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

Methods, figures and tables for the transcriptome part

2024-01-25

Methods

Magnetic resonance imaging of the prostate

Dear co-authors, it's your part as experts:)

ADC was compared between patient-matched measurements of benign prostate and malignancy by paired T test with Cohen's d effect size statistic.

Transcriptome analyses

The transcriptome analysis was performed with R version 4.2.3. Its details are presented in **Supplementary Methods**.

Transcriptome data sets

Six publicly available transcriptome data sets of primary prostate cancer samples, the TCGA prostate cancer cohort (1,2) (n = 493, RNA sequencing), GSE54460 (3) (n = 106, RNA sequencing), GSE70768 (4) (n = 125, microarray), GSE70769 (4) (n = 94, microarray), GSE220095 (5) (n = 176, RNA sequencing), DKFZ (6) (n = 118, RNA sequencing) were re-analyzed (**Supplementary Table S1**). Normalized gene expression data and clinical information were fetched from the cBioportal repository (DKFZ, TCGA) or Gene Expression Omnibus (GEO). Selection criteria of the data sets were: availability of Gleason scoring, information of biochemical relapse and relapse-free survival, and availability of expression data for 55 collagen-related genes covered by proteomic analysis as well (**Supplementary Table S2**). Donor-matched benign and malignant tissue samples were available for the GSE70768 (n = 73 tissue pairs) and TCGA cohort (n = 52 tissue pairs). Expression values were log_2-transformed prior to analyses. Non-immune cell content in cancer samples was estimated with the *MCP Counter* and *xCell* algorithms (7,8). single sample gene set enrichment scores (ssGSEA) for the Reactome pathway gene signatures retrieved from the MSig database version 7.5.1 were computed with the *GSVA* algorithm (9).

The collagen-related genes were retrieved from the published Collagen Signature (10) and the Reactome collagen formation pathway (R-HSA-1474290) and restricted to genes covered by the proteomic analyses (**Supplementary Table S2**).

Differential expression in benign and malignant prostate tissue and cancers stratified by Gleason score

Expression of the collagen-related genes of interest between patient-matched benign and malignant specimens was compared by the false discovery rate method (FDR) corrected paired two-tailed T test with Cohen's d effect size statistic. Differentially expressed genes were defined by pFDR < 0.05 and d ≥ 0.2 criteria indicative of significant, at least weak expression regulation (**Supplementary Table S3**).

Differential gene expression in cancer samples stratified by the ISUP risk strata (ISUP1: Gleason score 5 - 6, ISUP2: Gleason score 7 and ISUP3: Gleason score ≥ 8) was assessed by FDR-corrected one-way ANOVA with η^2 effect size statistic. Genes with pFDR < 0.05 and $\eta^2 \geq$ 0.02 were deemed differentially regulated (**Supplementary Table S4**).

Prediction of biochemical relapse-free survival by gradient boosted machine modeling

For survival analyses, log_2-transformed expression values of the collagen-related genes (**Supplementary Table S2**) were subjected to batch-adjustment with the ComBat algorithm (11). Subsequently, the ComBat-adjusted expression data sets of the GSE54460, GSE70768, GSE70769, GSE220095 cohorts were merged (further referred to as 'pooled GEO' cohort).

Biochemical relapse-free survival in the pooled GEO training cohort was modeled with normalized (Z-scores) expression values of the collagen-related genes as explanatory variables by the Gradient Boosted Machine (GBM) algorithm (12–15). Selection of the optimal set of modeling parameters was accomplished by tuning with 10-fold cross-validation and out-of-fold model deviance as performance statistic (**Supplementary Tables S15**). The Transcriptomic Collagen Score was computed as the linear predictor score of the GBM model. Importance of the explanatory variables for the predictive performance of the GBM model were measured with the relative influence method (12) and expressed as difference in the sum of squared errors (ΔSSE, **Figure 3A**).

