Prognostic and biological relevance of collagen-related genes in prostate cancer

Methods, figures and tables for the transcriptome part

2025-02-24

# Methods

Analyses of collagen-related transcriptome and survival in transcriptomic cohorts were done with R version 4.2.3 (R foundation). Their details are provided in **Supplementary Methods**.

## Transcriptome data sets

Seven publicly available transcriptome data sets of primary prostate cancer samples, the TCGA prostate cancer cohort (1,2) (n = 493 cancer samples, RNA sequencing), GSE16560 (3) (n = 281, microarray) GSE54460 (4) (n = 106, RNA sequencing), GSE70768 (5) (n = 125, microarray), GSE70769 (5) (n = 94, microarray), GSE220095 (6) (n = 176, RNA sequencing), and DKFZ (7) (n = 118, RNA sequencing) were re-analyzed. Selection criteria for the transcriptomic cohorts were: information of biochemical relapse and relapse-free survival or death and overall survival, and availability of expression data for collagen-related transcript investigated in our previous collagen project (8).

Normalized gene expression data and clinical information were fetched from the cBioportal repository (DKFZ, TCGA) or Gene Expression Omnibus (GEO). The microarray signals of RNA sequencing counts were analyzed as -transformed counts adjusted for differences in distribution between the cohorts with ComBat algorithm (9). ComBat-adjusted expression and clinical information for the GSE54460, GSE70768, GSE70769, and GSE220095 studies was merged into a ‘pooled GEO cohort’. In machine learning models of BCR-free survival, the pooled GEO cohort was used for training, and TCGA and DKFZ served as validation cohorts. The GSE16560 cohort split into a training and test subset in a 2:1 ratio was used for modeling of overall survival. Characteristic of the cohorts and a list of the collagen-related genes of interest are are provided in **Supplementary Tables S1 - S4**.

## Co-expression network analyses

## Univariable modeling of BCR-free survival

## Multi-parameter modeling of BCR-free and overall survival

## Data and code availability

Publicly available data sets were analyzed. Formatted data sets used for analyses will be made available upon request to the corresponding author. The R analysis pipeline for the transcriptome data is available from GitHub (<https://github.com/PiotrTymoszuk/BCR_collagen>).

