Prognostic and biologic relevance of collagen biosynthesis pathway in prostate cancer

Methods, figures and tables for the transcriptome part

2023-08-07

# Methods

## Magnetic resonance imaging of the prostate

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

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

## Transcriptome analyses

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

### Datasets

Five publicly available transcriptome datasets of primary prostate cancer samples, the TCGA prostate cancer cohort (1) (n = 495), GSE16560 (n = 281) (2), GSE70768 (n = 125) (3), GSE70769 (n = 94) (3), and GSE116918 (n = 248) (4), were re-analyzed. Normalized gene expression data were fetched from the GDC data portal for the TCGA or from the Gene Expression Omnibus for the remaining cohorts. Donor-matched benign and malignant tissue samples were available for the TCGA (n = 52 specimen pairs) and GSE70768 cohort (n = 74). Expression values were -transformed prior to analyses. Non-immune cell content in cancer samples was estimated with the *MCP-counter* and *xCell* algorithms (5,6). single sample gene set enrichment scores (ssGSEA) for the Reactome pathway gene signatures retrieved from the MSig database version 7.5.1 were computed with the *GSVA* algorithm (7).

The collagen pathway genes were retrieved from the published Collagen Signature (8) and the Reactome collagen formation pathway (R-HSA-1474290) and restricted to the gene set present in all investigated datasets (**Supplementary Table S2**).

### Differential expression in benign and malignant prostate tissue

Expression of the collagen pathway gene between patient-matched benign and malignant specimens was compared by the false discovery rate method (FDR) (9) corrected paired two-tailed T test with Cohen’s d effect size statistic (10).

### Collagen clusters

Cancer samples in the training TCGA cohort were clustered in respect to COMBAT-batch adjusted, normalized -transformed gene expression levels of the collagen pathway genes with the PAM algorithm with cosine distance between the samples (11,12). The algorithm choice was motivated by its high explained clustering variance and stability in 10-fold cross-validation as compared with several other procedures (13,14). Two sample clusters were defined (15): collagenhigh and collagenlow named after differences in expression of collagen subunit genes. Cancers samples in the GSE16560, GSE70768, GSE70769 and GSE116918 collectives were assigned to the collagen clusters by an inverse distance weighted k-nearest neighbor classifier (14).

Expression of the collagen pathway genes in the collagen clusters was compared by FDR-adjusted two-tailed T test with Cohen’s d effect size statistic. Differences in clinical variables between the collagen clusters were assessed by FDR-corrected Mann-Whitney test with r effect size statistic or FDR-corrected test with Cramer V effect size statistic.

### Gene set variation analysis and differential gene expression in the collagen clusters

ssGSEA scores of the Reactome signatures and -transformed gene expression levels were compared between the collagen clusters by FDR-corrected two-tailed T tests with Cohen’s d effect size statistic. Signatures and genes significantly regulated between the clusters were defined by the significance cutoff (pFDR < 0.05) and at least weak effect size of the differences between the clusters (d 0.2).

### Activity of signaling pathways and metabolic reactions

Modulation of activity of signaling pathways between the collagen clusters was investigated by the *SPIA* algorithm. Differentially activated or inhibited pathways were identified by the FDR-corrected combined p value for enrichment and pathway perturbation (pGFDR). The magnitude of signaling pathway modulation was measured with the tA parameter (16).

Modulation of metabolic reactions of the Recon2 human metabolism model based on whole-genome differential gene expression was evaluated with the *BiGGR* and [*biggrExtra*](https://github.com/PiotrTymoszuk/biggrExtra) R packages as described (17–19). Statistical significance of reaction activity estimates in the collagenhigh cluster as compared with the collagenlow subset was assessed by a Monte Carlo simulation with n = 1000 draws from normal distribution of gene expression. P values were corrected for multiple testing with the FDR method.

Enrichment of significantly activated or inhibited biochemical reactions within the Recon model metabolic subsystems (19) was investigated by FDR-corrected Fisher’s exact test with odds ratio (OR) effect size statistic. Significantly activated and inhibited metabolic subsystems were defined by the significance cutoff (pFDR < 0.05) and at least weak effect size of the enrichment (OR 1.44).

