microarray user is the existence of numerous options for analyzing the same data set, which lack adequate scientific vetting of their capabilities, implications, and limitations[20]. There is a pressing need to critically evaluate currently available methods with relevant and objective criteria. For example, reproducibility has seldom been, but in the future should be, used as a critical criterion to judge the performance of data analysis procedures. In addition, several differential gene expression profiling studies have demonstrated that the relative expression measures (*i.e.*, difference in transcript abundance between sample types) are typically more consistent than the absolute gene expression levels. The MAQC data set is expected to be widely utilized by the community in order to promote and reach consensus on the appropriate methods for analyzing microarray data.

Lists of differentially expressed genes selected solely by a statistical significance measure are irreproducible: The MAQC-I analyses demonstrated[7, 8] that the apparent lack of reproducibility reported in previous studies using microarray assays[1, 2] was likely caused, at least in part, by the common practice of ranking genes solely by a statistical significance measure, for example, *P* values derived from simple *t*-tests, and selecting differentially expressed genes with a stringent significance threshold, a result that is consistent with a previous report[24]. The gene lists in the MAQC study were much more concordant when fold change was used as the ranking criterion. In addition, widely used statistical methods such as ranking based on FDR values from SAM did not appear to improve interlaboratory or interplatform reproducibility compared to fold-change ranking. Importantly, non-reproducible gene lists could lead to inconsistent

biological interpretations, for example, in terms of enriched GO terms and pathways[8]. Fold-change ranking combined with a less-stringent *P*-value cutoff was found to yield more reproducible signature gene lists[7, 8].

The effect of various data normalization methods on the stability of lists of differentially expressed gene is greatly reduced when fold change is used for gene selection: Data normalization was identified as a major factor for differences when comparing results and data interpretations performed by VGDS (Voluntary Genomic Data Submission) sponsors and FDA reviewers[15]. It should be noted that, although there are many options for normalizing microarray data, when lists of differentially expressed genes are identified by the ranking of fold change, the results are much less susceptible to the impact of normalization methods. In fact, global scaling methods (*e.g*., median- or mean-scaling) do not change the relative rank-order of genes based on fold change; they do, however, significantly impact gene ranking by *P*-value[7, 8, 11].

The MAQC results suggest that microarray data analysis for the identification of reproducible lists of differentially expressed genes does not need be as complicated and confusing as it has been practiced, and consensus on data analysis appears to be attainable. However, concerns have been raised by some in the microarray community about the MAQC recommendations on the identification of differentially expressed genes[19, 25, 26] or on the MAQC project as a whole[27, 28], highlighting the importance for the microarray community to continue the debate in order to reach consensus on microarray data analysis and quality control (http://www.esi-topics.com/nhp/2007/march-07-LemingShi.html).

1.6 From MAQC-I to MAQC-II

The MAQC Phase I (MAQC-I) has demonstrated the technical reliability of microarray technology in detecting differential gene expression. However, questions remain regarding the reliability of the technology in clinical applications such as for disease diagnostics or prognostics, and for tailoring treatments based on gene expression profiles[3, 4, 5, 29]. To investigate the capabilities and limitations of microarrays in such practical applications, the MAQC Phase II (MAQC-II) has been launched to address technical and scientific issues involved in the development and validation of predictive models or classifiers. Invitation for participation in MAQC-II was announced in *Federal Register*, 71(77), 20707-8, April 21, 2006 (available at http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/docs/FederalRegister_MAQC_FollowUp.pdf). Multiple data sets will be collected and distributed to participating organizations for independent analyses. The results will normally be evaluated in three different levels: within a single data set via cross-validation, validation across one or more independent data sets from studies with the same (or similar) study objectives, and validation with blinded "prospective" samples. It is anticipated that the MAQC project, through the community's active participation, will foster development of "best practices" for the generation, analysis, and application of microarray data in the discovery, development, and review of FDA-regulated products.

2. Objectives of MAQC-II

The overall goal of MAQC-II is to comprehensively evaluate different approaches for the development and validation of predictive models or classifiers for clinical and preclinical (toxicogenomics) applications by applying the same set of approaches to a variety of data sets with diverse endpoints on which predictions are being developed.

2.1 Clinical Applications

The primary objectives are to characterize approaches to prediction using DNA microarrays for potential diagnostic, prognostic or therapeutic applications, as well as assist the FDA in understanding the performance characteristics and limitations of multigene clinical outcome predictors using RNA from clinical specimens. The MAQC-II Clinical Working Group was formed to systematically examine these issues, with the understanding that individual academic or industry sites may not possess the resources to independently address each of them. All predictions pertain to an individual patient endpoint.

1. Understand the behavior of various prediction rules and gene selection methods that may be applied to microarray data sets to produce clinical outcome predictors: (a) Examine the influence of the number of variables (probes or probe sets) on prediction accuracy and robustness of the prediction result (in cross-validation and in independent and "prospective" validation); (b) Examine the influence of prediction rules (algorithms) on prediction accuracy and the robustness of prediction results (in cross-validation and in independent validation); and (c) Examine robustness of prediction results in the face of increasing experimental and artificial noise.
2. Identify and characterize the sources of variability in multi-gene prediction results including (a) Impact of tissue acquisition (biopsy method) and sample preparation; (b) Inter- and intra-laboratory variation in prediction results (in replicate experiments on the same platform); and (c) Cross-platform performance of prediction results (in replicate experiments on different platforms).

2.2 Preclinical (Toxicogenomics) Applications

A primary goal is to assess the reliability of models for the prediction of toxicity of new chemicals based on microarray gene expression profiling. The entity to be predicted is the toxicological endpoint (e.g., the presence or absence of liver toxicity) for a chemical, and usually not for an individual animal. Note that in Clinical Applications, the entity to be predicted is usually outcome of a subject (patient).

3. Design of MAQC-II

3.1 Overview of MAQC-II Workflow

To investigate the capabilities and limitations of microarray technology in such practical applications, the MAQC Phase II (MAQC-II) has been launched to address technical and scientific issues involved in the development and validation of predictive models and classifiers (Figure 3). Multiple data sets will be collected and distributed (subject to Confidential Information Disclosure and Transfer Agreements) to participating organizations for independent analyses with available methodologies. The resulting models or analysis methods will be evaluated at three different levels: within a single data set via cross-validation, validation across independent data sets from studies with the same study objectives, and "prospective" validation with data from "prospective" samples. In addition, there may be related studies that are performed as pilot studies or to explore a specialized topic of general interest to the MAQC participants.

3.2 Four Working Groups

1. The Clinical Working Group (CWG) will focus on data sets related to clinical applications.
2. The Toxicogenomics Working Group (TGxWG) will focus on data sets related to toxicogenomic applications.
3. The Titrations Working Group (TitrationWG) will focus on data sets from MAQC titration samples (including the MAQC-I Pilot II data from 13 titration mixtures run by four platforms).
4. The Regulatory Biostatistics Working Group (RBWG) will provide recommendations to MAQC-II CWG and TGxWG on the process and criteria for evaluating the performance of predictive models and classifiers.

If you are interested in contributing to a particular WG, please contact the coordinators of the corresponding WG listed in Table 1, and notify Leming Shi (leming.shi@fda.hhs.gov) to ensure that you will be included in the MAQC mailing list. Leming Shi will coordinate the overall activities of the entire MAQC-II project.

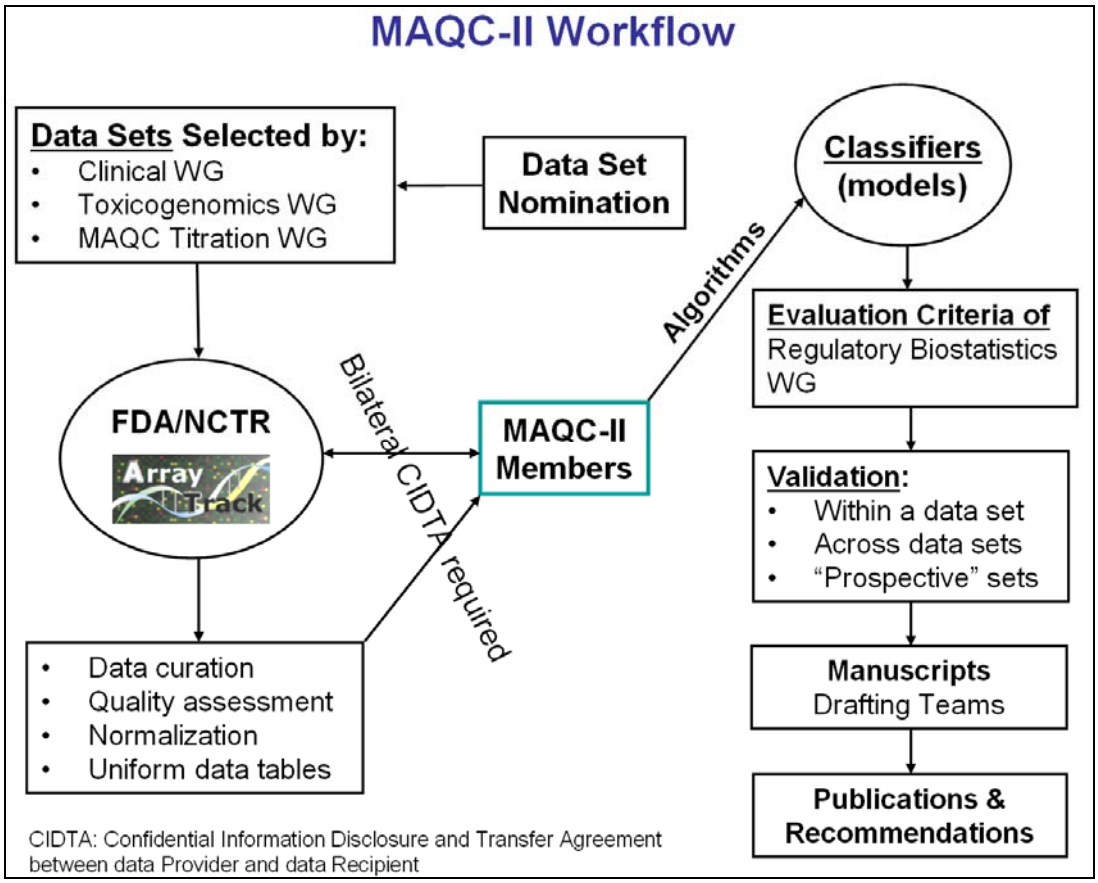


Figure 3a. A simplified overview of the workflow of MAQC-II on the development and validation of predictive models.

