# Meta-analysis of gene expression and survival data using the GXA framework: a new prognosis tool for breast cancer

Markus Schröder[1,2], Daniel Gusenleitner[1], Matthew Schwede[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]Computational Genomics, 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 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

## A new Prognosis Tool for Breast Cancer

Several excellent online resources for mining of gene expression data exist, including Oncomine, NextBio and the Gene Expression Atlas (GXA). However, none provide survival data in addition to significant gene rankings. We are building an online resource combining gene expression and survival data from multiple datasets to enable clinicians and biologists to assess the prognostic values of genes of interest.

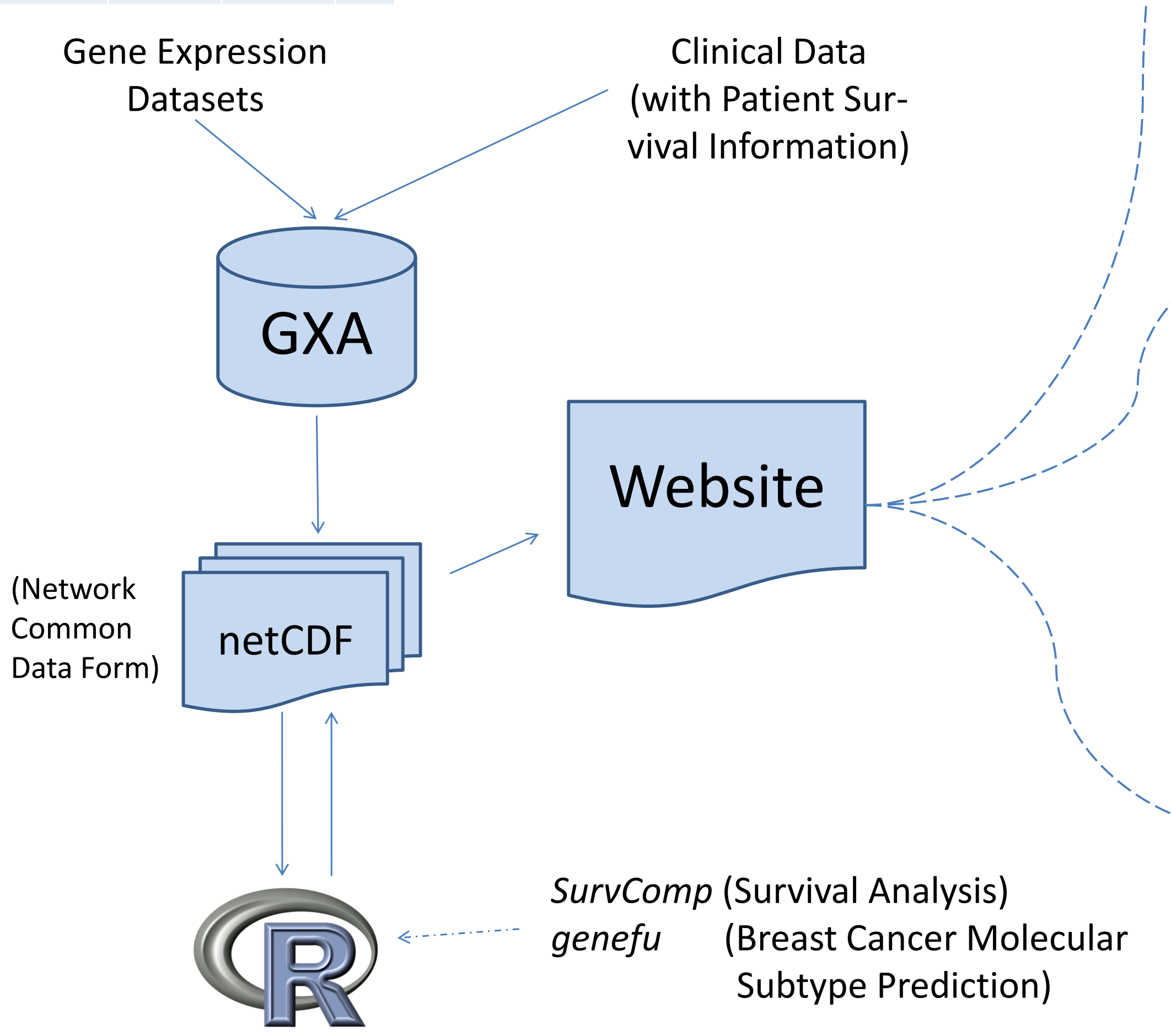
We extend the existing framework of the GXA in our pipeline. The GXA, maintained by the European Bioinformatics Institute, is an added-value database of gene expression sequencing data for different cell types, organism parts, developmental stages, disease states, sample treatments and other biological/experimental conditions.

The structure of the GXA has three layers: R Analytics, Database and Front End. It includes a pipeline to port data from ArrayExpress/GEO to GXA. Each dataset is annotated with a standard experimental factor ontology (EFO). We describe our new gene mining approach below.