### Data and code availability

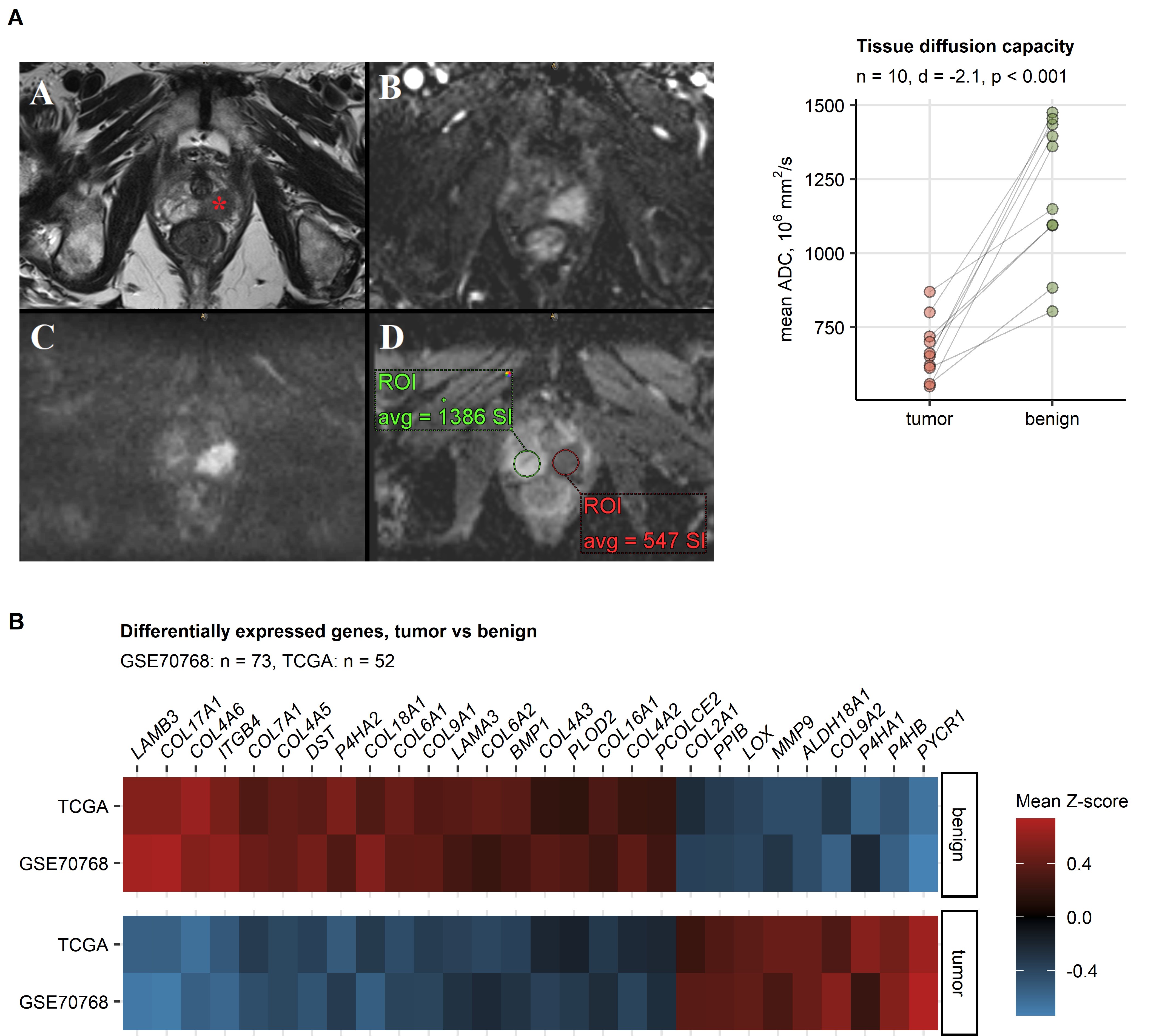
Publicly available data sets were analyzed. Cleared data sets used for analyses will be made available upon request to the corresponding author. The transcriptome R analysis pipeline is available from GitHub (<https://github.com/PiotrTymoszuk/collagen_pca>).

