Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Variance-based Decomposition of Inter-patient Transcriptomic Heterogeneity Reveals Recurrent Modes of Therapeutic Antigen Biology Across 33 Cancer Types

XIAOQI HU

sonia549527@gmail.com

Independent Researcher https://orcid.org/0009-0009-9829-6404

Research Article

Keywords: therapeutic antigen heterogeneity, variance decomposition, TANK, CLDN6, PRAME, CLDN18, oncofetal antigen, recurrent antigen modes, pan-cancer, predictive framework, single-cell validation, DepMap, precision oncology

Posted Date: March 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9183099/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Variance-based Decomposition of Inter-patient Transcriptomic Heterogeneity Reveals Recurrent Modes of Therapeutic Antigen Biology Across 33 Cancer Types

Xiaoqi Hu

Independent Researcher, Lower Hutt, New Zealand

sonia549527@gmail.com

Abstract

Background: Inter-patient expression heterogeneity is not merely statistical dispersion but a structured biological signal encoding distinct therapeutic antigen states. Current antigen discovery frameworks over-prioritize mean expression and systematically underinterpret this variance structure, potentially obscuring the most clinically actionable targets. We asked whether inter-patient transcriptomic heterogeneity could be operationalized as a decomposition framework to reveal recurrent, biologically interpretable modes of therapeutic antigen biology.

Methods: We developed TANK (Tumor Antigen prioritization by variance-based raNKing) as a heterogeneity decomposition framework - not primarily a ranking method, but an approach to resolving recurrent antigen modes from patient-level transcriptomic distributions. TANK was applied across 33 TCGA cancer types (n>11,000 patients, 60,656 genes). A 10-gene reference panel was predefined based on independent clinical development status. Non-randomness was confirmed against 1,000 random gene set controls (empirical $p<0.0001$). External validation was performed in two independent gastric cancer cohorts (GEO GSE26942, n=217; ACRG GSE66229, n=400). Single-cell validation was conducted across three cancer types totaling 28,617 annotated tumor epithelial cells. Mode 3 candidates were characterized by survival analysis, immune correlation, and DepMap CRISPR dependency. Beyond the reference panel, CLDN6 was examined as a showcase candidate across the same dimensions.

Results: TANK resolved four recurrent modes of therapeutic antigen heterogeneity - tumor-restricted rare activation (Mode 1: PRAME), lineage-dependent expression (Mode 2: CLDN18), tumor-enriched heterogeneous expression with functional stratification (Mode 3: MSLN/OLFM4/VSIG1/MUC16), and mean-dominant baseline (Mode 4: ERBB2/EGFR) - each with distinct implications for patient stratification, immune context, and translational modality. Concordance between TANK and MAD confirms the signal reflects a robust variance-associated structure rather than any single metric artifact. Beyond the reference panel, CLDN6 ranked top 0.22% (comparable to FDA-approved CLDN18 at top 0.16%), with significant adverse survival association ($p=0.0049$), immune-cold correlates (CD274 $r=-0.204$, $p<0.0001$), and directional CRISPR dependency across gastric (mean=-0.227), colorectal (mean=-0.242), and lung cancer cell lines (mean=-0.288).

Conclusions: Inter-patient transcriptomic variance defines a previously underutilized axis of antigen biology encoding recurrent therapeutic modes that are systematically inaccessible to mean-based approaches — CV-based ranking recovers 0/9 reference targets and DESeq2 fails to prioritize CLDN18 and CLDN6 within the top 20%, while variance decomposition places both in the top 0.22%. The four-mode framework enables prospective mapping of novel candidates to distinct therapeutic strategies from heterogeneity structure alone, without prior biological knowledge of the candidate. CLDN6 exemplifies this capacity: nominated solely from variance structure, its convergent multi-dimensional evidence positions it as a high-priority oncofetal antigen candidate for patients with limited therapeutic options.

