Meta-analysis of gene expression and survival data using the R packages *SurvComp* and *genefu*

Markus Schröder[1,2], Daniel Gusenleitner[1], Alexander Goesmann[2], Aedín C. Culhane[1], John Quackenbush[1] and Benjamin Haibe-Kains [1]

[1]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, USA
[2]Bioinformatics Research Facility (BRF), Centrum for Biotechnology (CeBiTec), University of Bielefeld, Bielefeld, Germany

Introduction to *SurvComp*

Gene expression profiling has generated unprecedented insight into our molecular understanding of cancer. In breast cancer, gene expression profiling studies have been widely employed and have not only advanced our understanding of disease, but have provided multi-gene predictive and prognostic tests including Oncotype DX, mammaprint, Veridex GGI and the Breast BioClassifier for breast cancer molecular subtypes.

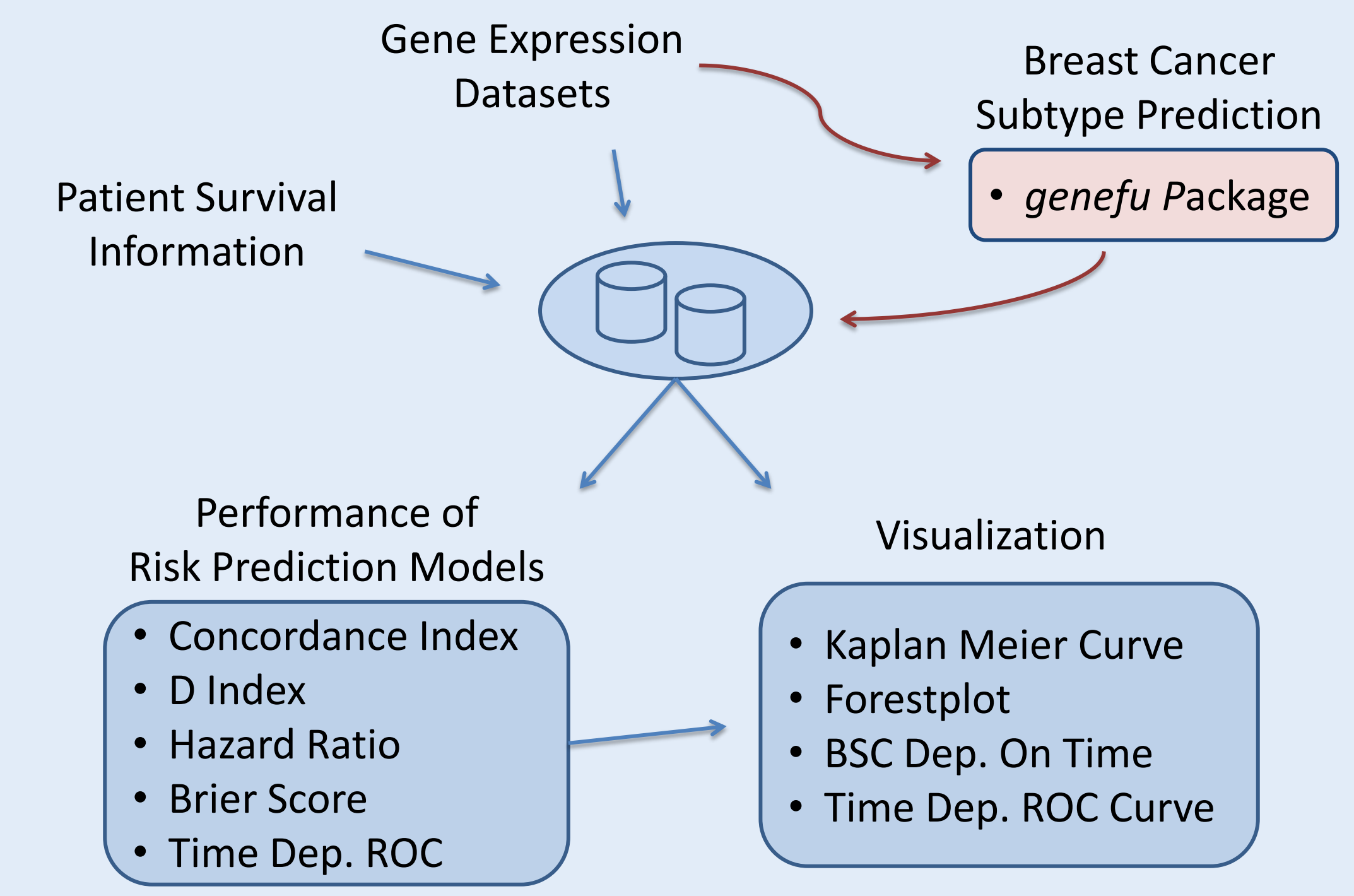
To identify new prognostic genes and gene signatures, several risk prediction models have been introduced recently.

