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

Supplementary material, transcriptome part

2024-01-31

Supplementary Methods

Software

The analysis was done with R version 4.2.3 (R Foundation).

Tabular data were handled with the packages *tidyverse* (1), *rlang* (2) and *trafo*. Text data were handled with *stringi* (3).

Import of the TCGA and DKFZ data sets from the cBioportal repository was accomplished with in-house-developed R scripts. Transcriptome datasets from the Gene Expression Omnibus were fetched with the *GEOquery* package (4). Gene and probe annotation was accomplished with the *AnnotationDbi* (5) and *org.Hs.eg.db* packages (6). For prediction of non-malignant cell counts and fractions in cancer samples, the R implementation of the *MCP Counter* and *xCell* algorithms provided by the *immunedeconv* package was used (7–9). Single sample gene set enrichment analysis scores (ssGSEA) were computed with the *GSVA* algorithm (10) implemented by the package *gseaTools*.

Semi-supervised clustering was done with the package *clustTools* employing algorithms from the *philentropy*, *cluster*, *factoextra* and *umap* packages (11–14).

For statistical hypothesis testing, effect size calculation, gene set variation analysis (GSVA) and differential gene expression analysis, the packages *rstatix* (15), *ExDA* and *microViz* were employed. Biological process gene ontology (GO) enrichment analysis was performed with the *goana* algorithm implemented by the *limma* (16) and *microViz* packages. Clustering of GO terms by their semantic similarity was accomplished with *GOSemSim* (17), *microViz* and *clustTools*. Differential modulation of transcriptional regulons and signaling pathways was analyzed with the *collecTRI* (18) and *PROGENy* (19) databases by linear modeling tools of the *decoupleR* package (20). Metabolic reaction activity modeling and metabolic subsystem enrichment analysis were performed with the packages *BiGGR* (21,22) and *biggrExtra*.

Multi-parameter modeling of biochemical relapse-free survival was done with the packages *glmnet* (23) (Ridge, Elastic Net and LASSO cox regression), *survivalsvm* (24) (Support Vector Machines [SVM]), *rfsrc* (25,26) (Random Forest), *gbm* (27) (Gradient Boosted Machines [GBM]), *survival* (28), *survminer* (29) and *coxExtensions* were used. Univariable analysis of relapse-free survival was accomplished with *kmOptimizer*.

For visualization of the results, the packages *ggplot2* (1) (scatter, stack, bubble and bar plots, heat maps), *ExDA* (violin, stack and ribbon plots), *survminer* and *coxExtensions* (Kaplan-Meier plots), and *microViz* (heat maps) were used. Figures and tables were created with the packages *cowplot* (30) and *flextable* (31).

Data sources and data import

Transcriptome cohorts for analysis in the current report were selected from studies deposited at cBioportal and Gene Expression Omnibus (GEO) with the following criteria: availability of information on Gleason scoring or ISUP risk strata (International Society of Urological Pathology), biochemical relapse and biochemical relapse-free survival, and expression data for 55 collagen-related genes investigated also in the proteomic part of the project (**Supplementary Table S1** and **S2**).

The TCGA prostate cancer data set (32,33) and DKFZ data set (34) consisted of normalized RNA sequencing data for 493 and 118 cancer samples, respectively, with accompanying clinical information and were obtained from the cBioportal repository with in-house-developed R scripts. The TCGA cohort included also 52 donor-matched samples of the benign prostate and prostate cancer tissue.
The GSE54460 (35) (RNA sequencing, n = 106 cancer samples), GSE70768 (36) (Illumina HumanHT-12 V4.0 expression beadchip microarray, n = 125), GSE70769 (36) (Illumina HumanHT-12 V4.0 expression beadchip microarray, n = 94), and GSE220095 (37) (RNA sequencing, n = 176) data sets included normalized whole-transcriptome and basic clinical information and were fetched from GEO with the `getGEO()` function from the *GEOquery* package. The GSE70768 data set (36) included also 73 donor-matched normal prostate and cancer tissue specimens. Demographic and clinical characteristic of the investigated transcriptome data sets is provided in **Supplementary Table S1**.

Gene expression levels were transformed with the $log_2(expression)$ and $log_2(count+1)$ function for the microarray and RNA sequencing data sets prior to further analyses. The *MCP Counter* and *xCell* estimates of non-malignant cell content in cancer samples were computed with the `deconvolute()` function from the *immunedeconv* package. Gene signatures of the Reactome pathways were obtained from the MSig database, version 7.5.1. ssGSEA scores were computed with the `calculate()` function from the *gseaTools* package. ISUP strata were defined based on Gleason Score information as follows: ISUP1: Gleason score 5 - 6, ISUP2: Gleason score 7, ISUP3+: Gleason score ≥ 8.

The investigated collagen-related genes included the published Collagen Signature (38) and the Reactome Collagen formation pathway genes (R-HSA-1474290) and were constrained to features covered by proteomic analyses of the current report (**Supplementary Table S2**).

Statistical hypothesis testing, effect size and statistical significance

Differences in numeric variables between two groups were investigated by paired and unpaired two-tailed T tests with Cohen's d effect size statistic or Mann-Whitney tests with r effect size statistic. Differences in numeric variables between three or more groups were

assessed by one-way ANOVA with η^2 effect size statistic. Differences in distribution of categorical variables were assessed by χ^2 test with Cramer's V effect size statistic. Odds ratio served as an effect size statistic in enrichment analyses. Effect sizes were interpreted as follows (39):

- Cohen's d, weak: 0.2 - 0.5, moderate: 0.5 - 0.8, large: ≥ 0.8
- r, weak: 0.1 - 0.3, moderate: 0.3 - 0.5, large: ≥ 0.5
- η^2: weak: 0.02 - 0.13, moderate: 0.13 - 0.26, large: ≥ 0.26
- Cramer's V, weak: 0.1 - 0.3, moderate: 0.3 - 0.5, large: ≥ 0.5
- OR, weak: 1.44 - 2.48, moderate: 2.48 - 4.27, large: ≥ 4.27

P values were corrected for multiple testing with the false discovery rate method (FDR) (40) within each analysis step and cohort. Effects were considered statistically significant for FDR-corrected p values < 0.05.

Comparison of gene expression between the benign and cancer tissue, and cancer samples stratified by the ISUP risk scheme

Differences in log_2-transformed expression levels of the collagen pathway genes between donor-matched cancer and benign prostate specimens were investigated with paired two-tailed T test with Cohen's effect size statistic in the GSE70768 and TCGA cohorts (**Figure 1B**, **Supplementary Figure S1** and **Supplementary Table S3**). Differences in gene expression between cancer samples stratified according to the ISUP risk scheme (ISUP1: Gleason Score 5 - 6, ISUP2: Gleason score 7, ISUP3+: Gleason Score ≥ 8) were explored by one-way ANOVA with η^2 effect size statistic (**Figure 1C**, **Supplementary Table S4**). Differentially regulated genes were defined by pFDR < 0.05 and at least weak effect size of the difference (d ≥ 0.2 or $\eta^2 \geq 0.02$). The analyses were done with the function `compare_variables()` from the *ExDA* package.

Collagen clusters of prostate cancer samples

Prostate cancer samples in the TCGA cohort were clustered by normalized log_2-transformed expression levels of the collagen-related genes of interest with the PAM (partition around medoids) algorithm with cosine distance between the samples (function `kcluster()`, package *clustTools*). The choice of the clustering algorithm was motivated by the following criteria (**Supplementary Figure S2A**):

- cluster separation measured by mean silhouette statistic (41)

- fraction of potentially misclassified observations (observations with negative silhouette widths) (41)
- explanatory performance gauged by fraction of explained clustering variance (ratio of the total between-cluster sum of squares to the total sum of squares)
- preservation of the nearest neighborhood (mean fraction of the five nearest neighbors placed in the same cluster) (42)
- cluster re-assignment accuracy in 5-fold cross-validation (43,44). Computation of those numeric statistics of clustering quality in a cross-validation setting was done with the methods `summary()` and `cv()` provided by the *clustTools* package. The cluster number choice was based on the bend of the peak of the average silhouette statistic (41) (method `plot()`, package *clustTools*, **Supplementary Figure S2B**). By this means, two clusters of cancer samples were identified: 'collagen high' and 'collagen low' cancers.

Assignment of cancers samples of the GSE54460, GSE70768, GSE70769, GSE220095, and DKFZ cohorts to the collagen clusters based on normalized log_2-transformed expression levels of the collagen-related genes was done with an inverse distance weighted k-nearest neighbor classifier (method `predict()`, package *clustTools*) (44). The quality of the cluster assignment was assessed by comparing the collagen cluster distribution and numeric statistics of clustering quality described above the training TCGA data set and the test cohorts (**Supplementary Figure S2CD**). In addition, separation of the clusters was assessed by a visual inspection of the UMAP layout plots (method `plot()`, package *clustTools*) and heat maps of the mean levels of the clustering factors in the collagen clusters (**Supplementary Figure S3**).

Differences in the log_2-transformed collagen pathway gene expression between the collagen clusters were investigated by two-tailed T test with Cohen's d effect size statistic (**Supplementary Table S5**).

Clinical characteristic of the collagen clusters

Differences in numeric clinical variables (age and PSA) between the collagen clusters were assessed by Mann-Whitney test with r effect size statistic (`test_two_groups()`, *microViz*). Differences in qualitative clinical variables between the collagen clusters were investigated by χ^2 test with Cramer's V effect size statistic (function `compare_variables()`, package *ExDA*). The demographic and clinical characteristic of the collagen clusters is presented in **Supplementary Figure S4** and **Supplementary Table S6**.

Non-malignant cell infiltration in the collagen clusters

Non-malignant cell counts and non-malignant cell fractions in the cancer samples were predicted by the *MCP Counter* (9) and *xCell* (8) algorithms, respectively. Differences in the predicted infiltration levels between the collagen clusters were investigated by Mann-Whitney test with r effect size statistic (function `compare_variables()`, package *ExDA*). Populations differing between the clusters with pFDR < 0.05 and at least weak effect size (r ≥ 0.1) were considered biologically relevant. The analysis results are shown in **Supplementary Figure S5** and **Supplementary Table S7**.

Gene set variance analysis of Reactome pathway gene signatures

ssGSEA scores (10) of the Reactome pathway gene signatures were compared between the collagen clusters by two-tailed T test with Cohen's d effect size statistic. Signatures found to be significantly regulated (pFDR < 0.05) with at least weak effect size (d ≥ 0.2) in at least five cohorts were further investigated. To get additional functional insight and for purposes of visualization, these common regulated signatures were classified by their co-expression patterns in the TCGA data set by unsupervised KMEANS clustering (function `kcluster()`, package *clustTools*). The analysis results are presented in **Supplementary Figure S6** and **Supplementary Table S8**.

Differential gene expression and GO enrichment analysis

Differences in whole-transcriptome gene expression between the collagen clusters were investigated by FDR-corrected two-tailed T test with Cohen's d effect size statistic (function `test_two_groups()`, package *microViz*). Genes significantly differentially expressed between the clusters were identified by the significance cutoff (pFDR < 0.05) and at least weak effect size of the expression differences (d ≥ 0.2). Genes found to be differentially regulated between the clusters are listed in **Supplementary Table S9**.

Biological process GO enrichment within genes upregulated in each of the collagen high and collagen low clusters was analyzed with the *limma* package *goana* algorithm (16), re-implemented by the development package *microViz* (function `GOana()`). Enrichment p values were adjusted for multiple testing with the FDR method. Odds ratio (OR) calculated for a GO term with the following formula served as a measure of effect size:

$$OR_{GO} = \frac{N_{GO \vee DGE} \times N_{total}}{N_{GO \vee total} \times N_{DGE}}$$

where $N_{GO \vee DGE}$ stands for the number of differentially regulated genes assigned to the GO term, N_{total} is the total number of investigated genes, $N_{GO \vee total}$ is the number of investigated genes assigned to the GO term, and N_{DGE} stands for the number of

differentially regulated genes. GO terms found to be significantly enriched with at least weak effect size (pFDR < 0.05, OR ≥ 1.44) in at least five cohorts were further investigated. For additional functional insight, such common enriched GO terms were subjected to multi-dimensional scaling and unsupervised KMEANS clustering with pairwise Wang distance as a measure of semantic similarity (functions `go_sem()` and `kcluster()`, packages *GOSemSim*, *microViz* and *clustTools*) (17). Results of the GO enrichment analysis are shown in **Supplementary Figure S6** and **Supplementary Table S10**.

Transcriptional regulons and signaling pathway activity in the collagen clusters

Modulation of transcriptional regulons, i.e. sets of genes whose expression is controlled by a common transcriptional factor, and activity of selected signaling pathways in collagen high cancers as compared with collagen low tumors were explored by linear modeling with the *collecTRI* (18) and *PROGENy* (19) databases, respectively. Liner models were constructed for whole-transcriptome effect sizes of gene regulation (Cohen's d, see: **Differential gene expression and GO enrichment analysis**) with the `run_ulm()` and `run_mlm()` tools from the *decoupleR* package (20). In each case, linear model score (LM score) served as a metric of activity magnitude. P values (LM score ≠ 0) were corrected for multiple testing with the FDR method. Regulons and signaling pathways found to be significantly activated or inhibited in at least five cohorts were further discussed. The analysis results are presented in **Supplementary Figure S7**,and **Supplementary Tables S11** and **S12**.

Activity of metabolic reactions in the collagen clusters

Rules of assignment of genes to biochemical reactions were retrieved from the Recon2 human metabolism model available via the BiGG database (22) and the R package *BiGGR* (21). Estimates and standard errors of log_2 fold-regulation of expression for all available genes between the collagen clusters were calculated by two-tailed T test as described in **Differential gene expression and GO enrichment analysis**. Estimates of biochemical pathway fold-regulation were computed by evaluation of the gene assignment rules in the Recon2 model. The '$gene_A$ OR $gene_B$' operator was interpreted as arithmetic mean of expression regulation estimates for the genes A and B. The '$gene_A$ AND $gene_B$' operator was interpreted as minimum of expression regulation estimates for the gene A and gene B (21). Standard deviation, 95% confidence intervals and p values for the predicted reaction regulation estimates were obtained by a Monte Carlo simulation with n = 3000 draws from normal distribution of gene expression regulation estimates (mean: gene expression regulation estimate, standard deviation: standard error of the gene expression regulation estimate) (21,45). P values were corrected for multiple testing with the FDR method. The analysis was done with the packages *BiGGR* (21) and *biggrExtra* (45) (function

`build_geneSBML()`). Biochemical reactions predicted to be significantly activated or inhibited in the collagen high cluster as compared with the collagen low cluster are listed in **Supplementary Table S13**.

Enrichment of significantly activated or inhibited biochemical reactions within the Recon model metabolic subsystems (22) was investigated by comparing the frequency of significantly activated or inhibited reactions in the given subsystem with 10000 random samples from the total reaction pool (function `suba()`, package *biggrExtra*). The magnitude of enrichment was assessed with OR defined for the i-th metabolic subsystem as follows:

$$OR_i = \frac{n_{regulated,i} \times n_{total}}{n_i \times n_{regulated}}$$

where $n_{regulated,i}$ denotes the number of significantly activated or inhibited biochemical reactions within the i-th metabolic subsystem, n_{total} denotes the total number of investigated reactions, n_i denotes the number of metabolic reactions within the i-th metabolic subsystem, and $n_{regulated}$ is the total number of significantly activated or inhibited biochemical reactions. Common significantly activated or inhibited metabolic subsystems were identified as metabolic subsystems significantly activated or inhibited in at least four cohorts, with at least weak effect size defined by OR ≥ 1.44 (**Supplementary Figure S8**, **Supplementary Table S14**).

