Prognostic and biologic relevance of collagen biosynthesis pathway in prostate cancer

Supplementary material, transcriptome part

2023-08-07

# 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*](https://github.com/PiotrTymoszuk/trafo). Text data were handled with *stringi* (3).

Import of the TCGA dataset was accomplished with the [*TCGA-Assembler-2* script](https://github.com/compgenome365/TCGA-Assembler-2/tree/master/TCGA-Assembler). 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 development package [*gseaTools*](https://github.com/PiotrTymoszuk/gseaTools). Expression data of collegen genes were adjusted for batch effects with the *COMBAT* algorithm implemented by the *sva* package (11).

Semi supervised clustering was accomplished with the development package [*clustTools*](https://github.com/PiotrTymoszuk/clustTools) employing algorithms provided by the *philentropy*, *cluster* and *factorextra* packages (12–15).

For statistical hypothesis testing, effect size calculation, gene set variation analysis (GSVA) and differential gene expression analysis, the packages *rstatix* (16), [*ExDA*](https://github.com/PiotrTymoszuk/ExDA) and [*microViz*](https://github.com/PiotrTymoszuk/microViz) were employed. Signaling pathway analysis was done with the package *SPIA* (17).  
Metabolic reaction activity modeling and metabolic subsystem enrichment analysis were performed with the packages *BiGGR* (18,19) and [*biggrExtra*](https://github.com/PiotrTymoszuk/biggrExtra).

For visualization of the results, the packages *ggplot2* (1) (scatter, bubble and bar plots, heat maps), [*ExDA*](https://github.com/PiotrTymoszuk/ExDA) (violin, stack and ribbon plots), [*microViz*](https://github.com/PiotrTymoszuk/microViz) (Forest plots) and [*biggrExtra*](https://github.com/PiotrTymoszuk/biggrExtra) (fold-regulation estimates of metabolic reactions). Figures and tables were created with the packages *cowplot* (20) and *flextable* (21).

## Data sources and data import

The TCGA prostate cancer dataset (22) consisting of 495 normalized expression (level 3) data for primary cancer samples and clinical information was imported in R from the GDC repository with the [*TCGA-Assembler-2* script](https://github.com/compgenome365/TCGA-Assembler-2/tree/master/TCGA-Assembler). The TCGA cohort included also 52 donor-matched samples of the benign prostate and prostate cancer tissue. The available clinical information for the TCGA cohort included age, pathological staging for the tumor, node and distant metastases, tumor grade, Gleason scoring, blood concentrations of prostate-specific antigen (PSA) and diagnosis, overall and biochemical relapse-free survival. The GSE16560, GSE70768, GSE70769 and GSE116918 normalized whole-transcriptome datasets deposited by the study authors were fetched from GEO with the *GEOquery* package (function getGEO()) (4). Transcriptomes of the GSE16560 primary cancer samples (n = 281) (23) were measured with a custom microarray developed by Illumina. The accompanying clinical information included age, Gleason score and overall survival. Transcriptome measurements in primary cancer samples (n = 125) in the GSE70768 cohort were done with the Illumina HumanHT-12 V4.0 expression beadchip microarray. The GSE70768 dataset (24) included also 74 donor-matched normal prostate and cancer tissue specimens. The clinical information for the GSE70768 dataset encompassed age, clinical and pathological staging of the tumor, lymph nodes and distant metastases, PSA concentrations at diagnosis, data on extra capsular extension and positive surgery margins, Gleason scoring as well as biochemical relapse-free survival. The GSE70769 dataset (24) consisted of 94 primary cancer specimens, whose transcriptomes were gauged with Illumina HumanHT-12 V4.0 expression beadchip. The clinical information for the GSE70769 dataset encompassed clinical and pathological staging of the tumor, lymph nodes and distant metastases, PSA concentrations at diagnosis, data on extra capsular extension and positive surgery margins, Gleason scoring as well as biochemical relapse-free survival. The GSE116918 dataset (25) included 248 prostate cancer biopsy specimens. The transcriptome measurements were performed with a custom-made microarray by Affymetrix (Almac Diagnostics Prostate Disease Specific Array). The accompanying clinical information consisted of age, blood PSA concentrations at diagnosis, pathological staging of the tumor, Gleason scoring and biochemical relapse-free survival. Clinical characteristic of the analyzed cohorts is provided in **Table 1**.

Gene expression levels were transformed with the and function for the microarray and RNA sequencing datasets prior to the analysis. Prior to semi-supervised clustering, inter-cohort variability in normalized -transformed expression of collagen pathway genes was reduced by with the ComBat() function from the package *sva* (11).

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](https://reactome.org/) were obtained from the [MSig database](https://www.gsea-msigdb.org/gsea/msigdb/index.jsp), version 7.5.1. ssGSEA scores were computed with the calculate() function from the [*gseaTools*](https://github.com/PiotrTymoszuk/gseaTools) package.

The collagen pathway genes investigated in the current report included the published Collagen Signature (26) and the Reactome Collagen formation pathway genes (R-HSA-1474290) and was constrained to 55 genes detected in all expression datasets (**Supplementary Table S1**).

## Statistical hypothesis testing, effect size and statistical significance

Differences in numeric variables 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 distribution of categorical variables were assessed by test with Cramer’s V effect size statistic. Statistical significance for enrichment of regulated metabolic reactions in metabolic subsystems was investigated by Fisher’s exact test with odds ratio (OR) as an effect size statistic. Effect sizes were interpreted as follows (27):

* 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
* 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) (28) within each analysis step and cohort. Effects were considered statistically significant for FDR-corrected p values < 0.05.

## Comparison of collagen pathway gene expression between the benign and cancer tissue

Differences in -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 (function compare\_variables(), package [*ExDA*](https://github.com/PiotrTymoszuk/ExDA)) in the GSE70768 and TCGA cohorts. The analysis results are presented in **Figure 1B**, **Supplementary Figure S1** and **Supplementary Table S2**.

## Collagen clusters of prostate cancer samples

Prostate cancer samples in the TCGA cohort were clustered by their COMBAT-adjusted normalized -transformed expression levels of the collagen pathway genes with the PAM (partition around medoids) algorithm with cosine distance between the samples (function kcluster(), package [*clustTools*](https://github.com/PiotrTymoszuk/clustTools)). The choice of the clustering algorithm was motivated by the trade-off between the high explanatory performance measured by ‘explained’ clustering variance (ratio of the total between-cluster sum of squares to the total sum of squares) and the optimal reproducibility in 10-fold cross-validation (mean rate of the correct cluster assignment in the folds) (29,30) as compared with hierarchical clustering and KMEANS procedures (**Supplementary Figure S2A**). For variance computation and cross-validation, the methods var() and cv() provided by the [*clustTools*](https://github.com/PiotrTymoszuk/clustTools) package were used. The cluster number choice was based on the bend of the curve of within-cluster sum of squares (12) and the peak of the average silhouette statistic (31) (method plot(), package [*clustTools*](https://github.com/PiotrTymoszuk/clustTools), **Supplementary Figure S2B**). By this means, two clusters of the samples were identified: ‘collagenhigh’ and ‘collagenlow’ cancers.

Assignment of cancers samples of the GSE16560, GSE70768, GSE70769 and GSE116918 collectives to the collagen clusters based on COMBAT-adjusted normalized -transformed expression levels of the collagen pathway genes was done with an inverse distance weighted k-nearest neighbor classifier (method predict(), package [*clustTools*](https://github.com/PiotrTymoszuk/clustTools)) (30). The choice of the nearest neighbor number k was motivated by the maximum value of the ‘explained’ clustering variance defined above ( GSE16560: k = 15, GSE70768: k = 19, GSE70769: k = 19, GSE116918: k = 27). The quality of the cluster assignment was assessed by comparing the collagen clusters distribution and ‘explained’ clustering variances between the training TCGA datasets and the test cohorts (**Supplementary Figure S2C**). Separation of the clusters was assessed by a visual inspection of the UMAP layout plots (method plot(), package [*clustTools*](https://github.com/PiotrTymoszuk/clustTools)) and heat maps of the mean levels of the clustering factors in the collagen clusters (**Figure 2AB**).

Differences in the -transformed collagen pathway gene expression between the collagen clusters were investigated by two-tailed T test with Cohen’s d effect size statistic. The comparison results are presented in **Supplementary Figure S3** and **Supplementary Table S3**.

## 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. Differences in qualitative clinical variables between the collagen clusters were investigated by test with Cramer’s V effect size statistic (function compare\_variables(), package [*ExDA*](https://github.com/PiotrTymoszuk/ExDA)). The clinical characteristics of the collagen clusters are presented in **Figure 2CD** and **Supplementary Table S4**).

## 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*](https://github.com/PiotrTymoszuk/ExDA)). The analysis results are shown in **Figure 3A**, **Supplementary Figures S4** - **S6** and **Supplementary Tables S5** - **S6**.

## 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 four cohort were further investigated. The analysis results are presented in **Figure 3B**, **Supplementary Figure S7** and **Supplementary Table S7**.

## Differential gene expression in the collagen clusters

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*](https://github.com/PiotrTymoszuk/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). Percentages of the investigated transcriptomes found to be significantly differentially expressed between the collagen clusters are displayed in **Supplementary Figure S8**. Expression regulation estimates for the strongest up- and downregulated genes in collagenhigh vs collagenlow cancers are presented in **Supplementary Figure S9**. Significantly differentially regulated genes are listed in **Supplementary Table S8**.

## Signaling pathway activity modulation in the collagen clusters

Modulation of activity of KEGG-listed signaling pathways was predicted by the *SPIA* algorithm (function spia()) (17) based on fold-regulation estimates of expression regulation for the significantly differentially regulated genes. The tA parameter served as a measure of the predicted pathway activity (inhibited: tA < 0, activated: tA > 0). Statistical significance of the activity modulation was assessed by the combined enrichment/perturbation p value corrected for multiple testing with the FDR method (pGFDR) (17). Common modulated signaling pathways were defined as signaling pathways predicted to be significantly activated or inhibited in at least four cohorts. The analysis results are presented in **Figure 3C** and **Supplementary Table S9**.