# Tables

Table 1: 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.

| **Variablea** | **GSE16560** | **GSE70768** | **GSE70769** | **GSE116918** | **TCGA** |
| --- | --- | --- | --- | --- | --- |
| Age at diagnosis, years | 74 [IQR: 69 - 79] range: 51 - 91 n = 281 | 62 [IQR: 57 - 67] range: 41 - 93 n = 124 |  | 68 [IQR: 63 - 72] range: 48 - 79 n = 248 | 61 [IQR: 56 - 66] range: 41 - 78 n = 495 |
| PSA at diagnosis |  | 8 [IQR: 6 - 12] range: 3.2 - 280 n = 123 | 8 [IQR: 5.9 - 11] range: 1.5 - 120 n = 90 | 18 [IQR: 11 - 27] range: 3.2 - 220 n = 248 | 0.1 [IQR: 0.03 - 0.11] range: 0 - 320 n = 438 |
| Clinical stage |  | T1: 55% (n = 61) T2: 30% (n = 33) T3: 15% (n = 16) T4: 0% (n = 0) n = 110 | T1: 46% (n = 41) T2: 44% (n = 39) T3: 10% (n = 9) T4: 0% (n = 0) n = 89 |  |  |
| Pathological tumor stage |  | T1: 0% (n = 0) T2: 31% (n = 34) T3: 68% (n = 76) T4: 0.9% (n = 1) n = 111 | T1: 0% (n = 0) T2: 53% (n = 48) T3: 47% (n = 42) T4: 0% (n = 0) n = 90 | T1: 23% (n = 51) T2: 34% (n = 76) T3: 41% (n = 92) T4: 1.8% (n = 4) n = 223 | T2: 38% (n = 176) T3: 62% (n = 283) n = 459 |
| Pathological node stage |  | N0: 91% (n = 82) N1: 8.9% (n = 8) n = 90 | N0: 100% (n = 18) N1: 0% (n = 0) n = 18 |  | N0: 82% (n = 344) N1: 18% (n = 78) n = 422 |
| Pathological metastasis stage |  | M0: 86% (n = 6) M1: 14% (n = 1) n = 7 | M0: 87% (n = 26) M1: 13% (n = 4) n = 30 |  | M0: 99% (n = 453) M1: 0.66% (n = 3) n = 456 |
| Gleason score | 7 [IQR: 6 - 8] range: 6 - 10 n = 281 | 7 [IQR: 7 - 7] range: 6 - 10 n = 122 | 7 [IQR: 7 - 7] range: 5 - 10 n = 91 | 7 [IQR: 7 - 8] range: 6 - 10 n = 248 | 7 [IQR: 7 - 9] range: 6 - 10 n = 495 |
| 6: 30% (n = 83) 7: 42% (n = 117) 8: 9.6% (n = 27) 9: 17% (n = 49) 10: 1.8% (n = 5) n = 281 | 6: 14% (n = 17) 7: 71% (n = 87) 8: 7.4% (n = 9) 9: 5.7% (n = 7) 10: 1.6% (n = 2) n = 122 | 5: 2.2% (n = 2) 6: 20% (n = 18) 7: 62% (n = 56) 8: 5.5% (n = 5) 9: 9.9% (n = 9) 10: 1.1% (n = 1) n = 91 | 6: 17% (n = 42) 7: 40% (n = 99) 8: 21% (n = 52) 9: 22% (n = 54) 10: 0.4% (n = 1) n = 248 | 6: 9.1% (n = 45) 7: 50% (n = 246) 8: 13% (n = 63) 9: 28% (n = 137) 10: 0.81% (n = 4) n = 495 |
| Positive surgical margins |  | 22% (n = 26) n = 119 | 45% (n = 42) n = 93 |  |  |
| Extracapsular extension |  | 69% (n = 77) n = 112 | 46% (n = 42) n = 91 |  |  |
| Death | 73% (n = 206) n = 281 |  |  |  | 1.6% (n = 8) n = 495 |
| Overall survival, months | 100 [IQR: 52 - 150] range: 6 - 270 n = 281 |  |  |  | 17 [IQR: 5.6 - 33] range: 0.033 - 150 n = 495 |
| Biochemical relapse |  | 17% (n = 19) n = 112 | 48% (n = 45) n = 93 | 23% (n = 56) n = 248 | 14% (n = 58) n = 427 |
| Biochemical relapse-free survival, months |  | 30 [IQR: 17 - 49] range: 1 - 65 n = 111 | 58 [IQR: 19 - 80] range: 0.36 - 100 n = 92 | 82 [IQR: 67 - 95] range: 10 - 120 n = 248 | 17 [IQR: 7.1 - 33] range: 0.92 - 150 n = 418 |
| aPSA: prostate-specific antigen. | | | | | |