Multi-parameter survival modeling and transcriptional collagen score

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 (46). 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 GBM algorithm (27,47–49). Selection of the optimal set of modeling parameters (number of trees, shrinkage factor, interaction depth, and minimal number of observations in the tree node) was accomplished by tuning with 10-fold cross-validation and out-of-fold model deviance as performance statistic (**Supplementary Table S15**). The tuning was performed with an in-house developed script. The GBM model was trained in the pooled GEO cohort with the optimal tuning parameter set with the `gbm()` function (package *gbm*). The Transcriptomic Collagen Score was defined as the linear predictor score of the GBM model and was computed for the best model iteration (function `gbm.perf(method = 'cv')`) with the `predict()` method and expression data from the training pooled GEO cohort and the TCGA and DKFZ test collectives. Importance of the explanatory variables for the predictive performance of the GBM model in the training

pooled GEO cohort were measured with the relative influence method (47) and expressed as gradient of the sum of squared errors attributed to particular variables (ΔSSE, **Figure 3A**).

Performance of the Transcriptional Collagen Score at prediction of biochemical relapse-free survival was assessed by uni-variable Cox regression (functions `coxph()` and `as_coxex()`, packages *survival* and *coxExtensions*) (28). The proportional hazard assumption of the Cox models was checked with the `summary(type = 'assumptions')` method, which implements the genuine `cox.zph()` algorithm (50). Harrell's concordance index (51) and integrated Brier score (52) were employed to investigate the overall concordance between the predicted and observed survival and to assess overall calibration (method `summary(type = 'fit')`, package *coxExtensions*, **Figure 3B** and **Supplementary Table S16**). Of note, as compared with several other machine learning algorithms (Ridge, Elastic Net, and LASSO Cox regression, SVM and Random Forest) (23–26), the GBM model demonstrated an excellent performance and the highest reproducibility at prediction of biochemical relapse-free survival in the investigated cohorts (**Supplementary Figure S9**, **Supplementary Table S16**). Statistical significance of differences in biochemical relapse-free survival between patients stratified by tertiles of the Transcriptional Collagen Score was determined by Peto-Peto test (functions `surv_fit()` and `surv_pvalue()`, package *survminer*, **Figure 3C**).

Survival analysis for single collagen-related transcripts

The univariable analysis of biochemical relapse-free survival for single collagen-related genes was done in the pooled GEO, TCGA and DKFZ data sets described above for the multi-parameter survival modeling. For each of the collagen-related genes, cancer patients were classified as high and low expressors by an expression cutoff corresponding to the largest difference in biochemical relapse-free survival measured by FDR-corrected Mentel-Henszel test (function `find_cutoff()`, package *kmOptimizer*, minimal expression strata size set to 25% of the observations). Hazard ratio (HR) with 95% confidence intervals for relapse risk in high expressors as compared with low expressors was calculated with the canonical *survival* package functions *coxph()*, *coef* and *confint()*. Genes associated with unfavorable prognosis were defined by pFDR < 0.05 and HR > 1 for the high vs low expressor comparison. Genes associated with favorable prognosis were defined by pFDR < 0.05 and HR < 1 (**Supplementary Figure S10**). Common survival markers were defined as genes significantly associated with survival in all cohorts (**Supplementary Figure S11** - **S14**, **Supplementary Table S17**).

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).

Supplementary Tables

Supplementary Table S1: Characteristic of the analyzed cohorts. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Qualitative variables are presented as percentages of categories within the complete observation set.

Variable[a]	GSE54460	GSE70768	GSE70769	GSE220095	TCGA	DKFZ
Age at diagnosis, years		62 [IQR: 57 - 67] range: 41 - 93 n = 124			61 [IQR: 56 - 66] range: 41 - 78 n = 493	48 [IQR: 46 - 49] range: 32 - 52 n = 118
PSA at diagnosis	7.2 [IQR: 5.5 - 13] range: 1.8 - 73 n = 103	8 [IQR: 6 - 12] range: 3.2 - 280 n = 123	8 [IQR: 5.9 - 11] range: 1.5 - 120 n = 90	8.2 [IQR: 5.6 - 14] range: 1 - 120 n = 176	0.1 [IQR: 0.03 - 0.11] range: 0 - 320 n = 436	8.1 [IQR: 5.9 - 23] range: 1.9 - 740 n = 116
Clinical tumor stage		T1: 56% (n = 62) T2: 30% (n = 33) T3: 14% (n = 16) n = 111	T1: 46% (n = 41) T2: 44% (n = 39) T3: 10% (n = 9) n = 89			
Pathological tumor stage	T1: 13% (n = 14) T2: 70% (n = 73) T3: 16% (n = 17) T4: 0.95% (n = 1) n = 105	T2: 31% (n = 34) T3: 68% (n = 76) T4: 0.9% (n = 1) n = 111	T2: 53% (n = 48) T3: 47% (n = 42) n = 90	T2: 66% (n = 117) T3: 26% (n = 46) T4: 7.4% (n = 13) n = 176	T1: 0% (n = 0) T2: 38% (n = 186) T3: 60% (n = 290) T4: 2.1% (n = 10) n = 486	T1: 0% (n = 0) T2: 64% (n = 74) T3: 30% (n = 35) T4: 6% (n = 7) n = 116
Pathological node stage		N0: 91% (n = 82) N1: 8.9% (n = 8) n = 90	N0: 100% (n = 18) n = 18	N0: 87% (n = 146) N1: 13% (n = 22) n = 168	N0: 81% (n = 342) N1: 19% (n = 78) n = 420	
Pathological metastasis stage		M0: 86% (n = 6)	M0: 87% (n = 26)			

Variable[a]	GSE54460	GSE70768	GSE70769	GSE220095	TCGA	DKFZ
		M1: 14% (n = 1) n = 7	M1: 13% (n = 4) n = 30			
Gleason score	5: 0.94% (n = 1) 6: 9.4% (n = 10) 7: 75% (n = 80) 8: 9.4% (n = 10) 9: 4.7% (n = 5) n = 106	1: 1.6% (n = 2) 6: 14% (n = 17) 7: 71% (n = 87) 8: 7.4% (n = 9) 9: 5.7% (n = 7) n = 122	1: 1.1% (n = 1) 5: 2.2% (n = 2) 6: 20% (n = 18) 7: 62% (n = 56) 8: 5.5% (n = 5) 9: 9.9% (n = 9) n = 91	1: 0.57% (n = 1) 6: 20% (n = 35) 7: 68% (n = 120) 8: 2.8% (n = 5) 9: 8.5% (n = 15) n = 176	7 [IQR: 7 - 9] range: 6 - 10 n = 493	6: 11% (n = 13) 7: 74% (n = 87) 8: 0.85% (n = 1) 9: 14% (n = 16) 10: 0.85% (n = 1) n = 118
	5 - 6: 10% (n = 11) 7: 75% (n = 80) 8+: 14% (n = 15) n = 106	5 - 6: 16% (n = 19) 7: 71% (n = 87) 8+: 13% (n = 16) n = 122	5 - 6: 23% (n = 21) 7: 62% (n = 56) 8+: 15% (n = 14) n = 91	5 - 6: 20% (n = 36) 7: 68% (n = 120) 8+: 11% (n = 20) n = 176	5 - 6: 9.1% (n = 45) 7: 50% (n = 245) 8+: 41% (n = 203) n = 493	5 - 6: 11% (n = 13) 7: 74% (n = 87) 8+: 15% (n = 18) n = 118
Surgical margins	negative: 60% (n = 61) positive: 40% (n = 40) n = 101	negative: 78% (n = 93) positive: 22% (n = 26) n = 119	negative: 55% (n = 51) positive: 45% (n = 42) n = 93			
Extracapsular extension			46% (n = 42) n = 91			
Death					2% (n = 10) n = 493	
Overall survival, months					30 [IQR: 17 - 48] range: 0.76 - 170 n = 493	
Biochemical	52% (n =	17% (n =	48% (n =	43% (n =	19% (n =	23% (n =

Variable[a]	GSE54460	GSE70768	GSE70769	GSE220095	TCGA	DKFZ
relapse	55) n = 106	19) n = 112	45) n = 93	75) n = 176	93) n = 493	24) n = 105
Biochemical relapse-free survival, months	49 [IQR: 18 - 77] range: 0 - 170 n = 106	30 [IQR: 17 - 49] range: 1 - 65 n = 111	58 [IQR: 19 - 80] range: 0.36 - 100 n = 92	75 [IQR: 46 - 110] range: 0.66 - 130 n = 176	26 [IQR: 14 - 45] range: 0.76 - 170 n = 493	36 [IQR: 13 - 49] range: 0.5 - 76 n = 105

[a]PSA: prostate-specific antigen.

Supplementary Table S2: Collagen-related genes and their classification.

Gene group	Gene symbol	Entrez ID
proline turnover	*ALDH18A1*	5832
	PEPD	5184
	PYCR1	5831
collagen modification	*LOX*	4015
	LOXL1	4016
	LOXL2	4017
	P4HA1	5033
	P4HA2	8974
	P4HB	5034
	PLOD1	5351
	PLOD2	5352
	PLOD3	8985
	PPIB	5479
ECM component	*COL11A1*	1301
	COL11A2	1302
	COL14A1	7373
	COL15A1	1306
	COL16A1	1307
	COL17A1	1308
	COL18A1	80781
	COL19A1	1310
	COL1A1	1277
	COL1A2	1278
	COL2A1	1280
	COL3A1	1281

Gene group	Gene symbol	Entrez ID
	COL4A1	1282
	COL4A2	1284
	COL4A3	1285
	COL4A5	1287
	COL4A6	1288
	COL5A1	1289
	COL5A2	1290
	COL6A1	1291
	COL6A2	1292
	COL6A3	1293
	COL7A1	1294
	COL9A1	1297
	COL9A2	1298
	COL9A3	1299
	LAMA3	3909
	LAMB3	3914
	LAMC2	3918
ECM processing	*ADAMTS2*	9509
	BMP1	649
	CTSS	1520
	MMP13	4322
	MMP7	4316
	MMP9	4318
	PCOLCE	5118

Gene group	Gene symbol	Entrez ID
	PCOLCE2	26577
	SERPINH1	871
adhesion	*CD151*	977
	DST	667
	ITGA6	3655
	ITGB4	3691

Supplementary Table S3: Expression of the collagen pathway genes in the malignant and benign tissue compared by paired T test with Cohen's d effect size statistic. P values were corrected for multiple testing with the false discovery rate method. log2-transformed expression values are presented as medians with interquartile ranges (IQR) and ranges. The table is available as a supplementary Excel file.

Supplementary Table S4: Expression of the collagen pathway genes in cancer samples stratified by the ISUP risk groups compared by one-way ANOVA with eta-square effect size statistic. P values were corrected for multiple testing with the false discovery rate method. log2-transformed expression values are presented as medians with interquartile ranges (IQR) and ranges. The table is available as a supplementary Excel file.

Supplementary Table S5: Expression of the cluster-defining collagen pathway genes in the collagen clusters of prostate cancer. Statistical significance was assessed by two-tailed T test with Cohen's d effect size statistic. P values were corrected for multiple testing with the false discovery rate method. log2-transformed expression values are presented as medians with interquartile ranges (IQR) and ranges. The table is available as a supplementary Excel file.

Supplementary Table S6: Clinical characteristic of the collagen clusters. Numeric variables are presented as medians with interquartile ranges (IQR) and ranges. Nominal variables are presented as percentages and counts of categories within the cluster.

Cohort	Variable[a]	Collagen low	Collagen high	Significance[b]	Effect size[b]
GSE54460	PSA at diagnosis	7.1 [IQR: 5.3 - 13] range: 1.8 - 40 n = 57	7.2 [IQR: 5.6 - 11] range: 1.8 - 73 n = 46	ns (p = 1)	r = 0.0085
	Pathological tumor stage	T1: 1.7% (n = 1) T2: 81% (n = 48) T3: 15% (n = 9) T4: 1.7% (n = 1) n = 59	T1: 28% (n = 13) T2: 54% (n = 25) T3: 17% (n = 8) T4: 0% (n = 0) n = 46	p = 0.0025	V = 0.41
	ISUP	ISUP1: 13% (n = 7) ISUP2: 87% (n = 47) n = 54	ISUP1: 11% (n = 4) ISUP2: 89% (n = 33) n = 37	ns (p = 1)	V = 0.032
	Surgical margins	negative: 61% (n = 34) positive: 39% (n = 22) n = 56	negative: 60% (n = 27) positive: 40% (n = 18) n = 45	ns (p = 1)	V = 0.0073
GSE70768	Age at diagnosis, years	62 [IQR: 56 - 65] range: 42 - 73 n = 62	63 [IQR: 58 - 69] range: 41 - 93 n = 62	ns (p = 0.17)	r = 0.17
	PSA at diagnosis	7.1 [IQR: 5.8 - 11] range: 4 - 19 n = 62	8.7 [IQR: 7 - 14] range: 3.2 - 280 n = 61	ns (p = 0.17)	r = 0.19
	Clinical tumor stage	T1: 57% (n = 35) T2: 30% (n = 18) T3: 13% (n = 8) n = 61	T1: 54% (n = 27) T2: 30% (n = 15) T3: 16% (n = 8) n = 50	ns (p = 0.9)	V = 0.044
	Pathological tumor stage	T2: 34% (n = 21) T3: 66% (n = 40) T4: 0% (n = 0) n = 61	T2: 26% (n = 13) T3: 72% (n = 36) T4: 2% (n = 1) n = 50	ns (p = 0.64)	V = 0.14
	Pathological node stage	N0: 89% (n = 42) N1: 11% (n = 5) n = 47	N0: 93% (n = 40) N1: 7% (n = 3) n = 43	ns (p = 0.9)	V = 0.064
	ISUP	ISUP1: 11% (n = 6) ISUP2: 89% (n = 50) n = 56	ISUP1: 26% (n = 13) ISUP2: 74% (n = 37) n = 50	ns (p = 0.17)	V = 0.2
	Surgical margins	negative: 81% (n = 50)	negative: 75% (n = 43)	ns (p = 0.9)	V = 0.063