Performance of the Transcriptomic Collagen Score at prediction of biochemical relapse-free survival in the training pooled GEO collective and validation data sets (TCGA and DKFZ) was assessed by univariable Cox regression with Harrell's concordance index (16) and integrated Brier score (17) as metrics of goodness of fit and calibration (**Figure 3B**, **Supplementary Table 16**). Of note, predictive performance of the GBM algorithm was excellent and reproducible in the investigated cohorts as compared with several other machine learning algorithms (e.g. Ridge Cox regression, Elastic Net or Support Vector

Machines, **Supplementary Figure S9**). Differences in biochemical relapse-free survival in cancer patients stratified by tertiles of the Transcriptomic Collagen Scores were assessed by false discovery rate (FDR) corrected Peto-Peto test (**Figure 3C**).

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 transcriptome R analysis pipeline is available from GitHub (https://github.com/PiotrTymoszuk/collagen_pca).

Figures

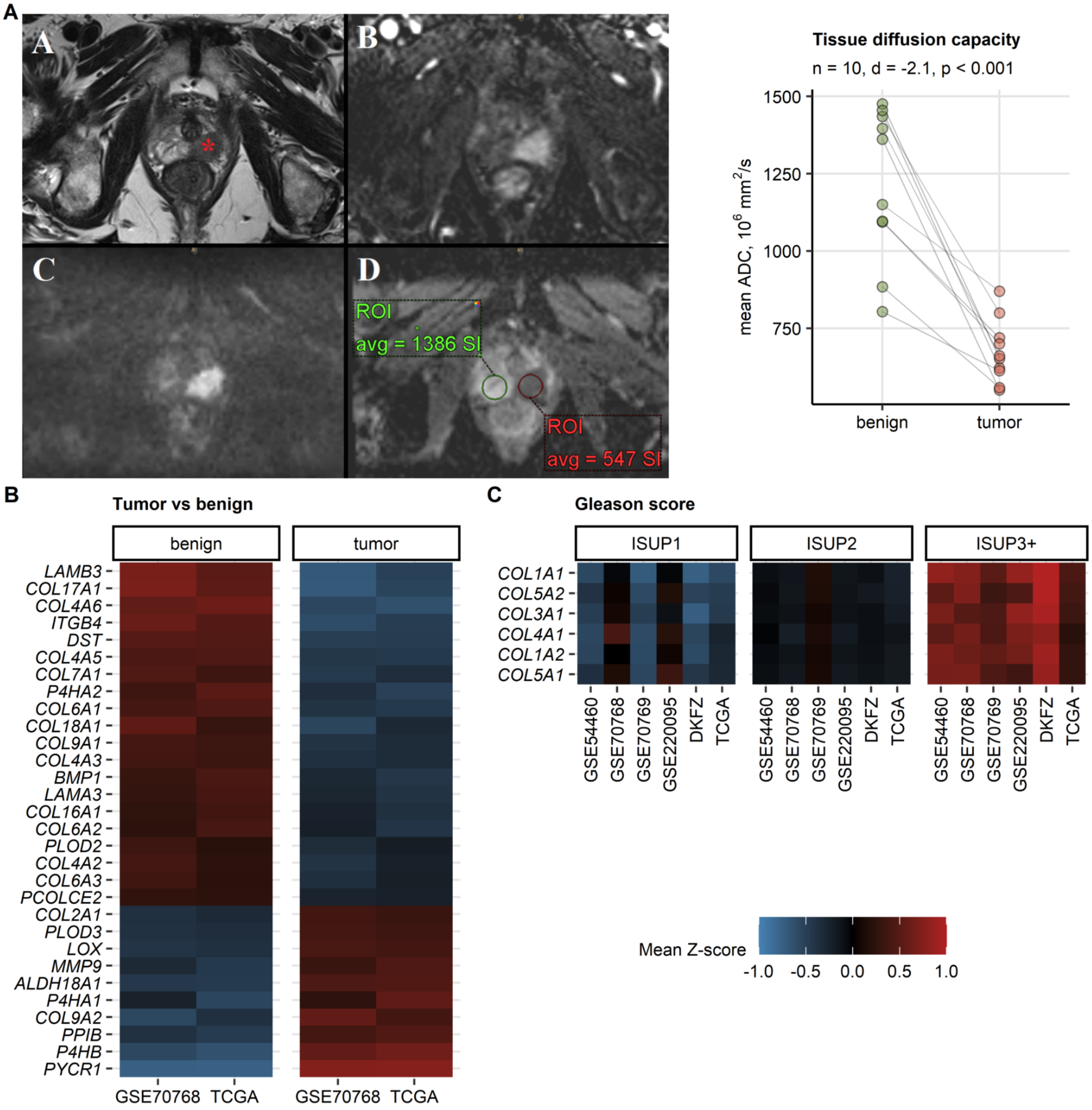

Figure 1. Differences in tissue diffusion capacity and expression of collagen-related genes between the prostate cancer and benign tissue. Expression of collagen-related genes in cancers stratified by Gleason scores.