## Workflow

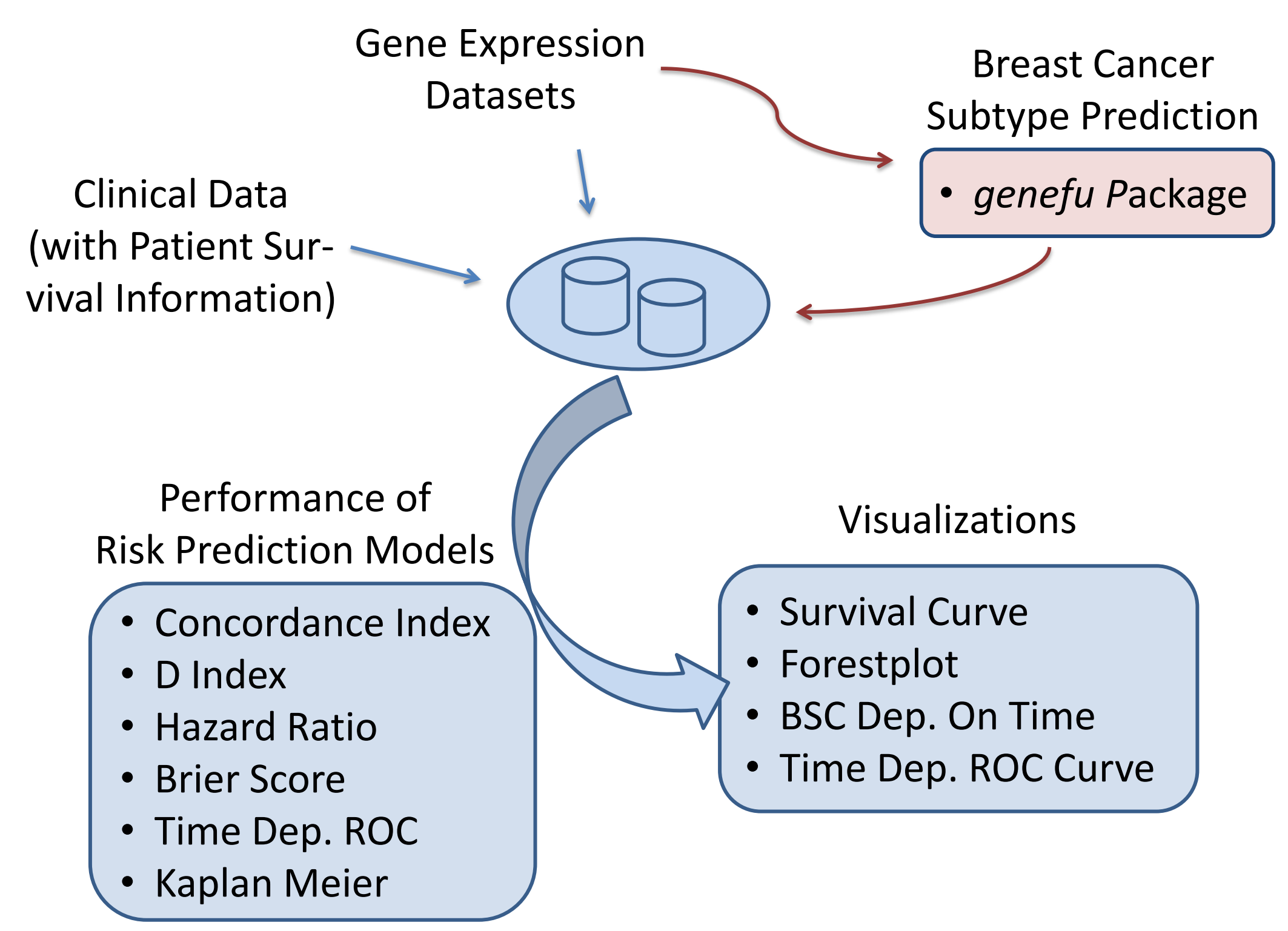
| | MAINZ_ BC6001 | MAINZ_ BC6002 | [...] |
|---|---|---|---|
| 1007_s_at | 11.514175 | 11.405460 | |
| 1053_at | 8.240630 | 8.161699 | [...] |
| 117_at | 7.691213 | 7.151237 | |
| | [...] | | |

| | Age | ER | Grade | e.dmfs | t.dmfs | [...] | |
|---|---|---|---|---|---|---|---|
| MAINZ_BC6001 | 54 | 0 | 2 | 1 | 2760 | | |
| MAINZ_BC6002 | 54 | 1 | 3 | 0 | 2130 | [...] | |
| MAINZ_BC6003 | 42 | 1 | 3 | 1 | 1740 | | |
| | | | [...] | | | | |


Gene Expression Datasets
Clinical Data (with Patient Survival Information)
GXA
Website
netCDF
(Network Common Data Form)

*SurvComp* (Survival Analysis)
*genefu* (Breast Cancer Molecular Subtype Prediction)

## Survival Analysis Workflow


Gene Expression Datasets
Breast Cancer Subtype Prediction
Clinical Data (with Patient Survival Information)

