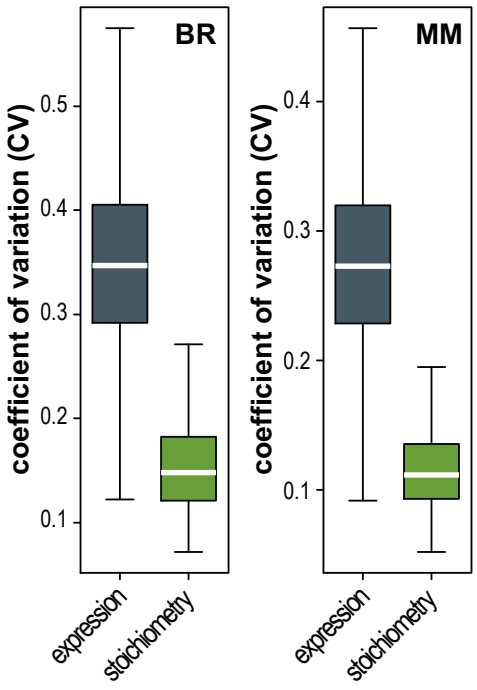
ANNEXE I. Predictive Gene Model based on specific set of Stoichiometry Signatures, and on data in open access (PGMSS open access)

Figure 1. Stoichiometry signatures reduce noise in gene expression feature space. One of a key problems of differential gene expression analysis stems from the noise as illustrated by a boxplot of coefficients of variations (CV) of library size normalized genes' transcripts copy-numbers - FPKM (Fragments Per Kilobase of transcript per Million mapped reads) (grey boxplots)[1]. To circumvent this, we developed a normalization free approach based on estimations of RNAs stoichiometries to generate noise-reduced model feature space. Estimates of gene expression stoichiometry signatures exhibit significantly lower (~2 times) intraindividual variability (green boxplots), which, in turn, facilitates training of predictive or diagnostic models. Data for breast cancer (BR) and multiple myeloma (MM) microarray profiles were taken from (Hess et al., 2006; Zhan et al., 2006).



[1] RNA-sequencing experiments suggest a quadratic mean-variance relationship for RNA copy-number:

[1] $\sigma_g^2 = \mu_g + \alpha_g \mu_g^2$,

where μ_g and σ_g^2 are mean and variance respectively of RNA counts or FPKMs for a given gene (g), and α_g is overdispersion coefficient. Squared coefficient of variation of RNA copy number is then:

[2] $cv_g^2 = \sigma_g^2/\mu_g^2 = \mu_g^{-1} + \alpha_g$,

and, for large values of μ_g ($\mu_g \gg 1$): $cv_g^2 \approx \alpha_g$.

α_g is also referred to as biological coefficient of variation (bcv_g^2) or, here, as gene noise. Variance stabilizing log transformation of RNA counts of FPKMs does not eliminate nor reduce gene noise. This follows from the Taylor expansion of the variance of log (here, natural log - *ln*) transformed variables:

[3] $\sigma_{\ln(g)}^2 \approx \sigma_g^2/\mu_g^2 + \sigma_g^4/(2\mu_g^4) = cv_g^2 + cv_g^4/2 \approx cv_g^2$,

and, for $\mu_g \gg 1$: $\sigma_{\ln(g)}^2 \approx bcv_g^2$.

This implies that feature space (a matrix of independent variables) represented by, for example, log transformed FPKMs and used for training of predictive or diagnostic models is inherently noisy. To that, the presence of biological noise in gene expression complicates the transfer of models trained on RNA-sequencing or microarray data to real-time quantitative PCR (RT-qPCR) or digital PCR (dPCR) platforms. The reason for this is that PCR always requires normalization of a target gene (g) to a control gene (n), which only increases the noise as:

[4] $\sigma_{\ln(g/n)}^2 = \sigma_{\ln(g)}^2 + \sigma_{\ln(n)}^2 \approx bcv_g^2 + bcv_n^2$.

Normalization-free estimates of RNAs stoichiometries a) reduce noise in models' features and b) make it, in principle, possible to utilize models trained on genome-wide transcriptomics data with other platforms, such as microarray or PCR.