(A) Representative MRI axial sequences of the prostate of an 83-year-old patient with prostate cancer (left panel). The T2-weighted images (a) show a 23mm suspicious hypo-intense lesion in the left peripheral zone (red star) of the prostate, which has an early enhancement in the contrast-enhanced T1-weighted image (b). The DWI scan (c) and ADC map (d) show a strong diffusion restriction (black and white character) of the tumor area suggesting a PI-RADS 5 lesion. Statistical significance for differences in diffusion capacity in patient-matched tumor and benign tissue measured by ADC was determined by paired T test with Cohen's d effect size statistic (right panel). Single ADC values are visualized as points, grey lines connect measurements of the same donors. The number of analyzed benign - tumor pairs, effect size and p value are displayed in the plot caption.

(B) Differences in log_2-transformed expression of 55 collagen-related genes between donor-matched tumor and benign prostate tissue samples were investigated by paired T test with Cohen's d effect size statistic in the GSE70768 (n = 73) and TCGA (n = 52 tissue pairs) cohorts. P values were corrected for multiple testing with the false discovery rate method. Mean normalized log_2-transformed expression levels in benign and cancer tissue for 30 genes significantly regulated in both cohorts with at least weak effect size (pFDR < 0.05, d ≥ 0.2) are presented in a heat map.

(C) Differences in log_2-transformed expression levels of 55 collagen-related genes between cancer samples stratified according to the ISUP risk system were assessed by one-way ANOVA with η^2 effect size statistic in the GSE54460 (ISUP1: n = 11, ISUP2: n = 80, ISUP3+: n = 15), GSE70768 (ISUP1: n = 19, ISUP2: n = 87, ISUP3+: n = 16), GSE70769 (ISUP1: n = 21, ISUP2: n = 56, ISUP3+: n = 14), GSE220095 (ISUP1: n = 36, ISUP2: n = 120, ISUP3+: n = 20), DKFZ (ISUP1: n = 13, ISUP2: n = 87, ISUP3+: n = 18), and TCGA cohort (ISUP1: n = 45, ISUP2: n = 245, ISUP3+: n = 203). Mean normalized log_2-transformed expression levels of genes significantly regulated between the Gleason score strata with at least weak effect size (pFDR < 0.05, $\eta^2 \geq$ 0.02) are presented in a heat map.

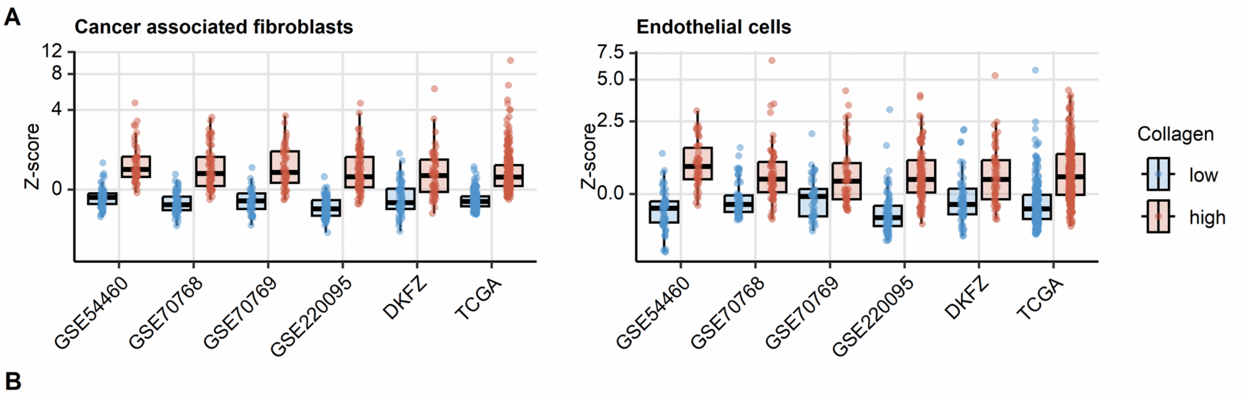

Figure 2. Expression of collagen-related genes in components of the tumor microenvironment.