- *genefu* Package

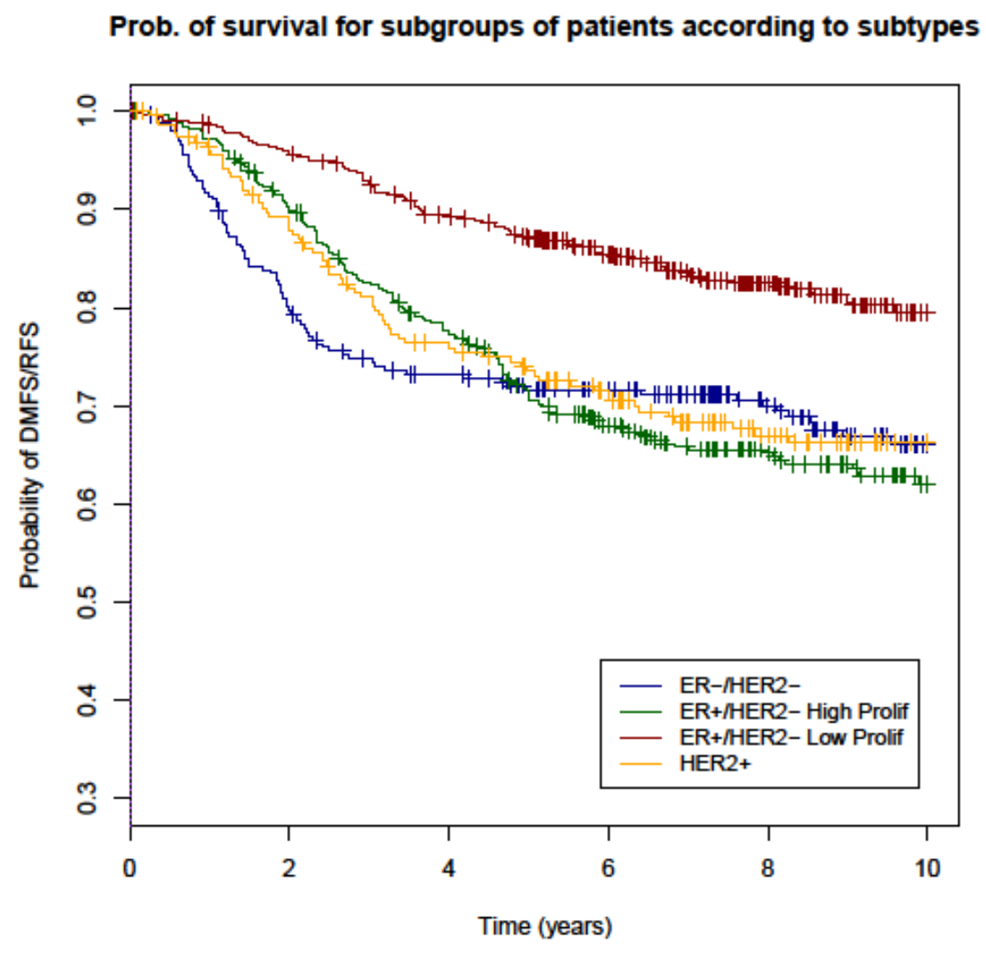
Performance of Risk Prediction Models

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

Visualizations

- Survival Curve
- Forestplot
- BSC Dep. On Time
- Time Dep. ROC Curve

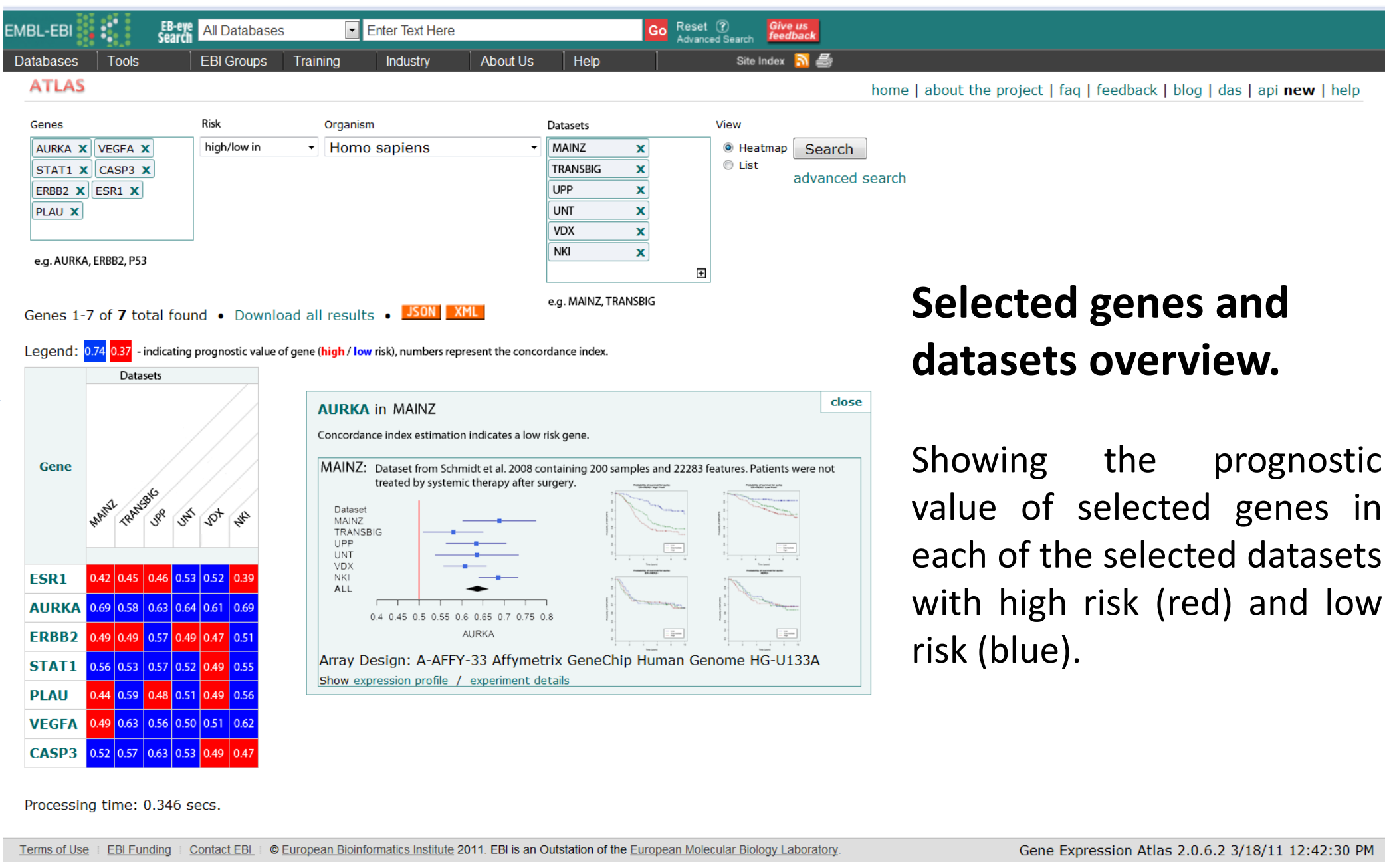
## Kaplan Meier Survival Curves



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

Kaplan Meier survival curve for a combination of subtype information and survival data for patients from six breast cancer datasets with a total of 1467 patients. The molecular subtypes are ER-/HER2-, ER+/HER2- High Proliferation, ER+/HER2- Low Proliferation and HER2+.

## Website Hierarchy



EMBL-EBI | All Databases | Enter Text Here | Go | Reset | Advanced Search | Give us feedback
Databases | Tools | EBI Groups | Training | Industry | About Us | Help | Site Index

ATLAS — home | about the project | faq | feedback | blog | das | api new | help

Genes: AURKA, VEGFA, STAT1, CASP3, ERBB2, ESR1, PLAU (e.g. AURKA, ERBB2, P53)
Risk: high/low in
Organism: Homo sapiens
Datasets: MAINZ, TRANSBIG, UPP, UNT, VDX, NKI (e.g. MAINZ, TRANSBIG)