## 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 (19) and the R package *BiGGR* (18). Estimates of differential expression for all available genes between the collagen clusters were calculated by two-tailed T test as describe above. Estimates of biochemical pathway fold-regulation were computed by evaluation of the gene assignment rules in the Recon2 model. The ‘geneA OR geneB’ operator was interpreted as arithmetic mean of expression regulation estimates for the genes A and B. The ‘geneA AND geneB’ operator was interpreted as minimum of expression regulation estimates for the gene A and gene B (18). Standard deviation, 95% confidence intervals and p values for the predicted reaction regulation estimates were obtained by a Monte Carlo simulation with n = 1000 draws from normal distribution of gene expression regulation estimates (mean: expression regulation estimate, standard deviation: standard error of the expression regulation estimate) (18,32). P values were corrected for multiple testing with the FDR method. The analysis was done with the package *BiGGR* (18) and the development package [*biggrExtra*](https://github.com/PiotrTymoszuk/biggrExtra) (32) (function build\_geneSBML()). Biochemical reactions predicted to be significantly modulated between the collagen clusters are listed in **Supplementary Table S10**.

Enrichment of significantly activated or inhibited biochemical reactions within the Recon model metabolic subsystems (19) was investigated by FDR-corrected Fisher’s exact test (function suba(), package [*biggrExtra*](https://github.com/PiotrTymoszuk/biggrExtra)). The magnitude of enrichment was assessed with OR defined for the -th metabolic subsystem as follows:

where denotes the number of significantly regulated (activated or inhibited) biochemical reactions within the -th metabolic subsystem, denotes the total number of investigated reactions, denotes the number of metabolic reactions within the -th metabolic subsystem, and is the total number of significantly regulated 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 as OR 1.44 (27) (**Figure 3D**, **Supplementary Table S11**).

## Data and code availability

Publicly available datasets were analyzed. Cleared datasets 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: Investigated collagen pathway genes and their classification.

| **Gene group** | **Gene symbol** | **Entrez ID** |
| --- | --- | --- |
| proline turnover | *ALDH18A1* | 5832 |
| *PYCR1* | 5831 |
| *PEPD* | 5184 |
| collagen modification | *PLOD2* | 5352 |
| *P4HA1* | 5033 |
| *P4HA2* | 8974 |
| *LOXL2* | 4017 |
| *LOXL1* | 4016 |
| *LOX* | 4015 |
| *PLOD3* | 8985 |
| *PLOD1* | 5351 |
| *P4HB* | 5034 |
| *PPIB* | 5479 |
| ECM component | *COL3A1* | 1281 |
| *COL2A1* | 1280 |
| *COL11A1* | 1301 |
| *COL11A2* | 1302 |
| *COL5A2* | 1290 |
| *COL5A1* | 1289 |
| *COL1A2* | 1278 |
| *COL1A1* | 1277 |
| *COL7A1* | 1294 |
| *LAMA3* | 3909 |
| *LAMC2* | 3918 |
| *LAMB3* | 3914 |
| *COL4A2* | 1284 |
| *COL4A1* | 1282 |
| *COL4A5* | 1287 |
| *COL4A3* | 1285 |
| *COL4A6* | 1288 |
| *COL18A1* | 80781 |
| *COL15A1* | 1306 |
| *COL14A1* | 7373 |
| *COL6A3* | 1293 |
| *COL6A1* | 1291 |
| *COL6A2* | 1292 |
| *COL17A1* | 1308 |
| *COL9A2* | 1298 |
| *COL9A1* | 1297 |
| *COL9A3* | 1299 |
| *COL16A1* | 1307 |
| *COL19A1* | 1310 |
| ECM processing | *BMP1* | 649 |
| *MMP13* | 4322 |
| *MMP9* | 4318 |
| *CTSS* | 1520 |
| *MMP7* | 4316 |
| *PCOLCE* | 5118 |
| *ADAMTS2* | 9509 |
| *SERPINH1* | 871 |
| *PCOLCE2* | 26577 |
| adhesion | *DST* | 667 |
| *ITGA6* | 3655 |
| *ITGB4* | 3691 |
| *CD151* | 977 |

Supplementary Table S2: 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 S3: 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 S4: 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** | **Variablea** | **Collagen low** | **Collagen high** | **Significanceb** | **Effect sizeb** |
| --- | --- | --- | --- | --- | --- |
| GSE16560 | Age at diagnosis, years | 74 [IQR: 70 - 79] range: 55 - 91 n = 129 | 74 [IQR: 69 - 79] range: 51 - 89 n = 152 | ns (p = 0.38) | d = 0.11 |
| Gleason score | 7 [IQR: 6 - 8] range: 6 - 10 n = 129 | 7 [IQR: 7 - 8] range: 6 - 10 n = 152 | ns (p = 0.064) | d = -0.24 |
| 5 - 6: 37% (n = 48) 7: 36% (n = 47) 8+: 26% (n = 34) n = 129 | 5 - 6: 23% (n = 35) 7: 46% (n = 70) 8+: 31% (n = 47) n = 152 | ns (p = 0.064) | V = 0.16 |
| GSE70768 | Age at diagnosis, years | 62 [IQR: 57 - 65] range: 42 - 73 n = 62 | 63 [IQR: 57 - 69] range: 41 - 93 n = 62 | ns (p = 0.26) | d = -0.3 |
| PSA at diagnosis | 7.2 [IQR: 5.8 - 11] range: 4 - 19 n = 62 | 8.7 [IQR: 6.6 - 14] range: 3.2 - 280 n = 61 | ns (p = 0.062) | d = -0.47 |
| Clinical stage | T1: 55% (n = 33) T2: 32% (n = 19) T3: 13% (n = 8) n = 60 | T1: 56% (n = 28) T2: 28% (n = 14) T3: 16% (n = 8) n = 50 | ns (p = 0.97) | V = 0.049 |
| 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 = 0.67) | 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 = 0.97) | V = 0.064 |
| Pathological metastasis stage | M0: 100% (n = 4) M1: 0% (n = 0) n = 4 | M0: 67% (n = 2) M1: 33% (n = 1) n = 3 | ns (p = 0.97) | V = 0.47 |
| Gleason score | 7 [IQR: 7 - 7] range: 6 - 9 n = 62 | 7 [IQR: 7 - 7] range: 6 - 10 n = 60 | ns (p = 0.52) | d = -0.22 |
| Positive surgical margins | 21% (n = 13) n = 62 | 23% (n = 13) n = 57 | ns (p = 0.98) | V = 0.022 |
| Extracapsular extension | 72% (n = 44) n = 61 | 65% (n = 33) n = 51 | ns (p = 0.82) | V = 0.08 |
| Gleason score | 5 - 6: 9.7% (n = 6) 7: 81% (n = 50) 8+: 9.7% (n = 6) n = 62 | 5 - 6: 18% (n = 11) 7: 62% (n = 37) 8+: 20% (n = 12) n = 60 | ns (p = 0.25) | V = 0.21 |
| GSE70769 | PSA at diagnosis | 7.3 [IQR: 5.1 - 11] range: 2.2 - 35 n = 42 | 8.4 [IQR: 6.4 - 12] range: 1.5 - 120 n = 48 | ns (p = 0.2) | d = -0.32 |
| Clinical 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 = 0.35) | V = 0.18 |
| Pathological tumor stage | T2: 64% (n = 28) T3: 36% (n = 16) n = 44 | T2: 43% (n = 20) T3: 57% (n = 26) n = 46 | ns (p = 0.2) | V = 0.2 |
| Pathological node stage | N0: 100% (n = 7) n = 7 | N0: 100% (n = 11) n = 11 | ns (p = 0.43) | V = Inf |
| Pathological metastasis stage | M0: 93% (n = 14) M1: 6.7% (n = 1) n = 15 | M0: 80% (n = 12) M1: 20% (n = 3) n = 15 | ns (p = 0.59) | V = 0.2 |
| Gleason score | 7 [IQR: 6 - 7] range: 5 - 9 n = 44 | 7 [IQR: 7 - 7.5] range: 6 - 10 n = 47 | p = 0.012 | d = -0.65 |
| Positive surgical margins | 41% (n = 18) n = 44 | 49% (n = 24) n = 49 | ns (p = 0.59) | V = 0.081 |
| Extracapsular extension | 36% (n = 16) n = 44 | 55% (n = 26) n = 47 | ns (p = 0.2) | V = 0.19 |
| Gleason score | 5 - 6: 32% (n = 14) 7: 61% (n = 27) 8+: 6.8% (n = 3) n = 44 | 5 - 6: 13% (n = 6) 7: 62% (n = 29) 8+: 26% (n = 12) n = 47 | p = 0.046 | V = 0.31 |
| GSE116918 | Age at diagnosis, years | 67 [IQR: 63 - 72] range: 48 - 79 n = 123 | 69 [IQR: 64 - 73] range: 51 - 78 n = 125 | ns (p = 0.44) | d = -0.13 |
| PSA at diagnosis | 18 [IQR: 11 - 27] range: 5.9 - 130 n = 123 | 16 [IQR: 10 - 27] range: 3.2 - 220 n = 125 | ns (p = 0.93) | d = -0.011 |
| Pathological tumor stage | T1: 23% (n = 25) T2: 33% (n = 36) T3: 45% (n = 49) T4: 0% (n = 0) n = 110 | T1: 23% (n = 26) T2: 35% (n = 40) T3: 38% (n = 43) T4: 3.5% (n = 4) n = 113 | ns (p = 0.41) | V = 0.14 |
| Gleason score | 7 [IQR: 7 - 8] range: 6 - 9 n = 123 | 7 [IQR: 7 - 9] range: 6 - 10 n = 125 | ns (p = 0.25) | d = -0.22 |
| 5 - 6: 19% (n = 23) 7: 42% (n = 52) 8+: 39% (n = 48) n = 123 | 5 - 6: 15% (n = 19) 7: 38% (n = 47) 8+: 47% (n = 59) n = 125 | ns (p = 0.5) | V = 0.084 |
| TCGA | Age at diagnosis, years | 61 [IQR: 56 - 66] range: 44 - 78 n = 234 | 62 [IQR: 57 - 66] range: 41 - 77 n = 261 | ns (p = 0.16) | d = -0.15 |
| PSA at diagnosis | 0.1 [IQR: 0.03 - 0.1] range: 0 - 320 n = 204 | 0.1 [IQR: 0.03 - 0.12] range: 0 - 37 n = 234 | ns (p = 0.47) | d = 0.091 |
| Pathological tumor stage | T2: 44% (n = 95) T3: 56% (n = 119) n = 214 | T2: 33% (n = 81) T3: 67% (n = 164) n = 245 | p = 0.044 | V = 0.12 |
| Pathological node stage | N0: 82% (n = 161) N1: 18% (n = 36) n = 197 | N0: 81% (n = 183) N1: 19% (n = 42) n = 225 | ns (p = 1) | V = 0.005 |
| Pathological metastasis stage | M0: 100% (n = 212) M1: 0.47% (n = 1) n = 213 | M0: 99% (n = 241) M1: 0.82% (n = 2) n = 243 | ns (p = 1) | V = 0.022 |
| Gleason score | 7 [IQR: 7 - 8] range: 6 - 10 n = 234 | 7 [IQR: 7 - 9] range: 6 - 10 n = 261 | p = 0.018 | d = -0.26 |
| 5 - 6: 10% (n = 24) 7: 55% (n = 128) 8+: 35% (n = 82) n = 234 | 5 - 6: 8% (n = 21) 7: 45% (n = 118) 8+: 47% (n = 122) n = 261 | ns (p = 0.06) | V = 0.12 |
| aPSA: prostate-specific antigen. | | | | | |
| bCategorical variables: χ² test with Cramer's V effec size statistic. Numeric variables: Mann-whitney test with r effect size statistic. P values corrected for multiple testing with the false discovery rate method. | | | | | |