Nature Biotechnology: doi: 10.1038/nbt.1665

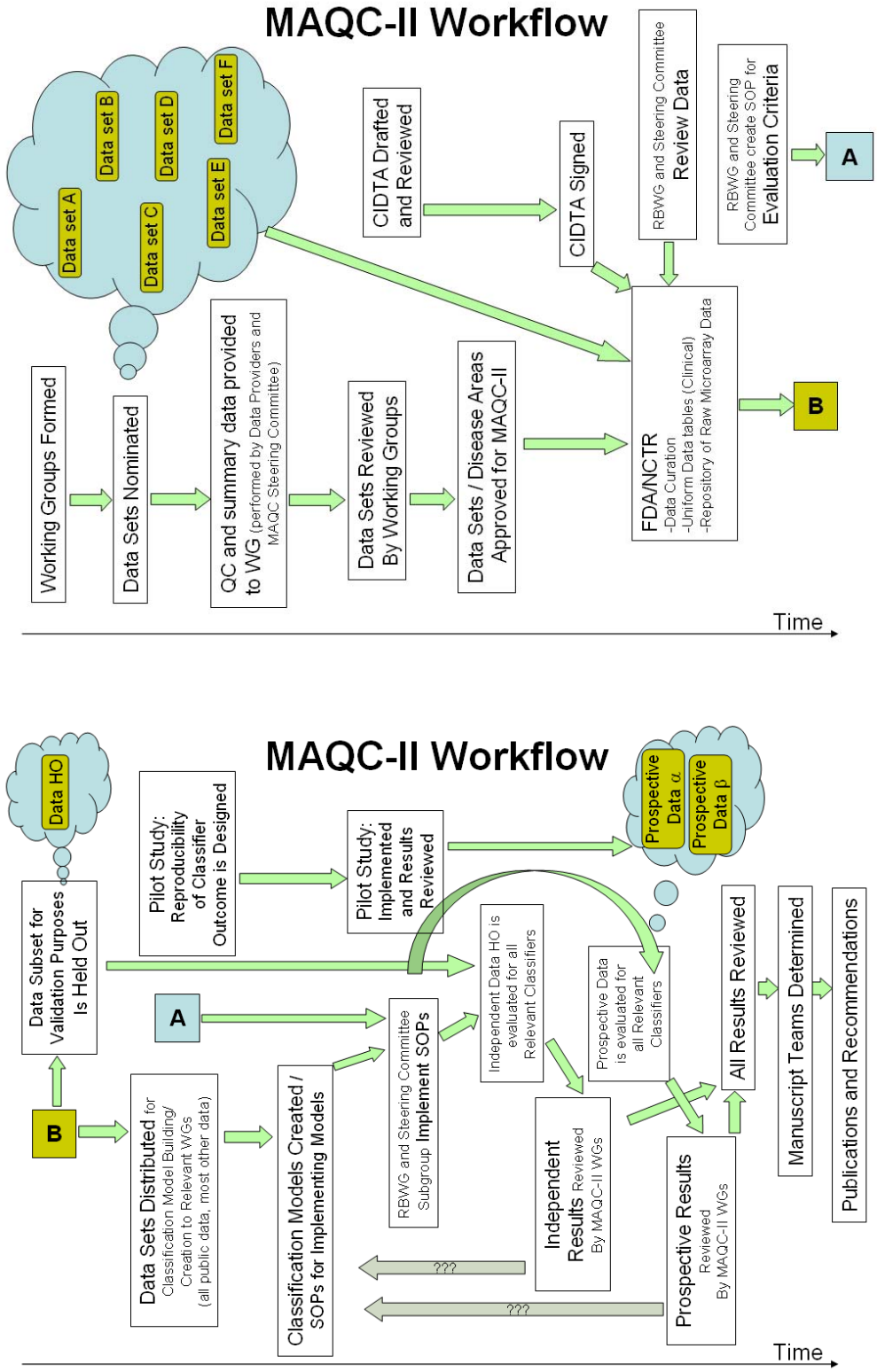


Figure 3b. A more detailed overview of the workflow of MAQC-II on the development and validation of predictive models. *(Courtesy of Dr. Wendell Jones, Expression Analysis, Inc.)*

Nature Biotechnology: doi: 10.1038/nbt.1665

Table 1. Coordinators of the four MAQC-II Working Groups

Working Group	Coordinator	E-mail
Clinical WG (CWG)	Uwe Scherf Wendell D. Jones Lajos Pusztai	uwe.scherf@fda.hhs.gov wjones@expressionanalysis.com lpusztai@mdanderson.org
Toxicogenomics WG (TGxWG)	Federico M. Goodsaid David J. Dix	federico.goodsaid@fda.hhs.gov dix.david@epa.gov
MAQC Titrations WG (TitrationWG)	Richard Shippy Roderick V. Jensen Russell D. Wolfinger	richard.shippy@ge.com roderick.jensen@umb.edu russ.wolfinger@sas.com
Regulatory Biostatistics WG (RBWG)	Gregory Campbell Lakshmi Vishnuvajjala Timothy S. Davison	greg.campbell@fda.hhs.gov lakshmi.vishnuvajjala@fda.hhs.gov tdavison@asuragen.com
MAQC Coordinator	*Leming Shi*	*leming.shi@fda.hhs.gov*

3.3 Data Sets for Clinical, Toxicogenomics, and Titration Applications

Data sets are being identified for the purposes of evaluating

a) the performance of predictive models and classifiers, and
b) the performance of different approaches and methodologies for algorithms commonly used in the development of predictive models and classifiers.

Data sets that were initially nominated early in the process were discussed during the 6th MAQC face-to-face meeting in Washington, DC and Silver Spring, MD, November 28-29, 2006; meeting agenda and summary are available at the MAQC web site: http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/docs/MAQC6_Nov-28and29-2006_Summary.pdf. New data sets may continue to be identified for the purposes of effectively conducting the three stages of validation of predictive models and classifiers (see Section 3.6).

1. **Data Sets for Clinical Working Group**: Four diseases, namely breast cancer (BR), multiple myeloma (MM), acute lymphoblastic leukemia (ALL), and neuroblastoma (NB), are being considered for more detailed examination (esp. "prospective" performance) for predictive modeling using microarray data in MAQC-II. The clinical data sets that were discussed during the 6th MAQC meeting or were recently identified for use by the CWG are summarized in Table 2. Additional data sets that are useful for MAQC-II may still be considered during the course of the MAQC-II. Most clinical data sets can be used for addressing different types of clinical applications: disease subtype classification, treatment outcome, response to therapy, and disease prognosis. These clinical endpoints are examples of "dependent variables" that can be predicted by the predictive models or classifiers. The CWG is responsible for finalizing the diseases and related outcomes and the corresponding data sets that will be analyzed by MAQC-II. All data will be reviewed for quality of sample collection and processing consistency, and quality of microarray and clinical data.

Nature Biotechnology: doi: 10.1038/nbt.1665

Table 2. Summary of clinical data sets being considered for MAQC-II

Data Source	Clinical Applications	Number of Samples	Additional Samples	Contact
Breast Cancer				
MD Anderson Cancer Center	Treatment outcome (Subtype classification)	133	Yes	Lajos Pusztai lpusztai@mdanderson.org
Jules Bordet Instituet (Brussels)	Prognosis Treatment outcome	198+ 61	Yes	Christos Sotiriou christos.sotiriou@bordet.be
University of North Carolina	Subtype classification	131		
NKI/Rosetta	Prognosis	97		
Multiple Myeloma				
University of Arkansas for Medical Sciences (UAMS)	Subtype classification; Prognosis; Treatment outcome	565	Yes (hundreds)	John D. Shaughnessy, Jr. shaughnessyjohn@uams.edu
Millennium	Subtype classification; Treatment outcome; (Prognosis)	264		George J. Mulligan george.mulligan@mpi.com
University of Heidelberg	Prognosis (Subtype classification)	112	Yes	Dirk Hose dirk.hose@med.uni-heidelberg.de
University of Milan	Subtype classification	102	Yes	Antonino Neri neri.a@policlinico.mi.it
Acute Lymphoblastic Leukemia (ALL)				
St. Jude Children's Research Hospital	Treatment outcome; Subtype classification; (Prognosis)	98 (360)	Yes	Meyling H. Cheok meyling.cheok@stjude.org
Erasmus University Medical Center	Treatment outcome; Subtype classification; (Prognosis)	173	Yes	
Neuroblastoma				
University of Cologne	Prognosis	251	Yes (>200)	André Oberthuer andre.oberthuer@uk-koeln.de

2. **Data Sets for Toxicogenomics Working Group**: The goal of the TGx WG is to develop and compare methods for deriving genomic signatures from gene expression data that diagnose or predict toxicity of compounds in animal models. It should be noted that the individual entities that will be predicted or classified are individual chemicals, not individual animals. Except for a few data sets, the initially nominated data sets were determined to be unsuitable for developing predictive classifiers due to the very limited number of compounds involved in a data set (Table 3). However, some of these small data sets might be useful during the validation process. Iconix nominated the largest TGx data sets for three distinct applications based on microarray gene expression profiles: (1) predicting non-genotoxic liver carcinogens from non-carcinogens; (2) predicting liver toxicants from non-toxicants; and (3) predicting kidney toxicants from non-toxicants. New data (and/or samples) from EPA's on-going ToxCast program and Hamner's mouse lung tumor study could serve as "prospective" validation.