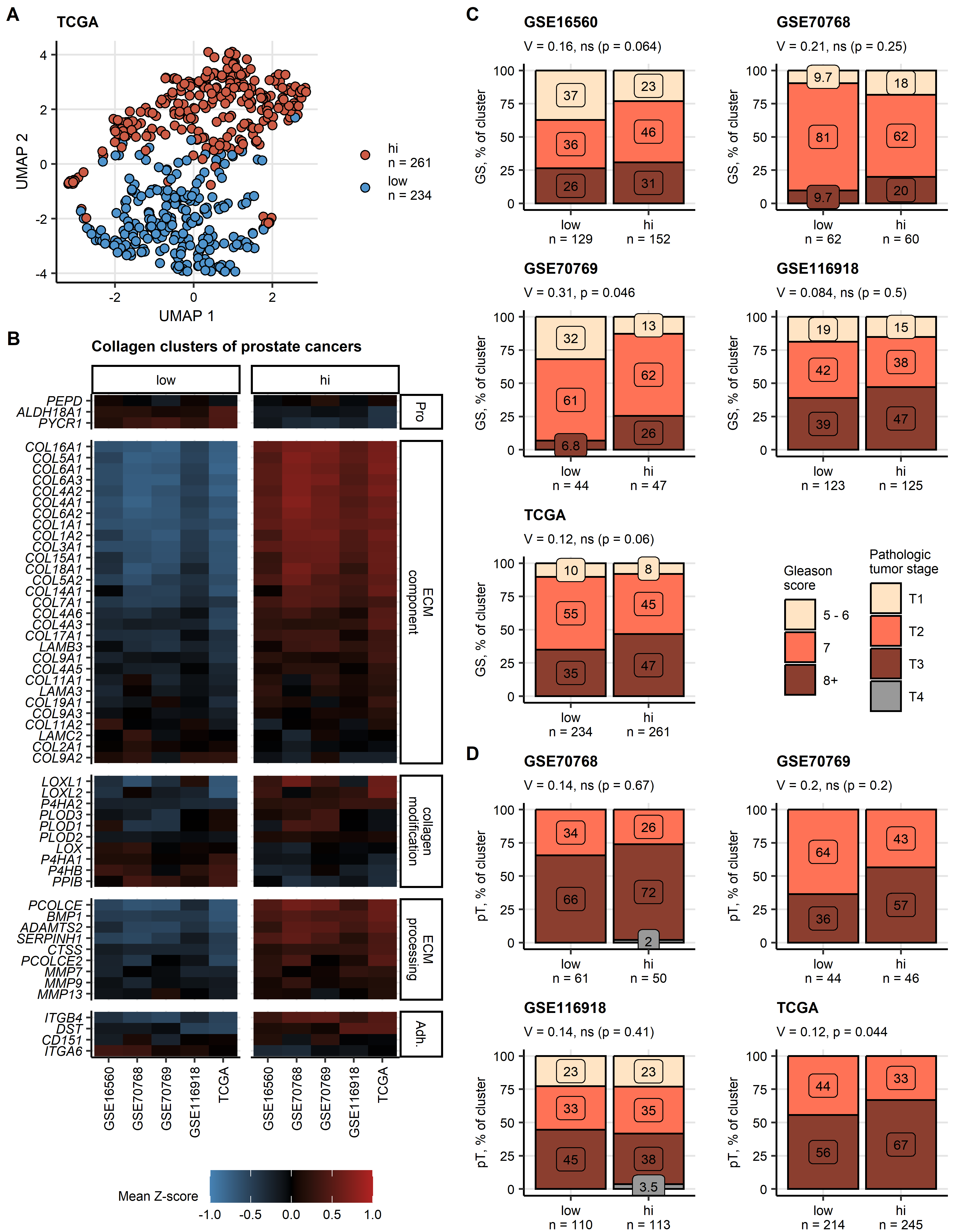
# Figures



**Figure 1. Differences in tissue diffusion capacity and expression of the collagen pathway genes between the prostate cancer and benign tissue.**

