Higher-order dependence in TCGA-SARC gene expression: a case study with RJdCov

Detecting higher-order-only dependence among immune, stromal, and proliferative programs in soft-tissue sarcoma

Motivation and data description. Soft-tissue sarcomas are a heterogeneous family of mesenchymal malignancies with substantial diversity in histology, tumor composition, and immune microenvironment [Cancer Genome Atlas Research Network, 2017]. In bulk RNA-seq each tumor sample reflects a mixture of tumor-intrinsic programs (e.g., proliferation) and microenvironment programs (e.g., immune infiltration, extracellular matrix remodeling), creating multi-axis constraints in which immune function may depend jointly on stromal context and tumor state rather than being well captured by any single pairwise relationship. Evidence that TGF-β–associated stroma contributes to immune exclusion [Mariathasan et al., 2018, Ganesh and Massagué, 2018] and that sarcoma immune landscapes are heterogeneous [Weng et al., 2022] motivates investigating a triplet structure spanning (i) cytotoxic immune activity, (ii) ECM/TGF-β–related stromal remodeling, and (iii) proliferation.

We obtained bulk RNA-seq data (STAR–Counts, primary tumors, open access) for the TCGA-SARC cohort via the `TCGAbiolinks` R/Bioconductor package. After removing genes with fewer than 10 counts in at least 10 samples, we applied TMM normalization (`edgeR`), computed $\log_2$-counts per million, and standardized each gene to zero mean and unit variance, yielding $n = 259$ samples across $\sim$19,000 genes. We defined three biologically motivated modules of six genes each:

- **CYTO (cytotoxic immune):** `TRAC, NKG7, KLRD1, PRF1, GZMB, GNLY`.
- **ECM (TGF-β/stromal):** `TGFB1, SERPINE1, COL1A1, FN1, ACTA2, TAGLN`.
- **PROLIF (proliferation):** `MKI67, TOP2A, CDK1, CCNB1, MCM2, UBE2C`.

Analysis procedure and methods. We tested all $6 \times 6 \times 6 = 216$ single-gene triplets (one gene per module) for pairwise independence and three-way dependence.

For every triplet the test battery comprised: (a) *pairwise independence screening* using both the rank-based distance covariance (RdCov; null pre-computed via Halton permutations, $B_{\text{null}} = 2{,}000$) and the classical distance covariance (dCov; bootstrap, $B = 500$), with Holm correction across the three pairs; (b) *three-way joint independence testing* using both the rank-based joint distance covariance (RJdCov; pre-computed null, $B_{\text{null}} = 2{,}000$) and the classical joint distance covariance (JdCov; permutation test, same B); and additionally the higher-order (Lancaster-type) variants RdCov and HodCov. A triplet is declared "higher-order dependence only" if all Holm-corrected pairwise p-values exceed 0.05 while the joint test (RJdCov or JdCov) rejects at level 0.05.

We treat RJdCov as the primary evidence for higher-order structure and JdCov as supportive, for reasons discussed below.

Results and interpretation. We identified 16 unique triplets exhibiting higher-order-only dependence: 11 detected by RJdCov, 11 by JdCov, with 6 found by both methods. Table 1 lists all 16 triplets with their joint-test p-values. The recurring pattern is biologically coherent: immune cytotoxic markers (`TRAC` appearing in 10 of 16 triplets, alongside `NKG7/PRF1/GZMB/GNLY`), stromal activation markers (`FN1/ACTA2/SERPINE1/COL1A1/TAGLN`), and canonical proliferation markers (`MKI67/TOP2A/CDK1/CCNB1/UBE2C`). This is consistent with a "regime-mixture" interpretation: across a heterogeneous cohort, cytotoxic activity can be high in immune-inflamed tumors yet remains low in tumors with strong ECM/TGF-β programs or in highly proliferative states. Such regime mixtures can produce weak pairwise associations while maintaining a strong three-way constraint, making higher-order dependence testing a principled tool for detecting these patterns.