# Figures

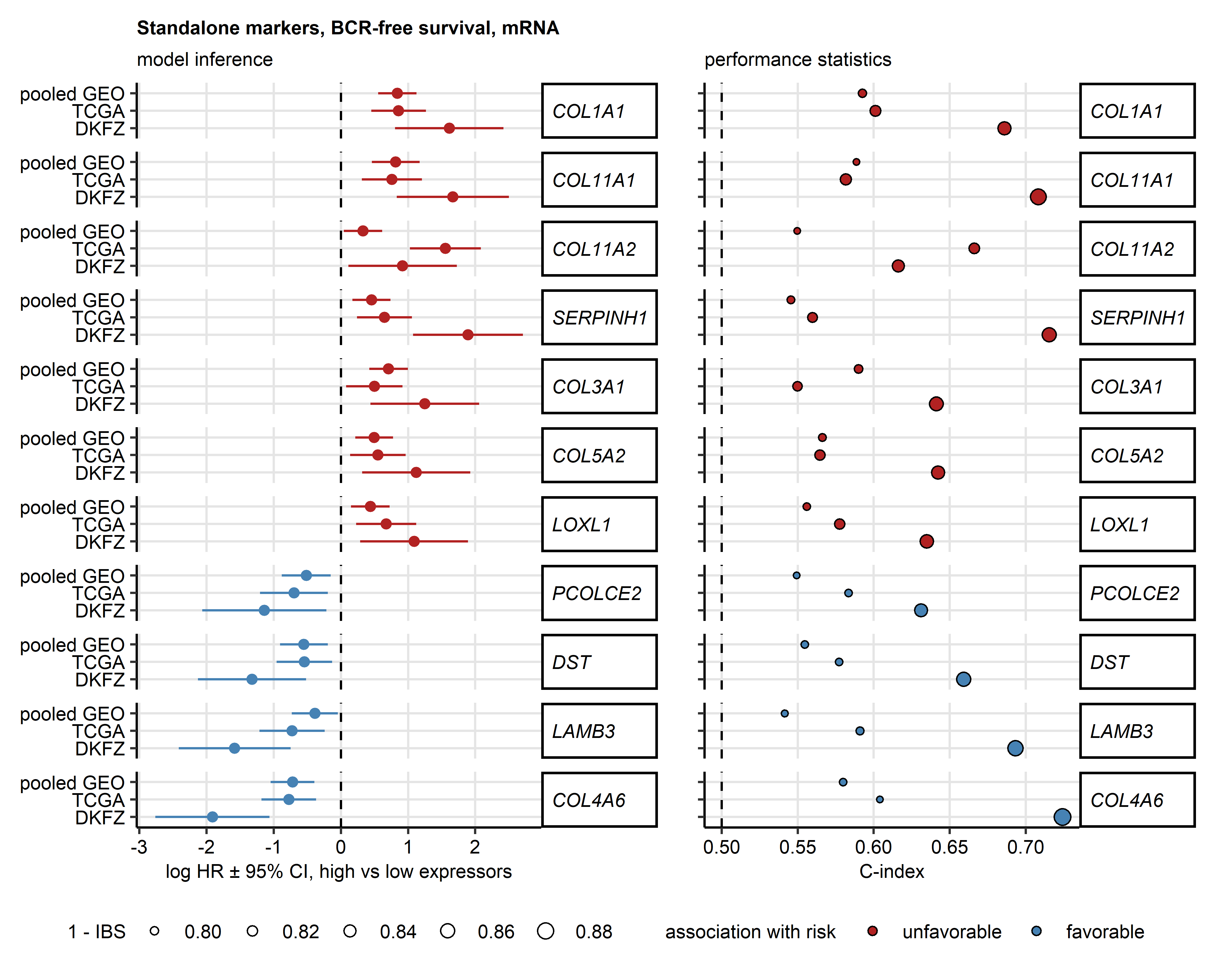


Figure 1: Univariable analysis of BCR-free survival with collagen-related transcripts as candidate standalone prognostic factors.

**Figure 1. Univariable analysis of BCR-free survival with collagen-related transcripts as candidate standalone prognostic factors.**

*Prostate cancer (PCA) patients in the pooled GEO (total: n = 485, biochemical relapse [BCR]: n = 194), TCGA (total: n = 493, BCR: n = 93), and DKFZ cohort (total: n = 105, BCR: n = 24) were classified as high and low expressors for each of the rlength(globals$genes)` collagen-related transcripts with the expression cutoffs corresponding to the largest differences in survival between the high and low expressors.* *BCR risk was compared between the high and low expressors for each transcript by univariable Cox proportional hazard regression. P values were corrected for multiple testing with the false discovery rate (FDR) method.* *Standalone BCR risk markers were defined as transcripts associated with unfavorable prognosis (high versus low expressors, hazard ratio [HR] > 1, pFDR < 0.05) or favorable prognosis (HR < 1, pFDR < 0.05) in all three cohorts.*

*HR with 95% confidence intervals for BCR risk in the high versus low expression strata are presented in a Forest plot (left). The model performance statistics, Harrell’s concordance index (C-index, measure of model accuracy), and Integrated Brier Scores (IBS, measure of overall model calibration, low values indicative of good calibration) are presented in a dot plot (right).*