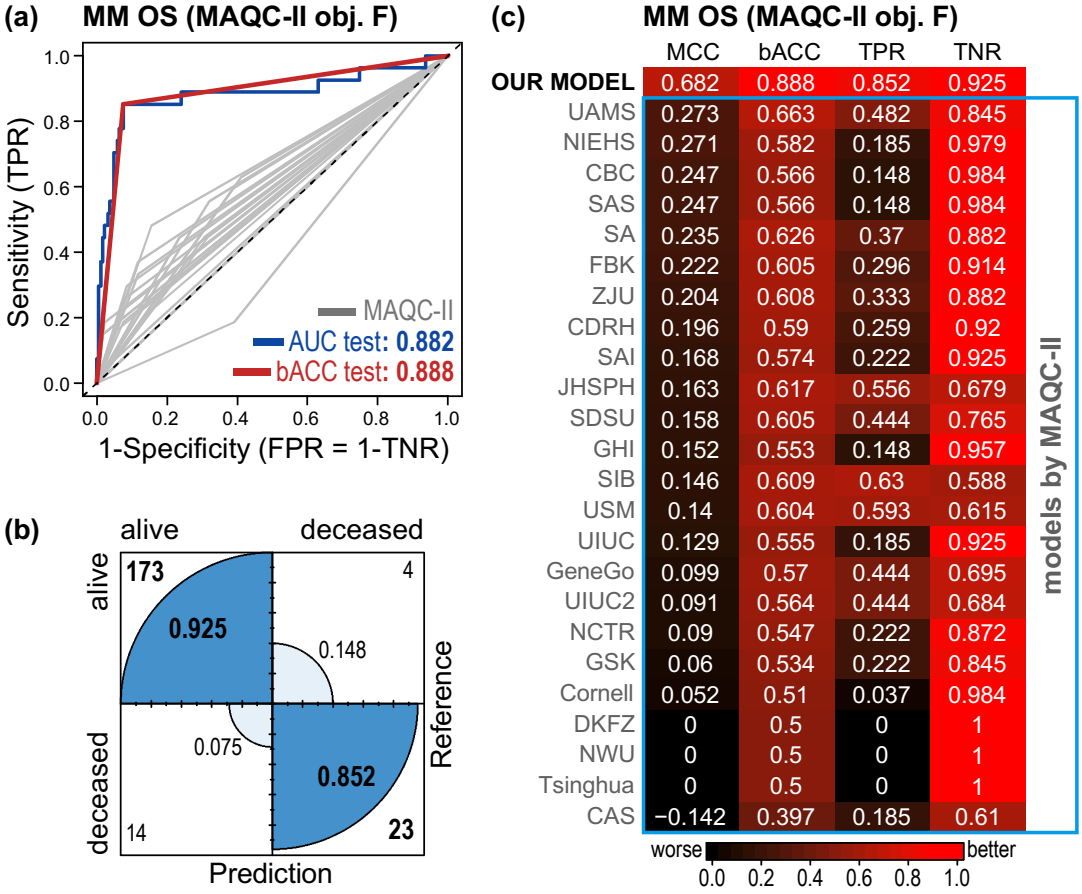


Figure 2. Comparison of our modelling approach using stoichiometry signatures with models developed by MicroArray Quality Control (MAQC)-II study (Shi et al., 2010). Multiple myeloma example, prediction of overall survival (OS).

(a) Receiver Operating Characteristic curves (ROC) curves for the stoichiometry signature model predicting post-treatment overall survival (OS) after 730 days for multiple myeloma patients (MM) (Zhan et al., 2006). The model has been trained and independently validated on the same train and test cohorts as in the MAQC-II study. ROC (blue line) and binary ROC (red line) curves are shown for the validation cohort. AUC (area under the ROC curve) and balanced accuracy (bACC, or binary AUC) are indicated for the test cohort. Binary ROC curves for models developed by MAQC-II participants are shown in grey.

(b) Model confusion matrix for replication (test) cohort. True negative (top-left), true positive (bottom-right) rates are indicated within circle along with the numbers of correctly and mis- classified individuals.