View: ● Heatmap ○ List — Search — advanced search

Genes 1-7 of 7 total found • Download all results JSON XML

Legend: 0.74 0.37 - indicating prognostic value of gene (high / low risk), numbers represent the concordance index.

| Gene | MAINZ | TRANSBIG | UPP | UNT | VDX | NKI |
|---|---|---|---|---|---|---|
| ESR1 | 0.42 | 0.45 | 0.46 | 0.53 | 0.52 | 0.39 |
| AURKA | 0.69 | 0.58 | 0.63 | 0.64 | 0.61 | 0.69 |
| ERBB2 | 0.49 | 0.49 | 0.57 | 0.49 | 0.47 | 0.51 |
| STAT1 | 0.56 | 0.53 | 0.57 | 0.52 | 0.49 | 0.55 |
| PLAU | 0.44 | 0.59 | 0.48 | 0.51 | 0.49 | 0.56 |
| VEGFA | 0.49 | 0.63 | 0.56 | 0.50 | 0.51 | 0.62 |
| CASP3 | 0.52 | 0.57 | 0.63 | 0.53 | 0.49 | 0.47 |

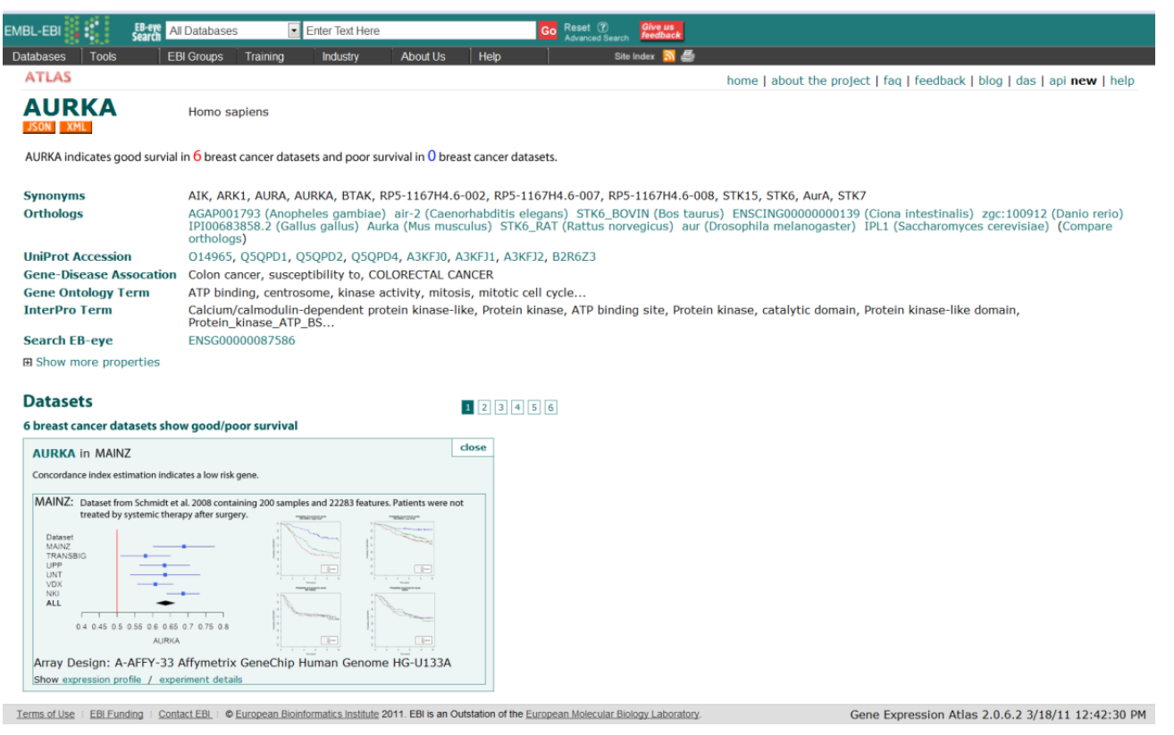
AURKA in MAINZ — close
Concordance index estimation indicates a low risk gene.
MAINZ: Dataset from Schmidt et al. 2008 containing 200 samples and 22283 features. Patients were not treated by systemic therapy after surgery.
Array Design: A-AFFY-33 Affymetrix GeneChip Human Genome HG-U133A
Show expression profile / experiment details

Processing time: 0.346 secs.