Keywords: therapeutic antigen heterogeneity, variance decomposition, TANK, CLDN6, PRAME, CLDN18, oncofetal antigen, recurrent antigen modes, pan-cancer, predictive framework, single-cell validation, DepMap, precision oncology

1. Introduction

Inter-patient expression heterogeneity is a defining feature of human cancers. The same tumor type exhibits dramatic variation in the expression of individual genes across patients - variation that drives differential responses to therapy and underlies the clinical rationale for patient stratification and companion diagnostics. Yet standard computational approaches for therapeutic antigen discovery treat this heterogeneity as noise to be averaged away, prioritizing genes with elevated mean expression across cohorts rather than genes with high inter-patient variability.

This methodological choice has direct clinical consequences. Companion diagnostic-driven therapies - including antibody-drug conjugates, CAR-T cell therapy, and TCR-T cell therapy - depend on identifying antigens expressed at high levels in a tractable patient subpopulation rather than uniformly across all patients. The clinical success of CLDN18.2-targeting therapy (Zolbetuximab, FDA-approved 2024) exemplifies this principle: CLDN18.2 is not elevated in mean expression across gastric cancer patients, but exhibits extreme inter-patient variability defining a companion diagnostic-stratified subpopulation. Conventional differential expression analysis, by design, would deprioritize this target.

We previously developed TANK (Tumor Antigen prioritization by variance-based raNKing) and demonstrated its ability to independently recover CLDN18.2 as a top-ranked outlier in gastric cancer. Here, we extend this work to ask a more fundamental question: is inter-patient transcriptomic variance itself a structured signal that can be decomposed to reveal recurrent, biologically interpretable modes of therapeutic antigen behavior? If so, variance-based prioritization would represent not merely an alternative ranking approach, but a previously underutilized axis of antigen biology that is systematically inaccessible to mean-based methods.

Through integrative analysis across 33 TCGA cancer types, three independent validation cohorts, 28,617 single-cell annotated epithelial cells from three cancer types, null model validation, survival analysis, immune correlation, and DepMap functional dependency data, and systematic comparison against mean-based and dispersion-normalized alternatives, we demonstrate that variance decomposition resolves four recurrent modes of therapeutic antigen heterogeneity that cannot be recovered by differential expression analysis or coefficient of variation ranking — and that this structure is sufficient to prospectively nominate clinically actionable candidates without prior target knowledge, exemplified by CLDN6, a clinically tractable oncofetal antigen in Phase 1/2 CAR-T development.

2. Methods

2.1 Data Sources

RNA-seq STAR count data for 33 TCGA cancer types were obtained from the UCSC Xena hub. Overall survival data were downloaded from the TCGA Pan-Cancer Clinical Data Resource. For external validation: GSE26942 (Illumina HumanHT-12 V4.0, n=217) and GSE66229 (Affymetrix HG-U133 Plus 2.0, n=400, ACRG cohort). For single-cell validation: GSE183904 (gastric cancer, n=31 patients), GSE132465 (colorectal cancer, n=23 patients), and GSE131907 (lung cancer, n=58 patients). Tumor purity estimates (ABSOLUTE method) from the TCGA

PanCanAtlas. DepMap Public 23Q4 CRISPR gene effect scores (Chronos) and cell line annotations from the DepMap portal. All datasets are publicly available.

2.2 TANK Heterogeneity Decomposition Framework

TANK operationalizes inter-patient transcriptomic variance as an entry point for heterogeneity decomposition rather than as a terminal ranking statistic. Gene expression variance was calculated across all patient samples using raw count values. Genes were ranked in descending order of variance. Sex-linked genes (XIST, RPS4Y1, KDM5D) and non-coding RNAs (MALAT1) were excluded. The 10-gene reference panel (CLDN18, ERBB2, EGFR, MSLN, MUC16, PRAME, MUC2, VSIG1, OLFM4, MUC5B) was defined prior to analysis based on independent clinical development status and was not modified after observing TANK results. TANK's concordance with MAD (equivalent enrichment performance, 77.8-fold) confirms that the observed signal reflects a robust variance-associated biological property rather than an artifact of any single statistical metric.

2.3 Non-randomness Validation

Given that variance is invariant to sample permutation by definition, non-randomness was assessed through random gene set controls. For each of 1,000 iterations, a gene set of equivalent size to the reference panel was sampled uniformly at random from the full transcriptome. The empirical p-value was calculated as the proportion of random sets achieving fold enrichment equal to or greater than observed TANK enrichment.