Our SurvComp package is providing functions to assess and to statistically compare the performance of these risk prediction (survival) models. It includes:

i. Implementation of state-of-the-art statistics developed to measure the performance of risk prediction models
ii. Combining these statistics estimated from multiple datasets using a meta-analytical framework
iii. Functions to visualize those measurements in a clear and compact way

SurvComp is available on Bioconductor.org

Survival Analysis Workflow



Gene Expression Datasets

Breast Cancer Subtype Prediction

- *genefu* Package

Patient Survival Information

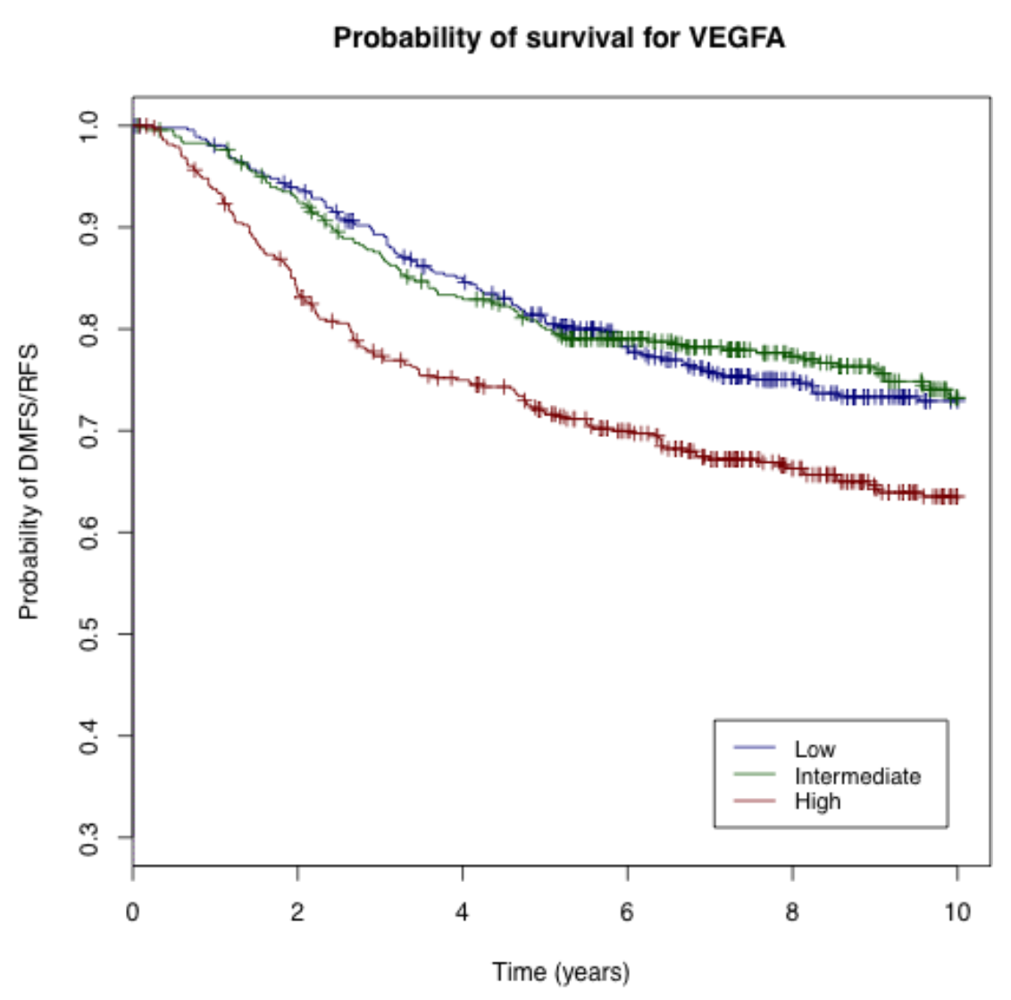
Performance of Risk Prediction Models

- Concordance Index
- D Index
- Hazard Ratio
- Brier Score
- Time Dep. ROC

Visualization

- Kaplan Meier Curve
- Forestplot
- BSC Dep. On Time
- Time Dep. ROC Curve

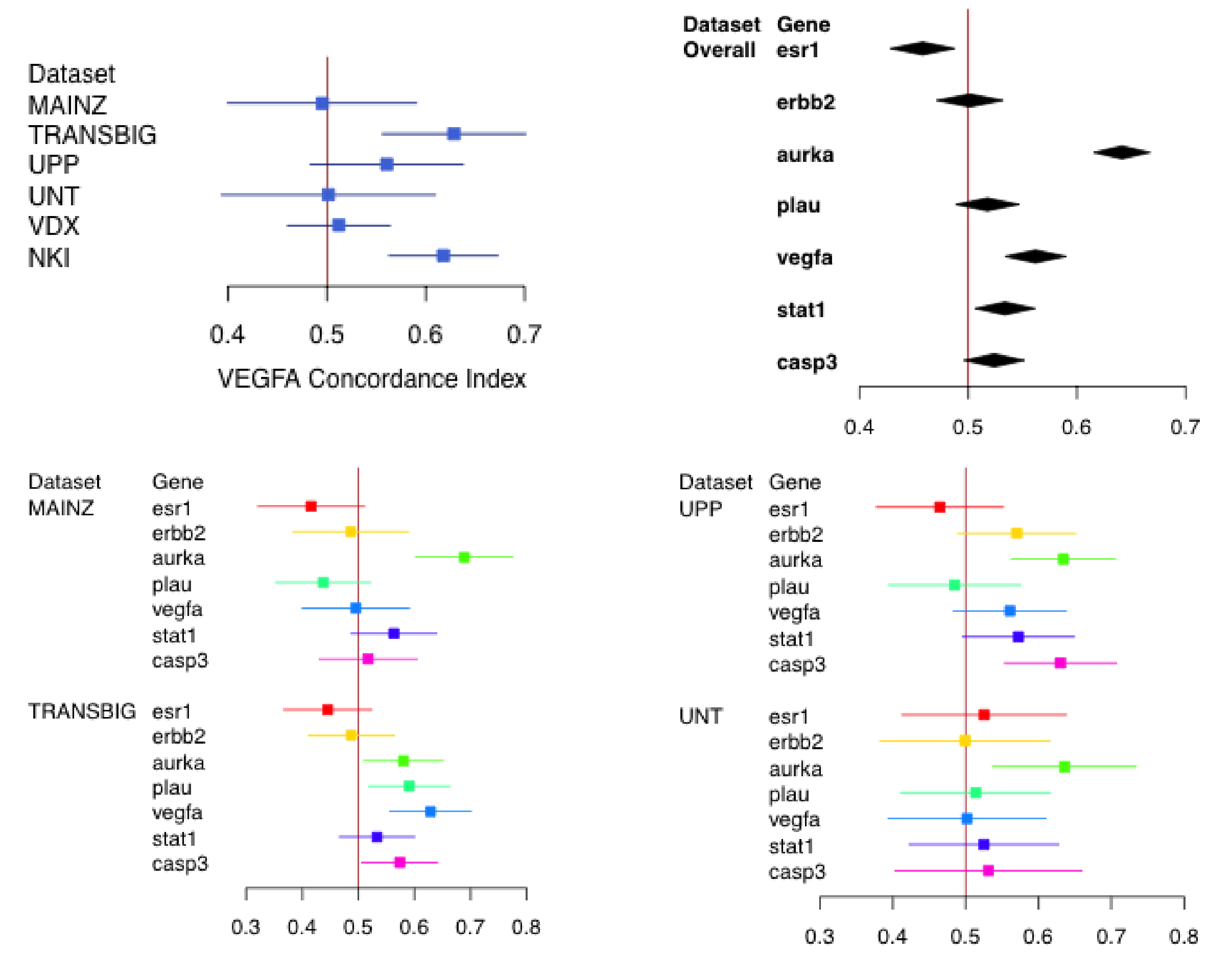
Kaplan Meier Survival Curves



The Kaplan-Meier curve: estimation of the survival expectancy for a group of patients with respect to time

Gene expression values are split using quantiles at 33% and 66%, leaving three groups: low (under 33%), intermediate (between 33% and 66%) and high (over 66%)