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

*(B) Differences in -transformed expression levels of 55 genes related to collagen metabolism between donor-matched tumor and benign prostate tissue were investigated by paired T test with Cohen’s d effect size statistic in the GSE70768 and TCGA cohort. P values were corrected for multiple testing with the false discovery method. Normalized mean -transformed expression levels in benign and cancer tissue for 28 genes significantly regulated in both cohorts are presented in a heat map. Numbers of tissue pairs are displayed in the plot caption.*



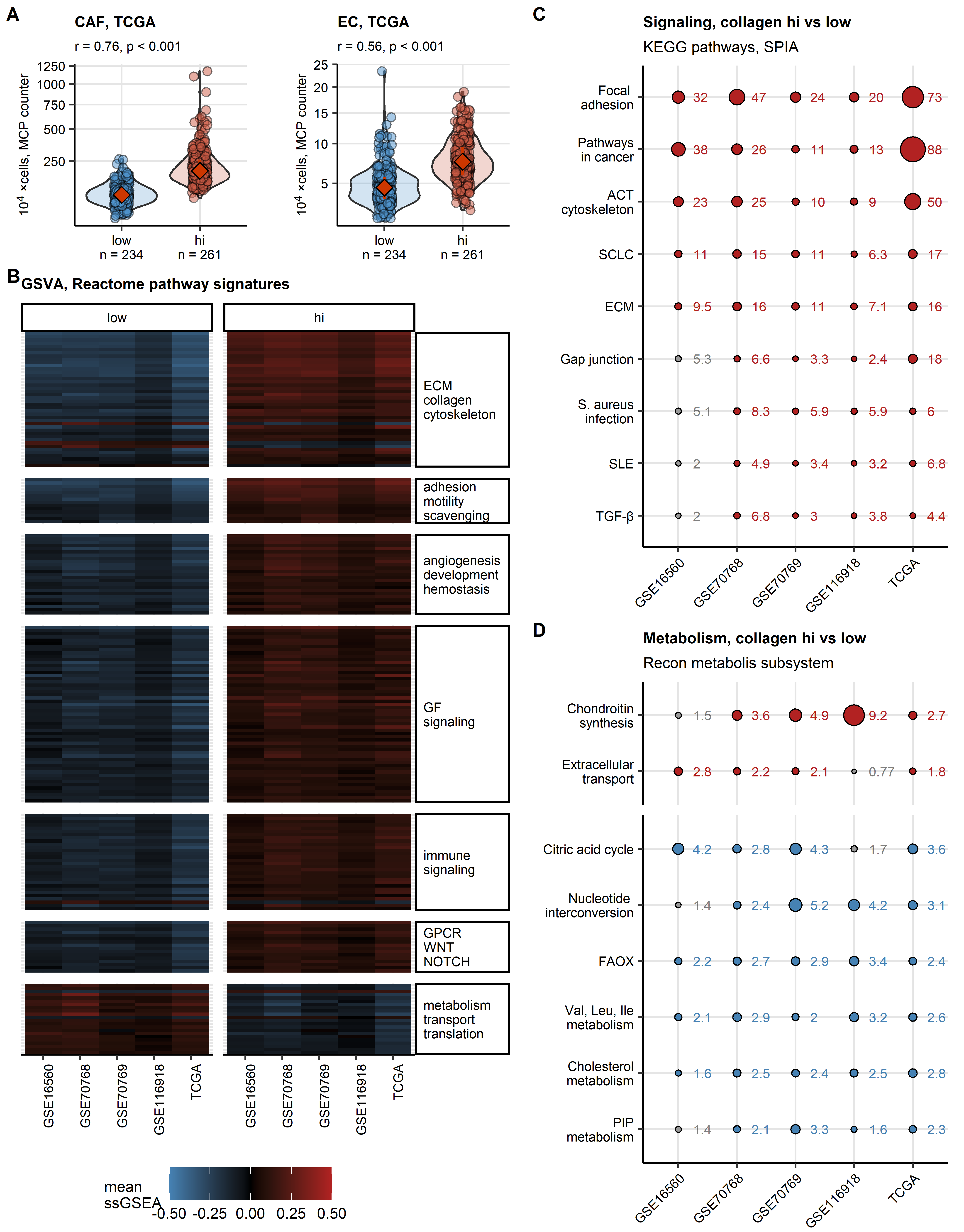
**Figure 2. Collagen clusters of prostate carcinoma and their clinical characteristic.**