Nature Biotechnology: doi: 10.1038/nbt.1665

Table 3. Summary of toxicogenomics data sets for MAQC-II

Data Source	TGx Applications	Number of Chemicals	Additional Chemicals	Contact
Liver Carcinogenicity				
Iconix	Non-genotoxic hepatocarcinogenicity	147		Mark Fielden mfielden@iconixbiosciences.com
EPA	Non-genotoxic hepatocarcinogenicity	?	Yes	David J. Dix dix.david@epa.gov
Lung Carcinogenicity (Mice)				
Hamner	Non-genotoxic chemical-induced hepatocarcinogenicity	13	5	Russell S. Thomas rthomas@thehamner.org
Liver Toxicity				
Iconix	Liver toxicity	22		Mark Fielden mfielden@iconixbiosciences.com
EPA	Liver toxicity	5+2+12	Yes	David J. Dix dix.david@epa.gov
NIEHS/Cogenics	Liver toxicity	8		Richard S. Paules paules@niehs.nih.gov
NIEHS/Cogenics	Acetaminophen treatment	1		Edward K. Lobenhofer elobenhofer@icoria.com
Kidney Toxicity				
Iconix	Kidney toxicity	75		Mark Fielden mfielden@iconixbiosciences.com
Miscellaneous				
MGH	Estrogenicity	6		Toshi Shioda shioda@helix.mgh.harvard.edu

3. **Data Sets for Titration Working Group**: The Titration Working Group's main objective is to provide a "positive control" study for evaluating the performance of classifiers by using the titration data sets generated by the MAQC-I main study (A, B, C, and D samples) and the MAQC-I Pilot II Titration with 13 titration mixtures from defined ratios of A and B (Table 4, non-public). Affymetrix, GE Healthcare, and Illumina submitted Pilot II titration data to MAQC. If needed, additional titration samples may be created and profiled.

Table 4. Summary of titration data sets for MAQC-II

No.	Sample B (%)	Sample A (%)	Number of Replicates
1	100	0	6
2	99.5	0.5	3 (or 6)
3	99	1	3
4	95	5	3
5	90	10	3
6	75	25	3
7	50	50	3
8	25	75	3
9	10	90	3
10	5	95	3
11	1	99	3
12	0.5	99.5	3 (or 6)
13	0	100	6
Total number of arrays per site (manufacturer)			**45 (or 51)**

Nature Biotechnology: doi: 10.1038/nbt.1665

3.4 Prediction and Classification Methods

Numerous algorithms (methods) have been reported in the literature for developing prediction models and classifiers based on microarray gene expression data. The Regulatory Biostatistics WG (RBWG) will conduct a literature survey (or participant survey) and suggest more commonly (and possibly appropriately) used methods to be evaluated with the MAQC-II data sets. Timothy Davison (Asuragen) has been compiling a list of modeling and classification methods and procedures for evaluation; this list could be used as a starting point for investigators and the RBWG to understand the performance characteristics of different methods for MAQC-II evaluation.

3.5 Criteria for Evaluating Model Performance

The Regulatory Biostatistics WG will be recommending a set of criteria for the objective evaluation of the performance of predictive models and classifiers. Although prediction accuracy, sensitivity and specificity should be the main criterion for evaluating the performance, the robustness and mechanistic relevance of the model/classifier are also important additional considerations (Figure 4). That is, when the prediction accuracy (sensitivity and specificity) is comparable, a model/classifier that offers a robust and reproducible outcome across data sets, is less sensitive to sporadic errors, or offers new insights to the biological problems should be given a higher priority. It is anticipated that a better understanding of the capabilities and limitations of microarrays in clinical and toxicogenomic applications could be reached and recommendations on the development and validation of predictive classifiers (signatures) may be put forward through MAQC-II.

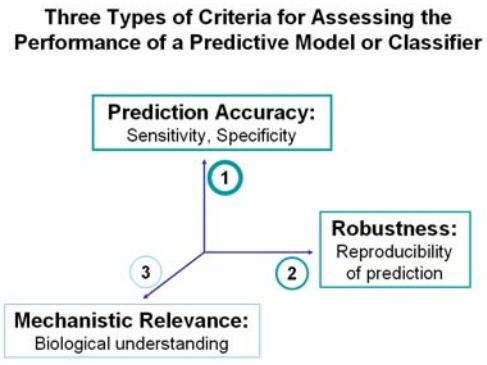


Figure 4. In addition to prediction accuracy, robustness and mechanistic relevance are desirable features for a predictive model.

During the 6th MAQC meeting, there were discussions on relevant performance evaluation criteria. For example, Richard Simon (NIH/NCI) highlighted some "guiding principles" on the development and evaluation of predictive models and classifiers and suggested that the validation should NOT involve (1) measuring overlap of gene sets used in classifiers developed from independent data; (2) statistical significance of individual gene expression levels or summary signatures in multivariate analysis; (3) confirmation of gene expression measurements on other platforms; and (4) demonstrating that the model/classifier or any of its components are "validated biomarkers of disease status". Instead, valid metrics for the validation of predictive models or classifiers should include (1) predictive accuracy; (2) reproducibility of outcome for individual patients; and (3) medical utility.

Gene Pennello (FDA/CDRH) stated that the value of MAQC-II should not be in evaluating whether particular prediction rules are better than others, *per se*, but in evaluating if strategies for validating a prediction rule are better than others. Validation strategies that work can be used to support approval of genomic signatures, and validation strategies that are least burdensome can shorten time to market. Strategies for evaluating models or classifiers should

include performance validation, algorithm stability, and reproducibility. The evaluation of strategies for developing models or classifiers is useful to the FDA because (1) the dissemination of good principles for models or classifier development can lead to the decreased likelihood of an approvable, but flawed model/classifier; and (2) the proper assessment of error rates is needed to properly determine the sample size for a Phase III or pivotal trial.

The RBWG SOP on data analysis includes suggested criteria for evaluating performance of predictive models and classifiers (Appendix 1: RBWG_SOP_DataAnalysis.doc).

3.6 Three Stages of Performance Validation of Predictive Models

An important objective of MAQC-II is to reach consensus on procedures for performance evaluation of different models or classifiers. To adequately evaluate the performance, we need to subject a predictive model or classifier to three stages of validation (Figure 5):

1. **Stage I - Initial Discovery (Internal Validation within One Data Set)**: A predictive model or classifier will typically be developed based on a single data set generated from a single institution. The performance may be assessed in an "internal validation" process such as a leave-n-out cross-validation using the same single data set.
2. **Stage II - Independent Validation (Cross-study/Data Set Validation)**: Prediction models or classifiers will be developed from one or more data sets and a separate data set will be used as an independent test data set for validating the performance of the models or classifiers trained on the initial data sets. The data sets should share the same clinical design, share common data characteristics, and be generated independently from multiple institutions or from the same institution but from different time periods or platforms.
3. **Stage III - Clinical Utility**: The clinical utility of a prediction model or classifier is validated (or challenged) by comparing its performance against traditional (e.g., non-transcriptomic profiling) clinical practices, preferably based on new data from "prospective" studies.

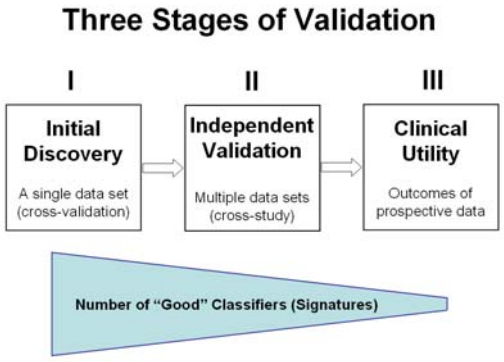


Figure 5. Validating predictive models in three stages

It is anticipated that the number of predictive models or classifiers that "survive" the three stages of validation will decrease dramatically as the stringency of validation increases from Stage I to Stage III validation. It is likely that some methods will be found more prone to over-fitting, thus limiting their practical utilities because of the lack of extrapolation power to new samples or new studies. Most of the same set of criteria defined in "Section 3.5: Criteria for Evaluating Model Performance" will be applied in each of the three stages to validate/evaluate the performance of predictive models/classifiers.

Data sets to be used in MAQC-II analysis should serve for the purposes of the three stages of validation for any given clinical or toxicogenomic application endpoint. MAQC-II may continue to seek additional data sets if there is a gap for conducting the three stages of validation.

3.7 Matrix of Performance Metrics

It is expected that a matrix of performance metrics (Table 5) will be created by applying various data analysis methods to different types of data sets and applications (toxicogenomic and clinical) under the three stages of the validation process (Section 3.6). The RBWG is expected to finalize a list of performance metrics to objectively assess model (method) performance. For meaningful meta-analysis of the matrix of performance metrics, each data analysis group will apply its analysis methods across different data sets and record the corresponding performance metrics.

An important goal of MAQC-II data analysis is to populate the matrix of performance metrics by applying various data analysis methods to multiple data sets in different stages of validation, and calculating different performance metrics for each method (model). Meta-analysis of the resulting performance matrix will provide important information on the appropriate procedures for the development and validation of predictive models based on microarray gene expression profiles.

Table 5. Populating the matrix of performance metrics

		Performance Metrics						
		1	2	3	.	.	.	*m*
Analysis Methods (Models)	1	$PM_{1,1}$	$PM_{1,2}$	$PM_{1,3}$	.	.	.	$PM_{1,m}$
	2	$PM_{2,1}$	$PM_{2,2}$	$PM_{2,3}$	.	.	.	$PM_{2,m}$
	3	$PM_{3,1}$	$PM_{3,2}$	$PM_{3,3}$	.	.	.	$PM_{3,m}$
	.							
	.							
	.							
	.							
	.							
	n	$PM_{n,1}$	$PM_{n,2}$	$PM_{n,3}$	.	.	.	$PM_{n,m}$

4. Participants

Participation in MAQC-II is voluntary, and each participant is expected to cover the costs associated with participating in the MAQC-II project. There is no “MAQC fund” to provide for any participant. Each participant agrees to the conditions and terms set in this Research Plan and the Confidential Information Disclosure and Transfer Agreement (CIDTA).