Cohort	Variable[a]	Collagen low	Collagen high	Significance[b]	Effect size[b]
		positive: 19% (n = 12) n = 62	positive: 25% (n = 14) n = 57		
GSE70769	PSA at diagnosis	6.9 [IQR: 5.1 - 11] range: 2.2 - 35 n = 42	8.6 [IQR: 6.4 - 12] range: 1.5 - 120 n = 48	ns (p = 0.34)	r = 0.19
	Clinical tumor stage	T1: 50% (n = 21) T2: 36% (n = 15) T3: 14% (n = 6) n = 42	T1: 43% (n = 20) T2: 51% (n = 24) T3: 6.4% (n = 3) n = 47	ns (p = 0.34)	V = 0.18
	Pathological tumor stage	T2: 61% (n = 27) T3: 39% (n = 17) n = 44	T2: 46% (n = 21) T3: 54% (n = 25) n = 46	ns (p = 0.34)	V = 0.16
	Pathological node stage	N0: 100% (n = 7) n = 7	N0: 100% (n = 11) n = 11	ns (p = 0.4)	V = Inf
	ISUP	ISUP1: 35% (n = 14) ISUP2: 65% (n = 26) n = 40	ISUP1: 19% (n = 7) ISUP2: 81% (n = 30) n = 37	ns (p = 0.34)	V = 0.18
	Surgical margins	negative: 57% (n = 25) positive: 43% (n = 19) n = 44	negative: 53% (n = 26) positive: 47% (n = 23) n = 49	ns (p = 0.88)	V = 0.038
	Extracapsular extension	39% (n = 17) n = 44	53% (n = 25) n = 47	ns (p = 0.34)	V = 0.15
GSE220095	PSA at diagnosis	8.4 [IQR: 5.7 - 14] range: 1 - 77 n = 78	8.1 [IQR: 5.4 - 14] range: 2.4 - 120 n = 98	ns (p = 0.83)	r = 0.016
	Pathological tumor stage	T2: 65% (n = 51) T3: 24% (n = 19) T4: 10% (n = 8) n = 78	T2: 67% (n = 66) T3: 28% (n = 27) T4: 5.1% (n = 5) n = 98	ns (p = 0.55)	V = 0.1
	Pathological node stage	N0: 91% (n = 68) N1: 9.3% (n = 7) n = 75	N0: 84% (n = 78) N1: 16% (n = 15) n = 93	ns (p = 0.55)	V = 0.1
	ISUP	ISUP1: 14% (n = 10) ISUP2: 86% (n = 62) n = 72	ISUP1: 31% (n = 26) ISUP2: 69% (n = 58) n = 84	ns (p = 0.079)	V = 0.2
TCGA	Age at diagnosis, years	61 [IQR: 56 - 66] range: 44 - 78	62 [IQR: 57 - 67] range: 41 - 77	ns (p = 0.24)	r = 0.075

Cohort	Variable[a]	Collagen low	Collagen high	Significance[b]	Effect size[b]
		n = 236	n = 257		
	PSA at diagnosis	0.1 [IQR: 0.03 - 0.11] range: 0 - 320 n = 206	0.1 [IQR: 0.03 - 0.12] range: 0 - 37 n = 230	ns (p = 0.88)	r = 0.018
	Pathological tumor stage	T2: 44% (n = 102) T3: 54% (n = 125) T4: 2.6% (n = 6) n = 233	T2: 33% (n = 84) T3: 65% (n = 165) T4: 1.6% (n = 4) n = 253	ns (p = 0.16)	V = 0.12
	Pathological node stage	N0: 82% (n = 161) N1: 18% (n = 36) n = 197	N0: 81% (n = 181) N1: 19% (n = 42) n = 223	ns (p = 0.98)	V = 0.0072
	ISUP	ISUP1: 17% (n = 26) ISUP2: 83% (n = 129) n = 155	ISUP1: 14% (n = 19) ISUP2: 86% (n = 116) n = 135	ns (p = 0.88)	V = 0.037
DKFZ	Age at diagnosis, years	48 [IQR: 45 - 49] range: 38 - 52 n = 61	48 [IQR: 46 - 49] range: 32 - 52 n = 57	ns (p = 0.7)	r = 0.056
	PSA at diagnosis	8.4 [IQR: 5.8 - 16] range: 1.9 - 740 n = 60	7.7 [IQR: 6 - 39] range: 3.2 - 150 n = 56	ns (p = 0.7)	r = 0.065
	Pathological tumor stage	T2: 68% (n = 41) T3: 32% (n = 19) T4: 0% (n = 0) n = 60	T2: 59% (n = 33) T3: 29% (n = 16) T4: 12% (n = 7) n = 56	ns (p = 0.073)	V = 0.26
	ISUP	ISUP1: 11% (n = 6) ISUP2: 89% (n = 49) n = 55	ISUP1: 16% (n = 7) ISUP2: 84% (n = 38) n = 45	ns (p = 0.7)	V = 0.069

[a]PSA: prostate-specific antigen; ISUP: International Society of Urological Pathology risk stratification system.

[b]Qualitative variables: χ^2 test with Cramer V effect size statistic. Numeric variables: Mann-Whitney test with r effect size statistic. P values corrected for multiple testing with the false discovery rate.

Supplementary Table S7: Non-malignant cell numbers predicted for the collagen clusters by the MCP Counter and xCell algorithms. Statistical significance was assessed by Mann-Whitney test with r effect size statistic. P values were corrected for multiple testing with the false discovery method. The table is available as a supplementary Excel file.

Supplementary Table S8: Gene set variation analysis with the Reactome pathway gene signatures. Differences in ssGSEA scores between collagen high and collagen low cancers were investigated by two-tailed T test with Cohen's d effect size statistic. Results for signatures significantly regulated with at least weak effect size (d at least 0.2) in at least five cohorts are presented. P values were corrected for multiple testing with the false discovery rate method (FDR). The table is available as a supplementary Excel file.

Supplementary Table S9: Genes differentially expressed in the collagen high cluster as compared with collagen low cancers were identified by two-tailed T test with the 1.25-fold regulation cutoff and the Cohen's d effect size statistic of 0.2 P values were corrected for multiple testing with the false discovery rate method (FDR). The table is available as a supplementary Excel file.

Supplementary Table S10: Biological process gene ontology (GO) term enrichment within genes differentially regulated in the collagen clusters. The enrichment analysis was performed with goana tool, enrichment p values were corrected for multiple testing with the false discovery rate (FDR) method. Significant enrichment was defined by pFDR < 0.05 and odds ratio (OR) for enrichment within differentially regulated genes of at least 1.44. OR for enrichment within genes upregulated in the collagen high and collagen low clusters are presented for significant GO terms shared by at least five cohorts. The table is available as a supplementary Excel file.

Supplementary Table S11: Activity of transcriptional regulons in the collagen high cluster as compared with the collagen low cluster predicted by the collecTRI model. Regulon activity was estimated with uni-parameter linear modeling with whole-transcriptome effect sizes of differential gene expression, p values were corrected for multiple testing with the false discovery rate (FDR) method. Linear model scores are presented for regulons significantly activated or inhibited in at least five cohorts. The table is available as a supplementary Excel file.

Supplementary Table S12: Activity of signaling pathways in the collagen high cluster as compared with the collagen low cluster predicted by the PROGENy model. Pathway activity was estimated with multi-parameter linear modeling with whole-transcriptome effect sizes of differential gene expression. P values were corrected for multiple testing with the false discovery rate (FDR) method, linear model scores serve as measures of pathway activity. The table is available as a supplementary Excel file.

Supplementary Table S13: Biochemical reactions predicted to be significantly activated in collagen high as compared with collagen low cancers. Statistical significance was determined by a Monte Carlo simulation. P values were corrected for multiple testing with the false discovery rate (FDR) method. The table is available as a supplementary Excel file.

Supplementary Table S14: Results of enrichment analysis for significantly activated and inhibited biochemical reactions within the Recon metabolism subsystem. Statistical significance was determined by random sampling from the entire reaction pool p values were corrected for multiple testing with the false discovery rate (FDR) method. Effect size of enrichment of the subsystem in significantly activated or inhibited reactions was measured by odds ratio (OR) statistic. The table is available as a supplementary Excel file.

Supplementary Table S15: Selection of the optimal parameters of machine learining survival models of biochemical relapse-free survival with expression of the collagen-related genes. The selection process was accomplished by cross-validation tuning in the pooled GEO cohort.

Algorithm[a]	**Selection criterion**[b]	**Parameter**	**Value**
Ridge Cox	minimal deviance, repeated 10-fold CV	λ	0.466
Elastic Net Cox	minimal deviance, repeated 10-fold CV	λ	0.0291
LASSO Cox	minimal deviance, repeated 10-fold CV	λ	0.016
SVM	maximal concordance index, repeated 10-fold CV	SVM model type	vanbelle1
		γ	0.01
		kernel	add_kernel
Random Forest	maximal concordance index, out-of-bag predictions	number of variables per try, mtry	20
		splitting rule	logrank
		number of splits	2
		minimal node size	5
GBM	minimal deviance, 10-fold CV	number of decision trees	1000
		shrinkage	0.035
		interaction depth	2
		minimal node size	5

[a]SVM: Support Vector Machines; GBM: Gradient Boosted Machines.

[b]CV: cross-validation.

Supplementary Table S16: Performance of machine learning models at prediction of biochemical relapse-free survival with expression of the collagen-related genes.

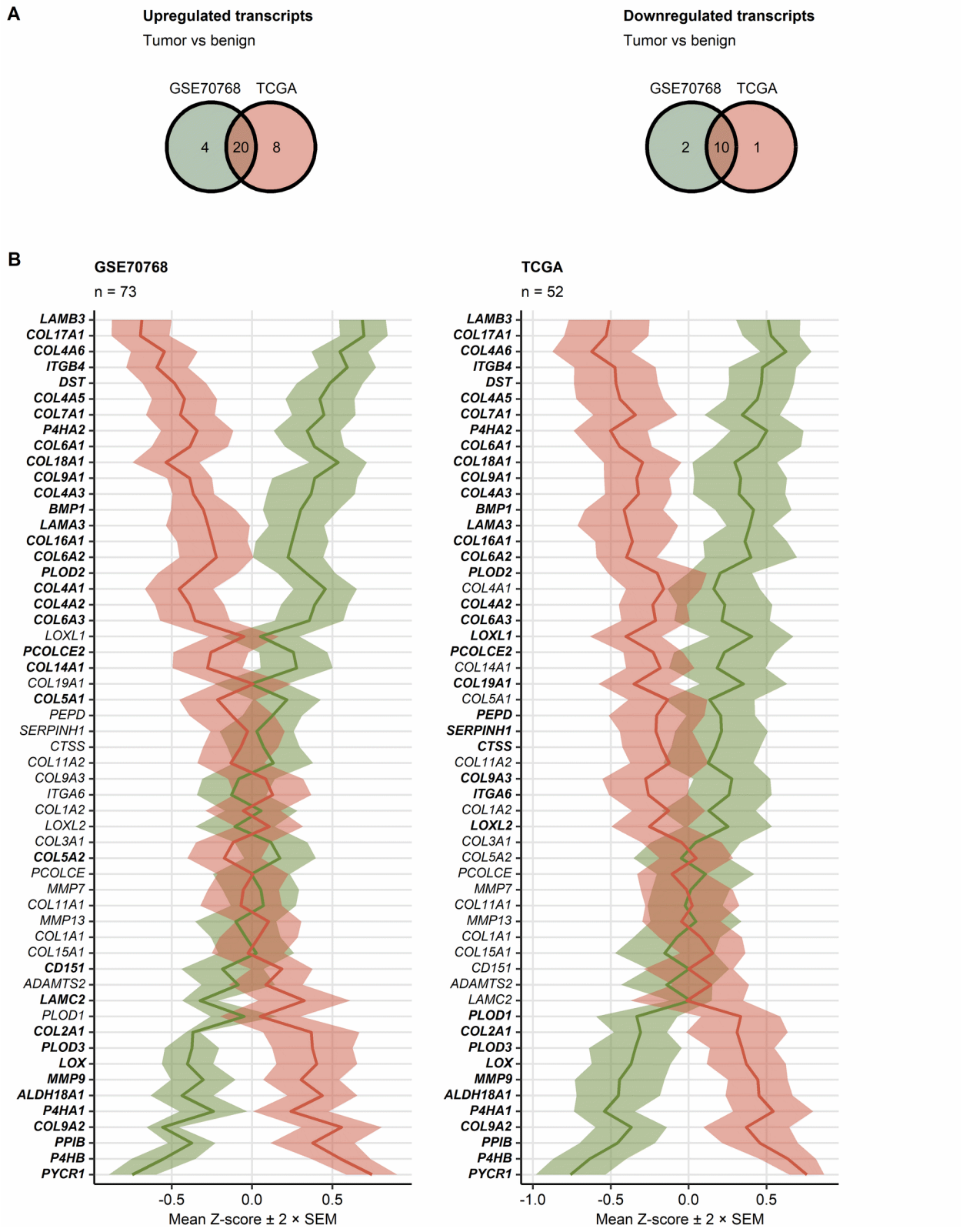
Algorithm[a]	Data set type	Cohort[b]	Concordance index	Integrated Brier score
Ridge Cox	training	pooled GEO	0.740	0.1770
	test	TCGA	0.656	0.1900
		DKFZ	0.791	0.1150
Elastic Net Cox	training	pooled GEO	0.747	0.1690
	test	TCGA	0.644	0.1910
		DKFZ	0.769	0.1210
LASSO Cox	training	pooled GEO	0.747	0.1680
	test	TCGA	0.646	0.1910
		DKFZ	0.774	0.1210
SVM	training	pooled GEO	0.632	0.1850
	test	TCGA	0.592	0.2000
		DKFZ	0.702	0.1380
Random Forest	training	pooled GEO	0.942	0.0439
	test	TCGA	0.660	0.2380
		DKFZ	0.820	0.3730
GBM	training	pooled GEO	0.878	0.0986
	test	TCGA	0.654	0.1850
		DKFZ	0.802	0.1190

[a]SVM: Support Vector Machines; GBM: Gradient Boosted Machines.

[b]pooled GEO: GSE54460, GSE70768, GSE70769, and GSE220095

Supplementary Table S17: Results of univariable analysis of biochemical relapse-free survival with expression of the collagen-related genes. Prostate cancer patients were stratified by expression cutoffs corresponding to the largest difference in survival assessed by Mentel-Henszel test. Genes found to be significantly associated with the survival in all analyzed cohorts (pooled GEO, TCGA and DKFZ) are presented. The table is available in a supplementary Excel file.