(c) The model summary statistics is given for validation cohort: MCC - Matthews correlation coefficient[2], bACC – balanced accuracy (or binary AUC)[3], TPR – true positive rate (sensitivity) and TNR – true negative rate (specificity). Summary statistics for MAQC-II models are outlined and our

[2] $MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$, where TP – true positive, TN – true negative, FP – false positive, FN – false negative.

[3] $bACC = \frac{TPR+TPN}{2}$, where $TPR = \frac{TP}{TP+FN}$ – true positive rate (sensitivity) and $TNR = \frac{TN}{TN+FP}$ – true negative rate (specificity). bACC also equals to binary AUC.

model is shown on the top of the table. For our model we used the same training and validation cohorts as in MAQC-II. Colours correspond to low (black) and high (red) model accuracy.

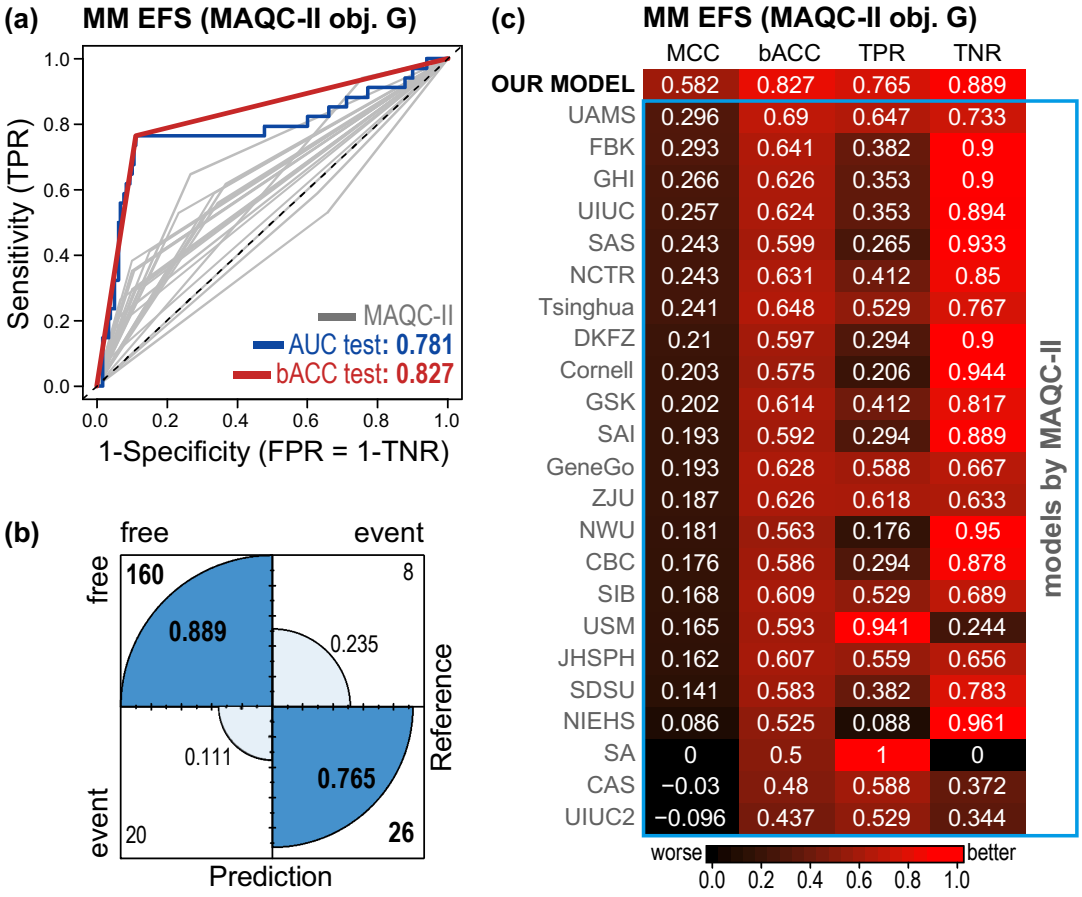


Figure 3. Prediction of event free survival (EFS) for multiple myeloma patients.