4.1 Data Providers

Data providers are organizations that provide either publicly available data and/or Confidential Information to the MAQC Data Warehouse. Such Confidential Information may be microarray data, clinical information, or both. Current data providers are listed in Tables 2 and 3. Additional

Nature Biotechnology: doi: 10.1038/nbt.1665

data providers may be identified and invited during the MAQC-II project if needed. Furthermore, some of the current data providers may acquire and provide additional biological samples for "prospective" validation of models/classifiers.

4.2 Data Analysis Sites

Data analysis sites are organizations that receive general information, curated public data, and/or Confidential Information from MAQC Data Warehouse and conduct data analysis under the scope of the MAQC-II project. Data analysis sites should report analysis results back to MAQC-II.

4.3 Platform Providers

MAQC-II may generate additional microarray data sets using biological samples from on-going "prospective" studies so that Stage III validation (Section 3.6) may be conducted. In addition, other pilot studies and specialized studies relevant to the goals of MAQC-II may be designed and executed. Microarray manufactures (Table 6) including Affymetrix, Agilent, Eppendorf, Illumina, PhalanxBiotech, and Telechem agreed in principle to supply substantial numbers of microarrays for the MAQC-II validation and "prospective" efforts as well as related studies. Separately, ABI, Gene Express, Panomics, and SuperArray have also pledged support with gene expression platforms other than microarrays. MAQC-II will decide, as a group, what additional data will be generated and on what platforms as we move forward.

It may not be feasible to process samples and perform testing/validation on multiple platforms for an individual study that was initially trained on one platform, especially for the clinical studies. However, if there is success of having accurate, reproducible (*i.e.*, more than one laboratory) results from a "prospectively" validated model/classifiers on one particular microarray platform with a separate platform for a different study (possibly run in parallel), then this would be considered in harmony with the overall MAQC effort. In addition, it may be possible to examine the robustness of a model/classifier designed on one platform when it is "ported" to another, given proper technological and methodological considerations. This examination may be appropriate for a pilot or specialized study.

A solid proposal is needed for the array and alternative platform manufacturers so that they can adequately plan and solicit funds for resources related to the MAQC-II efforts.

Table 6. Platform providers pledged support to MAQC-II

No.	Provider	Contact	E-mail
1	Affymetrix	Janet A. Warrington	janet_warrington@affymetrix.com
2	Agilent	Paul K. Wolber	paul_wolber@agilent.com
3	Applied Biosystems	Raymond R. Samaha	raymond.samaha@appliedbiosystems.com
4	Eppendorf Array Tech.	Francoise de Longueville	delongueville.f@eppendorf.be
5	Gene Express	James C. Willey	james.willey2@utoledo.edu
6	Illumina	Shawn C. Baker	scbaker@illumina.com
7	Panomics	Yuling Luo	yluo@panomics.com
8	PhalanxBiotech	Charles Ma	charlesma@phalanxbiotech.com
9	SuperArray	Jingping Yang	jpyang@superarray.net
10	TeleChem ArrayIt	Paul K. Haje	paul@arrayit.com

Nature Biotechnology: doi: 10.1038/nbt.1665

4.4 Reference Sites

Reference sites will be identified in the future to process biological samples from data providers in actual microarray experiments using microarrays (and reagents) from platform providers for "prospective" related studies identified by the MAQC-II.

4.5 Including or Excluding a Data Set

The decision to include a new data set after March 31, 2007 will be determined by the MAQC-II Steering Committee (Section 7); only data sets that significantly help achieve the overall objectives of the MAQC-II should be considered for inclusion. The Steering Committee may also decide to exclude a data set from MAQC-II analysis if the data set is found to be of limited use for the MAQC-II project.

4.6 Including or Excluding a Participant

The decision to include a new participant after March 31, 2007 will be determined by the Steering Committee. If approved, the Steering Committee (Section 7) and coordinators of the corresponding Working Groups will ensure that the new participant is adequately briefed and agrees to the conditions of the MAQC-II Research Plan and related documents. Participants who do not follow the MAQC-II Research Plan, as determined by the Steering Committee, may be excluded from future MAQC activities.

5. Participant's Responsibilities

1. Participant agrees to follow the general principles set in this Research Plan document and the RBWG SOP.
2. Participant agrees to the Confidentiality Terms (Section 6) before accessing the Confidential Information portion of the MAQC data sets, as defined in the MAQC CIDTA document.
3. Participant agrees to the publication and public deposition of the data sets and related analysis results at the time of acceptance of MAQC-II manuscript(s).
4. Data Provider agrees to submit data set(s) to the MAQC Data Warehouse as son as possible and provide sufficient background information to participants to understand the data set(s).
5. Data analysis site agrees to report analysis results to the MAQC in a timely fashion during face-to-face meetings, conference calls, and e-mail exchanges. Data analysis site also agrees to actively participate in conference call discussions and manuscripts preparation.
6. Participant agrees to cover the costs as a result of her/his involvement in MAQC-II. No fund is provided to any participant.
7. Each platform provider will provide consistent lots of arrays and kits to reference sites along with a standardized protocol on the generation of the "prospective" data, when needed.
8. Each platform provider will help ensure performance capabilities of the selected reference sites by providing arrays and reagents, and potential training.
9. Each reference site will be expected to be proficient in and execute the standardized protocol (with appropriate reagents and kits) to ensure competency and consistency with the protocol.
10. Each reference site should submit data to FDA/NCTR within 5 weeks of receiving the RNA.

Nature Biotechnology: doi: 10.1038/nbt.1665

6. Confidentiality Terms for Accessing MAQC-II Data Sets

1. Participant is required to sign and abide by the Confidential Information Disclosure and Transfer Agreement (CIDTA) with Data Provider before the data provider's data in the MAQC Data Warehouse can be made available to the participant.
2. Participant should not disseminate the MAQC-II data sets or results to others not bound to the respective CIDTA.
3. Prior to acceptance for publication of MAQC-II manuscript(s), public presentation or publication of the Confidential Information portion of the MAQC-II data, as defined in the MAQC CIDTA document, and results derived specifically from the Confidential Information portion of the MAQC-II data is prohibited.

7. MAQC Steering Committee

Whenever possible, consensus from the entire MAQC consortium will be sought before any important decision is made about the MAQC project. However, there may be situations when it becomes unfeasible to reach consensus among all members of the MAQC-II project. The MAQC Steering Committee (Table 7) has been established for resolving any remaining issues not addressed in this Research Plan document or whenever consortium consensus is not feasible. The Steering Committee is responsible for ensuring the delivery of the project outcomes. The Steering Committee consists of coordinators of the four Working Groups (Table 1) and representatives from US federal government agencies. It will be the responsibility of all members of the Steering Committee to abstain from participating in a particular activity of the Steering Committee if such participation would create a conflict of interest. Members of the MAQC may reasonably request that individual Steering Committee members abstain from participating in a particular decision-making.

Table 7. Members of the MAQC-II Steering Committee

No.	Name	E-mail	Organization
1	Gregory Campbell	greg.campbell@fda.hhs.gov	FDA/CDRH
2	Timothy S. Davison	tdavison@asuragen.com	Asuragen
3	David J. Dix	dix.david@epa.gov	EPA
4	Felix W. Frueh	felix.frueh@fda.hhs.gov	FDA/CDER
5	James C. Fuscoe	james.fuscoe@fda.hhs.gov	FDA/NCTR
6	Federico M. Goodsaid	federico.goodsaid@fda.hhs.gov	FDA/CDER
7	Roderick V. Jensen	roderick.jensen@umb.edu	Univ. Mass. Boston
8	Wendell D. Jones	wjones@expressionanalysis.com	Expression Analysis
9	Raj K. Puri	raj.puri@fda.hhs.gov	FDA/CBER
10	Lajos Pusztai	lpusztai@mdanderson.org	MD Anderson
11	Uwe Scherf	uwe.scherf@fda.hhs.gov	FDA/CDRH
12	Leming Shi	leming.shi@fda.hhs.gov	FDA/NCTR
13	Richard Shippy	richard.shippy@ge.com	GE Healthcare
14	Weida Tong	weida.tong@fda.hhs.gov	FDA/NCTR
15	Lakshmi R. Vishnuvajjala	lakshmi.vishnuvajjala@fda.hhs.gov	FDA/CDRH
16	Russell D. Wolfinger	russ.wolfinger@sas.com	SAS Institute

Nature Biotechnology: doi: 10.1038/nbt.1665

8. MAQC-II Procedures

8.1 Data Submission Procedures

1. For FDA IRB's record, Data Provider should sign a "banking" form addressed to Leming Shi, and contains the following language:

   "I am submitting the following data for inclusion into the MAQC project:
   (list specific data to be deposited)

   I am confirming that the above data were not collected specifically for the MAQC project. These data were collected under appropriate Institutional Review Board approval by our institution. Individuals agreed to have their data banked for further research, including genetic research. Confirmation of this consent for banking is available upon request.

   The data are coded so that the identity of the individuals is protected. Under no circumstances will FDA/NCTR or any MAQC recipient of these data be given access to either the key or to any information that may enable them to decipher the code."

2. Data Provider should submit its data set (microarray data and demographic/clinical information) to Leming Shi at the FDA/NCTR as part of the MAQC Data Warehouse in a timely manner upon request.
3. The microarray data should be submitted in a raw (original) data format; for example CEL file format for Affymetrix platforms and scanner FE output format in tab-delimited ASCII format for Agilent format.
4. Demographic/clinical information should be submitted in a separate Excel spreadsheet that links patient information to the microarray data file name unambiguously.
5. Data Provider is encouraged to provide quality assessment information to MAQC Data Warehouse, if available.
6. As each data set is submitted to MAQC Data Warehouse, it will be reviewed by a team at FDA/NCTR for completeness, quality, and errors. No other participants will be given access to the raw data until the Data Distributions Procedures (Section 8.2) are followed. High-level quality views or summaries of data, for quality review purposes, may be distributed to MAQC members ahead of time as the particular WG or Steering Committee sees fit.
7. Data should be submitted via FTP or in DVD to Leming Shi (contact information shown on page 1 of this Research Plan).