Supplementary Figures

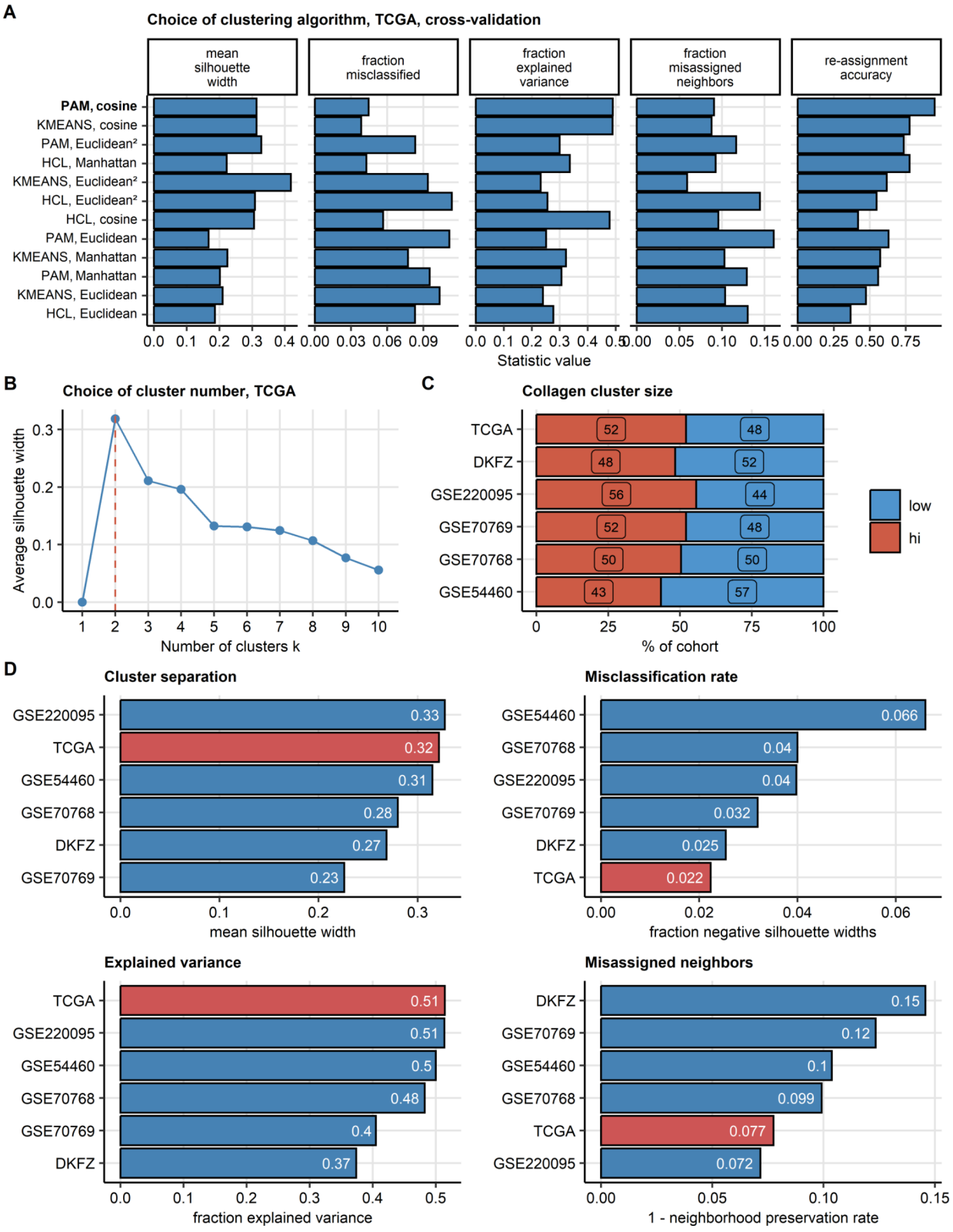

A
Upregulated transcripts
Tumor vs benign
GSE70768
TCGA
4
20
8
Downregulated transcripts
Tumor vs benign
GSE70768
TCGA
2
10
1
B
GSE70768
n = 73
TCGA
n = 52
LAMB3
COL17A1
COL4A6
ITGB4
DST
COL4A5
COL7A1
P4HA2
COL6A1
COL18A1
COL9A1
COL4A3
BMP1
LAMA3
COL16A1
COL6A2
PLOD2
COL4A1
COL4A2
COL6A3
LOXL1
PCOLCE2
COL14A1
COL19A1
COL5A1
PEPD
SERPINH1
CTSS
COL11A2
COL9A3
ITGA6
COL1A2
LOXL2
COL3A1
COL5A2
PCOLCE
MMP7
COL11A1
MMP13
COL1A1
COL15A1
CD151
ADAMTS2
LAMC2
PLOD1
COL2A1
PLOD3
LOX
MMP9
ALDH18A1
P4HA1
COL9A2
PPIB
P4HB
PYCR1
-0.5
0.0
0.5
-1.0
Mean Z-score ± 2 × SEM

Supplementary Figure S1. Expression of collagen-related genes in the normal prostate and prostate cancer tissue.

Differences in log_2-transformed expression of 55 collagen-related genes between donor-matched pairs of the prostate cancer and benign tissue were assessed by paired T test with Cohen's effect size statistic in the GSE70768 and TCGA cohorts. P values were corrected for multiple testing with the false discovery rate (FDR) method. Full analysis results are listed in Supplementary Table S3.

(A) Numbers of significantly up- and downregulated genes in the tumor tissue as compared with the benign tissue in the investigated cohorts presented in Venn plots.

(B) Mean normalized log_2 expression of the collagen-related genes presented as lines. Tinted regions represent the 2 × SEM (standard error of the mean) intervals. Significant effects are highlighted with bold font in the Y axis. Numbers of tissue pairs are displayed in the plot captions.


A
Choice of clustering algorithm, TCGA, cross-validation
mean silhouette width
fraction misclassified
fraction explained variance
fraction misassigned neighbors
re-assignment accuracy
PAM, cosine
KMEANS, cosine
PAM, Euclidean²
HCL, Manhattan
KMEANS, Euclidean²
HCL, Euclidean²
HCL, cosine
PAM, Euclidean
KMEANS, Manhattan
PAM, Manhattan
KMEANS, Euclidean
HCL, Euclidean
Statistic value
B
Choice of cluster number, TCGA
Average silhouette width
Number of clusters k
C
Collagen cluster size
TCGA
DKFZ
GSE220095
GSE70769
GSE70768
GSE54460
low
hi
% of cohort
D
Cluster separation
mean silhouette width
Misclassification rate
fraction negative silhouette widths
Explained variance
fraction explained variance
Misassigned neighbors
1 - neighborhood preservation rate

Supplementary Figure S2. Semi-supervised clustering of prostate cancer samples in respect to expression of collagen-related genes.

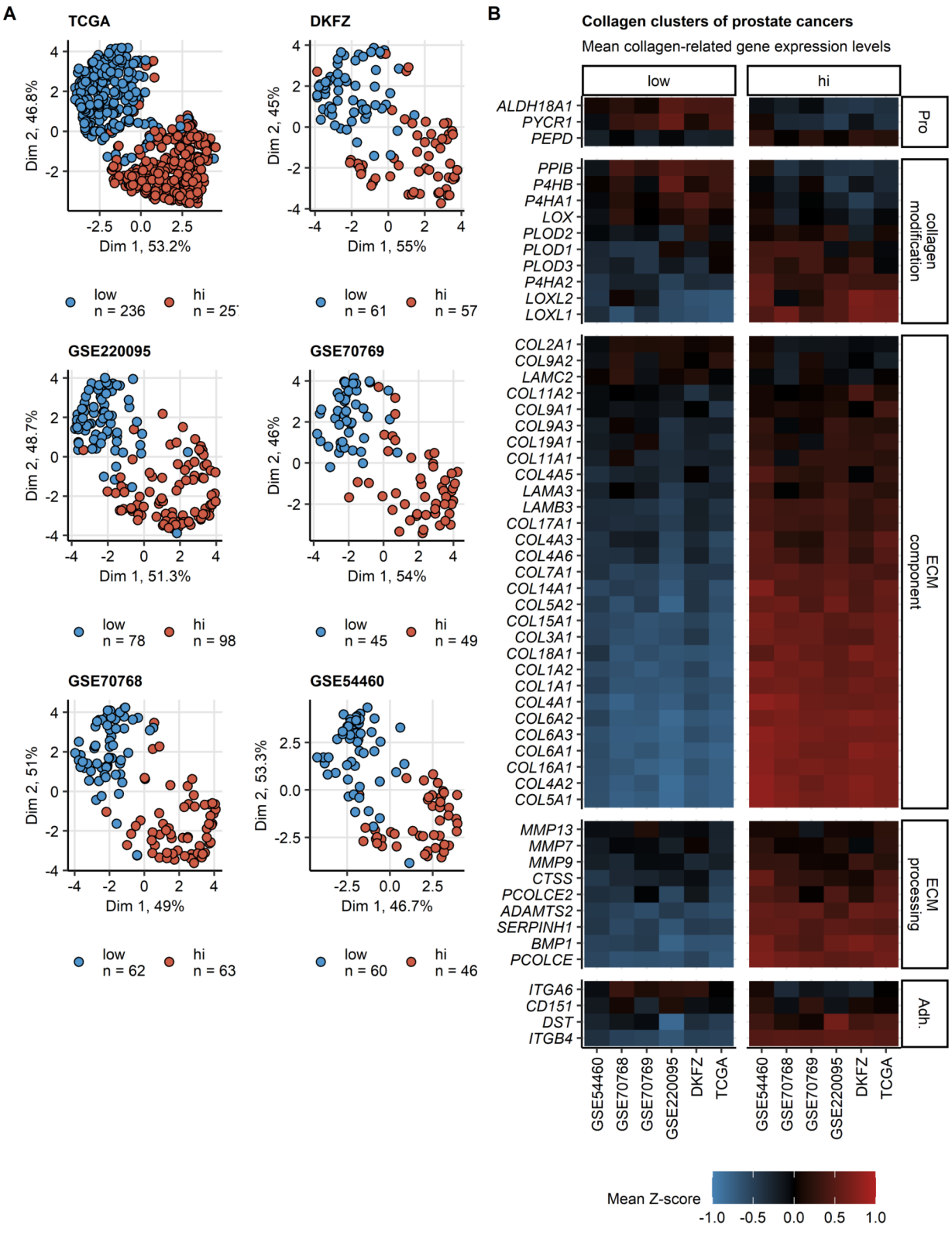
Cancer samples in the TCGA training cohort (n = 493) were clustered in respect to normalized, log_2-transformed expression levels of the collagen-related genes of interest with the PAM (partition around medoids) algorithm with cosine distance metric. Two clusters were defined: collagen low and collagen high. Assignment of the tumor samples in the training GSE54460 (n = 106), GSE70768 (n = 125), GSE70769 (n = 94), GSE220095 (n = 176), and DKFZ collective (n = 118) to the collagen clusters was accomplished by an inverse distance-weighted k-nearest neighbor classifier.

(A) Comparison of performance of several clustering algorithms (PAM: partition around medoids, HCL: hierarchical clustering/Ward D2 and KMEANS) and distance metrics (Euclidean, Manhattan, squared Euclidean and cosine) in 5-fold cross-validation of the TCGA data set gauged by mean silhouette width as a measure of cluster separation, fraction of observations with negative silhouette widths indicative of possible misclassification, explained clustering variance and fraction of misclassified 5-nearest neighbors, cluster assignment accuracy in the cross-validation folds. Note superior performance of the PAM/cosine algorithm used for definition of the collagen clusters.

(B) Choice of cluster number (k = 2) for PAM/cosine clustering of cancer samples in the TCGA cohort motivated by the peak mean silhouette width statistic.

(C) Distribution of sizes of the collagen clusters in the TCGA training cohort and the validation collectives. Percentages of samples assigned to the collagen clusters are presented in a stack plot.

(D) Quality of collagen clusters in the training cohort (TCGA) and validation collectives. Cluster separation was assessed by mean silhouette width, misclassification rate is expressed as fraction of observations with negative silhouette widths. Explanatory performance was gauged by fraction of explained clustering variance. Neighborhood preservation was assessed by mean fraction of 5-nearest neighbors placed in different clusters.

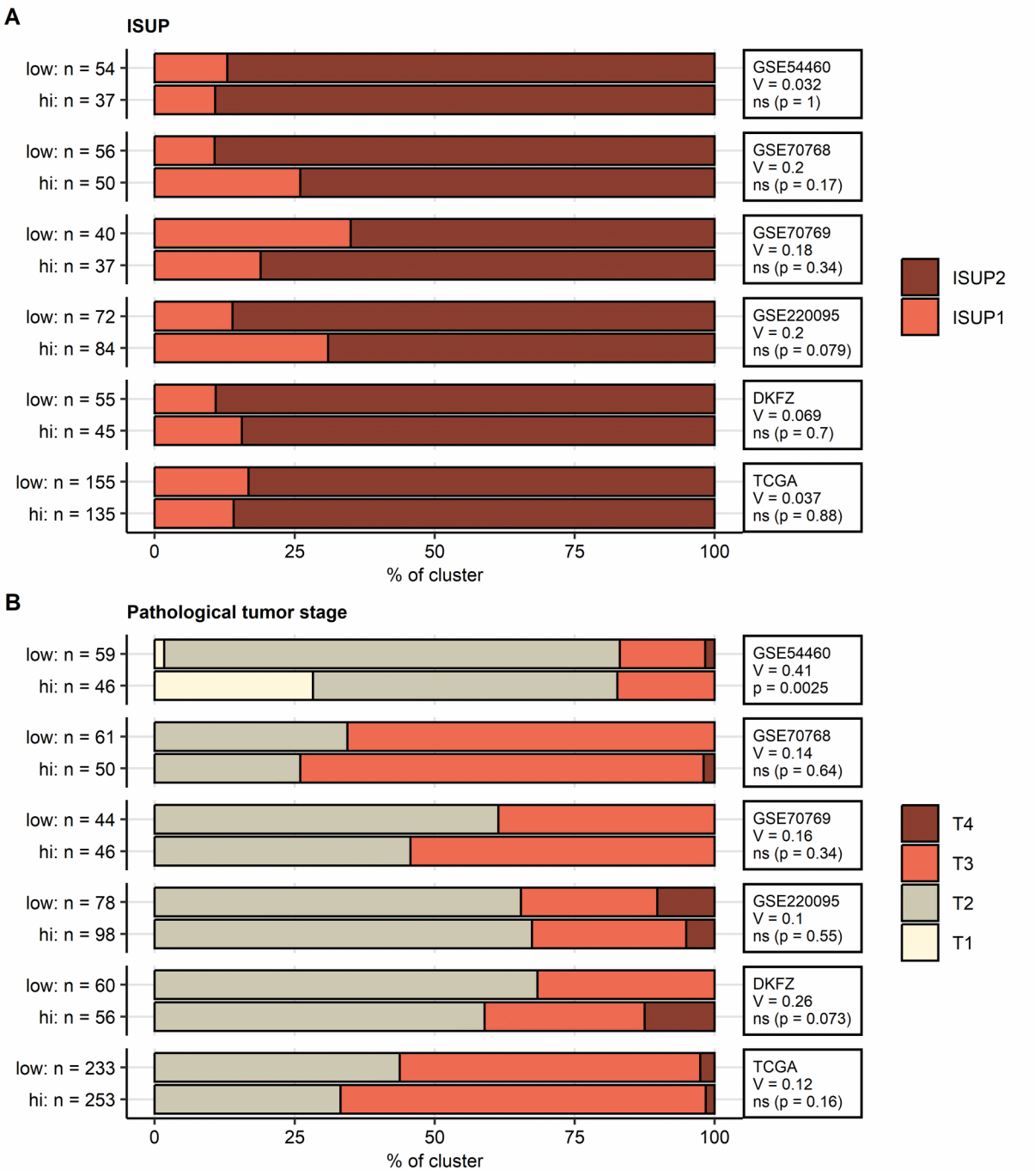

A
TCGA
Dim 2, 46.8%
Dim 1, 53.2%
low
n = 236
hi
n = 25
DKFZ
Dim 2, 45%
Dim 1, 55%
low
n = 61
hi
n = 57
GSE220095
Dim 2, 48.7%
Dim 1, 51.3%
low
n = 78
hi
n = 98
GSE70769
Dim 2, 46%
Dim 1, 54%
low
n = 45
hi
n = 49
GSE70768
Dim 2, 51%
Dim 1, 49%
low
n = 62
hi
n = 63
GSE54460
Dim 2, 53.3%
Dim 1, 46.7%
low
n = 60
hi
n = 46
B
Collagen clusters of prostate cancers
Mean collagen-related gene expression levels
low
hi
ALDH18A1
PYCR1
PEPD
Pro
PPIB
P4HB
P4HA1
LOX
PLOD2
PLOD1
PLOD3
P4HA2
LOXL2
LOXL1
collagen modification
COL2A1
COL9A2
LAMC2
COL11A2
COL9A1
COL9A3
COL19A1
COL11A1
COL4A5
LAMA3
LAMB3
COL17A1
COL4A3
COL4A6
COL7A1
COL14A1
COL5A2
COL15A1
COL3A1
COL18A1
COL1A2
COL1A1
COL4A1
COL6A2
COL6A3
COL6A1
COL16A1
COL4A2
COL5A1
ECM component
MMP13
MMP7
MMP9
CTSS
PCOLCE2
ADAMTS2
SERPINH1
BMP1
PCOLCE
ECM processing
ITGA6
CD151
DST
ITGB4
Adh.
GSE54460
GSE70768
GSE70769
GSE220095
DKFZ
TCGA
Mean Z-score
-1.0
-0.5
0.0
0.5
1.0

Supplementary Figure S3. Expression of the collagen-related genes in the collagen clusters of prostate cancers.