2.4 Recurrent Mode Classification

Four recurrent heterogeneity modes were operationally defined based on convergent evidence across multiple independent data layers: bulk variance rank, external cohort reproducibility, single-cell normal/tumor detection patterns, survival association, immune marker correlation, and DepMap dependency. Classification criteria: Mode 1 - normal epithelial detection <1%, tumor-specific expression in ≥2 cancer types, survival-independent, immune-orthogonal. Mode 2 - cancer-type-specific detection rate patterns reflecting tissue lineage programs. Mode 3 - significantly elevated tumor vs normal epithelial detection (Mann-Whitney $p<0.05$) in ≥2 cancer types, CV>2; subclassified into 3A (survival $p<0.05$), 3B ($p<0.1$), 3C ($p>0.1$ with immune marker correlation). Mode 4 - low CV, mean-dominant, higher DESeq2 than TANK rank. Genes may exhibit characteristics of multiple modes depending on cancer type context.

2.5 Method Comparison

Five heterogeneity metrics were compared: raw count variance (TANK), log2(count+1)-normalized variance, CV, MAD, and residual variance. DESeq2 differential expression was performed using pydeseq2 (tumor n=412 vs normal n=36). TANK's contribution lies in the decomposition framework and biological interpretation rather than in the statistical metric itself; concordance with MAD is interpreted as supporting signal robustness.

2.6 External Validation

TANK variance ranking was applied identically to GSE26942 and GSE66229 after platform-specific probe-to-gene-symbol mapping. Fisher's exact test assessed reference panel enrichment at top 1%, 5%, and 10% thresholds.

2.7 Single-cell RNA-seq Analysis

For GSE132465, published cell type annotations were used. For GSE183904 and GSE131907, epithelial cells were identified by co-expression of at least two markers (EPCAM, CDH1, KRT8, KRT18, KRT19, CLDN4 >0.5 counts). Detection rate: proportion of epithelial cells with expression >0 counts. Mann-Whitney U test compared tumor vs normal distributions.

2.8 Mode 3 Functional Characterization

Mode 3 candidates (MSLN, OLFM4, VSIG1, MUC16) were characterized by: cross-cohort TANK rank; Kaplan-Meier survival analysis in TCGA-STAD (n=390, 75th percentile cutoff, DECEASED event definition); Spearman correlation with CD8A, CD3D, IFNG, CD274, CD68, CD2; and single-cell epithelial detection rate across three cancer types.

2.9 CLDN6 Showcase Analysis

CLDN6 (ENSG00000184697) was examined as a reference-panel-external showcase candidate. TANK rank was calculated in TCGA-STAD. Survival analysis used the same protocol as Mode 3 candidates (n=390 matched, 75th percentile cutoff). Immune correlation: Spearman with CD8A, CD3D, IFNG, CD274. DepMap CRISPR dependency (Chronos) were extracted for Esophagus/Stomach (n=69), Bowel (n=63), and Lung (n=126) cell lines from DepMap Public 23Q4. Normal tissue expression was assessed via public reference resources; prior literature consistently indicates CLDN6 is largely silent in adult normal tissues and reactivated in cancer, consistent with its oncofetal antigen classification.

2.10 Tumor Purity Analysis

Spearman correlation between ABSOLUTE purity estimates and log2(count+1) expression across 294 matched TCGA-STAD samples. Genome-wide purity correlations were compared between TANK Top 1% genes and 10,000 randomly sampled genes using Mann-Whitney U test.