8.2 Data Distribution Procedures

1. If a participant (as a Data Recipient) is interested in analyzing a particular data set listed in Tables 2-4, s/he should contact the Data Provider directly.
2. Data Recipient and Data Provider sign a bilateral Confidential Information Disclosure and Transfer Agreement (CIDTA).
3. Data Provider sends a copy of the signed CIDTA by e-mail (in PDF format) or fax to Leming Shi (detailed contact information is listed on the cover page of this Research Plan).
4. Data Recipient sends an e-mail to Leming Shi to clearly state that (1) s/he has signed the CIDTA and wants to access the data and (2) s/he has carefully read and agrees to the MAQC-II Research Plan and the attached SOP document.
5. For FDA IRB's record, Data Recipient should sign a "withdrawal" form addressed to Leming Shi, and contains the following language:

   "I confirm that I have received the following data from the MAQC project:

(list specific data/specimens to be withdrawn).

I understand that under no circumstances will I be able to attain any identifying information about the individual data that I have received from the MAQC project.

I acknowledge that I will use this research material only in accordance with the conditions stipulated by the MAQC project. Any additional use of this material will require an approval by the FDA IRB and, where appropriate, by an IRB at the recipient site."

6. Leming Shi notifies Data Recipient by e-mail (cc Data Provider) that the data for which s/he has signed the CIDTA are ready for access (instructions will be provided separately).
7. Data Recipient calls Leming Shi (+1-870-543-7387) to get the FTP address and password.
8. Data Recipient conducts data analysis in accordance with the MAQC-II Research Plan and submits analysis results to the MAQC in a timely manner for consideration of inclusion in manuscripts. There is no restriction on the types of analyses that may be performed by each Data Recipient, but the results will be reviewed by other members of the MAQC.
9. Each Data Recipient accessing MAQC-II datasets must strictly fulfill its obligations of confidentiality as set forth in this Research Plan document and the CIDTA.
10. MAQC-II data sets will be submitted to a public repository (e.g., GEO) at the time of manuscript submission, tentatively scheduled for March, 2008.

8.3 Data Analysis Procedures

Participants conduct data analysis in accordance with the MAQC-II Research Plan and report results to MAQC-II. There is no restriction on the types of analyses that may be performed by each participant. However, each participant should report his/her results in sufficient details so that other MAQC members will have enough information to judge the validity of the analysis results. A separate document on the strategies of data analysis has been developed by RBWG and is attached at the end of this document (Appendix 1: RBWG_SOP_DataAnalysis.doc).

8.4 Conference Calls

Biweekly conference call will be up for each WG, and conference calls for the entire MAQC will be set up when needed so that participants will be updated about the progress of the project, and new information and ideas are exchanged in a timely fashion.

8.5 Face-to-face Meetings

Analysis results of the MAQC-II data sets will be extensively discussed during face-to-face project meetings. The frequency of face-to-face meetings is roughly once every three to six months. The location and duration of the face-to-face meetings will be decided by the Steering Committee and announced to the entire MAQC at least one month before the meeting dates.

8.6 Planning for Publication

The decisions about the content of the manuscripts that derive from MAQC-II are of significant importance. It is the intent of the MAQC-II Steering Committee to produce one manuscript that summarizes the general findings of the MAQC-II study. It is expected that additional manuscripts will be generated that may address data set-specific questions (e.g., one manuscript for each disease area). All the manuscripts will be produced by the MAQC members. Manuscript team leaders will be those with the domain expertise for a disease area and those who have contributed a significant amount of confidential information (data sets and/or analysis results) to the MAQC-II. It has been suggested that a manuscript on array QC assessment may be

developed based on on-going effort on the identification of potential outlying arrays from each and every data set before predictive models are developed. A "methodological" manuscript may be planned. Authorship will be determined based on an individual's actual contributions.

9. Checklist of Requirements before MAQC-II Data Distribution

1. Data Recipient is required to sign the Confidential Information Disclosure and Transfer Agreement (CIDTA) with data Provider.
2. Participants will agree to the scope of the MAQC-II project and accept the responsibilities of participants outlined within the Research Plan document.
3. Participants acknowledge familiarity with the contents of the MAQC-II Research Plan and the RBWG SOP on data analysis.

10. Timeline

1. **September 21, 2006**: Kickoff meeting at FDA/NCTR;
2. **November 28-29, 2006**: Data set review and model performance evaluation criteria meeting at FDA/CDER;
3. **December 15, 2006**: Data sets (raw data) submitted to FDA/NCTR;
4. **February 20, 2006**: Non-confidential data sets distributed;
5. **March 22, 2007**: Confidential Information Disclosure and Transfer Agreement (CIDTA) finalized and distributed to the entire MAQC-II mailing list of 290 people along with the Research Plan and RBWG SOP; Participants are expected to sign the CIDTA;
6. **March 31, 2007**: Confidential data sets distributed to participants who signed the CITDA;
7. **May 24-25, 2007**: The 7th face-to-face MAQC project meeting to be held at SAS Institute, Cary, North Carolina. Initial analysis results, data analysis strategies and questions to be discussed. Manuscripts topics to be proposed;
8. **June, 2007**: Manuscript teams to be assembled;
9. **October/November, 2007**: The 8th face-to-face MAQC project meeting on data analysis;
10. **March 31, 2008**: Manuscripts submitted;
11. **September, 2008**: MAQC-II results published;
12. **December, 2008**: MAQC-II recommendations on the development and validation of predictive models (classifiers).

11. Web Sites

1. The MAQC (MicroArray Quality Control) Project: http://www.fda.gov/nctr/science/centers/toxicoinformatics/maqc/index.htm
2. MAQC-I results were published in the September 2006 issue of *Nature Biotechnology*: http://www.nature.com/nbt/focus/maqc/index.html
3. FDA's Critical Path Initiative: http://www.fda.gov/oc/initiatives/criticalpath/
4. Genomics at FDA: http://www.fda.gov/cder/genomics/
5. ArrayTrack: http://www.fda.gov/nctr/science/centers/toxicoinformatics/ArrayTrack/
6. ERCC: http://www.cstl.nist.gov/biotech/ERCC/testplan.htm

Nature Biotechnology: doi: 10.1038/nbt.1665

12. References

1. Tan, P.K. et al. Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res* **31**, 5676-5684 (2003).
2. Marshall, E. Getting the noise out of gene arrays. *Science* **306**, 630-631 (2004).
3. Michiels, S., Koscielny, S. & Hill, C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* **365**, 488-492 (2005).
4. Ioannidis, J.P. Microarrays and molecular research: noise discovery? *Lancet* **365**, 454-455 (2005).
5. Ein-Dor, L., Zuk, O. & Domany, E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A* **103**, 5923-5928 (2006).
6. Dupuy, A. & Simon, R.M. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst* **99**, 147-157 (2007).
7. Shi, L. et al. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* **24**, 1151-1161 (2006).
8. Guo, L. et al. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol* **24**, 1162-1169 (2006).
9. Canales, R.D. et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol* **24**, in press (2006).
10. Shippy, R. et al. Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat Biotechnol* **24**, 1123-1131 (2006).
11. Patterson, T.A. et al. Performance comparison of one-color and two-color platforms within the Microarray Quality Control (MAQC) project. *Nat Biotechnol* **24**, 1140-1150 (2006).
12. Tong, W. et al. Evaluation of external RNA controls for the assessment of microarray performance. *Nat Biotechnol* **24**, 1132-1139 (2006).
13. Making the most of microarrays. *Nat Biotechnol* **24**, 1039 (2006).
14. Casciano, D.A. & Woodcock, J. Empowering microarrays in the regulatory setting. *Nat Biotechnol* **24**, 1103 (2006).
15. Frueh, F.W. Impact of microarray data quality on genomic data submissions to the FDA. *Nat Biotechnol* **24**, 1105-1107 (2006).
16. Dix, D.J. et al. A framework for the use of genomics data at the EPA. *Nat Biotechnol* **24**, 1108-1111 (2006).
17. Ji, H. & Davis, R.W. Data quality in genomics and microarrays. *Nat Biotechnol* **24**, 1112-1113 (2006).
18. Reid, L.H. & Warrington, J.A. A note on nomenclature. *Nat Biotechnol* **24**, ii (2006).
19. Strauss, E. Arrays of hope. *Cell* **127**, 657-659 (2006).
20. Eisenstein, M. Microarrays: quality control. *Nature* **442**, 1067-1070 (2006).
21. Couzin, J. Genomics. Microarray data reproduced, but some concerns remain. *Science* **313**, 1559 (2006).
22. Kiermer, V. Microarray quality in the spotlight again. *Nat Methods* **3**, 772 (2006).
23. Sage, L. Do microarrays measure up? *Anal Chem* **78**, 7358-7360 (2006).
24. Shi, L. et al. Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics* **6 Suppl 2**, S12 (2005). http://www.esi-topics.com/nhp/2007/march-07-LemingShi.html
25. Klebanov, L., Qiu, X., Welle, S. & Yakovlev, A. Statistical methods and microarray data. *Nat Biotechnol* **25**, 25-26 (2007).
26. Shi, L. et al. Reply to Statistical methods and microarray data. *Nat Biotechnol* **25**, 26-27 (2007).
27. Liang, P. MAQC papers over the cracks. *Nat Biotechnol* **25**, 27-28 (2007).
28. Shi, L. et al. Reply to MAQC papers over the cracks. *Nat Biotechnol* **25**, 28-29 (2007).
29. Simon, R. Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *J Natl Cancer Inst* **98**, 1169-1171 (2006).