The collagen clusters of prostate cancer samples were developed by semi-supervised PAM clustering as presented in Supplementary Figure S2. Results of comparison of levels of the cluster-defining factors between the collagen clusters are listed in Supplementary Table S5.

(A) Uniform manifold approximation and projection (UMAP) layout of normalized log_2-transformed expression values of the collagen-related genes utilized for definition of the collagen clusters. Each point represents a single cancer sample, cluster assignment is color-coded. Numbers of cancer samples in the collagen clusters are indicated in the figure legends.

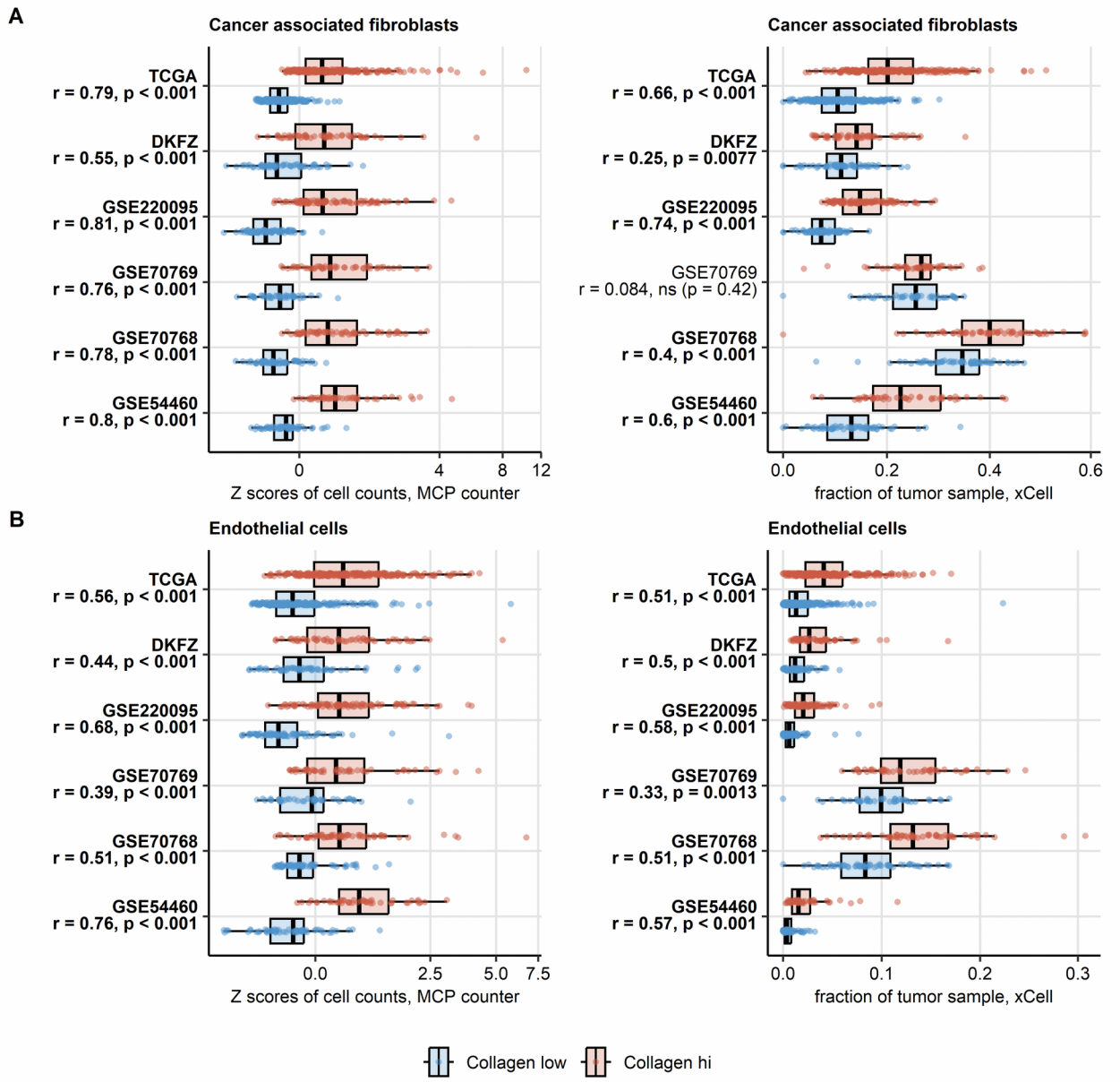
(B) Mean normalized log_2-transformed expression values of the cluster-defining collagen-related genes visualized in a heat map. The cluster-defining genes were classified by their biological function (Pro: proline metabolism, ECM: extracellular matrix, Adh.: adhesion).



Supplementary Figure S4. ISUP and pathological tumor stage in the collagen clusters.

Differences in distribution of ISUP risk strata (International Society of Urological Pathology: ISUP1, ISUP2 and ISUP3+) and pathological tumor stages (B) between the collagen clusters were investigated by χ^2 test with Cramer V effect size statistic. P values were corrected for multiple testing with the false discovery rate method. Percentages of samples assigned to the

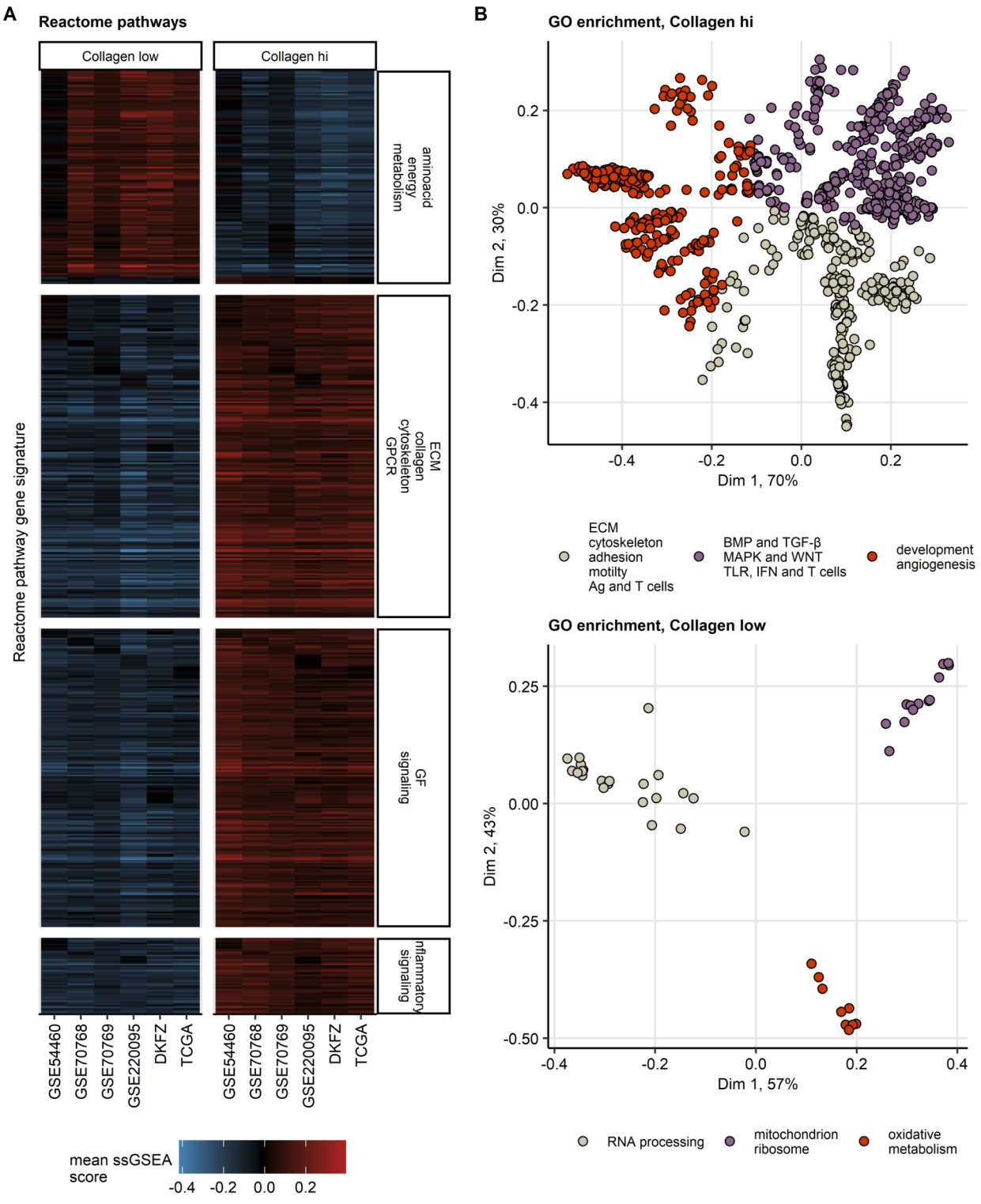
ISUP strata or tumor stages within each cluster are shown in stack plots. Effect sizes and p values are displayed in the plot facets. Numbers of complete observations in the clusters are indicated in the Y axes. Full clinical and pathological characteristic of the collagen clusters is shown in Supplementary Table S6.



Supplementary Figure S5. Cancer-associated fibroblasts and endothelial cell infiltration in the collagen clusters.

Levels of cancer-associated fibroblasts (A) and endothelial cells (B) in the collagen clusters were predicted by the MCP Counter and xCell algorithms. Statistical significance of differences between the clusters was assessed by Mann-Whitney test with r effect size statistic. P values were corrected for multiple testing with the false discovery rate method. Median infiltration levels with interquartile ranges are visualized as boxes, whiskers span over 150% of the interquartile ranges. Points represent single cancer samples. Effect sizes and p values are

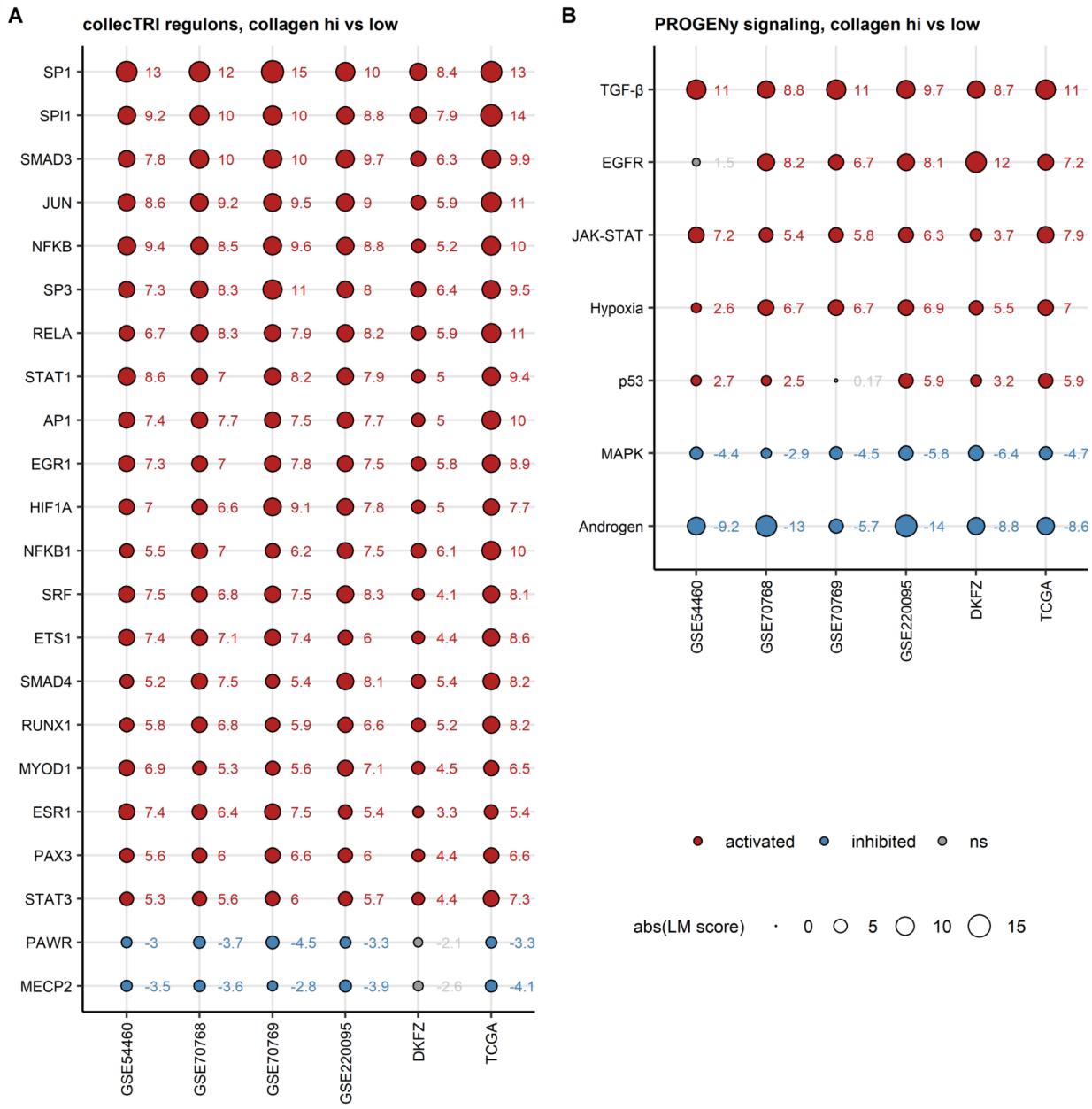
displayed in the Y axes, significant effects are highlighted in bold. GSE54460: Collagen low: n = 60, Collagen hi: n = 46, GSE70768: Collagen low: n = 62, Collagen hi: n = 63, GSE70769: Collagen low: n = 45, Collagen hi: n = 49, GSE220095: Collagen low: n = 78, Collagen hi: n = 98, DKFZ: Collagen low: n = 61, Collagen hi: n = 57, TCGA: Collagen low: n = 236, Collagen hi: n = 257. Results of the analysis for other cell populations are listed in Supplementary Table S7.


A
Reactome pathways
Collagen low
Collagen hi
aminoacid energy metabolism
ECM collagen cytoskeleton GPCR
GF signaling
inflammatory signaling
Reactome pathway gene signature
GSE54460
GSE70768
GSE70769
GSE220095
DKFZ
TCGA
mean ssGSEA score
-0.4 -0.2 0.0 0.2
B
GO enrichment, Collagen hi
Dim 2, 30%
Dim 1, 70%
ECM cytoskeleton adhesion motilty Ag and T cells
BMP and TGF-β MAPK and WNT TLR, IFN and T cells
development angiogenesis
GO enrichment, Collagen low
Dim 2, 43%
Dim 1, 57%
RNA processing
mitochondrion ribosome
oxidative metabolism

Supplementary Figure S6. Gene set variation analysis of Reactome pathway gene signatures and gene ontology term enrichment analysis for the collagen clusters.

(A) Gene set variation analysis (GSVA) with Reactome pathway gene signatures. Statistical significance for differences in signature single sample gene set enrichment analysis scores (ssGSEA) between the collagen clusters was determined by two-tailed T test with Cohen's d effect size statistic. P values were corrected for multiple testing with the false discovery rate method. Mean ssGSEA scores in the clusters and cohorts for signatures significantly regulated in at least five cohorts with at least weak effect size ($d \geq 0.2$) are presented in a heat map. Full analysis results are presented in Supplementary Table S8. The common regulated gene signatures were classified in respect their co-expression patterns by KMEANS clustering; signature classification is indicated in the Y axis facets of the heat map (ECM: extracellular matrix, GF: growth factor, GPCR: G protein-coupled receptor).