3. Results

3.1 Variance exposes non-random structure in therapeutic antigen heterogeneity

TANK identified clinically validated targets at 77.8-fold enrichment in the top 1% of variance-ranked genes (7/9 predefined targets, $p<0.0001$). None of 1,000 random gene sets of equivalent size recapitulated this enrichment (random mean 1.1x, empirical $p<0.0001$), confirming that clinically validated therapeutic targets are non-randomly distributed in variance space. TANK's concordance with MAD confirms that this signal reflects a robust variance-associated biological property rather than any single metric artifact. CV-based ranking recovered 0/9 reference targets; DESeq2 ranked CLDN18.2 in the 86th percentile while TANK placed it in the top 0.16%. Tumor purity analysis showed TANK Top 1% genes had lower purity correlation than the transcriptome-wide background ($r=0.004$ vs $r=0.045$, $p<0.0001$). [Supplementary Figure 2, Figures 9-13]

3.2 Therapeutic antigens occupy recurrent and biologically distinct heterogeneity modes

Integration of bulk variance analysis, single-cell RNA-seq across three cancer types, survival analysis, immune correlation, and DepMap functional dependency revealed that TANK-prioritized antigens occupy recurrent biological modes with distinct implications for patient stratification, immune context, and translational modality. These modes were operationally

defined based on convergent evidence across independent data layers. [Figure 1, Figure 3, Table 1]

Mode 1 - Tumor-restricted rare activation (PRAME). PRAME ranked top 1% in 28/33 TCGA cancer types. Single-cell analysis confirmed near-zero normal epithelial detection (0-0.4%) with tumor-specific expression across all three cancer types ($p<0.0001$ in lung). Survival-independent ($p=0.31$), immune-orthogonal ($r\sim0$ with all four immune markers). GTEx confirmed restricted normal expression (95.6% of tissues <1 TPM). Consistent with Phase 3 SUPRAME trial (ORR 50-67%), validating Mode 1 as the computational signature of TCR-T-tractable cancer-testis antigens. [Figures 2-9]

Mode 2 - Lineage-dependent expression (CLDN18). CLDN18 showed cancer-type-specific profiles: gastric high detection in both tumor (43.0%) and normal (43.1%) epithelium; colorectal tumor-restricted (1.1% vs 0.0%, $p=0.0005$); lung dominant normal expression (50.5% vs 9.8% tumor, $p<0.0001$). This cancer-type-specific biology explains the FDA-approved gastric but not lung cancer indication, demonstrating that TANK-identified heterogeneity reflects genuine tissue lineage programs. [Figures 14-15]

Mode 3 - Tumor-enriched heterogeneous expression with functional stratification (MSLN, OLFM4, VSIG1, MUC16). All four Mode 3 candidates ranked in the TANK top 1% across three independent cohorts. Survival analysis defined three subtypes: Mode 3A (MSLN, $p=0.0137$) - tumor-intrinsic heterogeneity with clinically relevant prognostic association. Mode 3B (OLFM4, $p=0.0752$) - extreme intra-tumor heterogeneity without clear survival coupling. Mode 3C (MUC16, VSIG1, $p>0.29$) - survival-independent but immune-correlated (MUC16-CD274 $r=0.285$, $p<0.0001$), consistent with microenvironment-driven variance. This stratification demonstrates that variance identifies candidates, but biological integration distinguishes tumor-intrinsic from context-dependent signals. [Figures 16-17]

Mode 4 - Mean-dominant baseline comparator (ERBB2, EGFR). Low CV values and preferential DESeq2 identification confirm these as broadly overexpressed targets for which conventional differential expression is appropriate, validating that TANK and DESeq2 capture orthogonal biological dimensions. [Figure 9]

Table 1. Recurrent therapeutic antigen modes: classification criteria and translational implications

Mode	Exemplar	Multi-layer classification criteria	Survival/Immune profile	Predicted translational modality
1: Tumor-restricted rare activation	PRAME	Normal <1%; tumor-specific ≥2 types; survival-independent; immune-orthogonal	$p=0.31$; $r\sim0$ immune	TCR-T / cancer-testis targeting
2: Lineage-dependent expression	CLDN18	Cancer-type-specific detection; high CV in lineage of origin	Survival-independent	Companion diagnostic-driven organ-specific therapy
3A: Tumor-intrinsic, clinically relevant	MSLN	Tumor-enriched ≥2 types; CV>2;	$p=0.014$	ADC with prognostic stratification

		survival-associated		
3B: Tumor-intrinsic, biologically active	OLFM4	Extreme intra-tumor CV; tumor-enriched; survival-marginal	p=0.075	ADC candidate; biomarker development
3C: Microenvironment-linked	MUC16, VSIG1	Tumor-enriched; immune marker-correlated; survival-independent	CD274 r=0.285	Combination immunotherapy
4: Mean-dominant baseline	ERBB2, EGFR	Low CV; mean-elevated; higher DESeq2 than TANK rank	Variable	Conventional DE approach appropriate