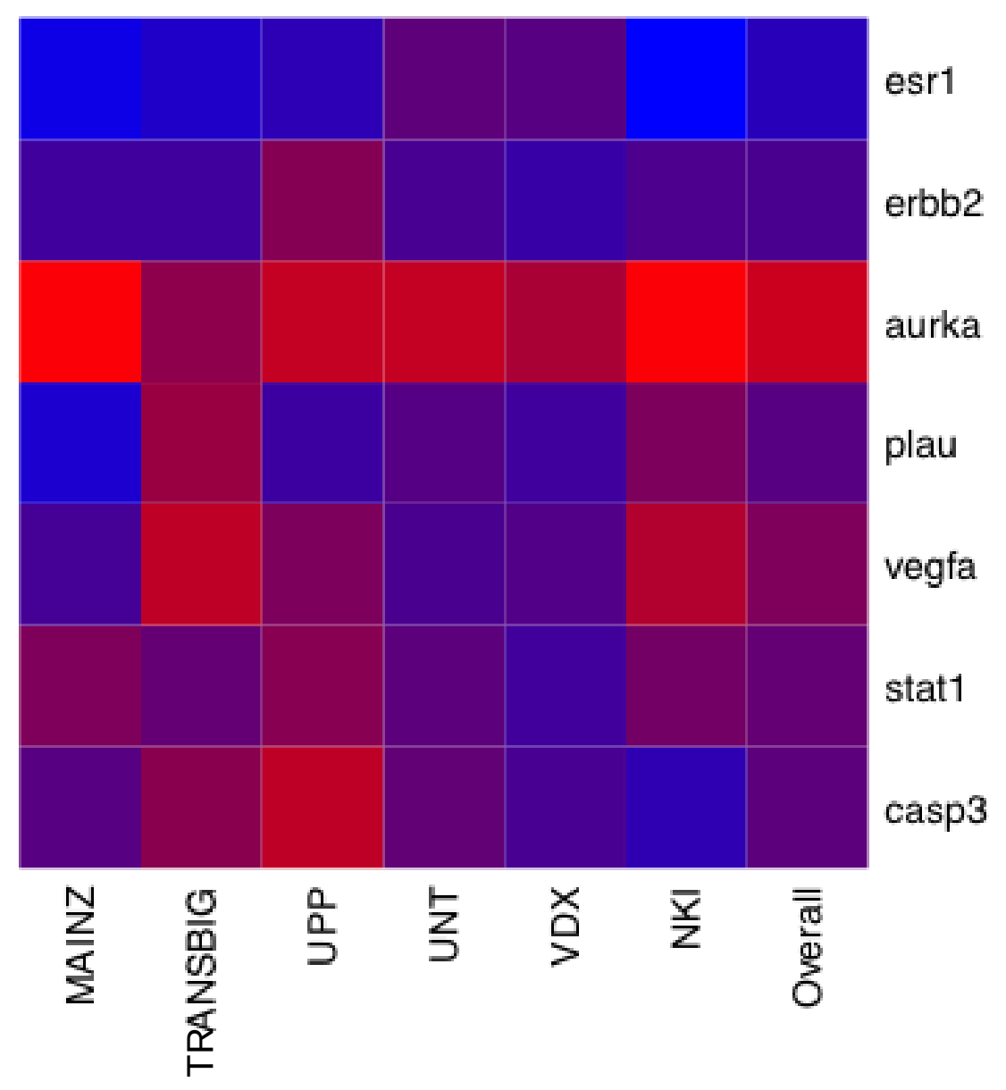
Forestplots of Concordance Indices

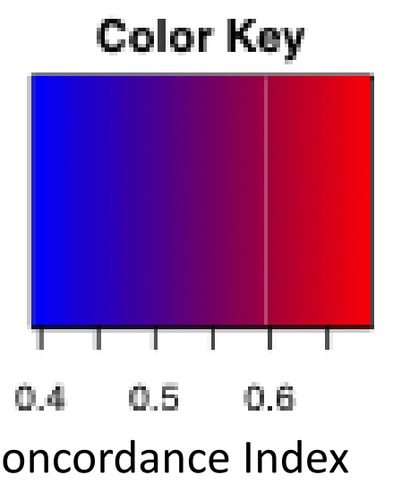


Bottom: Concordance indices for seven genes (introduced by Desmedt et al. 2008) in different breast cancer datasets. The genes are ESR1 (representing the ER signaling), ERBB2 (HER2 signaling), AURKA (proliferation), PLAU (tumor invasion/metastasis), VEGFA (angiogenesis), STAT1 (immune response) and CASP3 (apoptosis phenotypes).
Top left: Concordance indices for the gene VEGFA from six datasets. The six datasets include five Affymetrix platforms (HGU133A and/or B) and one Agilent platform (rosetta).
Top right: Concordance indices for seven genes, using six datasets, combined to one estimation for each gene.

The Forestplot: plot of estimations with their 95% confidence interval and informations about these estimations.

Heatmap of Concordance Indices

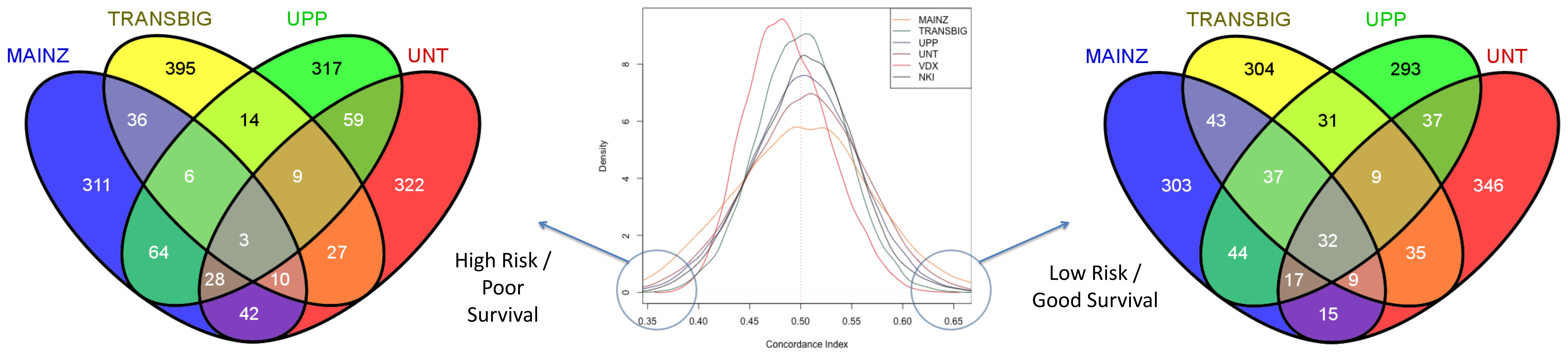




Data contains seven genes (rows) and six datasets (columns). The concordance indices in the heatplot are represented as blue (low value) and red (high value). AURKA stands out from the remaining six genes with high concordance indices, wheras ESR1 stands out with low concordance indices.