(B) Genes differentially expressed between the collagen clusters were identified by false discovery rate (FDR) corrected two-tailed T test with Cohen's d effect size statistic. Differentially expressed genes were defined by the $pFDR < 0.05$ and $d \geq 0.2$ cutoffs (Supplementary Table 9). Biological process gene ontology (GO) enrichment within genes significantly unregulated in the collagen high and the collagen low clusters was investigated with the goana algorithm. Enrichment p values were adjusted for multiple testing with the FDR method, odds ratio of enrichment over the whole-transcriptome GO frequency served as an effect size statistic. GO terms significantly enriched with at least weak effect size ($OR \geq 1.44$) in at least five cohorts were subjected to multi-dimensional scaling and unsupervised KMEANS clustering in respect to their semantic similarity. Multi-dimensional scaling layouts of Wang distances between the common enriched GO terms are presented. Each point represents a single GO term, GO cluster assignment is color coded. Complete results of the GO enrichment analysis are listed in Supplementary Table S10.



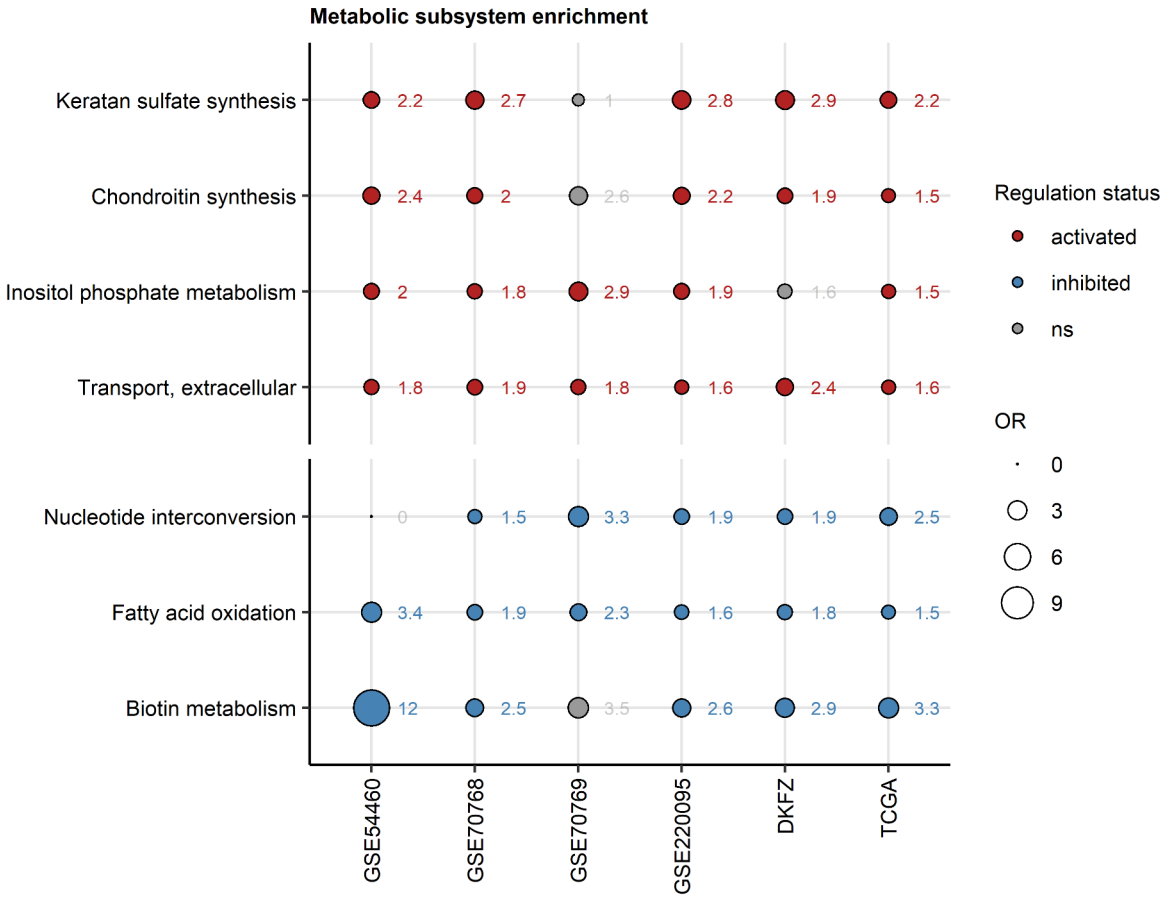
Supplementary Figure S7. Transcriptional regulons and signaling pathway activity in the collagen clusters.

Differences in activity of transcriptional regulons (gene sets controlled by a common transcription factor, A) and signaling pathways (B) in the collagen high cluster as compared with the collagen low cluster were assessed by linear modeling with the collecTRI and PROGENy databases implemented by the decoupler algorithm. Linear model score (LM score) served as a measure of regulon or pathway activity, p values for significant modulation of

activity (LM score ≠ 0) were corrected for multiple testing with the false discovery rate method. Full analysis results are listed in Supplementary Tables S11 and S12.

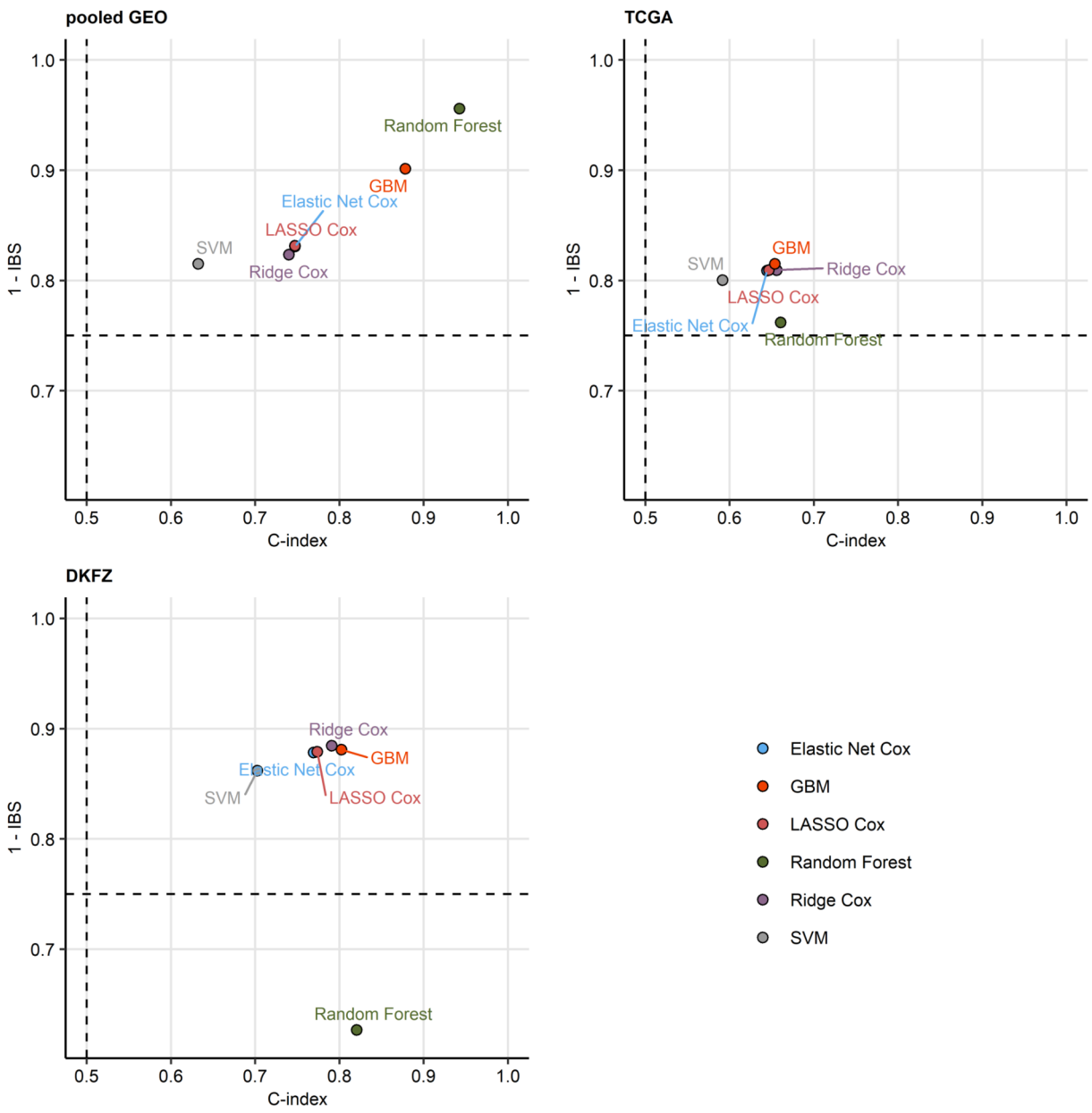
(A) LM scores of top modulated transcriptional regulons found to be significantly activated or inhibited in at least five cohorts visualized as bubble plots. Point color codes for the regulation sign, point size codes for absolute value of the LM score. LM scores for regulon activity in particular cohorts are presented next to the data points.

(B) LM scores of signaling pathways found to be significantly activated or inhibited in at least five cohorts visualized as bubble plots. Point color codes for the regulation sign, point size codes for absolute value of the LM score. LM scores for regulon activity in particular cohorts are presented next to the data points.



Supplementary Figure S8. Metabolism in the collagen clusters.

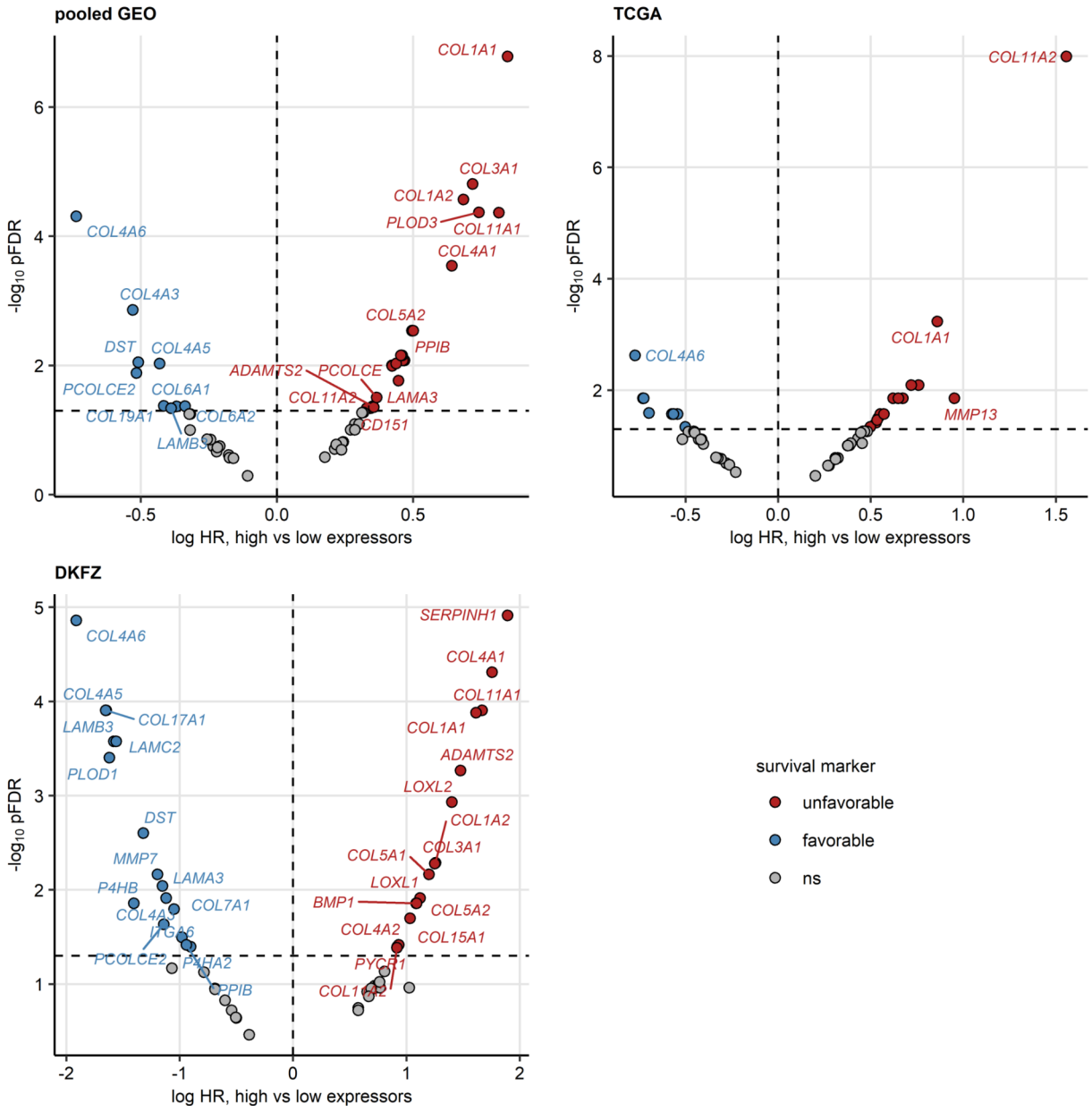
Modulation of Recon2 model metabolic reactions in collagen high tumors as compared with collagen low cancers was predicted by Monte Carlo simulation implemented by the BiGGR and biggrExtra algorithms based on whole-genome differential gene expression estimates (Supplementary Table 13). Enrichment of significantly activated and significantly inhibited reactions in the Recon metabolic subsystems was assessed by comparing the observed frequency of the subsystem's significantly regulated reactions with 10000 random draws from the entire reaction pool. Enrichment p values were corrected for multiple testing with the false discovery rate method. Enrichment odds ratios (OR) for significantly enriched metabolic subsystems with at least weak effect size (OR ≥ 1.44) shared by at least five cohorts are presented in a bubble plot. Point color codes for reaction modulation status, point sizes represent absolute values of OR. Points are labeled with their OR values. Full analysis results are presented in Supplementary Table S14.



Supplementary Figure S9. Comparison of performance of various machine learning algorithms at prediction of biochemical relapse-free survival with expression of the collagen-related genes.