Both JdCov and RJdCov are grounded in distance-based dependence ideas [Székely et al., 2007], but RJdCov is particularly appealing in TCGA bulk RNA-seq because it is rank/OT-based and therefore more robust to common sources of nuisance variation. TCGA expression measurements across heterogeneous tumors are often heavy-tailed and susceptible to outliers driven by varying tumor purity, extreme infiltration, and batch/normalization effects. In such settings, raw-value distance statistics (JdCov) can be disproportionately influenced by a small number of extreme samples, whereas rank-based procedures dampen outlier leverage and are invariant (or near-invariant) to monotone rescalings that arise from different reasonable preprocessing choices (e.g., log-CPM vs. variance-stabilized transforms). The OT-rank framework further provides a distribution-free calibration strategy and robustness properties that are explicitly emphasized in the RJdCov methodology [Niu and Bhattacharya, 2022]. The fact that six triplets are flagged by both the rank-based (RJdCov) and raw-value (JdCov) procedures provides additional confidence that the signal reflects genuine biological structure rather than a methodological artifact. We note two caveats: (i) the 216 gene-level p-values have not been adjusted for multiplicity and should be interpreted as exploratory; (ii) some portion of the triadic signal may reflect compositional or subtype-mixing effects inherent in bulk RNA-seq, which could be further investigated by stratifying by histology or adjusting for tumor purity.

References

Cancer Genome Atlas Research Network. Comprehensive and integrated genomic characterization of adult soft tissue sarcomas. *Cell*, 171(4):950–965.e28, 2017. doi: 10.1016/j.cell.2017.10.014.

Karuna Ganesh and Joan Massagué. Tgf-β inhibition and immunotherapy: Checkmate. *Immunity*, 48(4):626–628, 2018. doi: 10.1016/j.immuni.2018.03.037.

Sanjeev Mariathasan, Shannon J. Turley, et al. Tgf-β attenuates tumour response to PD-L1 blockade by contributing to exclusion of T cells. *Nature*, 554(7693):544–548, 2018. doi: 10.1038/nature25501.

Ziang Niu and Bhaswar B. Bhattacharya. Distribution-free joint independence testing and robust independent component analysis using optimal transport. *arXiv*, 2022.

Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. doi: 10.1214/009053607000000505.

Weiwei Weng, Lin Yu, Zhang Li, Cong Tan, Jiaojie Lv, I Weng Lao, Wenhuo Hu, et al. The immune subtypes and landscape of sarcomas. *BMC Immunology*, 23:46, 2022. doi: 10.1186/s12865-022-00522-3.

Table 1: Gene-level triplets with higher-order-only dependence. Each triplet passed the pairwise gatekeeping screen (all Holm-corrected pairwise $p > 0.05$) and was rejected by RJdCov, JdCov, or both at level 0.05. The "Detected by" column indicates which joint test(s) flagged the triplet.

CYTO gene	ECM gene	PROLIF gene	p_{RJdCov}	p_{JdCov}
Detected by both RJdCov and JdCov				
TRAC	ACTA2	MKI67	0.014	0.008
TRAC	ACTA2	TOP2A	0.021	0.008
TRAC	FN1	MKI67	0.037	0.026
TRAC	SERPINE1	UBE2C	0.039	0.002
PRF1	FN1	MKI67	0.044	0.046
GNLY	COL1A1	CCNB1	0.037	0.020
Detected by RJdCov only				
TRAC	SERPINE1	MKI67	0.050	0.008
TRAC	SERPINE1	TOP2A	0.030	0.012
NKG7	FN1	CDK1	0.021	0.024
NKG7	FN1	CCNB1	0.025	0.048
GZMB	FN1	MKI67	0.044	0.080
Detected by JdCov only				
TRAC	ACTA2	CDK1	0.079	0.026
TRAC	TAGLN	TOP2A	0.072	0.038
TRAC	TAGLN	CCNB1	0.044	0.038
GNLY	TAGLN	CCNB1	0.013	0.032
GZMB	COL1A1	TOP2A	0.012	0.006