(a) ROC curves for the stoichiometry signature model predicting post-treatment event free survival (-EFS) for multiple myeloma patients (MM) (Zhan et al., 2006). The model has been trained and independently validated on the same train and test cohorts as in the MAQC-II study. ROC (blue line) and binary ROC (red line) curves are shown for the validation cohort. AUC (area under the ROC curve) and balanced accuracy (bACC, or binary AUC) are indicated for the test cohort. Binary ROC curves for models developed by MAQC-II participants are shown in grey.

(b) Model confusion matrix for replication (test) cohort. True negative (top-left), true positive (bottom-right) rates are indicated within circle along with the numbers of correctly and mis-classified individuals.

(c) The model summary statistics is given for the validation cohort. Summary statistics for MAQC-II models are outlined and our model is shown on the top of the table. For our model we used the same training and validation cohorts as in MAQC-II. Colours correspond to low (black) and high (red) model accuracy.

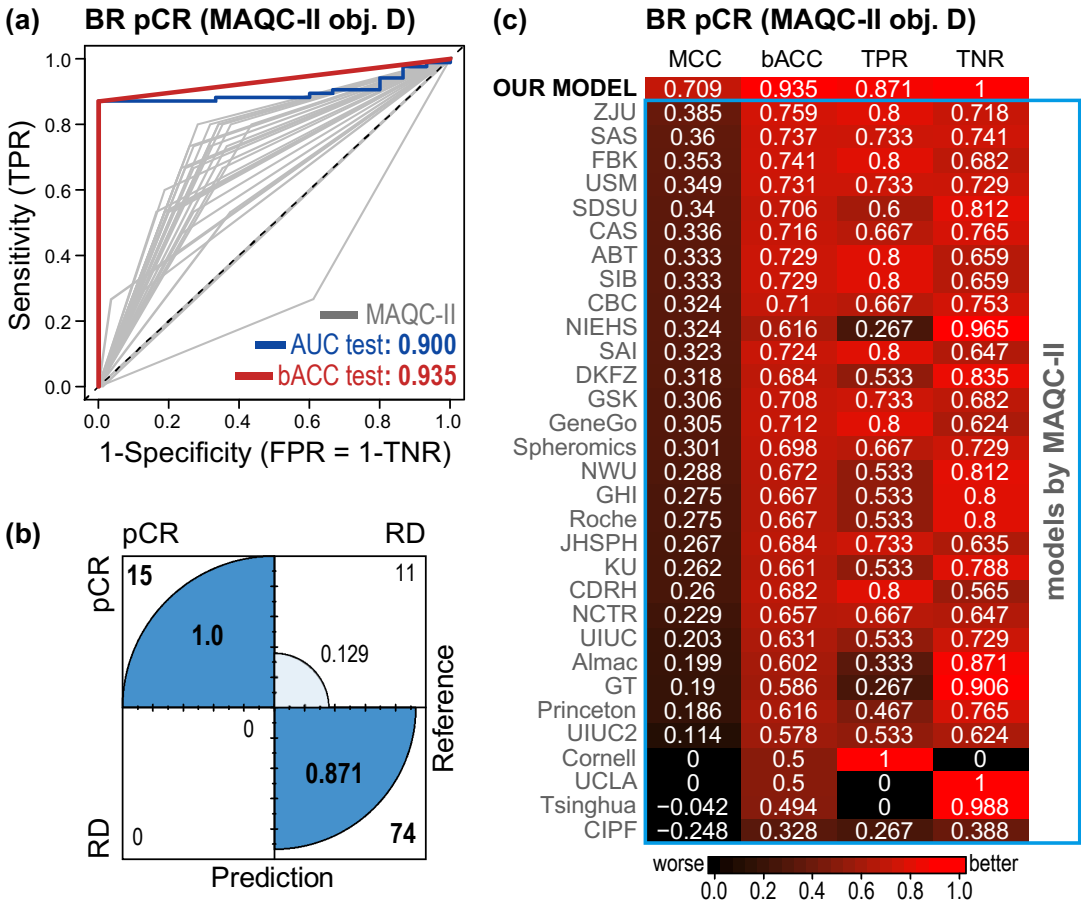


Figure 4. Comparison of our modelling approach using stoichiometry signatures with models developed by MicroArray Quality Control (MAQC)-II study (Shi et al., 2010). Breast cancer example, prediction of pathologic complete response (pCR).