Biochemical relapse-free survival was modeled with normalized, log_2-transformed and ComBat-adjusted expression levels of the collagen-related genes by various machine learning algorithms (Ridge Cox regression, Elastic Net Cox regression, LASSO Cox regression, Support Vector Machines [SVM], Random Forest, and Gradient Boosted Machines [GBM]). Their performance at prediction of biochemical relapse-free survival in the training pooled GEO

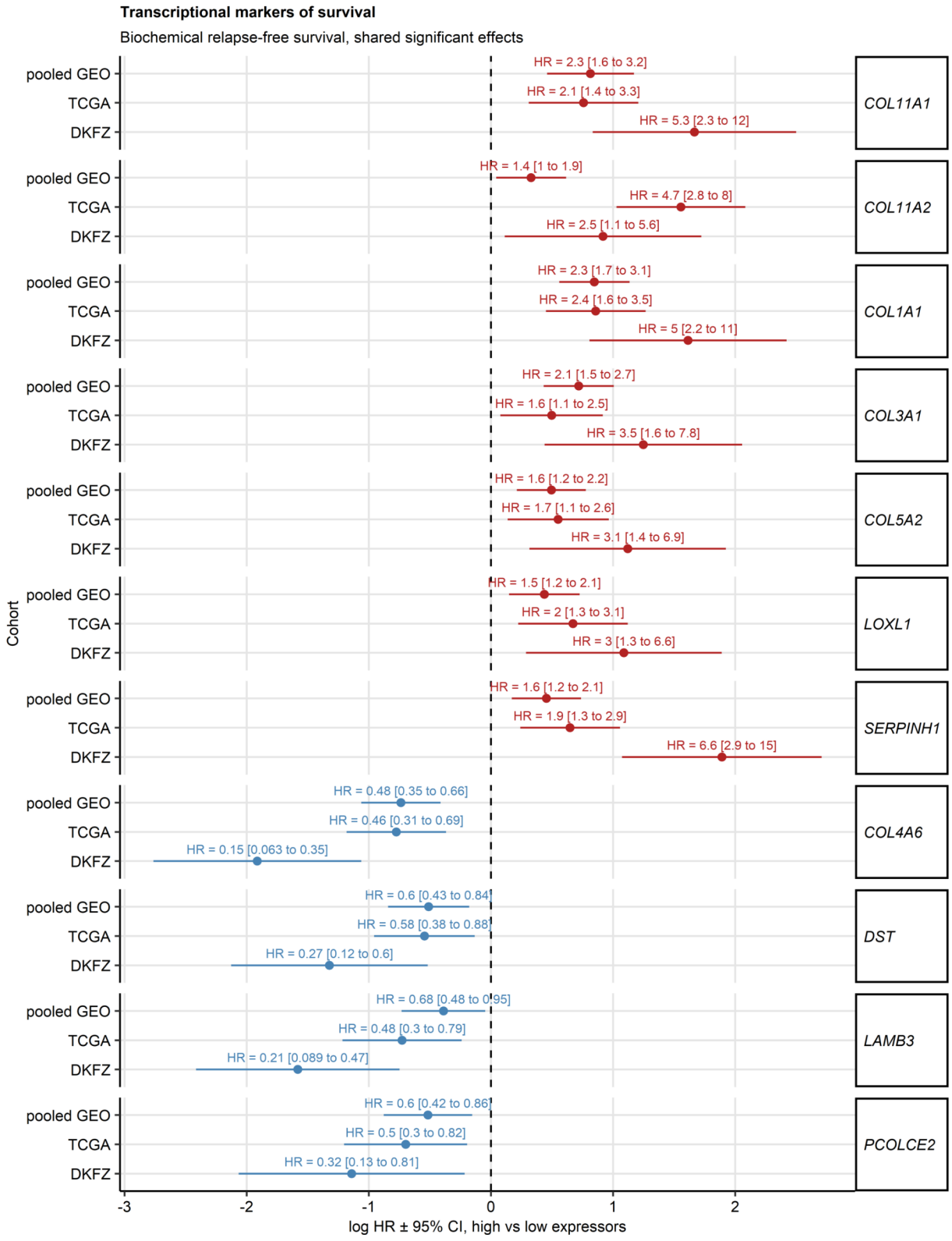
cohort, and the test TCGA and DKFZ collectives was assessed by Harrell's concordance index (C-index) and integrated Brier scores (IBS). High C-index values indicate good concordance between the observed and predicted survival. Low values of IBS indicate good model calibration. C-index and IBS values are presented in scatter plots. Each point represents a single algorithm. Values of C-index and IBS expected for a model predicting survival at random are visualized as dashed lines.



Supplementary Figure S10. Differences in biochemical relapse-free survival between high and low expressors of the collagen-related genes.

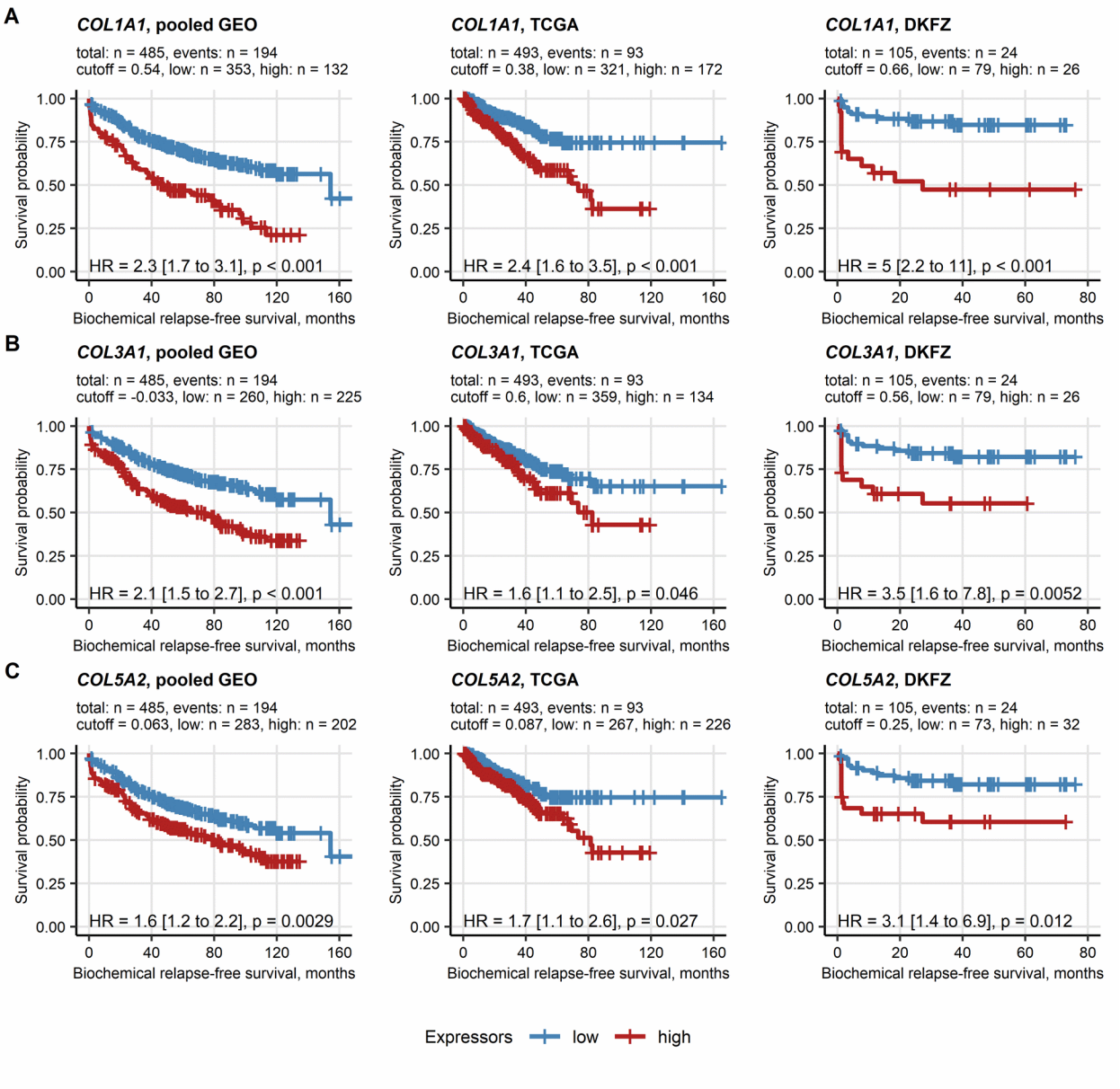
Cancer patients were stratified by cutoffs of normalized expression for each collagen-related genes which corresponded to the largest differences in biochemical relapse-free survival assessed by false discovery rate (FDR) corrected Mentel-Henszel test into high and low expressors for particular genes. Relative risk of biochemical relapse in high expressors as compared with low expressors was modeled by Cox proportional hazard regression and expressed as hazard ratio (HR). Unfavorable survival markers were defined by pFDR < 0.05

and HR > 1. Favorable survival markers were defined by pFDR < 0.05 and HR < 1. HR and pFDR values are presented in Volcano plots. Each point represents a single collagen-related gene. Significance and sign of association with the relapse risk are color coded. Significant genes are labeled with their symbols. Dashed lines indicate the significance and regulation cutoffs.


Transcriptional markers of survival
Biochemical relapse-free survival, shared significant effects
Cohort
pooled GEO
TCGA
DKFZ
COL11A1
HR = 2.3 [1.6 to 3.2]
HR = 2.1 [1.4 to 3.3]
HR = 5.3 [2.3 to 12]
COL11A2
HR = 1.4 [1 to 1.9]
HR = 4.7 [2.8 to 8]
HR = 2.5 [1.1 to 5.6]
COL1A1
HR = 2.3 [1.7 to 3.1]
HR = 2.4 [1.6 to 3.5]
HR = 5 [2.2 to 11]
COL3A1
HR = 2.1 [1.5 to 2.7]
HR = 1.6 [1.1 to 2.5]
HR = 3.5 [1.6 to 7.8]
COL5A2
HR = 1.6 [1.2 to 2.2]
HR = 1.7 [1.1 to 2.6]
HR = 3.1 [1.4 to 6.9]
LOXL1
HR = 1.5 [1.2 to 2.1]
HR = 2 [1.3 to 3.1]
HR = 3 [1.3 to 6.6]
SERPINH1
HR = 1.6 [1.2 to 2.1]
HR = 1.9 [1.3 to 2.9]
HR = 6.6 [2.9 to 15]
COL4A6
HR = 0.48 [0.35 to 0.66]
HR = 0.46 [0.31 to 0.69]
HR = 0.15 [0.063 to 0.35]
DST
HR = 0.6 [0.43 to 0.84]
HR = 0.58 [0.38 to 0.88]
HR = 0.27 [0.12 to 0.6]
LAMB3
HR = 0.68 [0.48 to 0.95]
HR = 0.48 [0.3 to 0.79]
HR = 0.21 [0.089 to 0.47]
PCOLCE2
HR = 0.6 [0.42 to 0.86]
HR = 0.5 [0.3 to 0.82]
HR = 0.32 [0.13 to 0.81]
-3
-2
-1
0
1
2
log HR ± 95% CI, high vs low expressors

Supplementary Figure S11. Collagen-related transcriptional markers of biochemical relapse risk shared by all investigated cohorts.

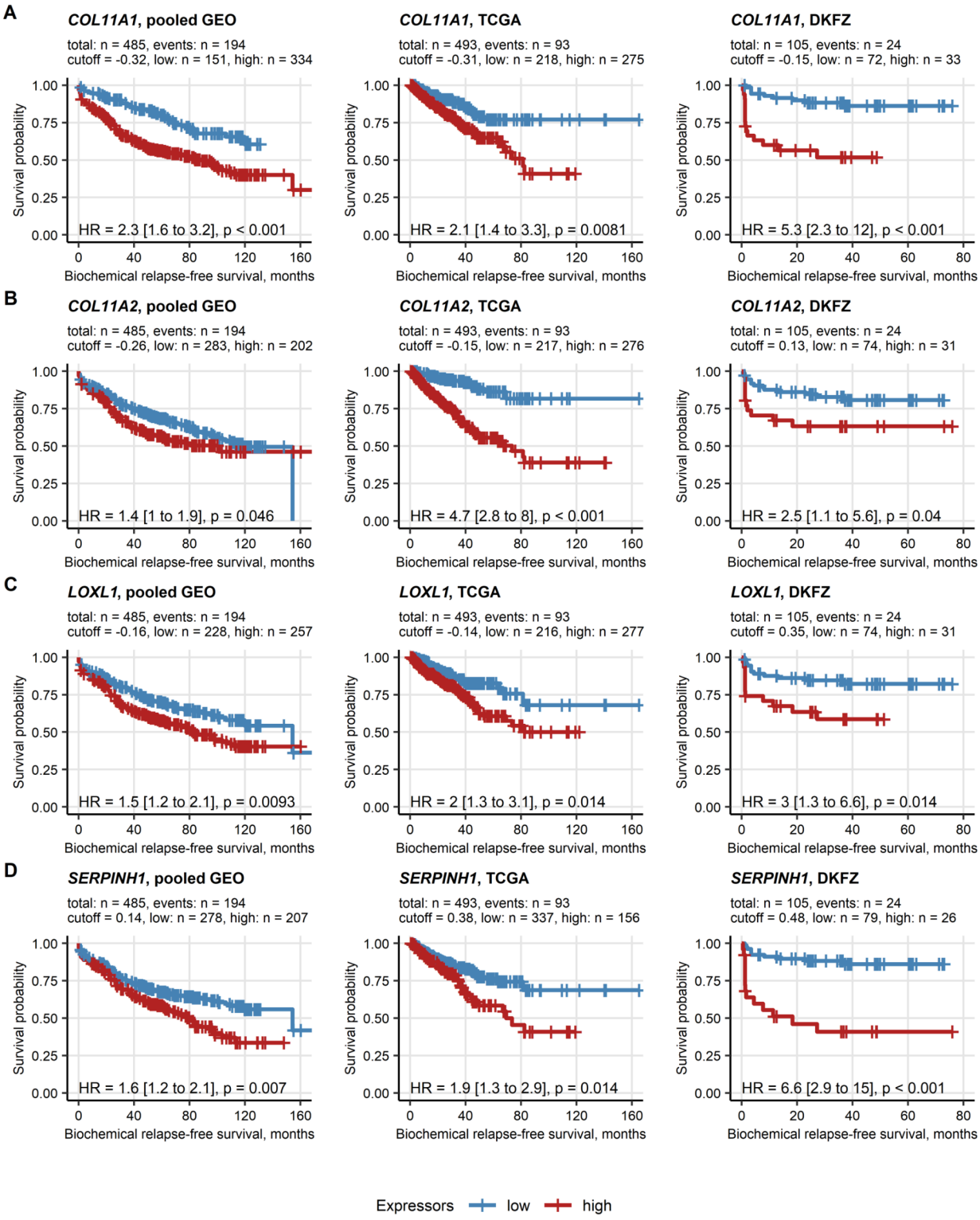
Differences in biochemical relapse-free survival in high expressors versus low expressors of each collagen-related gene were compared by Cox modeling and false discovery rate (FDR) corrected Mentel-Henszel test as presented in Supplementary Figure S10. Unfavorable survival markers were defined by pFDR < 0.05 and hazard ratio (HR) > 1. Favorable survival markers were defined by pFDR < 0.05 and HR < 1. HR with 95% confidence intervals (CI) for unfavorable and favorable markers shared by all investigated cohorts are presented in a Forest plot.



Supplementary Figure S12. Differences in biochemical relapse-free survival between high and low gene expressors: *COL1A1*, *COL3A1*, and *COL5A2*.