Supplementary Table S5: Non-malignant cell numbers predicted for the collagen clusters by the MCP counter algorithm. 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 S6: Non-maignant cell fractions predicted for the collagen clusters by the xCell algorithm. 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 S7: 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 four 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 S8: 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 at least 0.2 (at least weak effect size). 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: Signaling pathway activity in the collagen clusters investigated by the SPIA algorithm. Results for signaling pathways significantly activated or inhibited in at least four cohorts are shown. The table is available as a supplementary Excel file.

|  |
| --- |
|  |

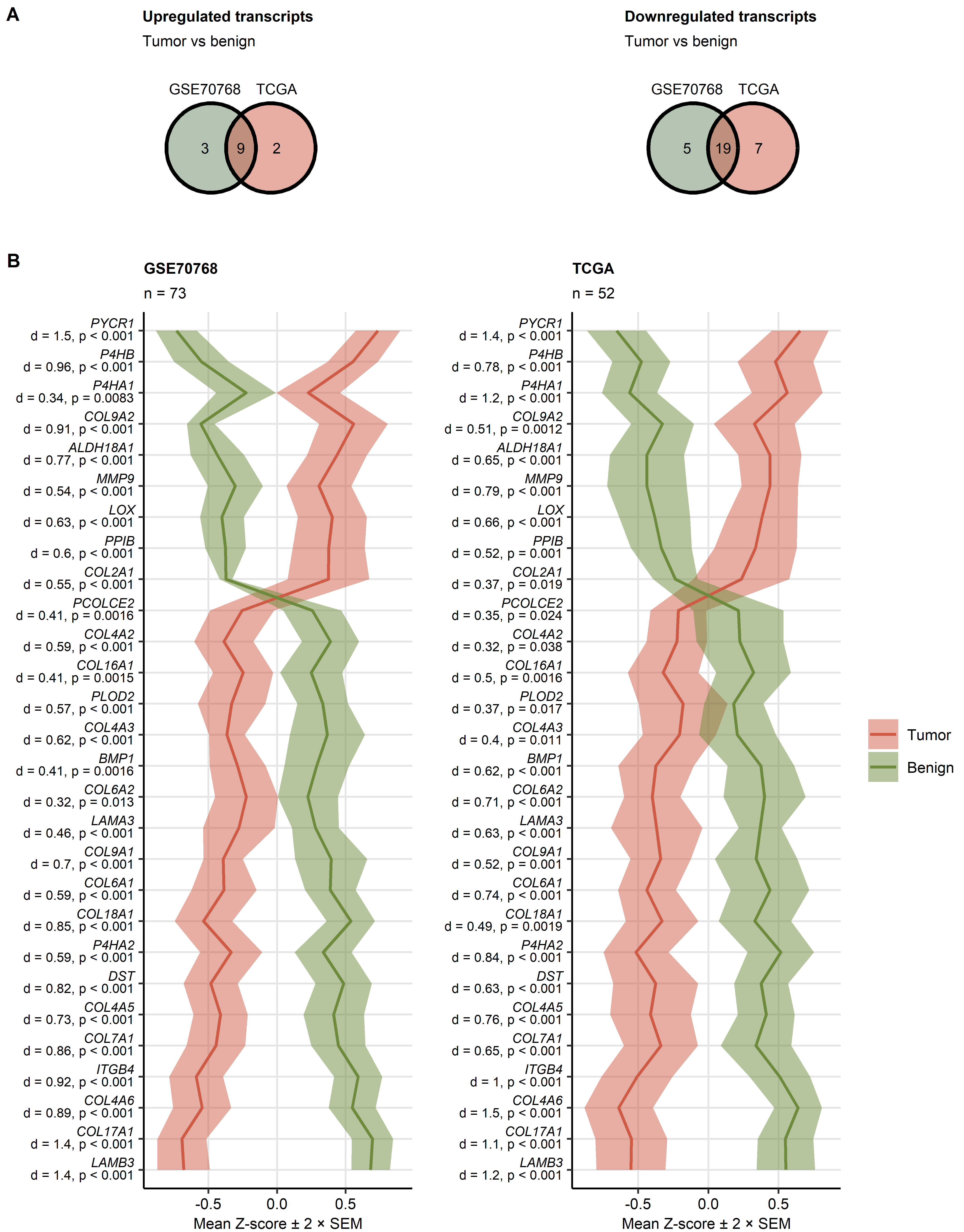
Supplementary Table S10: Biochemical reactions predicted to be significantly activated in the collagen high as compared with collagen low cancers. Statistical significance was determined by a Monte Carlo simulation. P values were corrected for multiple testing with the false discovery rate method. The table is available as a supplementary Excel file.

|  |
| --- |
|  |

Supplementary Table S11: Results of enrichment analysis for significantly activated and inhibited biochemical reaction within the Recon metabolism subsystem. Statistical significance was determined by Fisher's exact test corrected for multiple testing with the false discovery rate method (FDR). The table is available as a supplementary Excel file.

|  |
| --- |
|  |

# Supplementary Figures

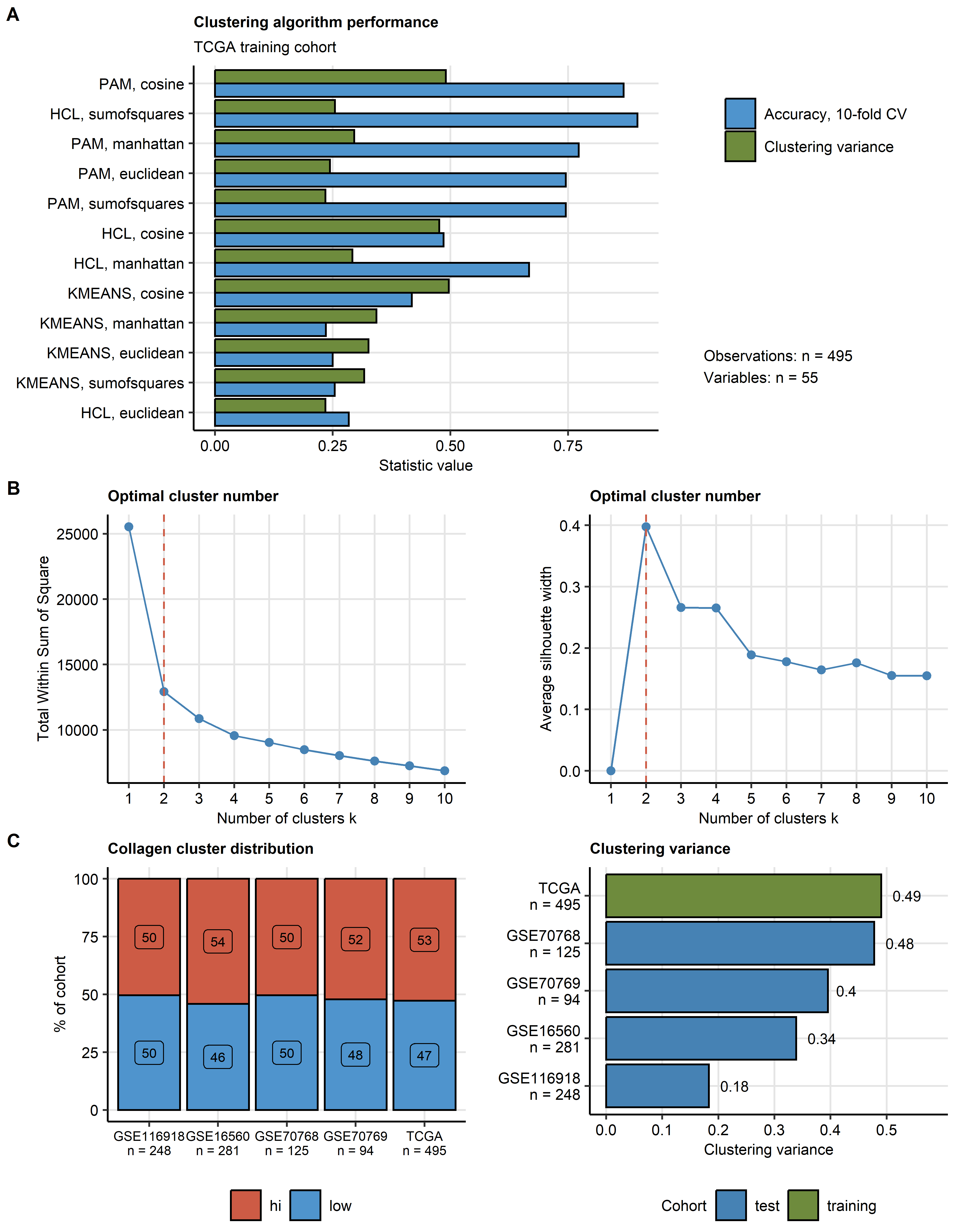


**Supplementary Figure S1. Expression of collagen pathway genes in the normal prostate and prostate cancer tissue.**

*Differences in -transformed expression of 55 genes related to collagen metabolism 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.*