The D Index: an estimate of the log hazard ratio comparing two equal-sized prognostic groups. This is a natural measure of separation between two independent survival distributions under the proportional hazards assumption.
The Hazard Ratio: estimation of relative risk, a ratio of the probability of the event occuring in an exposed group versus a non exposed group.

Genome Scale Overview of Concordance Indices from Multiple Datasets



High Risk / Poor Survival

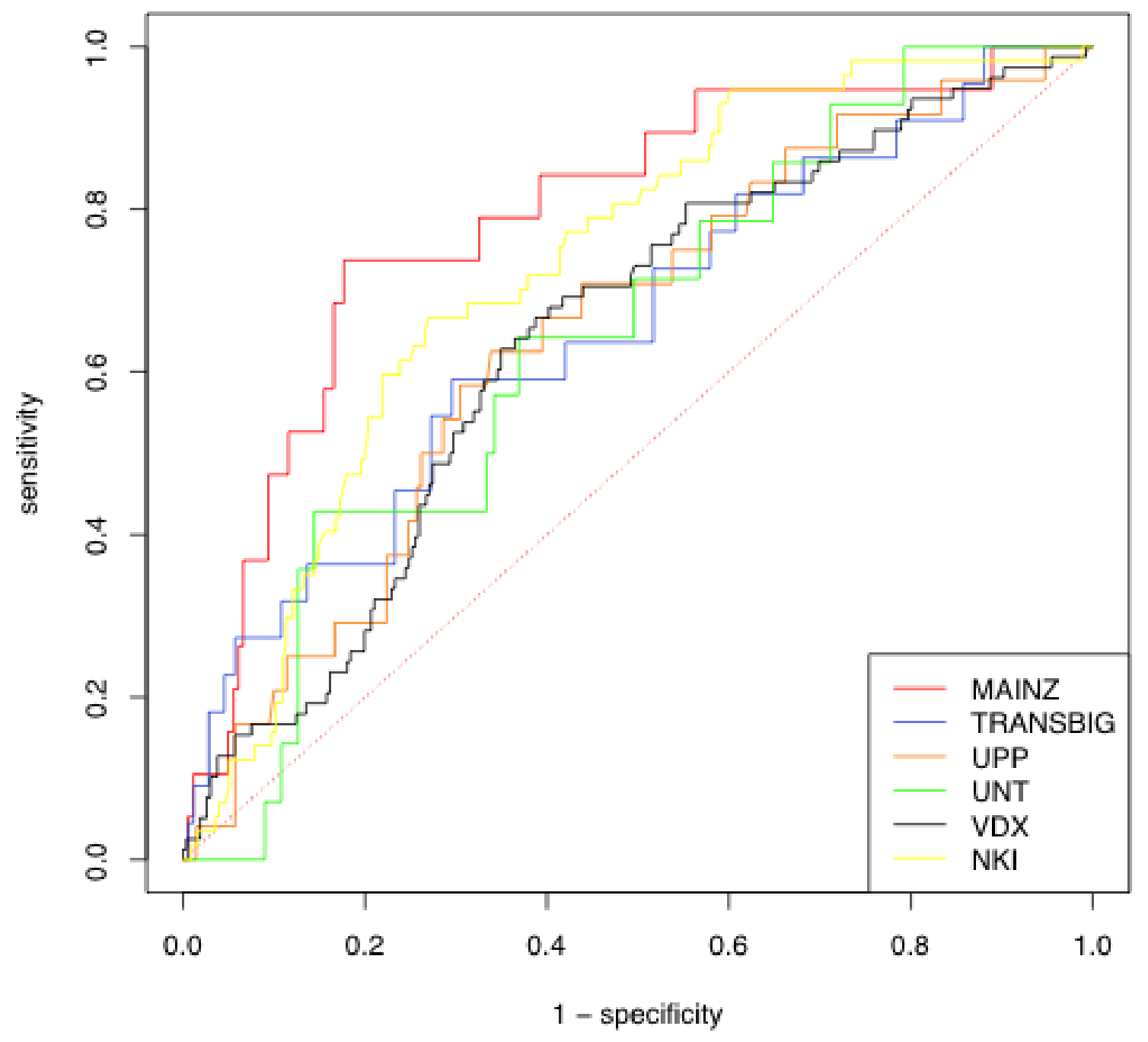
Low Risk / Good Survival

Representations of the 500 lowest (**left**) and 500 highest (**right**) concordance indices from 4 different datasets. High concordance indices indicate a good survival / low risk for patients, low concordance indices indicate poor survival / high risk for patients.

In the **middle** figure, the concordance indices from six different datasets are shown as a genome scale density plot, providing a overview of the concordance index distribution in each dataset.

The Concordance Index: probability that, for a pair of randomly chosen comparable samples, the patient with the higher risk prediction will experience an event before the other patient.