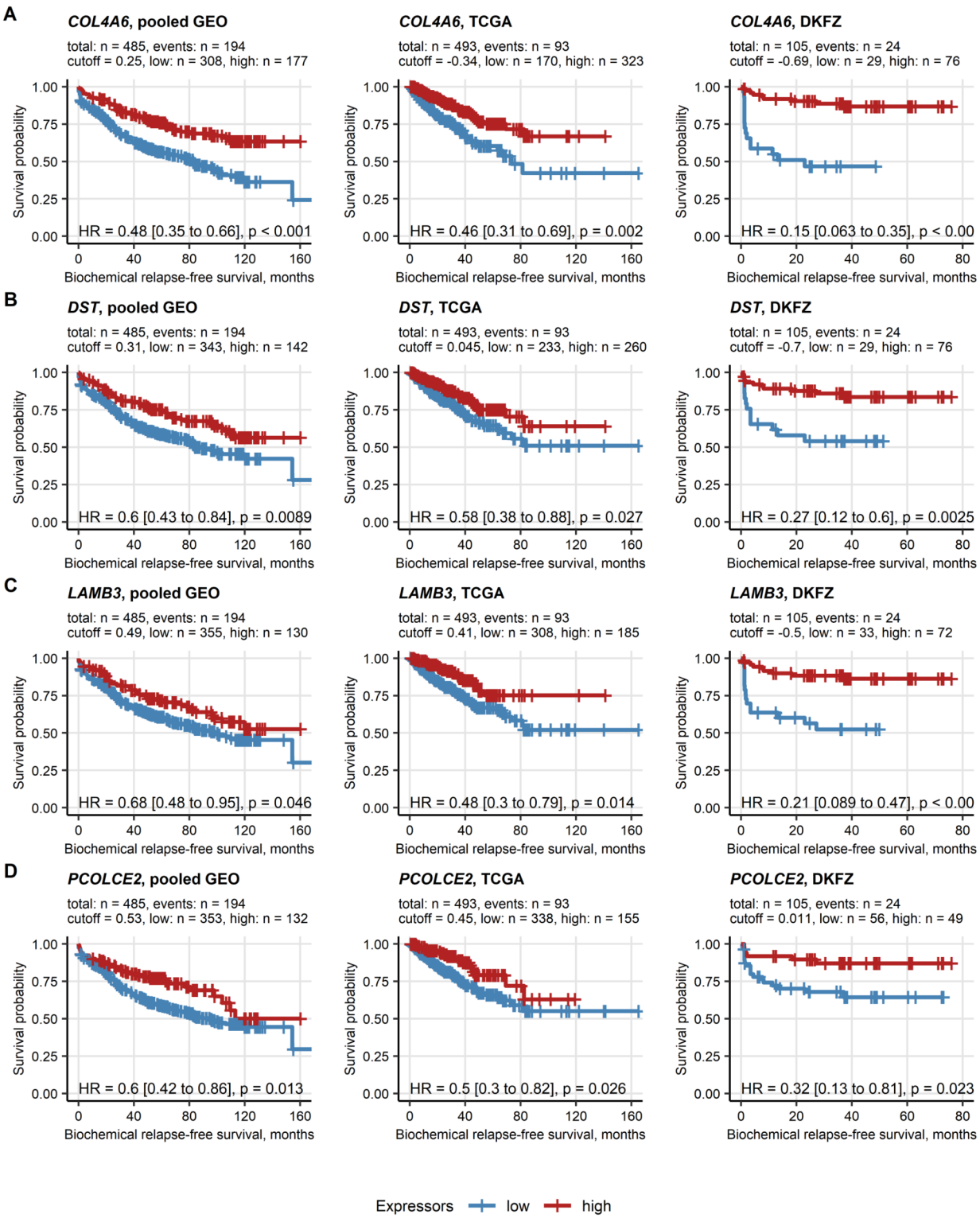
Differences in biochemical relapse-free survival in high expressors versus low expressors of each collagen-related gene were compared by Cox modeling and false discovery rate (FDR) corrected Mentel-Henszel test as presented in Supplementary Figure S10. Unfavorable survival markers were defined by pFDR < 0.05 and hazard ratio (HR) > 1. Differences in biochemical relapse-free survival for the COL1A1, COL3A1, and COL5A2 genes identified as unfavorable markers in all investigated cohorts are shown in Kaplan-Meier plots. Numbers of complete cases and relapses, expression cutoff values, and numbers of high and low expressors

are displayed in the plot captions. HR with 95% confidence intervals and pFDR obtained in Mentel-Henszel test are indicated in the plots.


A
COL11A1, pooled GEO
total: n = 485, events: n = 194
cutoff = -0.32, low: n = 151, high: n = 334
HR = 2.3 [1.6 to 3.2], p < 0.001
COL11A1, TCGA
total: n = 493, events: n = 93
cutoff = -0.31, low: n = 218, high: n = 275
HR = 2.1 [1.4 to 3.3], p = 0.0081
COL11A1, DKFZ
total: n = 105, events: n = 24
cutoff = -0.15, low: n = 72, high: n = 33
HR = 5.3 [2.3 to 12], p < 0.001
B
COL11A2, pooled GEO
total: n = 485, events: n = 194
cutoff = -0.26, low: n = 283, high: n = 202
HR = 1.4 [1 to 1.9], p = 0.046
COL11A2, TCGA
total: n = 493, events: n = 93
cutoff = -0.15, low: n = 217, high: n = 276
HR = 4.7 [2.8 to 8], p < 0.001
COL11A2, DKFZ
total: n = 105, events: n = 24
cutoff = 0.13, low: n = 74, high: n = 31
HR = 2.5 [1.1 to 5.6], p = 0.04
C
LOXL1, pooled GEO
total: n = 485, events: n = 194
cutoff = -0.16, low: n = 228, high: n = 257
HR = 1.5 [1.2 to 2.1], p = 0.0093
LOXL1, TCGA
total: n = 493, events: n = 93
cutoff = -0.14, low: n = 216, high: n = 277
HR = 2 [1.3 to 3.1], p = 0.014
LOXL1, DKFZ
total: n = 105, events: n = 24
cutoff = 0.35, low: n = 74, high: n = 31
HR = 3 [1.3 to 6.6], p = 0.014
D
SERPINH1, pooled GEO
total: n = 485, events: n = 194
cutoff = 0.14, low: n = 278, high: n = 207
HR = 1.6 [1.2 to 2.1], p = 0.007
SERPINH1, TCGA
total: n = 493, events: n = 93
cutoff = 0.38, low: n = 337, high: n = 156
HR = 1.9 [1.3 to 2.9], p = 0.014
SERPINH1, DKFZ
total: n = 105, events: n = 24
cutoff = 0.48, low: n = 79, high: n = 26
HR = 6.6 [2.9 to 15], p < 0.001
Survival probability
Biochemical relapse-free survival, months
Expressors
low
high

Supplementary Figure S13. Differences in biochemical relapse-free survival between high and low gene expressors: *COL11A1*, *COL11A2*, *LOXL1*, and *SERPINH1*.

Differences in biochemical relapse-free survival in high expressors versus low expressors of each collagen-related gene were compared by Cox modeling and false discovery rate (FDR) corrected Mentel-Henszel test as presented in Supplementary Figure S10. Unfavorable survival markers were defined by pFDR < 0.05 and hazard ratio (HR) > 1. Differences in biochemical relapse-free survival for the COL11A1, COL11A2, LOXL1, and SERPINH1 genes identified as unfavorable markers in all investigated cohorts are shown in Kaplan-Meier plots. Numbers of complete cases and relapses, expression cutoff values, and numbers of high and low expressors are displayed in the plot captions. HR with 95% confidence intervals and pFDR obtained in Mentel-Henszel test are indicated in the plots.


A
COL4A6, pooled GEO
total: n = 485, events: n = 194
cutoff = 0.25, low: n = 308, high: n = 177
HR = 0.48 [0.35 to 0.66], p < 0.001
COL4A6, TCGA
total: n = 493, events: n = 93
cutoff = -0.34, low: n = 170, high: n = 323
HR = 0.46 [0.31 to 0.69], p = 0.002
COL4A6, DKFZ
total: n = 105, events: n = 24
cutoff = -0.69, low: n = 29, high: n = 76
HR = 0.15 [0.063 to 0.35], p < 0.00
B
DST, pooled GEO
total: n = 485, events: n = 194
cutoff = 0.31, low: n = 343, high: n = 142
HR = 0.6 [0.43 to 0.84], p = 0.0089
DST, TCGA
total: n = 493, events: n = 93
cutoff = 0.045, low: n = 233, high: n = 260
HR = 0.58 [0.38 to 0.88], p = 0.027
DST, DKFZ
total: n = 105, events: n = 24
cutoff = -0.7, low: n = 29, high: n = 76
HR = 0.27 [0.12 to 0.6], p = 0.0025
C
LAMB3, pooled GEO
total: n = 485, events: n = 194
cutoff = 0.49, low: n = 355, high: n = 130
HR = 0.68 [0.48 to 0.95], p = 0.046
LAMB3, TCGA
total: n = 493, events: n = 93
cutoff = 0.41, low: n = 308, high: n = 185
HR = 0.48 [0.3 to 0.79], p = 0.014
LAMB3, DKFZ
total: n = 105, events: n = 24
cutoff = -0.5, low: n = 33, high: n = 72
HR = 0.21 [0.089 to 0.47], p < 0.00
D
PCOLCE2, pooled GEO
total: n = 485, events: n = 194
cutoff = 0.53, low: n = 353, high: n = 132
HR = 0.6 [0.42 to 0.86], p = 0.013
PCOLCE2, TCGA
total: n = 493, events: n = 93
cutoff = 0.45, low: n = 338, high: n = 155
HR = 0.5 [0.3 to 0.82], p = 0.026
PCOLCE2, DKFZ
total: n = 105, events: n = 24
cutoff = 0.011, low: n = 56, high: n = 49
HR = 0.32 [0.13 to 0.81], p = 0.023
Survival probability
Biochemical relapse-free survival, months
Expressors
low
high

Supplementary Figure S14. Differences in biochemical relapse-free survival between high and low gene expressors: *COL4A6*, *DST*, *LAMB3*, and *PCOLCE2*.

Differences in biochemical relapse-free survival in high expressors versus low expressors of each collagen-related gene were compared by Cox modeling and false discovery rate (FDR) corrected Mentel-Henszel test as presented in Supplementary Figure S10. Unfavorable survival markers were defined by pFDR < 0.05 and hazard ratio (HR) > 1. Differences in biochemical relapse-free survival for the COL4A6, DST, LAMB3, and PCOLCE2 genes identified as favorable markers in all investigated cohorts are shown in Kaplan-Meier plots. Numbers of complete cases and relapses, expression cutoff values, and numbers of high and low expressors are displayed in the plot captions. HR with 95% confidence intervals and pFDR obtained in Mentel-Henszel test are indicated in the plots.

References

1. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. Welcome to the Tidyverse. *Journal of Open Source Software* (2019) 4:1686. doi: 10.21105/joss.01686

2. Henry L, Wickham Hadley. rlang: Functions for Base Types and Core R and 'Tidyverse' Features. (2022) https://cran.r-project.org/web/packages/rlang/index.html

3. Gagolewski M, Tartanus B. Package 'stringi'. (2021) https://cran.r-project.org/web/packages/stringi/index.html http://cran.ism.ac.jp/web/packages/stringi/stringi.pdf

4. Sean D, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics (Oxford, England)* (2007) 23:1846–1847. doi: 10.1093/BIOINFORMATICS/BTM254

5. Pagès H, Carlson M, Falcon S, Li N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. (2022) doi: 10.18129/B9.bioc.AnnotationDbi

6. Carlson M. org.Hs.eg.db: Genome wide annotation for Human. (2022) doi: 10.18129/B9.bioc.org.Hs.eg.db

7. Sturm G, Finotello F, List M. Immunedeconv: An R Package for Unified Access to Computational Methods for Estimating Immune Cell Fractions from Bulk RNA-Sequencing Data. *Methods in molecular biology (Clifton, NJ)* (2020) 2120:223–232. doi: 10.1007/978-1-0716-0327-7_16

8. 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

9. 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

10. 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

11. Kassambara A, Mundt F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. (2020) https://cran.r-project.org/web/packages/factoextra/index.html

12. Schubert E, Rousseeuw PJ. Faster k-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*. Springer (2019). p. 171–187 doi: 10.1007/978-3-030-32047-8_16

13. Drost H-G. Philentropy: Information Theory and Distance Quantification with R. *Journal of Open Source Software* (2018) 3:765. doi: 10.21105/joss.00765

14. Konopka T. umap: Uniform Manifold Approximation and Projection. (2022) https://cran.r-project.org/web/packages/umap/index.html

15. Kassambara A. rstatix: Pipe-Friendly Framework for Basic Statistical Tests. (2021) https://cran.r-project.org/package=rstatix

16. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology* (2010) 11:R14. doi: 10.1186/gb-2010-11-2-r14

17. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* (2010) 26:976–978. doi: 10.1093/BIOINFORMATICS/BTQ064

18. Müller-Dott S, Tsirvouli E, Vazquez M, Ramirez Flores RO, Badia-i-Mompel P, Fallegger R, Türei D, Lægreid A, Saez-Rodriguez J. Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities. *Nucleic acids research* (2023) 51:10934–10949. doi: 10.1093/NAR/GKAD841

19. Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, Garnett MJ, Blüthgen N, Saez-Rodriguez J. Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nature Communications 2017 9:1* (2018) 9:1–11. doi: 10.1038/s41467-017-02391-6

20. Badia-I-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, Taus P, Dugourd A, Holland CH, Ramirez Flores RO, et al. decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* (2022) 2: doi: 10.1093/BIOADV/VBAC016

21. Gavai AK, Supandi F, Hettling H, Murrell P, Leunissen JAM, Van Beek JHGM. Using Bioconductor Package BiGGR for Metabolic Flux Estimation Based on Gene Expression Changes in Brain. *PLOS ONE* (2015) 10:e0119016. doi: 10.1371/JOURNAL.PONE.0119016

22. King ZA, Lu J, Dräger A, Miller P, Federowicz S, Lerman JA, Ebrahim A, Palsson BO, Lewis NE. BiGG Models: A platform for integrating, standardizing and sharing genome-scale models. *Nucleic Acids Research* (2016) 44:D515–D522. doi: 10.1093/NAR/GKV1049

23. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* (2010) 33:1–22. doi: 10.18637/jss.v033.i01

24. Fouodo CJK, König IR, Weihs C, Ziegler A, Wright MN. Support vector machines for survival analysis with R. *R Journal* (2018) 10:412–423. doi: 10.32614/RJ-2018-005

25. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *https://doiorg/101214/08-AOAS169* (2008) 2:841–860. doi: 10.1214/08-AOAS169

26. Ishwaran H, Kogalur UB. randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). (2022) https://cran.r-project.org/web/packages/randomForestSRC/index.html

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

28. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. 1st ed. New York: Springer Verlag (2000).

29. Kassambara A, Kosinski M, Biecek P. survminer: Drawing Survival Curves using 'ggplot2'. (2016) https://cran.r-project.org/package=survminer

30. Wilke CO. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 1st ed. Sebastopol: O'Reilly Media (2019).

31. Gohel D. flextable: Functions for Tabular Reporting. (2022) https://cran.r-project.org/web/packages/flextable/index.html

32. 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

33. 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

34. 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

35. 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

36. 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

37. 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

38. 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

39. Cohen J. Statistical Power Analysis for the Behavioral Sciences. *Statistical Power Analysis for the Behavioral Sciences* (2013) doi: 10.4324/9780203771587

40. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

41. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* (1987) 20:53–65. doi: 10.1016/0377-0427(87)90125-7

42. Venna J, Kaski S. Neighborhood preservation in nonlinear projection methods: An experimental study. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (2001) 2130:485–491. doi: 10.1007/3-540-44668-0_68

43. Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. *Neural Computation* (2004) 16:1299–1323. doi: 10.1162/089976604773717621

44. Leng M, Wang J, Cheng J, Zhou H, Chen X. Adaptive semi-supervised clustering algorithm with label propagation. *Journal of Software Engineering* (2014) 8:14–22. doi: 10.3923/jse.2014.14.22

45. Pichler R, Siska PJ, Tymoszuk P, Martowicz A, Untergasser G, Mayr R, Weber F, Seeber A, Kocher F, Barth DA, et al. A chemokine network of T cell exhaustion and metabolic reprogramming in renal cell carcinoma. *Frontiers in Immunology* (2023) 14:1208. doi: 10.3389/FIMMU.2023.1095195/BIBTEX

46. 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

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

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

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

50. Grambsch PM, Therneau TM. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. *Biometrika* (1994) 81:515. doi: 10.2307/2337123

51. 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

52. 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