(a) ROC curves for the stoichiometry signature model predicting pathologic complete response (pCR) or residual disease (RD) after preoperative chemotherapy in patients with stage I-III breast cancer (BR) (Hess et al., 2006). The model has been trained and independently validated on the same train and test cohorts as in the MAQC-II study. ROC (blue line) and binary ROC (red line) curves are shown for the validation cohort. AUC (area under the ROC curve) and balanced accuracy (bACC, or binary AUC) are indicated for the test cohort. Binary ROC curves for models developed by MAQC-II participants are shown in grey.

(b) Model confusion matrix for replication (test) cohort. True negative (top-left), true positive (bottom-right) rates are indicated within circle along with the numbers of correctly and mis-classified individuals.

(c) The model summary statistics is given for the validation cohort. Summary statistics for MAQC-II models are outlined and our model is shown on the top of the table. For our model we used the same training and validation cohorts as in MAQC-II. Colours correspond to low (black) and high (red) model accuracy.

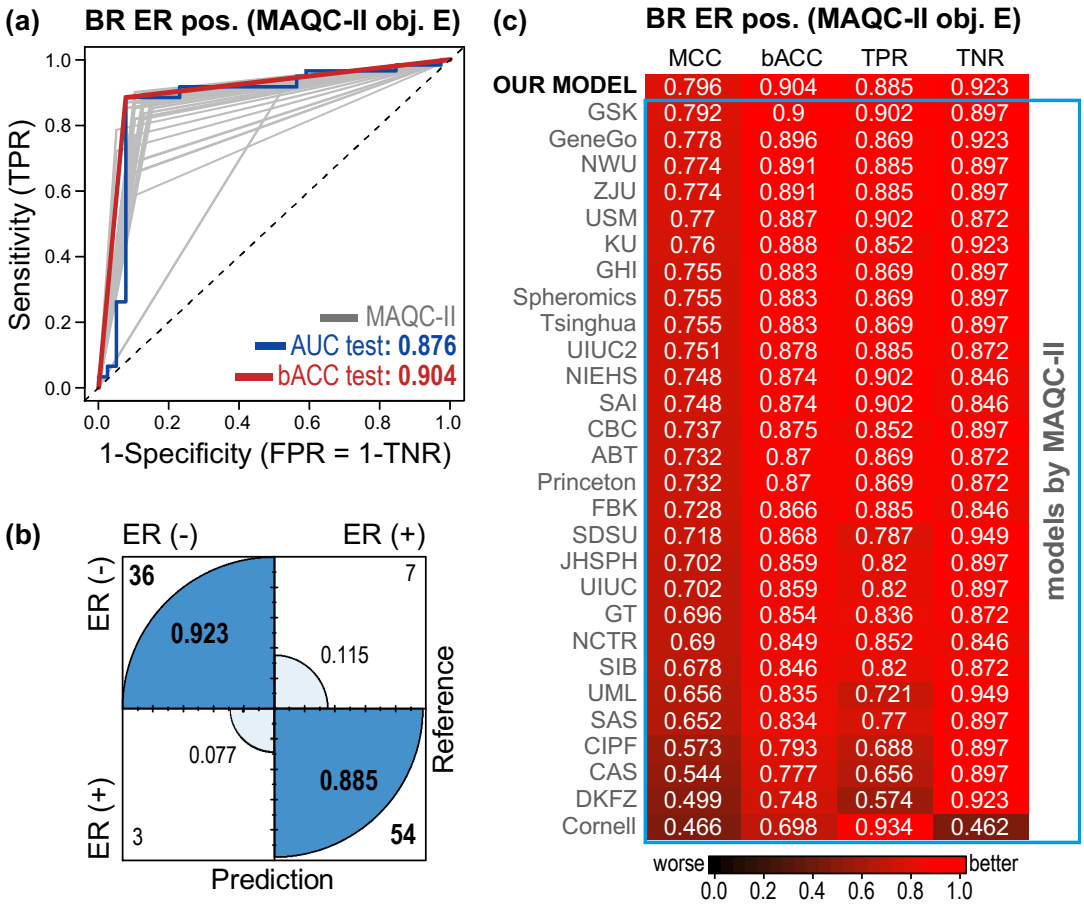


Figure 5. Prediction of estrogen receptor status (ER positive/negative) in patients with stage I-III breast cancer.