(A) Cancer samples in six published cohorts were classified as collagen low and collagen high tumors in respect to 55 collagen-related transcripts by semi-supervised PAM clustering (GSE54460: Collagen low: n = 60, Collagen hi: n = 46, GSE70768: Collagen low: n = 62, Collagen hi: n = 63, GSE70769: Collagen low: n = 45, Collagen hi: n = 49, GSE220095: Collagen low: n = 78, Collagen hi: n = 98, DKFZ: Collagen low: n = 61, Collagen hi: n = 57, TCGA: Collagen low: n = 236, Collagen hi: n = 257). Counts of cancer-associated fibroblasts and endothelial cells were predicted by the MCP Counter algorithm and compared between the collagen clusters by Mann-Whitney test with r effect size statistic. Median normalized cell counts with interquartile ranges are presented as boxes, whiskers span over 150% of the interquartile ranges. Values for single cancer samples are depicted as points. All presented effects were significant (pFDR < 0.05) with at least weak (endothelial cells, r > 0.21) or moderate effect size (fibroblasts, r > 0.55).

(B) Placeholder for the single cell RNA seq data.

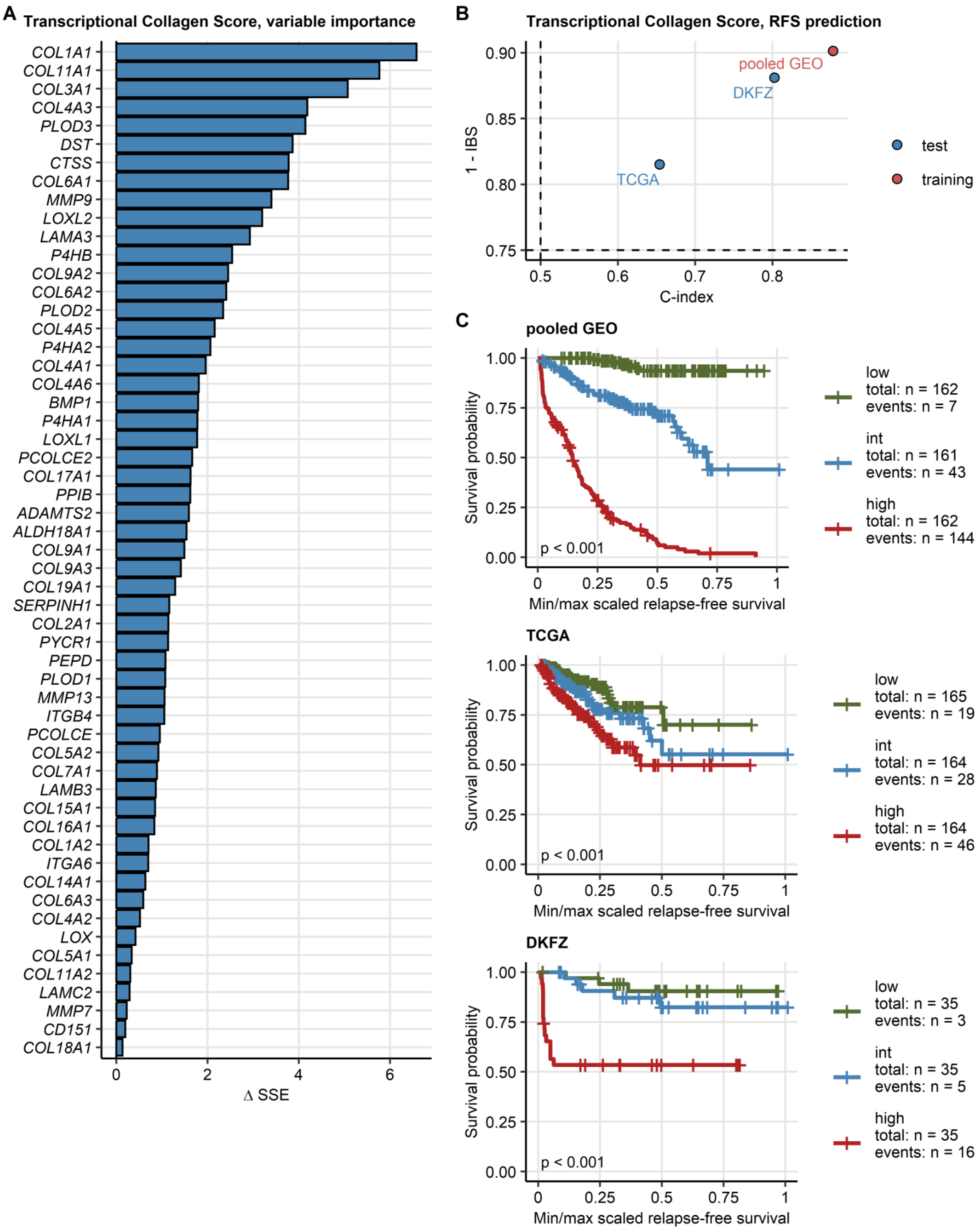

Figure 3. Transcriptional Collagen Model of biochemical relapse-free survival.

Transcriptional Collagen Score was established in the pooled GEO training cohort by multi-parameter Gradient Boosted Machine (GBM) modeling of biochemical relapse-free survival as a function of normalized log_2-transformed and ComBat-adjusted expression levels of 55 collagen-related genes.

(A) Importance of explanatory variables measured by the gradient of sum of squared errors (ΔSSE) during the training process attributed to particular collagen-related transcript levels.

(B) Performance of the Elastic Net Cox model at prediction of biochemical relapse-free survival in the training TCGA cohort and test collectives assessed by concordance index (C-index, high values indicate high concordance between the predicted and observed survival) and integrated Brier score (IBS, low values indicate good model fit and proper calibration). Values of C-index and IBS expected for random survival prediction are depicted as dashed lines.

(C) Fractions of biochemical relapse-free patients in low, intermediate and high tertiles of Transcriptional Collagen Score in the training TCGA cohort, and the GSE70768 and DKFZ validation collectives presented in Kaplan-Meier plots. Differences in survival between the tertiles were investigated by Peto-Peto test. Total numbers of patients and numbers of biochemical relapses are displayed in the plot captions. P values are shown in the plots.