13. Appendix 1: RBWG SOP on Data Analysis

The MAQC RBWG (Regulatory Biostatistics Working Group) prepared an SOP on data analysis, covering general procedures for developing predictive models/classifiers and criteria

for assessing the performance of the developed models/classifiers (Appendix 1: RBWG_SOP_DataAnalysis.doc). Data analysis sites should follow the SOP in the development and validation of predictive models.

Nature Biotechnology: doi: 10.1038/nbt.1665

Supplementary Document 5:

Standard Operating Procedures (SOPs), Methods and Analysis for MAQC-II

By the MAQC Regulatory Biostatistics Working Group (RBWG)

March 22, 2007

Address comments and questions regarding the SOP to the RBWG coordinators:
Gregory Campbell (greg.campbell@fda.hhs.gov)
Lakshmi Vishnuvajjala (lakshmi.vishnuvajjala@fda.hhs.gov)
Timothy S. Davison (timothy.davison@almacgroup.com)
and co-authors of the document:
Gene Pennello (gene.pennello@fda.hhs.gov)
Samir Lababidi (samir.lababidi@fda.hhs.gov)

The overall objective of MAQC-II is to characterize approaches for development and validation of classifiers on DNA microarray data for the purpose of diagnostic, prognostic, or therapeutic application. A specific regulatory focus for MAQC-II is to identify study designs and performance measures for the evaluation of microarray technology and processes for establishing choice of classifier algorithm, choice of validation strategy, choice of normalization method, and handling of missing data.

The FDA has solicited gene expression datasets from DNA microarray studies as well as proposals to analyze these datasets in order to evaluate the impact of different analysis protocols on the selection of genes and their associated signatures for biomarker pattern development. Although this project is being coordinated by FDA, there are no regulatory rights conferred to anyone by the participation of FDA personnel in this project. Although FDA personnel are involved in this project, the views expressed here in this document are not FDA guidance and do not necessarily represent FDA policy.

1. Executive Summary

For each of the datasets in MAQC-II, it is anticipated that there will be a number of Analysis Groups (AGs) that will each undertake the task of building classifiers. Each Analysis Group (AG) will provide to the MAQC-II Regulatory Biostatistics Working Group (RBWG) a specific Statistical Analysis Plan (SAP) describing methodology for the development and validation of classifier(s). It is strongly recommended that this Analysis Plan be submitted before any analysis is undertaken. The SAP should include procedures for external validation, data normalization, assessment of quality of the microarray data, feature selection, (optional) internal cross validation, selection of algorithms for prediction, evaluation of performance (and its variability), including comparison with existing clinical predictors, (optional) evaluation of reproducibility, any hypothesis testing, and the treatment of missing data. A checklist in the Appendix

Nature Biotechnology: doi: 10.1038/nbt.1665

provides some framework about the scope of the SAP. This document on Standard Operating Procedures, Methods and Analysis provides guidance on the procedures and essential characteristics of an SAP, but is not intended to exclude the use of other accepted or novel methods provided they are supported by peer-reviewed publications or explained with sufficient detail in the SAP. Additional guidance on the development and validation of classifiers is available in Dupuy and Simon (2007, *JNCI*, 99, 147-57).

2. **Quality Assessment**
   One way that the quality of microarray data can be assessed is using a quantitative (or possibly qualitative) measure describing the quality of each sample in a dataset but a more general notion is the quality of the system (assessed in part through repeatability and reproducibility of the array platform) as well as the quality of the experiment from which the data came.
   a. **Available quality control metrics:** It is anticipated that quality control metrics will be made available for every sample within each dataset by members of the different Working Groups and will accompany the data package distributed with each dataset. Metrics to describe the quality of each sample will be available at the following levels where appropriate to the specific dataset: (1) at the level of sample procurement and/or extraction; (2) array-specific level; (3) experiment-wide level. Quality control is of particular importance to assess the quality of samples collected at different times or sites or processed by different methods. A particular quality measure for arrays can then be used to support the justification for inclusion in, or exclusion from, of each sample from the analysis. It is expected that this will be done prior to development of any classifier. It is expected that sufficient information to support sample quality will be provided to, or generated, by the MAQC-II prior to data distribution. Different Analysis Groups for the same dataset may choose to share their assessments of quality of the data.
   b. **Quality control metrics for any new platforms:** Should other new platforms be included in the MAQC-II project, an equivalent and comprehensive set of quality control metrics will be proposed to the MAQC-II Steering Committee for review prior to inclusion of any data in the MAQC-II for distribution to the Analysis Groups.
   c. **Review of quality control metrics prior to data release:** It is expected that data quality review committees (DQRC) will be formed by the different Working Groups to assess the quality of all samples and arrays included in the MAQC-II project. Attempts will be made by the DQRC to establish general thresholds or methods which facilitate the identification of samples or arrays that may be excluded from the MAQC-II project on the grounds of insufficient quality.
   d. **Definition of insufficient quality for microarrays:**
      i. Insufficient quality of sample procurement and/or extraction procedures may be identified for samples with missing or inconsistent information.
      ii. Insufficient quality of each array will be defined by failure to meet quality control metrics thresholds defined by the DQRC.

Nature Biotechnology: doi: 10.1038/nbt.1665

e. **Use of poor quality microarray data:** Samples having insufficient microarray quality as identified by the DQRC may be included in classifier development by Analysis Groups at their own discretion.
f. **Repeatability and reproducibility:** Is there enough information in the dataset to understand the variability from array to array or from time to time or from site to site?
g. **Overall quality of experimental design and the associated clinical data:** It is vitally important to assess the quality of the overall experiment in which the microarray data arose and, in particular, to assess the quality of the clinical (non-array) data. Either a poorly designed experiment or a well-designed experiment with poor quality control of the clinical data is likely pose insurmountable challenges to its use in MAQC-II.
h. **Review and use of quality control metrics by Analysis Groups post data release:** Use of quality control metrics and identification of samples of insufficient quality may not be used to exclude samples from consideration after classifier development unless it is a valid component of the Statistical Analysis Plan (see Appendix) or an integral component of the classification algorithm (c.f. §5) as outlined by the Analysis Group prior to receiving the corresponding data package.
i. **Quality control review of prospective data**: Data received by the MAQC-II subsequent to the initial release of retrospective data will require a quality evaluation by the DQRC of the same type used for the initial data prior to its release to MAQC members and the appropriate Analysis Groups.

3. **Data Normalization**

Observed expression levels can include many sources of variability that have different effects on the data. This includes variations due to sample preparation, manufacturing of the arrays, and the processing of the arrays (labeling, hybridization, and scanning). Each Analysis Group may consider normalizing the arrays and it is expected that how that will be accomplished will be described by each Analysis Group in its Statistical Analysis Plan.

a. If studying the impact of normalization on performance of classification models is the goal of an Analysis Group, the Analysis Group may compare many different methods available for normalization of gene expression data, and then use the preferred method in the independent dataset (retrospective hold-out data or external prospective data). (Ref.: “Comparison of Affymetrix GeneChip expression measures”, by Irizarry et al., 2006. See Table 1 for many different methods of normalizations together with the corresponding references).
b. In the case of multi-array normalization, the method and parameter estimates to be used for new subjects or chemicals should be included in the experimental design report provided to the biostatistician. Here, the new subjects or chemicals are part of the prospective (external) validation data.
c. For normalization methods that include background correction, the Experiment Plan should indicate if the background correction is global and/or probe-specific.
d. It is not appropriate to use the entire dataset for the multi-array normalization if part of the data is being held out for the test set validation. Normalizing across the

Nature Biotechnology: doi: 10.1038/nbt.1665

entire dataset means that the training samples are being used in part to classify the test samples, an approach that may introduce bias in the estimates of performance from the test data. Data normalization should not be part of internal cross validation (with repeated train/test).

e. If a normalization is implemented, the method has to be adaptable to normalizing one additional array (such as a prospective one) that does not alter the values of the previous or reference arrays; i.e., the current summarized value from an array in the training set is not and should not depend on the arrays that are yet to be created.

4. **Feature Selection**

Feature selection (e.g., of genes or gene variants) can be an integral step in classifier development. If there is not some attempt to focus on a smaller subset of features, there is a concern by some that, with so many possible features on a microarray and so relatively few subjects, it may be very easy to find a classification model that achieves complete separability between the groups but may have absolutely no chance of being validated. Feature selection may also be required to make computation feasible. All feature selection algorithms are expected to be outlined within the Statistical Analysis Plan and either supported by peer-reviewed literature or described in detail.

a. Feature selection algorithms/workflows can be grouped into three classes: filter, wrapper and embedded. A filter is a feature selection method that selects genes based on individual performance. A wrapper is a feature selection method, such as cross-validation, embedded within a training dataset, which selects combinations of features that give the best performance. A wrapper has the advantage of being able to find synergies between genes but can require more training data than a filter. Embedded methods select features for use in the process of learning. Other strategies include multiple random validations within a training sample to select genes with the highest frequency of being selected. (cf. Stuart G Baker and Barnett S Kramer, BMC Bioinformatics, Identifying genes that contribute most to good classification in microarrays, 2006, 7:407)

b. Feature selection may need to be part of any cross-validation of a classifier. A frequent error in cross-validation is to select features based on the entire dataset, and then cross-validate only the model built on those selected features. Simon et al (JNCI, 2003, 95, 14-18) demonstrate that cross-validated estimates of performance can be overstated tremendously if the cross-validation does not include feature selection. A very important aspect of this project is to validate the process of predictive modeling using genomic data and not to validate specific probes or transcripts that may be optimal for a given predictive model. With so many variables to choose from, the challenge may be to determine how to reduce the dimensionality of the problem so that one may have a reasonable expectation of validating any classification model.

c. Given that several genes may be on the same pathway, the correlation structures among the genes can be utilized in feature selection. More formally, refined feature selection methods may be based on the joint distribution of the gene expression measures. This also has the potential of reducing the number of features selected for the classifier.