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2011. EBI is an Outstation of the European Molecular Biology Laboratory. Gene Expression Atlas 2.0.6.2 3/18/11 12:42:30 PM

### Selected genes and datasets overview.

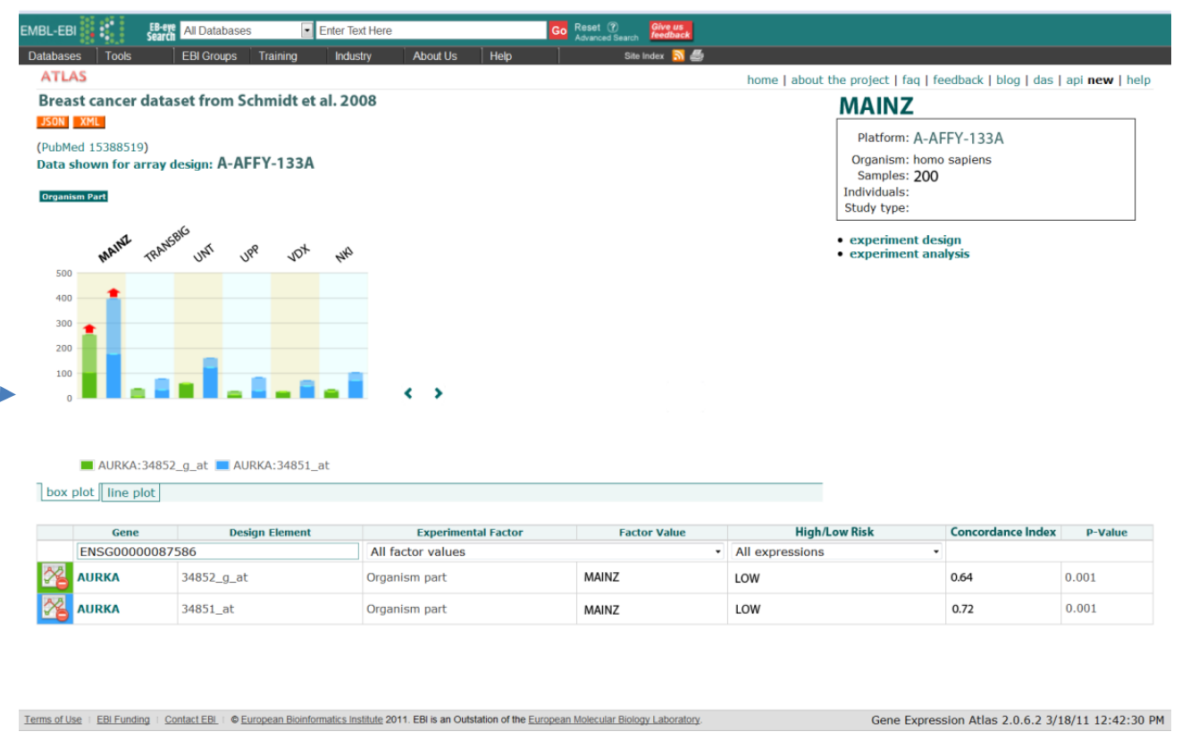
Showing the prognostic value of selected genes in each of the selected datasets with high risk (red) and low risk (blue).



ATLAS — AURKA — Homo sapiens
AURKA indicates good survial in 6 breast cancer datasets and poor survival in 0 breast cancer datasets.
Synonyms: AIK, ARK1, AURA, AURKA, BTAK, RP5-1167H4.6-002, RP5-1167H4.6-007, RP5-1167H4.6-008, STK15, STK6, AurA, STK7
Orthologs: AGAP001793 (Anopheles gambiae) air-2 (Caenorhabditis elegans) STK6_BOVIN (Bos taurus) ENSCING00000000139 (Ciona intestinalis) zgc:100912 (Danio rerio) IPI00683858.2 (Gallus gallus) Aurka (Mus musculus) STK6_RAT (Rattus norvegicus) aur (Drosophila melanogaster) IPL1 (Saccharomyces cerevisiae) (Compare orthologs)
UniProt Accession: O14965, Q5QPD1, Q5QPD2, Q5QPD4, A3KFJ0, A3KFJ1, A3KFJ2, B2R6Z3
Gene-Disease Assocation: Colon cancer, susceptibility to, COLORECTAL CANCER
Gene Ontology Term: ATP binding, centrosome, kinase activity, mitosis, mitotic cell cycle...
InterPro Term: Calcium/calmodulin-dependent protein kinase-like, Protein kinase, ATP binding site, Protein kinase, catalytic domain, Protein kinase-like domain, Protein_kinase_ATP_BS...
Search EB-eye: ENSG00000087586
Show more properties
Datasets — 6 breast cancer datasets show good/poor survival

### Gene view

Each gene has an annotation page with synonyms, orthologs, UniProt accessions, gene-disease association, GO terms, InterPro terms and links to public repositories. Also, each dataset that includes this gene is shown at the bottom.