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

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

3. 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

4. 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

5. 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

6. 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

7. Aran D, Hu Z, Butte AJ. xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* (2017) 18:220. doi: 10.1186/s13059-017-1349-1

8. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, Selves J, Laurent-Puig P, Sautès-Fridman C, Fridman WH, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biology* (2016) 17:218. doi: 10.1186/s13059-016-1070-5

9. Hänzelmann S, Castelo R, Guinney J. GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* (2013) 14:7. doi: 10.1186/1471-2105-14-7

10. Kocher F, Tymoszuk P, Amann A, Sprung S, Salcher S, Daum S, Haybaeck J, Rinnerthaler G, Huemer F, Kauffmann-Guerrero D, et al. Deregulated glutamate to pro-collagen conversion is associated with adverse outcome in lung cancer and may be targeted by renin-angiotensin-aldosterone system (RAS) inhibition. *Lung Cancer* (2021) 159:84–95. doi: 10.1016/j.lungcan.2021.06.020

11. 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

12. Friedman JH. Greedy function approximation: A gradient boosting machine. *https://doiorg/101214/aos/1013203451* (2001) 29:1189–1232. doi: 10.1214/AOS/1013203451

13. Greenwell B, Boehmke B, Cunningham J, Developers G. gbm: Generalized Boosted Regression Models. (2022) https://cran.r-project.org/package=gbm

14. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics* (2013) 7:63623. doi: 10.3389/FNBOT.2013.00021/BIBTEX

15. Friedman JH. Stochastic gradient boosting. *Computational Statistics & Data Analysis* (2002) 38:367–378. doi: 10.1016/S0167-9473(01)00065-2

16. Harrell FE, Lee KL, Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* (1996) 15:361–387. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4

17. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine* (1999) 18:2529–2545. doi: 10.1002/(sici)1097-0258(19990915/30)18:17/18<2529::aid-sim274>3.0.co;2-5