Nature Biotechnology: doi: 10.1038/nbt.1665

d. It is recommended that feature selection be approached taking into account the covariates. Potential features for a classifier may include not only genomic features but also existing clinical predictors. When gene-environment interactions are present, including both types of features in the classifier may significantly improve its predictive accuracy. For such classifiers, the genomic features should demonstrate added value over an optimal classification model using only the clinical predictors (c.f. § 10). Adjusting for such covariates may lead to the selection of features that improve the classifier's validated performance.

5. **General Guidance on Choice of Classification Algorithm(s)**
The choice of classification algorithm(s) is at the discretion of the Analysis Group. It is expected that it would be included in the Statistical Analysis Plan that would be reviewed by a committee of RBWG. A detailed description of the classification algorithm's implementation and model development sufficient to enable independent verification by a separate group must be provided (see Appendix). This would include any transformations of the data which are also expected to be detailed in the Statistical Analysis Plan (c.f. Appendix)
    a. **Commonly used algorithms**: A list of some of the commonly used classification algorithms may be made available upon request to Analysis Groups for reference.
    b. **Use of novel and/or proprietary classification algorithms:** It is anticipated that some AGs may employ novel and/or proprietary classification algorithms.
        i. It is expected that any such algorithm will be made available to other groups for the purpose of independent validation of performance in prospective samples.
        ii. Confidentiality of intellectual property protecting the provider of the algorithm will need to be worked out by the groups that share these algorithms with the advice of the Steering Committee.
    c. **Inclusion of covariate data in classification models:** The use or development of classification algorithms capable of including covariate data as input features are required to conform to the criteria set forth in §7.a-c. It is expected that comparisons will be made to classification models developed strictly on the covariate data.

6. **Performance Measures**
An essential outcome of the MAQC-II project will be the ability to assess the performance of classification models not only in the context of internal cross-validation, but also in prospective validation through the use of independent datasets derived from equivalent tissue, disease or challenged samples or through the use of new samples processed according to an initial experimental design. In order to achieve this outcome, performance measures must be established.
    a. **Selection of Performance measure(s):** The choice of performance measures is at the discretion of the Analysis Group subject to approval by the RBWG Statistical Analysis Plan review committee (c.f. Appendix).
    b. **Accounting for indeterminate or invalid results.** The fraction of samples for which the classifier yields an invalid or indeterminate result is an important performance measure. For example, missing data on an array can hinder some

Nature Biotechnology: doi: 10.1038/nbt.1665

classifiers more than others. Classification algorithms that provide diagnostic confidence of prediction in addition to the prediction itself may include a binary decision in the model development to identify whether each sample is deemed classifiable (regardless of whether the classifier result is valid or not).

c. **Use of novel and/or proprietary performance measures:** It is anticipated that novel and/or proprietary performance measures may be implemented in the MAQC-II. It is expected that the performance measure will be made available for the purpose of independent validation of performance.
d. **Assessment of statistical significance of performance measures:** Statistical measures (such as variance, confidence intervals, etc.) should accompany performance measures.
e. **General Guidance on appropriate use of performance measures.** Care should be given to the interpretation of performance measures in the presence of spectrum bias; i.e., when the data set that is being evaluated does not adequately represent the target population of interest in terms of the disease categories.

7. **Data Partitioning and Methods for <u>Internal</u> Validation**

In gene expression data, the number of genes that can be used in the classifier is often much larger than the number of subjects available for analysis. Therefore, having a classifier that accurately differentiates between classes in the same dataset that are used in the development is by no means guaranteed to be a good classifier in the independent validation phase. Internal validation can be used to weed out some of these ineffective classifiers. This process of internal validation can facilitate a better understanding of the misclassification error for the model.

a. Two types of internal validation are: (i) split-sample (retrospective) validation in which one portion of the data is held out (test dataset) while the classifier is being developed on the remaining data (training dataset) and then is tested on the test dataset. (ii) cross-validation in which the data is repeatedly divided into training and test datasets where the classifier is built on the training set and tested on the test set. This is done many times and assessment of classification error rate is then estimated. (Ref: “Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers”, by R. Simon, JCO, 2005).
b. In the split-sample validation procedure, a method for dividing the original dataset should be provided. One could choose to divide the data using either a random split or a split based on chronology or sites, for example. Note that withholding, for example, only 10% out for a dataset of moderate size to then use as a test set may not be sufficient. A larger test dataset may be required.
c. In the cross validation procedure, it is recommended that the uncertainty of the classification error rate be assessed
d. In both internal validation approaches, it is recommended that all steps used for building the classifier be done on each training separate dataset. The resulting classification method is then tested on each corresponding test dataset.
e. Examples of cross validation approaches include:
    i. Leave-One-Out CV (LOOCV)
    ii. K-fold CV (K varies, e.g. K=2, 3,4….)
    iii. Multiple random cross validations MRCV (partition size varies).

Nature Biotechnology: doi: 10.1038/nbt.1665

f. Alternate methods of partitioning or sampling the dataset (for the purpose of training and testing the classification model) in order to report estimates of performance should be supported by peer-reviewed publication(s) or a detailed description in the Statistical Analysis Plan sufficient for review by the RBWG.

8. Sources and Procedures for the Use and Identification of Independent (External) Validation Data

It is expected that each Analysis Group will identify a mechanism to provide independent validation for any classifier. In most cases this will be through the use of prospectively acquired data.

a. Sources of external validation under consideration include:
   i. Prospective data with the same disease area from the same institution processed by the same lab
   ii. Prospective data with the same disease area form the same institution processed by a different lab with the same kind of microarray
   iii. Prospective data within the same disease area but at different and/or multiple institutions
b. The method for normalization of external data with respect to training datasets is to be specified in the Statistical Analysis Plan (c.f. Appendix)
c. All external (prospective) validation data will be unavailable until the final lock-down of the developed classifier is established and verified by an RBWG committee. Validation data are to be used once and only once for the purpose estimating performance of the classifier; for example, after using the validation data to test classifier performance, the classifier may not be modified and tested again.
   i. Any prospective data will be delivered to all groups simultaneously with the following exception: Analysis groups that have not yet completed the development of their classifier(s) will not receive the validation datasets until development is completed.
   ii. Should any suitable dataset(s) become available for validation after the classifier is built, all appropriate Analysis Groups will be informed and provided with the corresponding microarray files, clinical covariates and quality control metrics.
   iii. If one or more publicly available datasets are used for training and validation of classification models, the dataset that will be used for validation should be specified in the Statistical Analysis Plan.
d. Validation dataset performance may be verified by one or more groups that are independent of the Analysis Group. Each Analysis Group that submits a classifier for validation is expected to be willing to validate not only their classifier but at least one other AG's one as well.
e. An approach that is less preferable is to rely on another independent but retrospective data set for the validation. In such a case it is vitally important that the dataset chosen for validation not be in the public domain. This dataset would be released to the Analysis Group only after the classifier has been built.
f. It is possible but very challenging to try to use a portion of an existing (retrospective) data set for a quasi-independent validation. In such cases, it is

Nature Biotechnology: doi: 10.1038/nbt.1665

absolutely essential that a decision to hold-out data would be made before any analysis of any sort (including normalization, selection of features, etc.) was attempted. This would need to be conveyed to the Regulatory Biostatistics Working Group before the data are electronically transferred or very soon thereafter. In such cases, it is important that when the data are split off for the validation set that a random split not be used. A split based on chronology of the samples or chronology of the processing of the arrays or based on test sites would be recommended to be able to address the criticism that a random split would generate; namely, that the classification model is fitting to the idiosyncrasies of the particular data set and therefore may not be generalizable beyond it.

9. Missing Data

Missing data are common in microarray experiments. The Statistical Analysis Plan should pre-specify a plan for handling missing data both at the development and validation stages of the classifier. A plan for handling missing data needs to anticipate missing data in the training set as well as the validation set, in the clinical as well as microarray data, and in the outcome variable. One approach to handling missing data may be imputation. It is quite likely that a plan for dealing with missing data will be needed at the time of classifier development as well as at validation, depending on the integrity of the dataset under consideration.

a. Impact on performance: If missing data in a sample are such that classifier cannot produce a valid result, then a pre-specified *Intent-To-Diagnose* analysis plan could be used to include the result as a positive or negative, or incorrect test result, whichever is more appropriate with regard to the anticipated management of patients in these cases. The Intention-to-Diagnose Principle dictates that a result cannot be ignored or dropped if the intention was to diagnose that individual. Although an analysis of only evaluable subjects that excludes unsatisfactory results can sometimes be appropriate to handle missing data, it may not fully represent the performance of the classifier. If too many results are invalid, the test may be too impractical for clinical practice.
b. Missing data in the clinical variables: When comparing a genomic signature to existing clinical predictors, a plan for the handling of missing data in the clinical predictors as well as in the genomic features needs to be pre-specified.
c. An assessment of the missing values and the types of missing data should be considered in the Statistical Analysis Plan and it is expected that the SAP will have a pre-specified approach for the missing data. For example, sometimes a missing gene may have clinical importance and imputation of its expression value may not be appropriate. Missing data may be missing for different reasons. The assessment of the missing values and the types of missing data should be made. For example, if the data are not missing at random (MAR), multiple imputations based on MAR assumptions are not appropriate. Sometimes a missing gene may have clinical importance and imputation of its expression value may not be appropriate, e.g., when the missing gene is an existing clinical predictor to which you are comparing the genomic signature.