ATLAS — Breast cancer dataset from Schmidt et al. 2008 (PubMed 15388519)
Data shown for array design: A-AFFY-133A
MAINZ — Platform: A-AFFY-133A; Organism: homo sapiens; Samples: 200; Individuals: ; Study type:
- experiment design
- experiment analysis

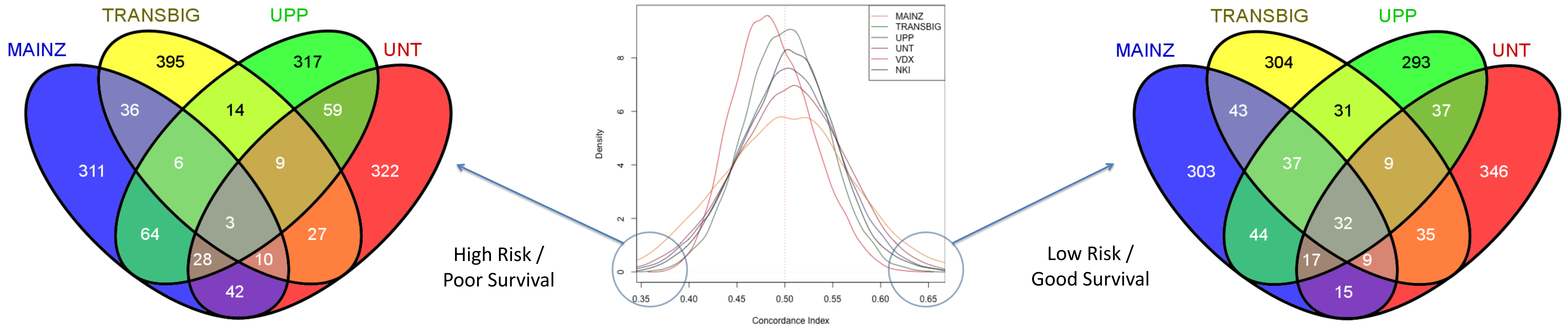
AURKA:34852_g_at AURKA:34851_at — box plot | line plot

| | Gene | Design Element | Experimental Factor | Factor Value | High/Low Risk | Concordance Index | P-Value |
|---|---|---|---|---|---|---|---|
| | ENSG00000087586 | | All factor values | | All expressions | | |
| | AURKA | 34852_g_at | Organism part | MAINZ | LOW | 0.64 | 0.001 |
| | AURKA | 34851_at | Organism part | MAINZ | LOW | 0.72 | 0.001 |

### Dataset view

Each dataset has an overview page with information about the patients and their subtypes, the microarray platform that was used and more detailed information about the gene expression data.

## Genome Scale Overview of Gene (n >20.000) Concordance Indices from Multiple Datasets



Representations of the 500 lowest (**left**) and 500 highest (**right**) concordance indices from 4 different datasets. High concordance indices for genes indicate a good survival / low risk for patients, low concordance indices for genes indicate poor survival / high risk for patients. In the **middle** figure, the concordance indices for genes from six different datasets are shown as a genome scale density plot, providing an 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.

## Analysis of Concordance Indices from Six Datasets

Low Concordance Indices / High Risk

| Gene Symbol | Gene Description |
|---|---|
| TXNIP | thioredoxin interacting protein |
| PKP2 | plakophilin 2 |
| PDLIM5 | PDZ and LIM domain 5 |
| LM04 | LIM domain only 4 |
| USP34 | ubiquitin specific peptidase 34 |
| HPS1 | Hermansky-Pudlak syndrome 1 |
| MUC5AC | mucin 5AC, oligomeric mucus/gel-forming |
| SIVA1 | SIVA1, apoptosis-inducing factor |
| CCND2 | cyclin D2 |
| SERPINA5 | serpin peptidase inhibitor, clade A member 5 |

We took the union of the genes from six datasets, which resulted in 19768 unique genes. Those genes were ranked according to their concordance index and summed over the six datasets with leaving the highest rank out, e.g. a gene with the lowest concordance index in all datasets would have score five since it has rank one in all datasets. We removed the highest rank for each gene over the six datasets to be more sensitive to outliers.

The **left** table shows the genes with the lowest scores strongly related to high risk patients.

The **right** table shows the genes with the lowest scores strongly related to low risk patients.