*(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) Expression of the collagen pathway genes significantly regulated in both analyzed cohorts. Mean normalized -transformed expression values are presented as lines. Tinted ribbons represent the 2 SEM (standard error of the mean) intervals. Effect sizes and p values are indicated in the Y axis.*



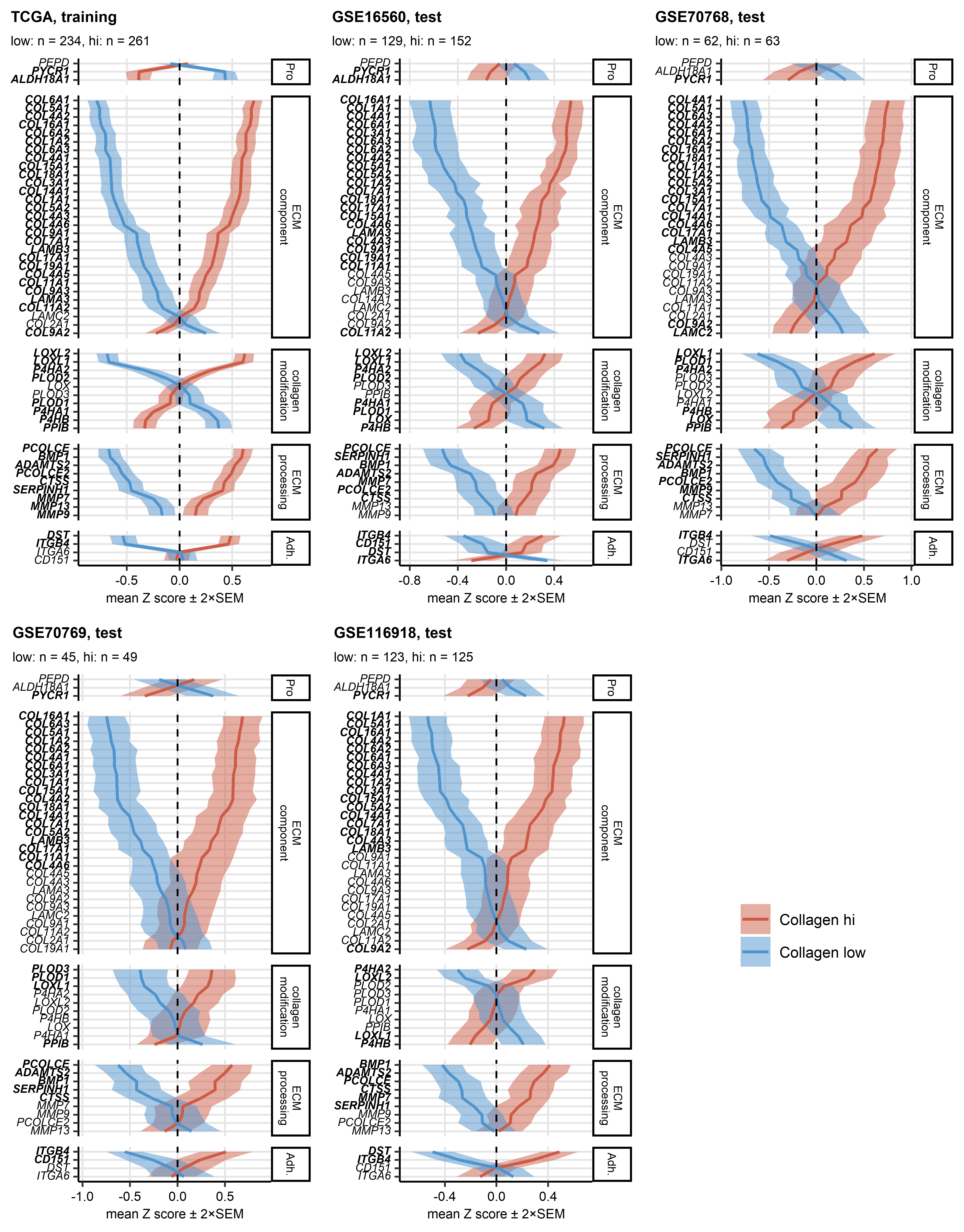
**Supplementary Figure S2. Semi-supervised clustering of prostate cancer samples in respect to expression of the collagen pathway genes.**

*Tumor samples in the TCGA training cohort were clustered in respect to normalized, -transformed expression levels of the collagen pathway genes with the PAM (partition around medoids) algorithm with cosine distance metric. Two clusters were defined: collagenlow (low) and collagenhigh (hi). Assignment of the tumor samples in the training GSE16560, GSE40272, GSE70768 and GSE70769 collectives 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, sum-of-squares and cosine) in the training TCGA cohort was measured by the amount of ‘explained’ clustering variance (ratio of the between-cluster sum of squares to the total sum of squares) and rate of correct cluster assignment in 10-fold cross-validation (CV). Number of complete observations and the clustering factors are shown next to the plot.*

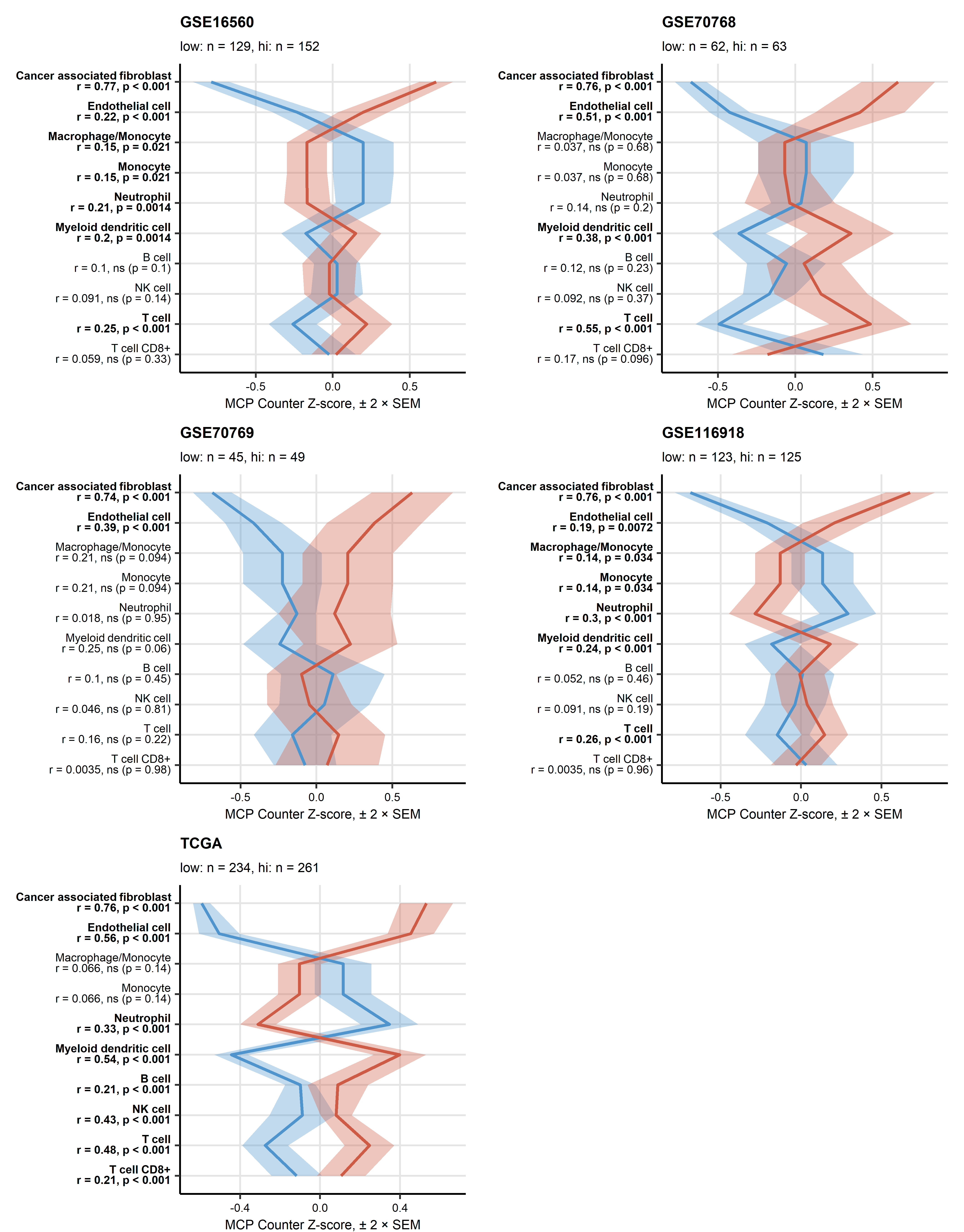
*(B) Determination of the cluster number for PAM/Manhattan distance algorithm in the TCGA cohort by the bend of the within-cluster sum of squares (WSS) (B) and the peak of the mean silhouette statistic (C).*