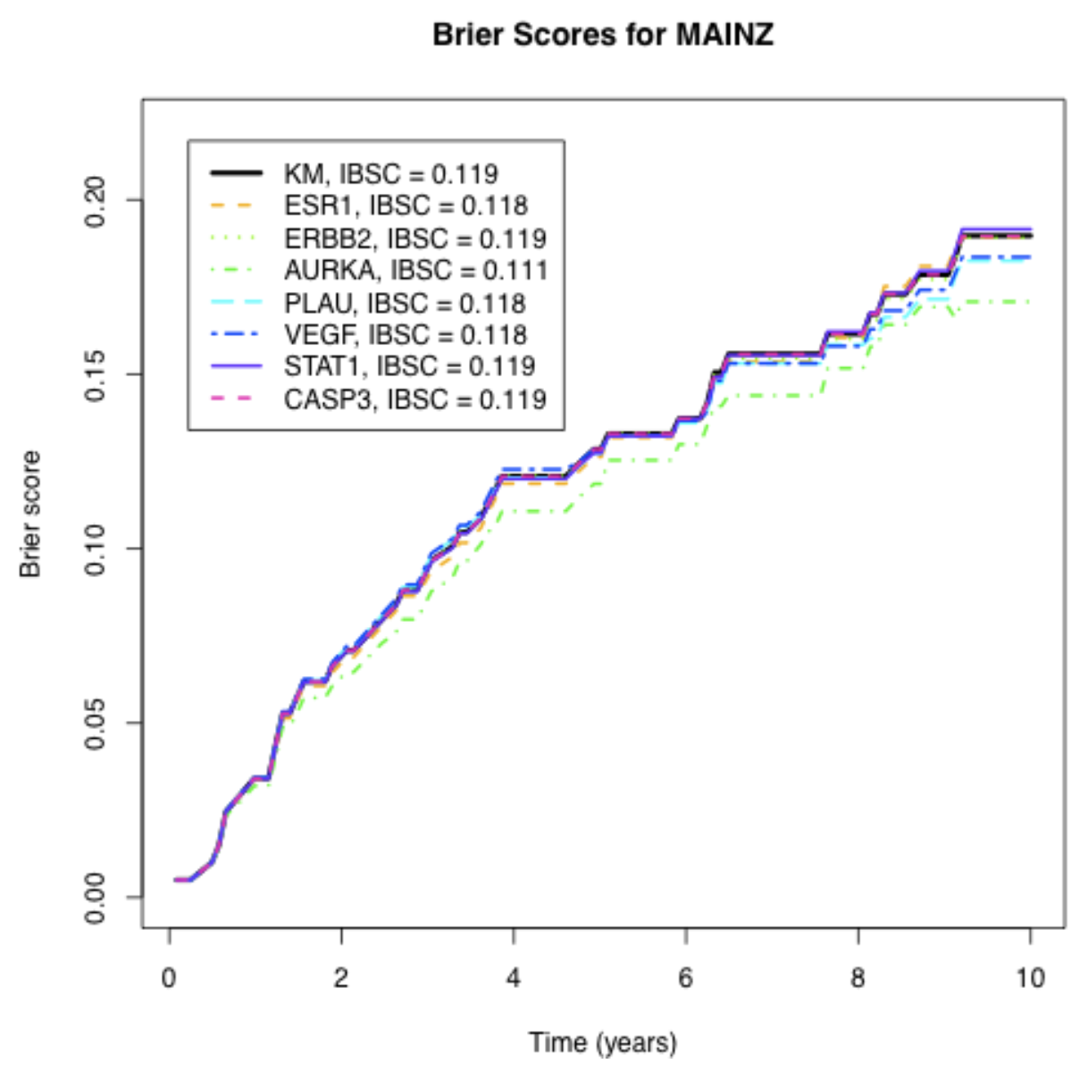
Time Dependent ROC Curve for AURKA at Three Years



Time dependent ROC curve for the gene AURKA (related to proliferation) after three years in six different datasets.

The time-dependend receiver operating characteristic curve: a plot of sensitivity versus 1-specificity for all the possible cutoff values of the continous variable as estimated at a specific time point.

Brier Score Depending On Time



Brier Score: measures the accuracy of a set of probability assessments. It measures the average squared deviation between predicted probabilities for a set of events and their outcomes. Lower Brier Scores represent higher accuracy.