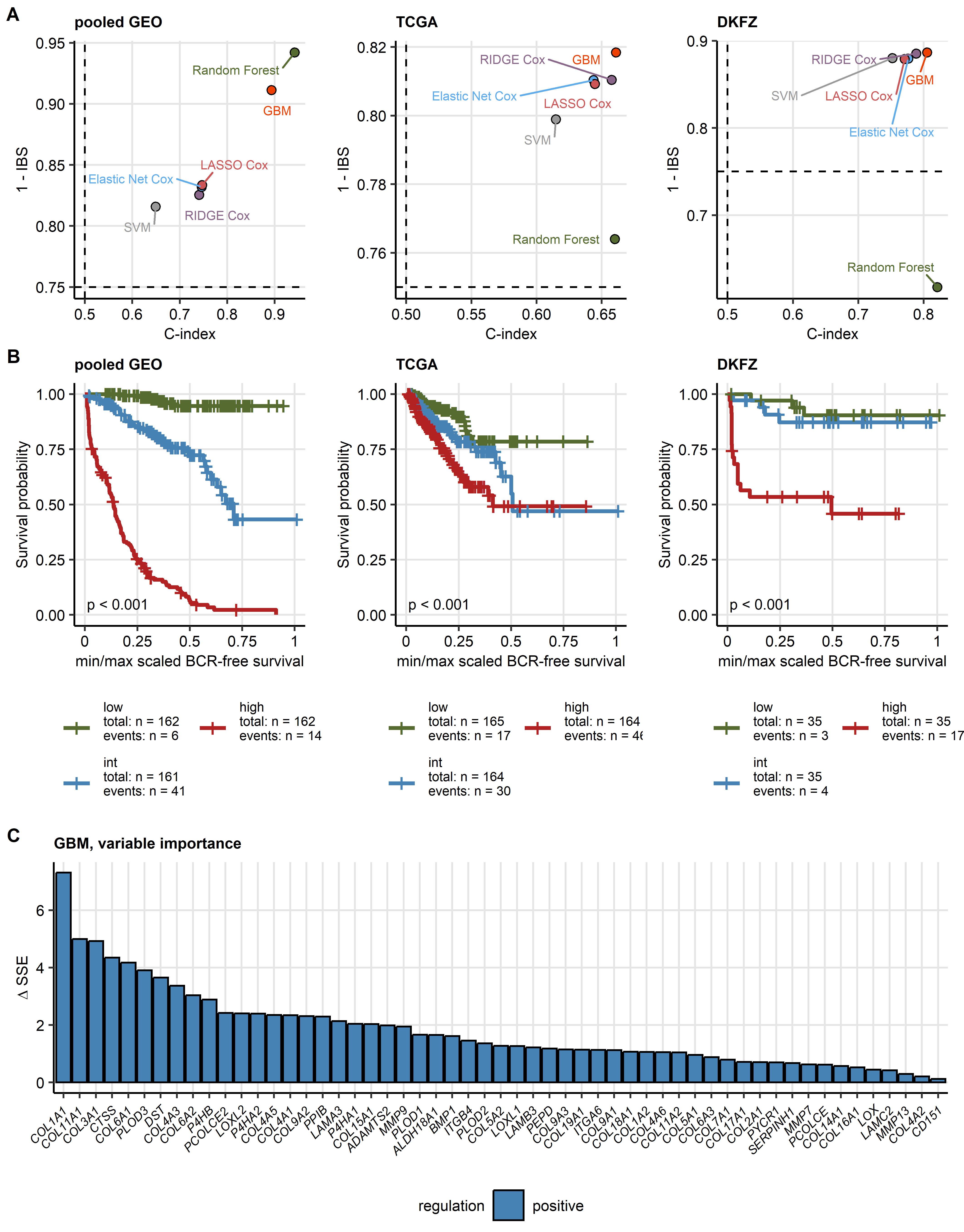


Figure 2: Machine learning modeling of BCR-free survival with expression levels of the collagen-related transcripts as explanatory factors.

**Figure 2. Machine learning modeling of BCR-free survival with expression levels of the collagen-related transcripts as explanatory factors.**

*Risk of biochemical relapse (BCR) in prostate cancer (PCA) was modeled by multi-parameter machine learning models with ComBat-processed -transformed expression levels of 55 collagen-related transcripts as explanatory factors.* *The models were trained in the pooled GEO cohort with six machine learning algorithms (RIDGE Cox regression, Elastic Net Cox regression, LASSO Cox regression, survival Support Vector Machines [SVM], survival Random Forest, and survival Gradient Boosted Machines [GBM]).* *Performance of the models was evaluated in the pooled GEO training cohort (total: n = 485, BCR: n = 194), and the TCGA (total: n = 493, BCR: n = 93) and DKFZ (total: n = 105, BCR: n = 24) validation collectives.*

*(A) Metrics of performance of the survival models, Harrell’s concordance index (C-index, measure of model accuracy) and Integrated Brier Score (IBS, measure of overall model calibration, low values are characteristic for good calibration), are presented in dot plots. Dashed lines represent C-index and IBS values expected for random BCR risk prediction. Note the superior accuracy and calibration of the GBM model in the validation collectives.*