(a) ROC curves for the stoichiometry signature model predicting estrogen receptor status (ER pos./neg.) in patients with stage I-III breast cancer (Hess et al., 2006). The model has been trained and independently validated on the same train and test cohorts as in the MAQC-II study. ROC (blue line) and binary ROC (red line) curves are shown for the validation cohort. AUC (area under the ROC curve) and balanced accuracy (bACC, or binary AUC) are indicated for the test cohort. Binary ROC curves for models developed by MAQC-II participants are shown in grey.

(b) Model confusion matrix for replication (test) cohort. True negative (top-left), true positive (bottom-right) rates are indicated within circle along with the numbers of correctly and mis-classified individuals.

(c) The model summary statistics is given for the validation cohort. Summary statistics for MAQC-II models are outlined and our model is shown on the top of the table. For our model we used the same training and validation cohorts as in MAQC-II. Colours correspond to low (black) and high (red) model accuracy.

Table listing all data science teams participated in MicroArray Quality Control (MAQC)-II study (Shi et al., 2010).

Org. Code	Organization Name
ABT	Abbott Laboratories
Almac	Almac Diagnostics, UK
CAS	Chinese Academy of Sciences, China
CBC	CapitalBio Corporation, China
CDRH	Center for Devices and Radiological Health, FDA
CIPF	Centro de Investigacion Principe Felipe, Spain
Cornell	Weill Medical College of Cornell University
DKFZ	German Cancer Research Center, Germany
EPA	U.S. Environmental Protection Agency
FBK	Fondazione Bruno Kessler, Italy
GeneGo	GeneGo Inc.
GHI	Golden Helix Inc.
GSK	GlaxoSmithKline
GT	Georgia Institute of Technology – Emory University
JHSPH	Johns Hopkins Bloomberg School of Public Health
KU	University of Kansas
Ligand	Ligand Pharmaceuticals
NCTR	National Center for Toxicological Research, FDA
NIEHS	National Institute of Environmental Health Sciences
NWU	Northwestern University
Princeton	Princeton University
Roche	Roche Palo Alto LLC
SA	SABioscience Corporation
SAI	Systems Analytics Inc.
SAS	SAS Institute Inc.
SDSU	South Dakota State University
SIB	Swiss Institute of Bioinformatics, Switzerland
Spheromics	Spheromics, Finland; University of Umeå, Sweden
Tsinghua	Tsinghua University, China
UAMS	University of Arkansas for Medical Sciences
UCLA	Cedars-Sinai Medical Center of UCLA
UIUC	University of Illinois at Urbana-Champaign
UIUC2	University of Illinois at Urbana-Champaign
UML	University of Massachusetts Lowell
USM	University of Southern Mississippi
ZJU	Zhejiang University, China

References

Hess, K.R., Anderson, K., Symmans, W.F., Valero, V., Ibrahim, N., Mejia, J.A., Booser, D., Theriault, R.L., Buzdar, A.U., Dempsey, P.J.*, et al.* (2006). Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. Journal of clinical oncology : official journal of the American Society of Clinical Oncology *24*, 4236-4244.

Shi, L., Campbell, G., Jones, W.D., Campagne, F., Wen, Z., Walker, S.J., Su, Z., Chu, T.M., Goodsaid, F.M., Pusztai, L.*, et al.* (2010). The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nature biotechnology *28*, 827-838.

Zhan, F., Huang, Y., Colla, S., Stewart, J.P., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B.*, et al.* (2006). The molecular classification of multiple myeloma. Blood *108*, 2020-2028.