3.3 TANK enrichment is reproducible across independent cohorts and platforms

TANK was applied to two independent gastric cancer microarray cohorts: GEO GSE26942 (Illumina, n=217, 40-fold, $p<0.0001$) and ACRG GSE66229 (Affymetrix, n=400, 44-fold, $p<0.0001$). Platform-associated attenuation is consistent with reduced microarray dynamic range. Targets reproducibly prioritized across all three cohorts: CLDN18, MSLN, VSIG1, OLFM4. [Figure 11]

3.4 Single-cell validation confirms intra-tumor epithelial heterogeneity

Across 28,617 annotated epithelial cells from three cancer types, TANK-prioritized targets showed consistently high intra-tumor CV (range 2.7-95.0). PRAME demonstrated the clearest Mode 1 pattern. OLFM4 showed extreme intra-tumor heterogeneity (gastric CV=13.4). The CLDN18 cancer-type-specific pattern confirms TANK-identified bulk variance reflects genuine tissue lineage biology. [Figures 14-15]

3.5 CLDN6: TANK nominates a high-priority oncofetal antigen candidate from heterogeneity structure

To demonstrate TANK's capacity for de novo antigen nomination beyond the predefined reference panel, we systematically examined the top-ranked transcriptome-wide outputs. CLDN6 (Claudin-6) ranked 132/60,656 (top 0.22%) in TCGA-STAD, comparable to FDA-approved CLDN18 (top 0.16%), yet was not included in the predefined reference panel due to earlier-stage clinical development status at study initiation.

CLDN6 is classified as an oncofetal antigen: public reference resources and prior literature consistently indicate that CLDN6 expression is largely silent in adult normal tissues and reactivated in cancer. Survival analysis in TCGA-STAD demonstrated significant adverse overall survival association (log-rank $p=0.0049$, n=390). Immune correlation analysis revealed consistent negative associations with CD8A ($r=-0.162$, $p=0.001$), CD3D ($r=-0.126$, $p=0.011$), IFNG ($r=-0.140$, $p=0.005$), and CD274 ($r=-0.204$, $p<0.0001$), indicating CLDN6-high tumors exhibit an immune-cold, immune-excluded microenvironment. This pattern suggests that CLDN6-high patients may specifically benefit from adoptive cellular therapies that do not require pre-existing immune infiltration.

DepMap CRISPR dependency analysis across 258 cell lines revealed directional negative scores across gastric (mean=-0.227, 26% <-0.3, n=69), colorectal (mean=-0.242, 38% <-0.3, n=63), and lung cancer (mean=-0.288, 40% <-0.3, n=126) cell lines. This subset-specific functional dependency pattern is consistent with TANK's heterogeneity-driven nomination logic: CLDN6 is not pan-essential, but represents a context-specific vulnerability in a tractable patient subpopulation.

Taken together, CLDN6's combination of high heterogeneity, adverse prognosis, immune exclusion, and cross-lineage directional dependency positions it at the intersection of Mode 1 (oncofetal silencing in normal tissue) and Mode 3A (tumor-enriched with clinical relevance) - a mode assignment that directly predicts its therapeutic modality as a CAR-T target in immune-cold tumors, consistent with BNT211 (BioNTech, Phase 1/2, Nature Medicine 2023). [Figure 18]

3.6 Integrative Therapeutic Antigen Value Matrix

The Therapeutic Antigen Value Matrix integrates TANK heterogeneity score, clinical survival association, and scRNA-seq tumor specificity, spatially separating the four modes. CLDN6, with TANK score comparable to CLDN18 but stronger adverse survival association, positions in the high-value target region, confirming its nomination as the top reference-panel-external candidate. [Figure 1]