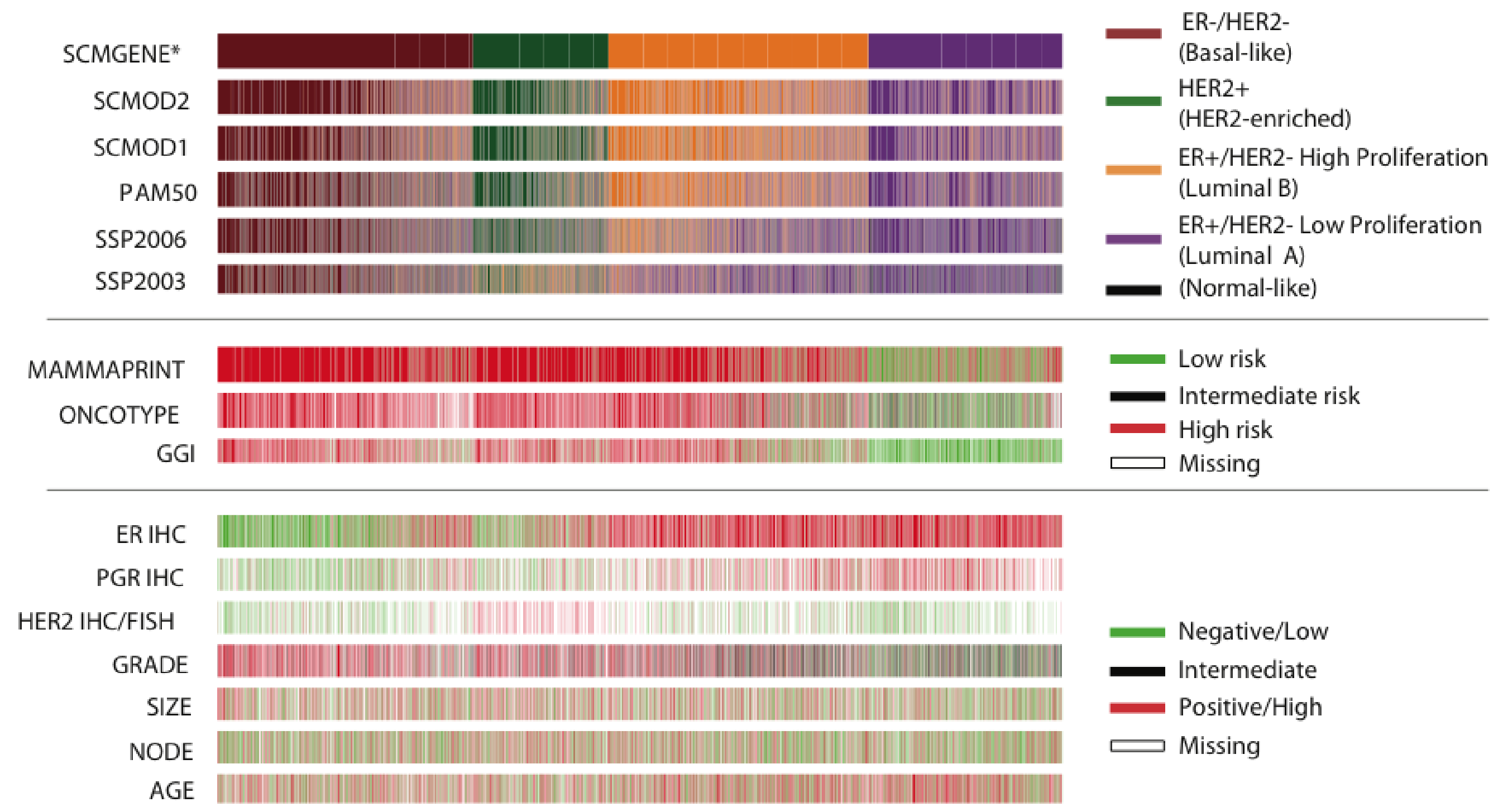
Integrated Brier score with respect to the time. IBSC for the seven genes in the MAINZ dataset.

Breast Cancer Molecular Subtype Prediction with *genefu*

Gene expression studies have well established that breast cancer (BC), in addition to be clinically diverse, is also a molecular heterogeneous disease with several clinically relevant molecular subtypes.

Our analysis pipeline, implemented in the *genefu* package (available on CRAN), comprises methods to robustly identify the breast cancer molecular subtypes and compute the prevalence of transcripts with respect to these subtypes.