*(C) Distribution of the collagen clusters (left panel) and performance of the PAM/cosine distance algorithm at semi-supervised clustering in the training cohort and the test collectives measured by the ‘explained’ clustering variance statistic (right panel). Numbers of complete observations are presented in the plot axes.*



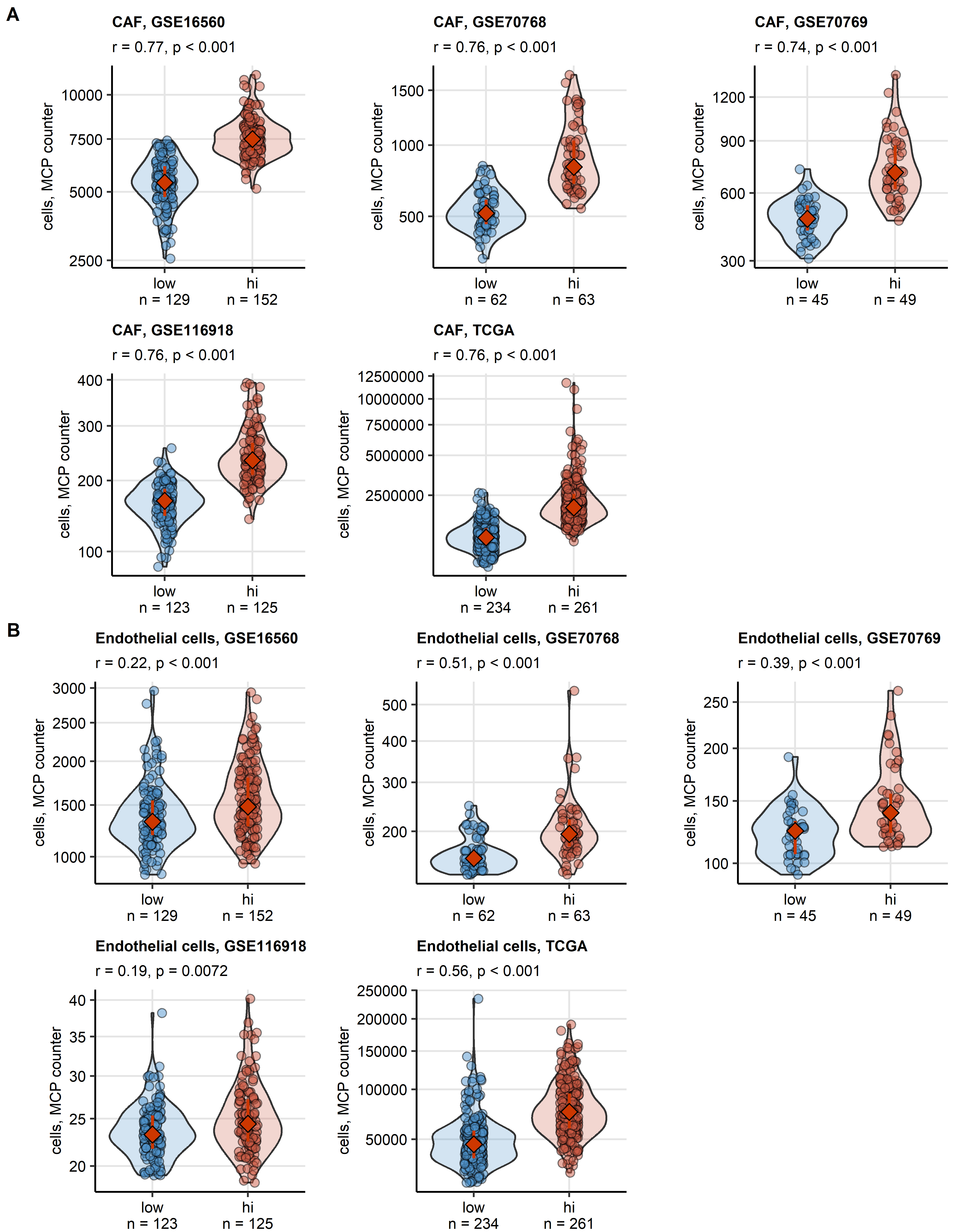
**Supplementary Figure S3. Expression of the collagen pathway genes in the collagen clusters of prostate cancers.**

*Tumor samples were subjected to semi-supervised clustering in respect to the collagen pathway gene expression with the PAM algorithm with cosine distance metric (training: TCGA, test cohorts: GSE16560, GSE40272, GSE70768 and GSE70769). Statistical significance of differences in expression of normalized, -transformed expression levels (Z-scores) between the collagen clusters were investigated 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 expression Z scores in the collagenlow (low) and collagenhigh (hi) clusters are presented as thick lines. The tinted ribbons represent 2 SEM (standard error of the mean) intervals. The collagen pathway genes were grouped according to their biological function as proline turnover (Pro), collagen modification, extracellular matrix (ECM) component, ECM processing and adhesion (Adh.) protein genes. Significantly regulated genes are highlighted in bold. Numbers of samples in the clusters are indicated in the plot captions.*



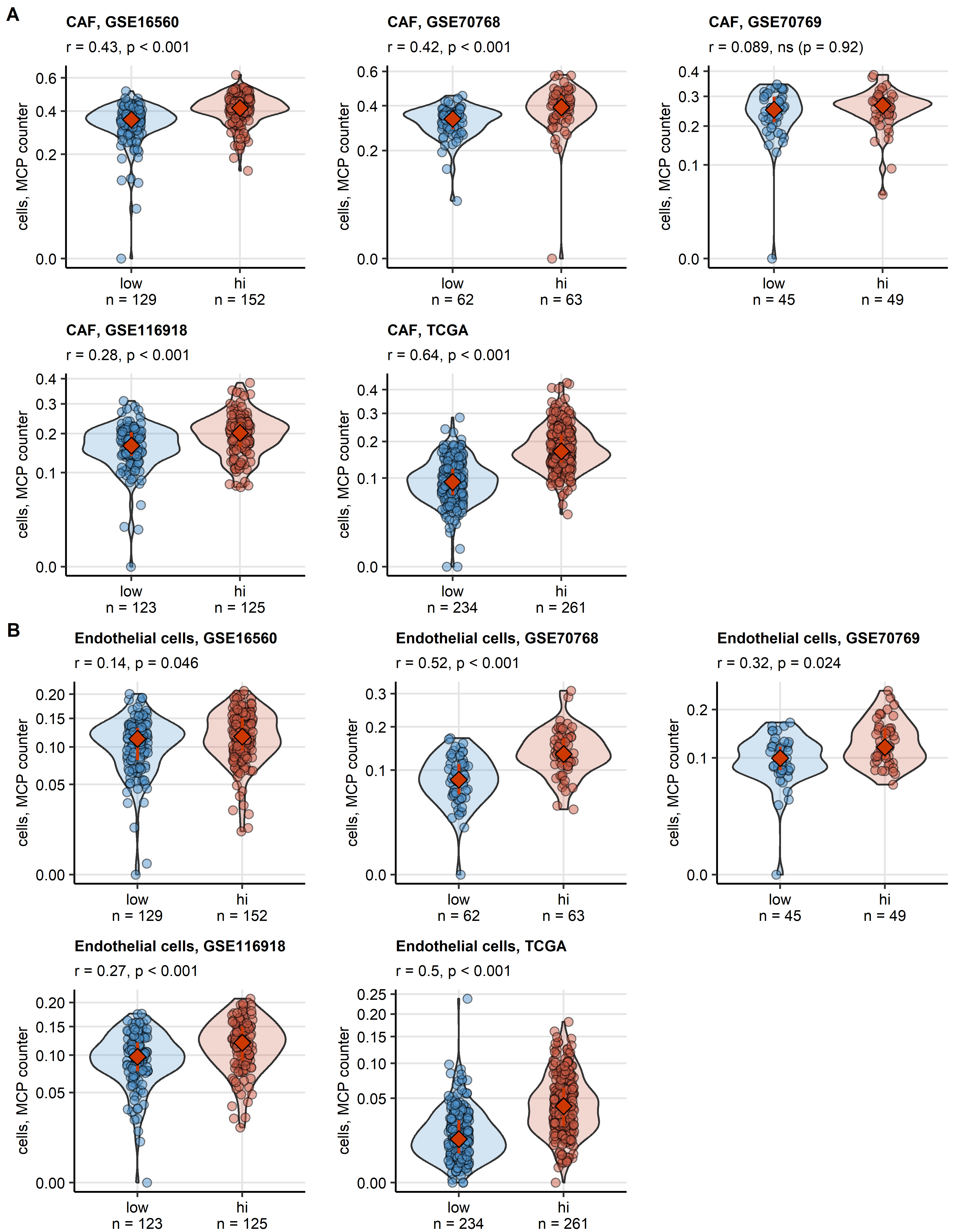
**Supplementary Figure S4. MCP Counter estimates of non-malignant cell content in the collagen clusters.**

*Counts of non-malignant cells (CAF: cancer-associated fibroblast, EC: endothelial cells, T cells, CD8+ T cells, NK cells: natural killer cells, B cells, macrophages, monocytes, neutrophils and mDC: myeloid dendritic cells) in the prostate cancer samples were predicted by the MCP counter algorithm. Differences in cell content between the collagenlow (low) and collagenhigh (hi) clusters were analyzed with Mann-Whitney test with r effect size statistic. P values were corrected for multiple testing with the false discovery rate method. Means of normalized cell counts in the collagen clusters are depicted as thick lines. Tinted ribbons represent the 2 SEM (standard error of the mean) intervals. Effect sizes and p values are shown in the Y axes. Significant effects are highlighted in bold. Numbers of samples assigned to the clusters are indicated in the plot captions.*