*(B) BCR-free survival of PCA patients stratified by tertiles of the predictor scores of the best performing GBM model. Statistical significance of differences between the predictor score tertiles was determined by Peto-Peto test corrected for multiple testing wit the false discovery rate method. Fractions of BCR-free patients and minimum/maximum scaled time after diagnosis are visualized in Kaplan-Meier plots. Numbers of patients and BCR in the predictor score tertiles are indicated in the legends. P values are displayed in the plots.*

*(C) Importance of the explanatory variables for BCR-risk prediction by the GBM model was gauged by difference in sum of squared error () attributed to inclusion of a particular collagen-related transcript in the GBM learner ensemble. High values are characteristic for highly influential variables. values for the variables that contributed substantially to the BCR risk prediction ( > 0) are presented in a bar plot.*

# References

1. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, Kovatich AJ, Benz CC, Levine DA, Lee AV, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* (2018) 173:400–416.e11. doi: [10.1016/J.CELL.2018.02.052](https://doi.org/10.1016/J.CELL.2018.02.052)

2. Abeshouse A, Ahn J, Akbani R, Ally A, Amin S, Andry CD, Annala M, Aprikian A, Armenia J, Arora A, et al. The Molecular Taxonomy of Primary Prostate Cancer. *Cell* (2015) 163:1011–1025. doi: [10.1016/j.cell.2015.10.025](https://doi.org/10.1016/j.cell.2015.10.025)

3. Sboner A, Demichelis F, Calza S, Pawitan Y, Setlur SR, Hoshida Y, Perner S, Adami HO, Fall K, Mucci LA, et al. Molecular sampling of prostate cancer: A dilemma for predicting disease progression. *BMC Medical Genomics* (2010) 3: doi: [10.1186/1755-8794-3-8](https://doi.org/10.1186/1755-8794-3-8)

4. Long Q, Xu J, Osunkoya AO, Sannigrahi S, Johnson BA, Zhou W, Gillespie T, Park JY, Nam RK, Sugar L, et al. Global transcriptome analysis of formalin-fixed prostate cancer specimens identifies biomarkers of disease recurrence. *Cancer research* (2014) 74:3228–3237. doi: [10.1158/0008-5472.CAN-13-2699](https://doi.org/10.1158/0008-5472.CAN-13-2699)

5. Ross-Adams H, Lamb A, Dunning M, Halim S, Lindberg J, Massie C, Egevad L, Russell R, Ramos-Montoya A, Vowler S, et al. Integration of copy number and transcriptomics provides risk stratification in prostate cancer: A discovery and validation cohort study. *EBioMedicine* (2015) 2:1133–1144. doi: [10.1016/j.ebiom.2015.07.017](https://doi.org/10.1016/j.ebiom.2015.07.017)

6. Schimmelpfennig C, Rade M, Füssel S, Löffler D, Blumert C, Bertram C, Borkowetz A, Otto DJ, Puppel SH, Hönscheid P, et al. Characterization and evaluation of gene fusions as a measure of genetic instability and disease prognosis in prostate cancer. *BMC cancer* (2023) 23: doi: [10.1186/S12885-023-11019-6](https://doi.org/10.1186/S12885-023-11019-6)

7. Gerhauser C, Favero F, Risch T, Simon R, Feuerbach L, Assenov Y, Heckmann D, Sidiropoulos N, Waszak SM, Hübschmann D, et al. Molecular Evolution of Early-Onset Prostate Cancer Identifies Molecular Risk Markers and Clinical Trajectories. *Cancer cell* (2018) 34:996–1011.e8. doi: [10.1016/J.CCELL.2018.10.016](https://doi.org/10.1016/J.CCELL.2018.10.016)

8. Heidegger I, Frantzi M, Salcher S, Tymoszuk P, Martowicz A, Gomez-Gomez E, Blanca A, Lendinez Cano G, Latosinska A, Mischak H, et al. Prediction of Clinically Significant Prostate Cancer by a Specific Collagen-related Transcriptome, Proteome, and Urinome Signature. *European urology oncology* (2024) doi: [10.1016/J.EUO.2024.05.014](https://doi.org/10.1016/J.EUO.2024.05.014)

9. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* (2012) 28:882. doi: [10.1093/BIOINFORMATICS/BTS034](https://doi.org/10.1093/BIOINFORMATICS/BTS034)