*Prostate cancers in the training TCGA cohort were clustered in respect to normalized -transformed expression levels of the collagen pathway genes of interest by the PAM algorithm with cosine distance between the observations. Two clusters, collagenlow (low) and collagenhigh (hi) were defined. Cancer samples from the test cohorts (GSE16560, GSE70768, GSE70769, GSE116918) were assigned to the collagen clusters with an inverse distance weighted k-nearest neighbor classifier.*

*(A) UMAP (uniform manifold approximation and projection) of the normalized -transformed dataset of the collagen gene expression in the training TCGA cohort. Single cancer samples are visualized as points, the collagen cluster assignment is color coded. Numbers of samples in the clusters are indicated in the plot legend.*

*(B) Mean normalized -transformed expression levels of the collagen pathway genes in the collagen clusters and cohorts presented in a heat map. ECM: extracellular matrix, Pro: proline turnover, Adh. adhesion.*

*(C, D) Distribution of Gleason scores (GS, C) and pathological tumor stages (pT, D) in the collagen clusters. Statistical significance was determined by test with Cramer’s V effect size statistic. P values were corrected for multiple testing with the false discovery rate method. Gleason score and tumor stage distributions are shown in stack plots. Effect sizes and p values are displayed in the plot captions. Numbers of observations in the clusters are indicated in the X axes.*



**Figure 3. Infiltration, biological processes, signaling and metabolic pathways differentially regulated in the collagen clusters.**

*(A) Counts of cancer-associated fibroblasts (CAF) and endothelial cells (EC) estimated by the MCP counter algorithm were found to differ significantly between the collagen clusters in all investigated collectives. Statistical significance was assessed by Mann-Whitney test with r effect size statistic. P values were corrected for multiple comparison with the false discovery rate method. Counts of fibroblasts and endothelial cells in the collagen clusters in the TCGA cohort are shown in violin plots. Red diamonds with whiskers represented medians with interquartile ranges. Single tumor samples are visualized as points. Effect sizes and p values are displayed in the plot captions. Numbers of observations in the clusters are indicated in the X axes. Full infiltration analysis results are presented in Supplementary Table S5.*

*(B) 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 four cohorts with at least weak effect size (d 0.2) are presented in a heat map. Full analysis results are presented in Supplementary Table S7. ECM: extracellular matrix, GF: growth factor, GPCR: G protein-coupled receptor.*

*(C) Modulation of KEGG-listed signaling pathways in collagenhigh (hi) tumors as compared with collagenlow cancers was predicted by the SPIA algorithm based on differential gene expression estimates. Predicted magnitude of pathway modulation (tA parameter) for pathways significantly modulated in at least four cohorts was visualized in a bubble plot (tA < 0: inhibition, blue; tA > 0, activation, red, ns: not significant, gray). Points are labeled with their tA values. Full analysis results are presented in Supplementary Table S9. ECM: extracellular matrix, ACT: actin, SCLC: small cell lung carcinoma, S. aureus: Staphylococcus aureus, SLE: systemic lupus erythematosus, TGF-: transforming growth factor-beta..*

*(D) Modulation of Recon2 model metabolic reactions in collagenhigh (hi) tumors as compared with collagenlow cancers was predicted by the BiGGR and biggrExtra algorithms based on whole-genome differential gene expression estimates. Enrichment of significantly activated and significantly inhibited reactions in the collagenhigh within the Recon model metabolic subsystems was investigated by Fisher’s exact test. 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 four cohorts are presented in a bubble plot (inhibited: blue, activated: red, not significant: gray). Points are labeled with their odds ratio values. Full analysis results are presented in Supplementary Table S11. Val: valine, Leu: leucine, Ile: isoleucine, FAOX: fatty acid oxidation, PIP: phosphoinositol phosphate.*

# References

1. 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)

2. 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)

3. 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)

4. 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)

5. 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)

6. 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)

7. 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)

8. 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)

9. 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)

10. 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)

11. 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)

12. 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)

13. 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)

14. 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)

15. 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)

16. 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)

17. 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)

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)