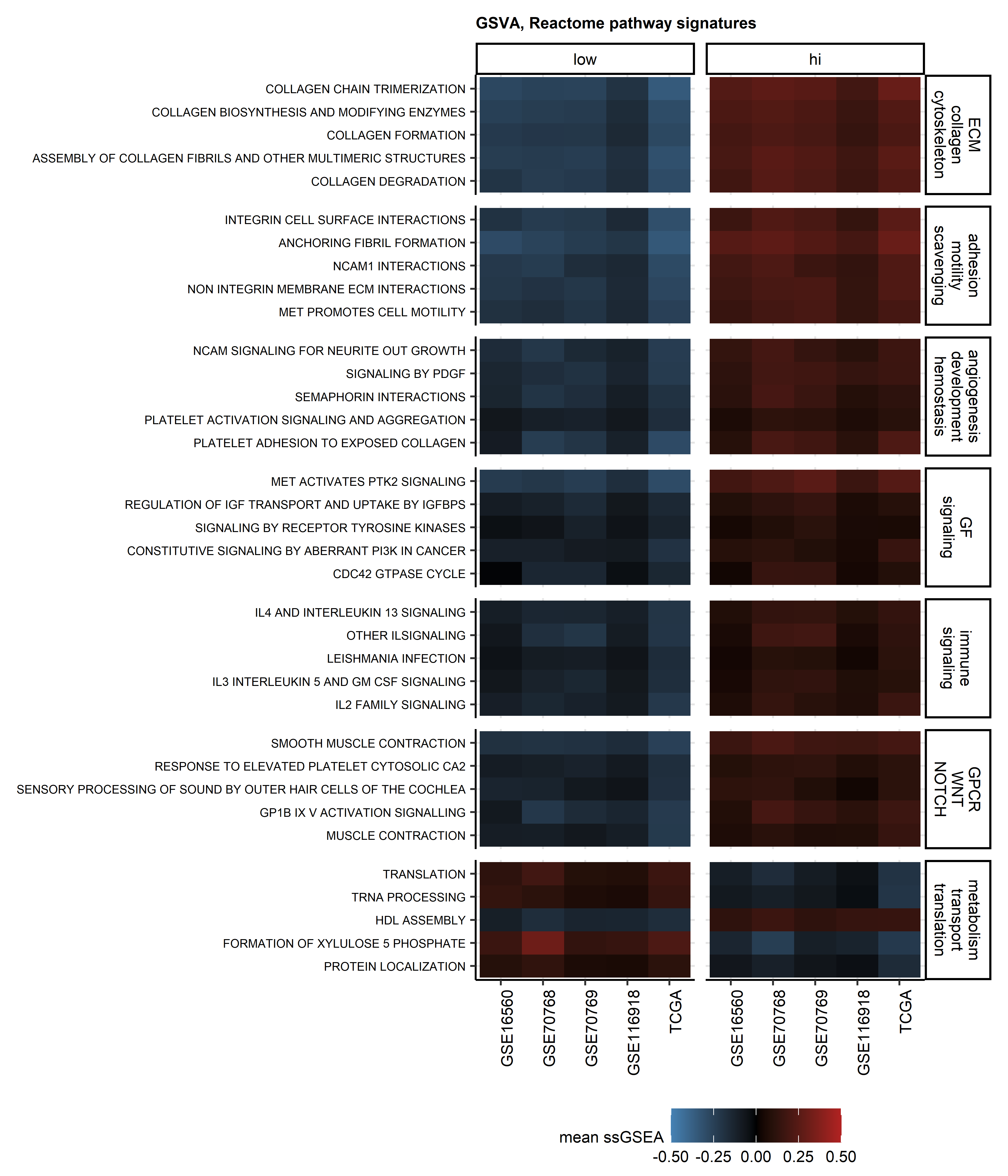
**Supplementary Figure S5. Counts of cancer-associated fibroblast and endothelial cell in tumor samples in the collagen clusters predicted by the MCP counter algorithm.**

*Counts of cancer associated fibroblasts (CAF, A) and endothelial cells (EC, B) in collagenlow (low) and collagenhigh (hi) prostate cancers were predicted by the MCP counter algorithm. Differences between the collagen clusters were assessed by Mann-Whitney test with r effect size statistic. Cell counts were presented in violin plots with single observations visualized as points. Red diamonds and whiskers represent medians with interquartile ranges. Effect sizes and p values are displayed in the plot captions. Numbers of samples in the clusters are indicated in the X axes.*



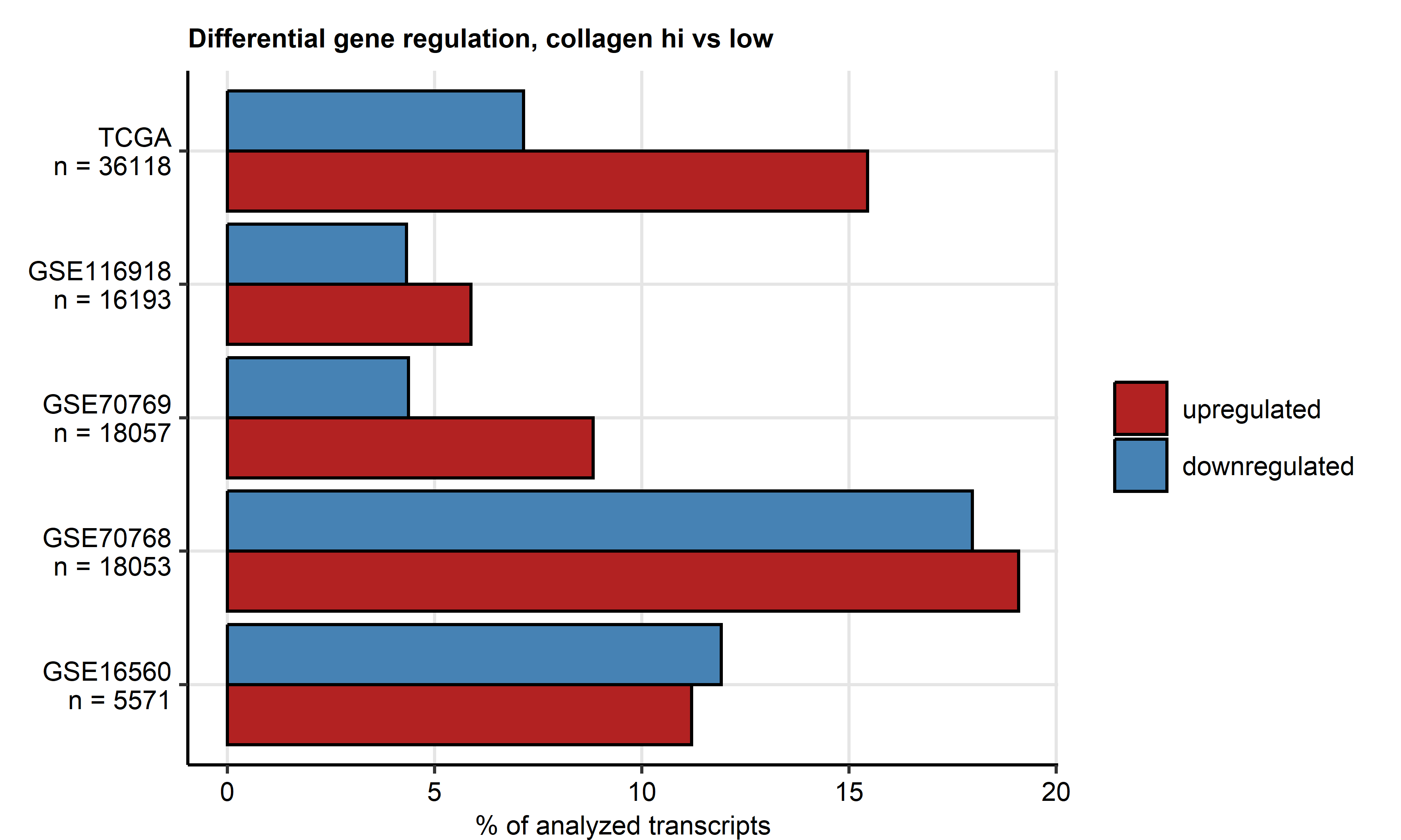
**Supplementary Figure S6. Fractions of cancer-associated fibroblast and endothelial cell in tumor samples in the collagen clusters predicted by the xCell algorithm.**

*Fractions of cancer associated fibroblasts (CAF, A) and endothelial cells (EC, B) in the malignant tissue of collagenlow (low) and collagenhigh (hi) prostate cancers were predicted by the xCell counter algorithm. Differences between the collagen clusters were assessed by Mann-Whitney test with r effect size statistic. Cell counts were presented in violin plots with single observations visualized as points. Red diamonds and whiskers represent medians with interquartile ranges. Effect sizes and p values are displayed in the plot captions. Numbers of samples in the clusters are indicated in the X axes.*