10. Statistical Inference

Nature Biotechnology: doi: 10.1038/nbt.1665

For a classifier to have clinical utility, a necessary but usually not sufficient requirement is that has some ability to discriminate subjects with the target condition (phenotype) being diagnosed from those without it. A classifier with discriminatory ability is called an *informative classifier*. To take an extreme example, consider classifying subjects according to the toss of a coin that has an 80% chance of turning up heads (i.e., of being test +). Because the chance of testing positive is the same for subjects with the phenotype (80%) as subjects without it (80%), the coin classifier has no ability to discriminate between the two groups of subjects. Note that the coin classifier has sensitivity 80% and specificity 20%, which may appear to be reasonable, but in fact is not: it has no clinical utility because it is completely useless as a discriminator.
The question of whether a classifier is informative is a subtle but fundamentally important measure of performance in the context of clinical utility and will be addressed as a focus of the MAQC-II project. The following subsections provide further context to this issue and should be considered by Analysis Groups as a challenge to the performance of the classification model in terms of clinical utility in a regulatory setting:

a. An important necessary feature of any classifier is that it is informative. Mathematically, a classifier can be shown to be informative for diagnosing a two-state phenotype if one of the following conditions hold: (1) sensitivity + specificity > 1 (this is sometimes expressed as sensitivity (Se) > 1 – specificity (Sp), or True Positive Rate > False Positive Rate), (2) (Positive Predictive Value (PPV) + Negative Predictive Value (NPV) > 1, (3) PPV > prevalence, (4) NPV< 1-prevalence, (5) LR+ > 1, (6) LR – < 1, (7) odds ratio > 1, or (8) for classifiers dichotomizing a semi-quantitative or continuous variable, the area under the ROC curve (AUC) > 0.5. Here LR+ = Se/(1-Sp) is the positive likelihood ratio, LR– = (1-Se)/Sp is the negative likelihood ratio, odds ratio = LR+ / LR– is the odds of testing positive in subjects with the phenotype over the odds of testing positive in subjects without it. Furthermore, since different statistical inferences are associated with each of the above conditions, it is absolutely crucial that the particular criterion be identified in the Statistical Analysis Plan, before the data are analyzed.
b. Accurate but non-informative classification models: A common misconception that has circulated in the microarray class prediction literature is that a classifier is informative if its predictive accuracy (probability of correct classification) is greater than 50%. On the contrary, non-informative classifiers can have predictive accuracy greater than 50% and informative classifiers can have predictive accuracy less than 50%. The reason is that predictive accuracy depends on the prevalence of the phenotype that is being diagnosed. As an example of the former, a classifier with sensitivity 80% and specificity 20% for diagnosing a phenotype with prevalence 90% has predictive accuracy of 74%, yet is useless (e.g., does not meet condition (1) above). As an example of the latter, a classifier with sensitivity 80% and specificity 40% for a phenotype with prevalence 10% has predictive accuracy of 44%, yet is informative (e.g., meets condition (1) above). Therefore, predictive accuracy alone is not an adequate evaluation of classifier performance.
c. A classifier can be shown to be informative for a condition (e.g., in an external validation study) with a hypothesis test or confidence interval. For example, to

Nature Biotechnology: doi: 10.1038/nbt.1665

show statement (2) above, a statistical analysis needs to demonstrate that, statistically, sensitivity is significantly greater than one minus specificity.

d. To have clinical utility, a genomic classifier should provide added value to existing clinical predictors. A genomic signature has added value if either (i) it is superior to existing clinical predictors of the phenotype or (ii) the combination of it with the clinical predictors is superior the clinical predictors alone. Again, it is expected that the Statistical Analysis Plan will clearly indicate which demonstration of added value, (i) or (ii), is the approach of the investigators.
e. The ability to show added value over the clinical predictors may depend on the quality of the model built using the clinical information alone. Analysis Groups for the same dataset are encouraged to compare their models based on the clinical information alone.
f. Conditions for which a diagnostic test is superior to another diagnostic test are given in Biggerstaff ("Comparing diagnostic tests: A simple graphic using likelihood ratios", *Statistics in Medicine*, 2000, 19: 649-663).
g. The particular test statistic that is to be used to demonstrate added value needs to be identified. In some cases statistical inference based on logistic regression may be appropriate for demonstrating added value of a genomic signature to clinical predictors. For classifiers based on an underlying semi-quantitative or continuous variable, ROC regression or other regression techniques may be appropriate for demonstrating added value.

11. **Performance Variability Measures**

It is important that the variability of the predictive model be assessed. One reason is that it is not sufficient merely to establish numerical superiority (as for example that sensitivity is numerically greater than one minus specificity or that the area under the ROC curve is numerically greater than 0.5). The statistical challenge is to show that for the performance measure that has been selected a priori, the claims can be demonstrated statistically taking the variation into account.

a. Variability of future predictions: Is enough understood about variation that the variability associated with a single future unit (patient) can be well characterized? There are in fact many sources of variability: for example, from variability of the arrays to the choice of normalization, the choice of features, the choice of how many and which units to use for the training set, the obvious variability associated with the size and choice of the test set. Some might propose a strategy that after the test set is used once, then the training and test sets are recombined and a different training and test set are selected. There is a danger here that one may not be able to sufficiently reproduce all the steps in the new training set, as observed in terms of the lack of reproducibility of bootstrapping by R. Simon (ref.)
b. Estimators of variance: The existence of an unbiased estimator of the variance for cross-validation is still the focus of debate. Consequently this should be considered as an active topic under consideration for the MAQC-II. Alternate methods for estimation of the variance of performance measures should include (but not be limited to) generalized or smoothed cross validation and the use of the delta-d-jack-knife method as an alternative to fold cross-validation.

Nature Biotechnology: doi: 10.1038/nbt.1665

12. **Statistical Analysis Plan**
    It is expected that each Analysis Group will prepare and submit a Statistical Analysis Plan of how that group intends to analyze the data. It is recommended that this will be submitted before any analysis commences. This Statistical Analysis Plan will be reviewed by a committee of the Regulatory Biostatistics Working Group (RBWG) that does not include any statisticians on the AG. In particular, a specific Statistical Analysis Plan describing methodology for each AG will be provided to the RBWG for review prior to the validation of classifier(s). An Analysis Group runs the very real risk that any analyses performed on data based on an inadequate or flawed Statistical Analysis Plan are unlikely to be accepted for comparison or publication with the MAQC-II results. It is expected that the Statistical Analysis Plan is will include the Checklist in Appendix 1.

13. **Revisions**
    Any revisions to the Statistical Analysis Plan should be documented and submitted to the committee of the RBWG. No validation set will be released to any Analysis Group until the Statistical Analysis Plan has been submitted and reviewed by a committee of the Regulatory Biostatistics Working Group.

14. **Documentation**
    It is expected that each AG will keep the equivalent of a laboratory notebook detailing on a regular basis as it happens what procedures were followed, how the features and the classifiers were selected, and how they were evaluated. In addition, at the time of manuscript preparation, each AG should be prepared to provide this equivalent of a laboratory notebook to the RBWG committee. The reason for this is to be able to provide assurance that good scientific method has been followed (it also has the benefit of recording how the AG evolved in its thinking and arrived at its final model). This is something that most bench scientists take for granted and we should expect no less in the MAQC project.

Nature Biotechnology: doi: 10.1038/nbt.1665

Appendix: Statistical Analysis Checklist

This checklist is to be submitted with each Statistical Analysis Plan or a Revision to it.

Check the appropriate line. If YES provide details when applicable.

	YES	NO
a. A complete list of all members included in the AG		
b. Objectives and specific methods	_____	_____
c. A brief description of both dataset(s) and the experimental design from which the data were generated. This includes list of all assays, the sites (clinical and laboratory), the time and site at which the specimen was obtained, the time and site at which the assay was performed, inclusion/exclusion criteria, and an accounting of all specimens (including specimens for which deviations from the SAP occurred, specimens with missing results by reason, specimens with invalid results by reason, etc.)	_____	_____
d. Spectrum bias evaluation, e.g., a comparison of the dataset and the target population of interest on the distribution patient characteristics	_____	_____
e. Specific discretization procedures of continuous outcome (if any)	_____	_____
f. Definition of class labels	_____	_____
g. Reasons for sample exclusion (if applicable)	_____	_____
h. Quality control assessed	_____	_____
i. Reproducibility of arrays	_____	_____
j. Description of normalization method to be used (if any)	_____	_____
k. Data transformation (if any)	_____	_____
l. Description of missing data and any imputation (if applicable)	_____	_____
m. A plan for the development (training) of classifier(s)	_____	_____
n. Classification method(s) to be considered:		
i. Criteria for feature selection (if any)	_____	_____
ii. Model development	_____	_____
iii. Classification rule	_____	_____
o. A validation plan for developed classifier(s):		
i. Prospective validation	_____	_____
ii. Retrospective external validation (training-test split sample)	_____	_____

Nature Biotechnology: doi: 10.1038/nbt.1665

a. Chronological split: sample processing (preferred) _____ _____
b. Chronological split: sample collection (preferred) _____ _____
c. Split by sample processing site (preferred) _____ _____
d. Split by sample collection site (preferred) _____ _____
e. Other non-random split (preferred) _____ _____
f. Random split (not recommended) _____ _____

iii. Internal cross-validation
a. Feature selection _____ _____
b. Model development _____ _____
c. Classification rule _____ _____
d. Missing data imputation (if applicable) _____ _____

p. Choice (and methods) for selection of partitioning for internal and hold-out validation dataset(s). This should include methods for sample size and should be made before any analysis is conducted _____ _____
q. A complete list of algorithms involved in development of classification model (including feature selection, normalization, missing data imputation, and model selection) _____ _____
r. A list of statistical hypotheses to be tested at validation stages _____ _____
s. A list of endpoints in validation phases (e.g., ROC AUC, Se, Sp, PPV, NPV, LR+, LR –, and time-to-event) _____ _____
t. Identification of the test statistic for the inference _____ _____
u. A list of clinical covariates and their distributions _____ _____
v. A list of clinical predictors to compare against genomic classifier(s) or to be included as part of classifier(s) _____ _____
w. Analysis plan for testing if genomic classifier(s) adds significant value over clinical predictors _____ _____
x. Description and source of all code and random number seed(s) included _____ _____

i. Others _____ _____
ii. Comments

Nature Biotechnology: doi: 10.1038/nbt.1665