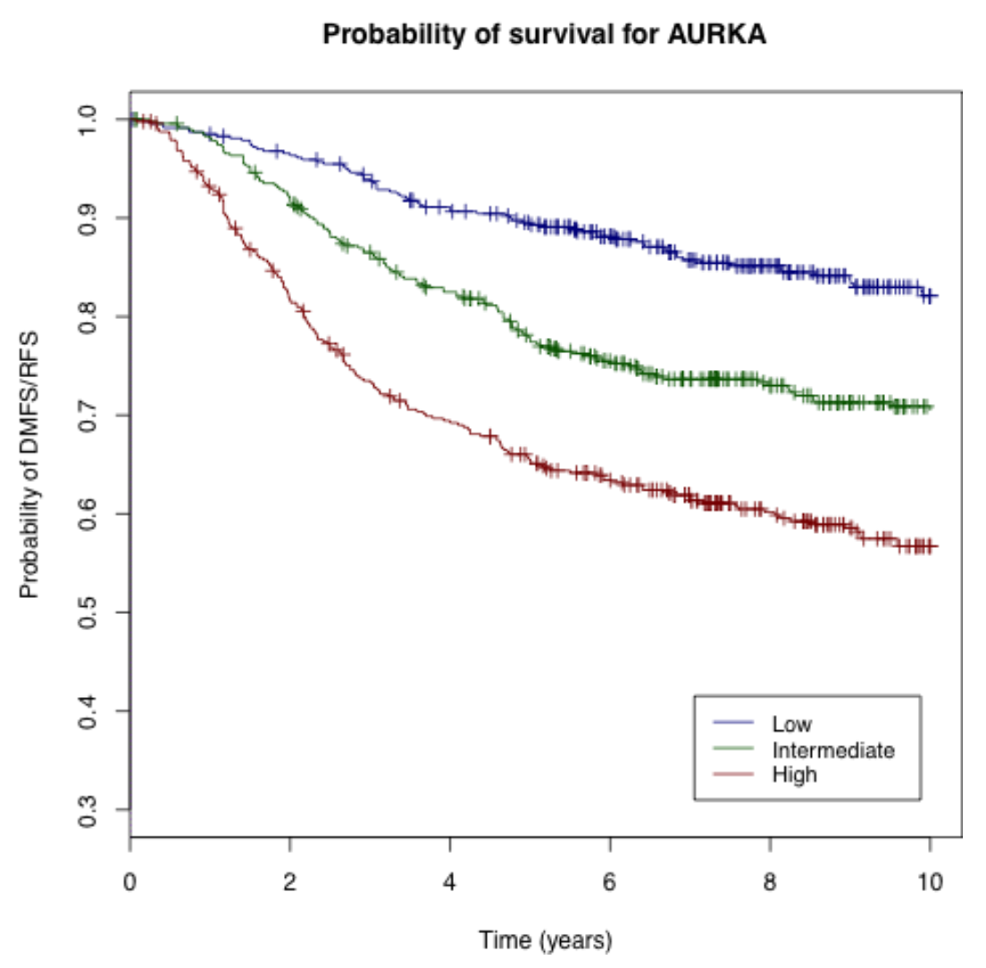
SCMGENE compared favorably to the other published classification models and yielded high concordance with existing prognostic gene signatures and traditional clinical parameters.

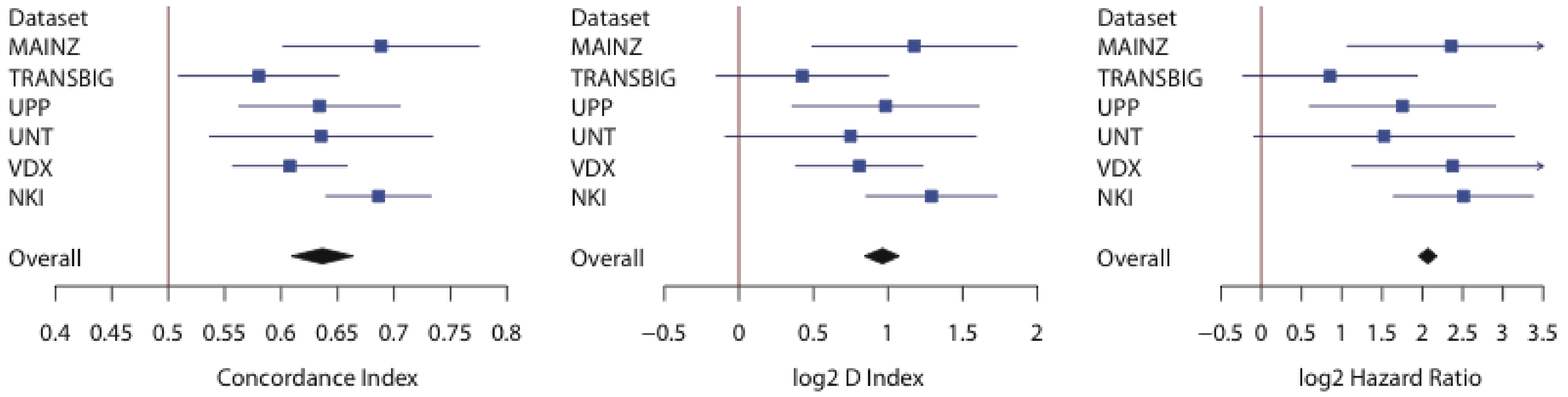


A Meta-Analysis Case Study of a Proliferation-Related Gene: AURKA

Left: Kaplan Meier curve for the gene AURKA. Gene expression and survival data is a combination of six breast cancer datasets with over 1400 patients.

Forestplots: show three different measurements for the performance of risk prediction models for the gene AURKA in six different datasets. An overall estimation for each measurement is included.





References

• Schröder et al. (2011), SurvComp: an R/Bioconductor package for performance assessment and comparison for survival analysis, in preparation
• Haibe-Kains et al. (2008), A comparativ study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?, *Bioinformatics*, **24**, 2200-2208
• Desmedt et al. (2008), Biological processes associated with breast cancer clinical outcome depend on the molecular outcome. *Clin Cancer Res*, **14**, 5158-5165