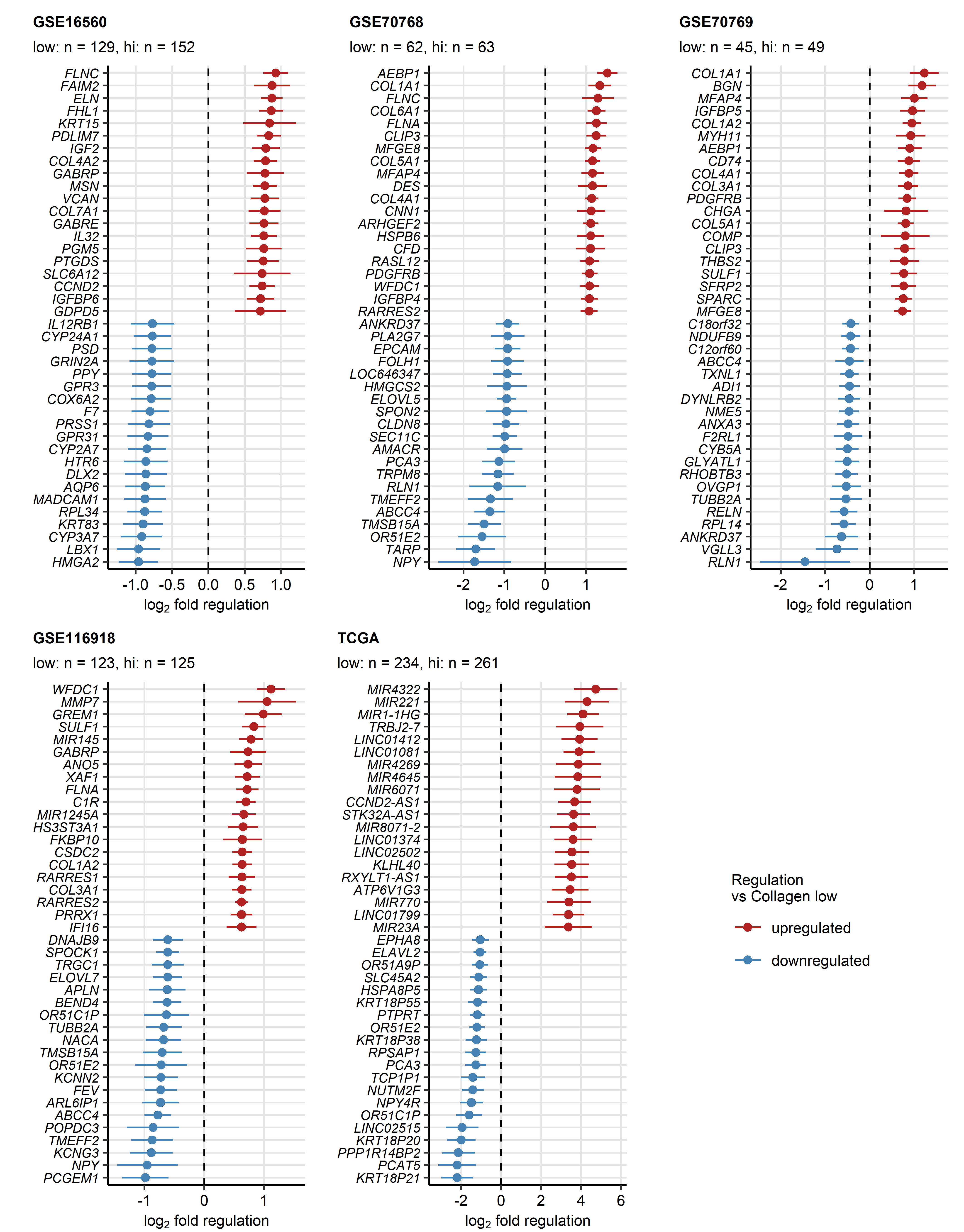
**Supplementary Figure S7. Gene set variation analysis results for representative Reactome pathway gene signatures differentiating between the collagen clusters.**

*Gene set variation analysis of the collagen clusters was performed with the 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. Common regulated signatures were defined as signatures significantly regulated with at least weak effect size (d 0.2) in at least four cohorts (Supplementary Table S7). Mean ssGSEA scores in the clusters and cohorts for the top five strongest regulated signatures for each functional category are presented in a heat map.* *ECM: extracellular matrix, GF: growth factor, GPCR: G protein-coupled receptor.*



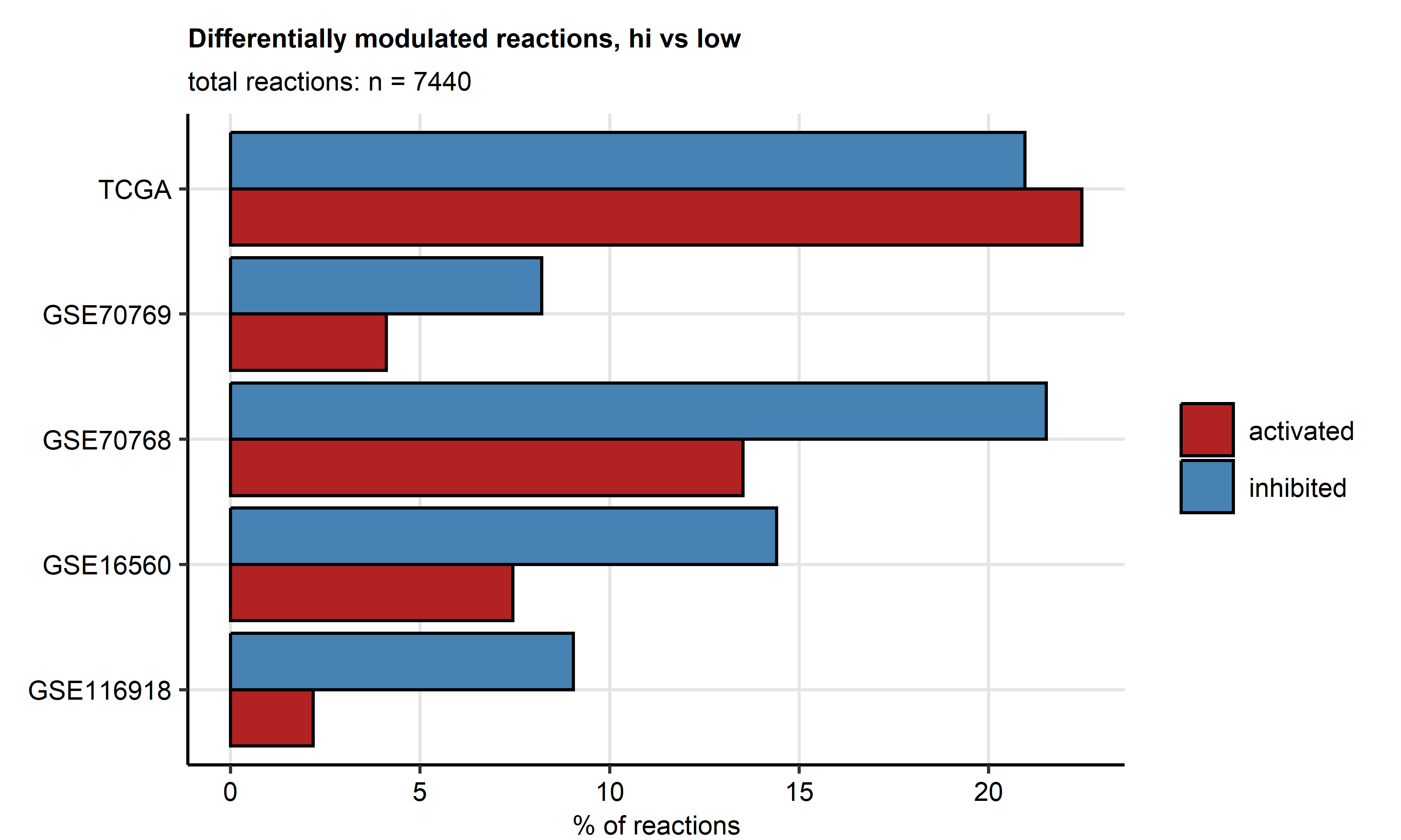
**Supplementary Figure S8. Percentages of the analyzed transcriptome significantly up- and downregulated in collagen high vs collagen low cluster cancers.**

*Genes significantly differentially regulated in collagenhigh cancers (hi) as compared with collagenlow prostate tumors (low) were identified by false discovery rate corrected two-tailed T test (pFDR < 0.05) with Cohen’s d effect size statistic (d 0.2). Percentages of significantly up- and downregulated genes are presented in bar plots. Total numbers of analyzed genes are indicated in the Y axes.*



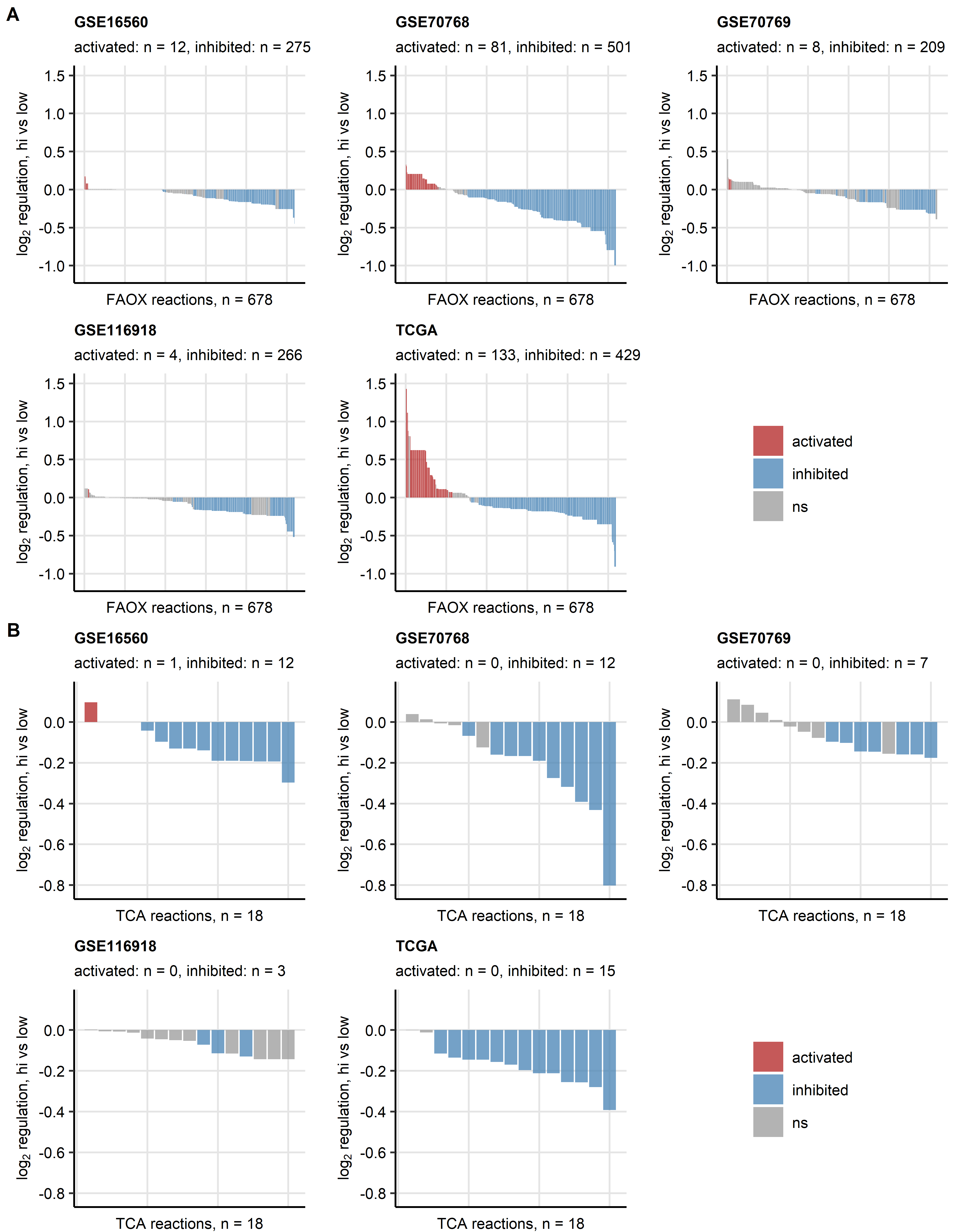
**Supplementary Figure S9. Top strongest differentially expressed genes in the collagen high cluster as compared with collagen low cancers.**