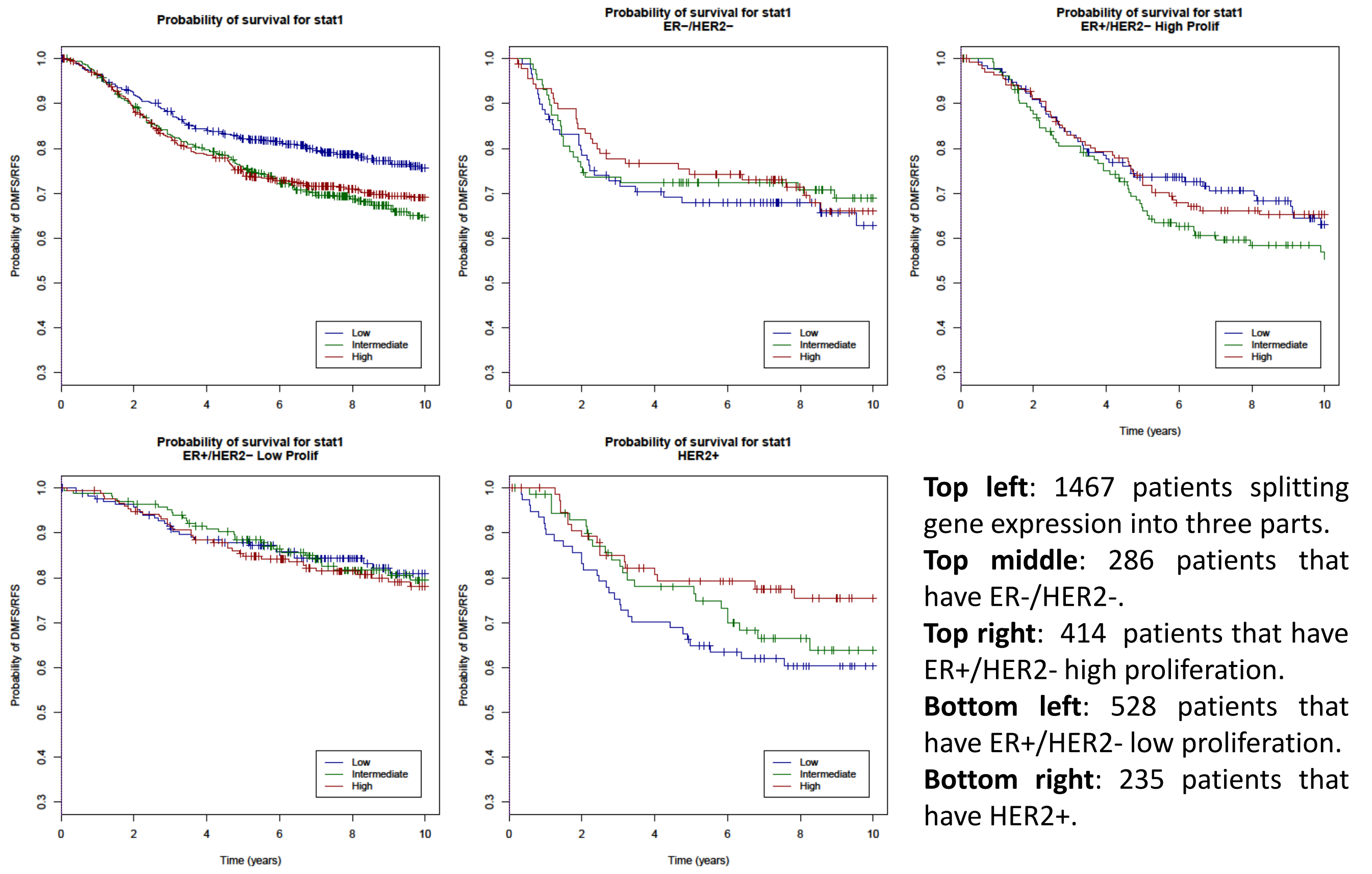
High Concordance Indices / Low Risk

| Gene Symbol | Gene Description |
|---|---|
| TRO | trophinin |
| B2M | beta-2-microglobulin |
| DDX47 | DEAD (Asp-Glu-Ala-Asp) box polypeptide 47 |
| TMEM45A | transmembrane protein 45A |
| CCNT2 | cyclin T2 |
| NFIB | nuclear factor I/B |
| TMEM132A | transmembrane protein 132A |
| RAB3IL1 | RAB3A interacting protein (rabin3)-like 1 |
| AGBL2 | ATP/GTP binding protein-like 2 |
| PION | pigeon homolog (Drosophila) |

## Gene Expression Datasets Included in Prototype

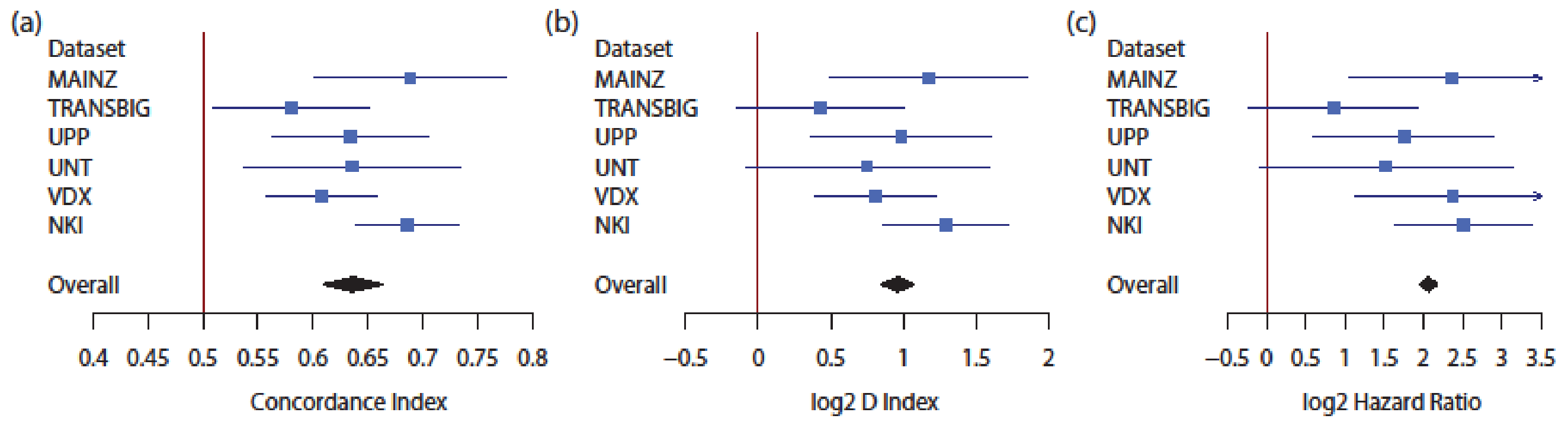
| Dataset | Patients [#] | ER+ [#] | HER2+ [#] | Age [years] | Grade [1/2/3] | Platform |
|---|---|---|---|---|---|---|
| MAINZ | 200 | 155 | 23 | 25-90 | 29/136/35 | HGU133A |
| TRANSBIG | 198 | 123 | 35 | 24-60 | 30/83/83 | HGU133A |
| UPP | 251 | 175 | 46 | 28-93 | 67/128/54 | HGU133AB |
| UNT | 137 | 94 | 21 | 24-73 | 32/51/29 | HGU133AB |
| VDX | 344 | 186 | 57 | 26-83 | 7/42/148 | HGU133A |
| NKI | 337 | 212 | 53 | 26-62 | 79/109/149 | Rosetta |
| **Overall** | **1467** | **945** | **235** | **24-93** | **244/549/498** | **Affy/Agilent** |