*fold-regulation estimates of differences in expression between collagenhigh (hi) and collagenlow (low) cancers with 95% confidence intervals for top 20 strongest up- and downregulated genes were presented in Forest plots. Numbers of samples in the clusters are displayed in the plot captions.*



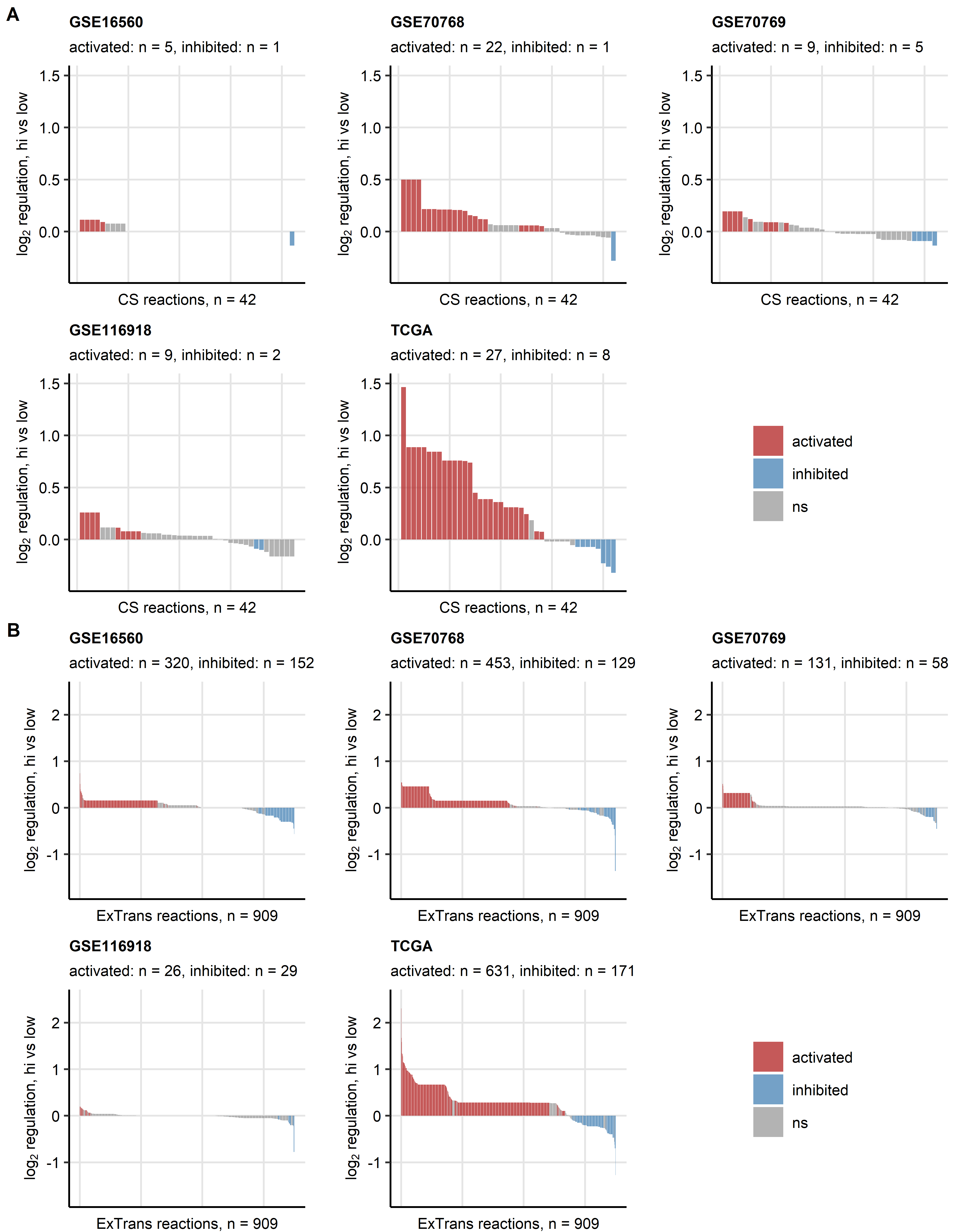
**Supplementary Figure S10. Numbers of significantly activated and inhibited biochemical reactions predicted for the collagen high cluster as compared with collagen low cancers.**

*Modulation of Recon2 model metabolic reactions in collagenhigh cancers (hi) as compared with collagenlow tumors (low) was predicted by the BiGGR and biggrExtra algorithms based on whole-genome differential gene expression estimates. Statistical significance was determined by Monte Carlo simulation with n = 1000 draws from normal distributions of the gene expression regulation estimates. P values were corrected for multiple testing with the false discovery rate method. Percentages of predicted significantly activated and inhibited metabolic reactions are shown in bar plots. The total reaction number is displayed in the plot caption.*



**Supplementary Figure S11. Activity of fatty acid oxidation and citric acid cycle reactions predicted for the collagen high cluster as compared with collagen low cancers.**

*Regulation of reactions of the fatty acid oxidation (FAOX, A) and citric acid cycle (tri-carboxylic acid cycle/TCA, B) Recon subsystems in collagenhigh cancers (hi) as compared with collagenlow tumors (low) was predicted by the BiGGR and biggrExtra algorithms based on whole-genome differential gene expression estimates. Statistical significance was determined by Monte Carlo simulation with n = 1000 draws from distributions of the gene expression estimates. P values were corrected for multiple testing with the false discovery rate method. Estimates of fold-regulation of reaction activity are presented in bar plots. Activity status is color coded. Numbers of significantly activated and inhibited reactions are displayed in the plot captions. Total reaction numbers in the investigated metabolic subsystem are indicated in the X axes.*



**Supplementary Figure S12. Activity of chondroitin synthesis and extracellular transport reactions predicted for the collagen high cluster as compared with collagen low cancers.**

*Regulation of reactions of the chondroitin synthesis (CS, A) and extracellular transport (ExTrans, B) Recon subsystems in collagenhigh cancers (hi) as compared with collagenlow tumors (low) was predicted by the BiGGR and biggrExtra algorithms based on whole-genome differential gene expression estimates. Statistical significance was determined by Monte Carlo simulation with n = 1000 draws from distributions of the gene expression estimates. P values were corrected for multiple testing with the false discovery rate method. Estimates of fold-regulation of reaction activity are presented in bar plots. Activity status is color coded. Numbers of significantly activated and inhibited reactions are displayed in the plot captions. Total reaction numbers in the investigated metabolic subsystem are indicated in the X axes.*

# 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1186/1471-2105-14-7)

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](https://doi.org/10.1093/BIOINFORMATICS/BTS034)

12. 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>

13. 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](https://doi.org/10.1007/978-3-030-32047-8_16)

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

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

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

17. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R. A novel signaling pathway impact analysis. *Bioinformatics* (2009) 25:75–82. doi: [10.1093/bioinformatics/btn577](https://doi.org/10.1093/bioinformatics/btn577)

18. 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](https://doi.org/10.1371/JOURNAL.PONE.0119016)

19. 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](https://doi.org/10.1093/NAR/GKV1049)

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

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

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

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

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

25. Jain S, Lyons CA, Walker SM, McQuaid S, Hynes SO, Mitchell DM, Pang B, Logan GE, McCavigan AM, O’Rourke D, et al. Validation of a Metastatic Assay using biopsies to improve risk stratification in patients with prostate cancer treated with radical radiation therapy. *Annals of oncology : official journal of the European Society for Medical Oncology* (2018) 29:215–222. doi: [10.1093/ANNONC/MDX637](https://doi.org/10.1093/ANNONC/MDX637)

26. 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](https://doi.org/10.1016/j.lungcan.2021.06.020)

27. Cohen J. Statistical Power Analysis for the Behavioral Sciences. *Statistical Power Analysis for the Behavioral Sciences* (2013) doi: [10.4324/9780203771587](https://doi.org/10.4324/9780203771587)

28. 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](https://doi.org/10.1111/j.2517-6161.1995.tb02031.x)

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

30. 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](https://doi.org/10.3923/jse.2014.14.22)

31. 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](https://doi.org/10.1016/0377-0427(87)90125-7)

32. 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](https://doi.org/10.3389/FIMMU.2023.1095195/BIBTEX)