## Prognostic Value of STAT1 Using Subtypes

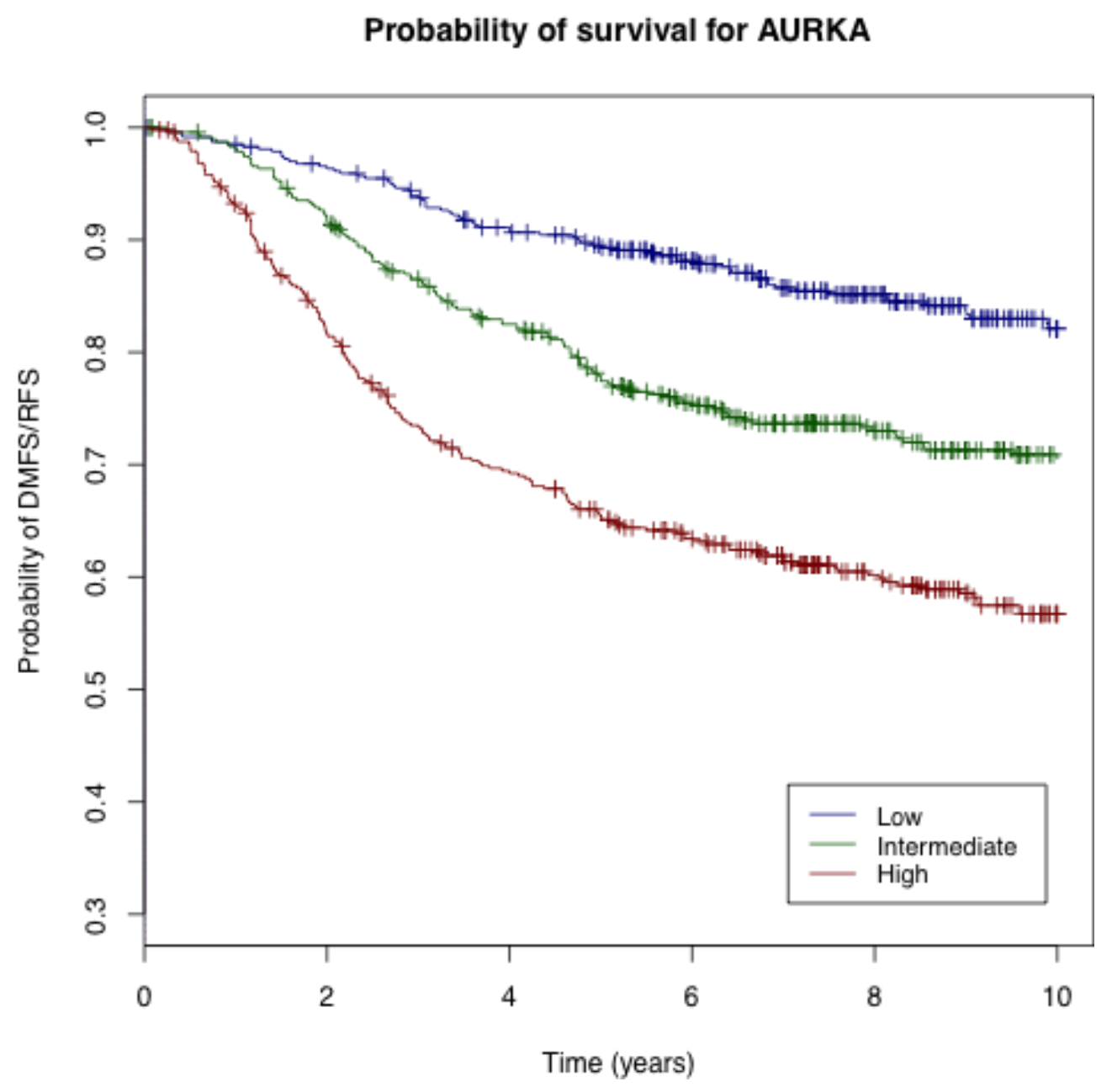


**Top left**: 1467 patients splitting gene expression into three parts.
**Top middle**: 286 patients that have ER-/HER2-.
**Top right**: 414 patients that have ER+/HER2- high proliferation.
**Bottom left**: 528 patients that have ER+/HER2- low proliferation.
**Bottom right**: 235 patients that have HER2+.

## A Meta-Analysis Case Study of AURKA



**Top:** 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. (a) concordance index, (b) log2 D index, (c) log2 hazard ratio.



**Right:** Kaplan Meier survival curve for the gene AURKA. Gene expression and survival data is a combination of six breast cancer datasets with over 1400 patients. The gene expression is split into 3 parts using quantiles (33% and 66%)

## 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 comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?, *Bioinformatics*, **24**, 2200-2208

Kapushesky et al. (2010), Gene Expression Atlas at the European Bioinformatics Institute, Bioinformatics, **38**, D690-D698