4. Discussion

The central contribution of this study is not the introduction of variance as a new statistical metric, but the demonstration that inter-patient transcriptomic variance encodes a structured biological space — one that is systematically inaccessible to mean-based or dispersion-normalized approaches, and that organizes therapeutic antigen candidates into recurrent modes with distinct clinical implications. This inaccessibility is not a theoretical claim: CV-based ranking recovers 0/9 predefined reference targets in the top 1%, DESeq2 ranks CLDN18 in the 86th percentile while variance decomposition places it in the top 0.16%, and no random gene set of equivalent size recapitulates the observed 77.8-fold enrichment across 1,000 permutations. The modes are not imposed by the framework — they are recovered by it. TANK’s concordance with MAD confirms that the signal resides in the variance structure of the transcriptome itself, not in the choice of dispersion statistic: the biological organization is real, and multiple metrics that capture inter-patient variance will find it, while metrics that normalize or average it away will not.

The non-randomness validation establishes a statistical foundation: none of 1,000 random gene sets recapitulated TANK's 77.8-fold enrichment ($p<0.0001$), confirming that clinically validated therapeutic targets are non-randomly distributed in variance space. This finding demonstrates that high inter-patient expression variance is a biologically structured property preferentially marking antigens with established therapeutic relevance - a principle that holds across multiple development modalities from TCR-T to ADC to CAR-T.

Beyond descriptive categorization, the four-mode framework provides a predictive mapping from heterogeneity patterns to translational strategies. Given a novel candidate gene, its placement within the TANK-derived heterogeneity space - defined by variance magnitude, normal tissue detection, immune correlation, survival association, and functional dependency - enables inference of its likely therapeutic modality. Genes exhibiting tumor-restricted rare activation with immune-orthogonal profiles are predicted to be compatible with TCR-T approaches. Lineage-dependent high-variance genes require companion diagnostic-driven stratification. Variance signals coupled with immune marker correlation suggest microenvironment-associated biology rather than tumor-intrinsic antigen programs, directing toward combination immunotherapy rather than single-agent targeting. Variance signals combined with adverse prognosis and

immune exclusion, as in CLDN6, predict suitability for adoptive cellular therapies in immune-cold tumors. In this sense, TANK functions not only as a discovery framework, but as a decision layer linking transcriptomic heterogeneity to translational prioritization.

The CLDN6 nomination exemplifies this predictive capacity. CLDN6 was not included in the predefined reference panel, yet variance decomposition placed it in the same tier as FDA-approved CLDN18. Its convergent evidence profile - adverse prognosis, immune-cold microenvironment, oncofetal expression, cross-lineage directional DepMap dependency in 26-40% of cancer cell lines per lineage - was predicted entirely from heterogeneity structure before any mechanistic investigation. The immune-cold association is mechanistically informative: CLDN6-high tumors with low CD8A, CD3D, IFNG, and CD274 represent a population that may specifically benefit from CAR-T therapy, consistent with BNT211's clinical rationale. This positions CLDN6 as a candidate at the intersection of oncofetal antigen biology and immune-excluded tumor microenvironment - a combination that TANK's heterogeneity decomposition framework is uniquely positioned to identify.

Critically, not all variance-derived signals correspond to clinically actionable targets. The Mode 3C subclass illustrates that some high-variance signals arise from microenvironmental composition rather than tumor cell-intrinsic programs. Variance identifies candidates; biological and clinical integration distinguishes actionable targets from context-dependent signals. This interpretive framework, requiring multiple orthogonal evidence layers for nomination, is essential for translating computational variance signals into therapeutic development priorities.

Limitations include reliance on bulk RNA-seq as the primary discovery platform, incomplete single-cell coverage across all 33 cancer types, absence of direct experimental validation for CLDN6 and other newly nominated candidates, and lack of prospective validation of the mode classification framework in independent cohorts. Future work incorporating functional genomics and spatial transcriptomics will provide additional resolution.

5. Conclusion

Variance-based decomposition of inter-patient transcriptomic heterogeneity resolves four recurrent modes of therapeutic antigen biology across 33 cancer types, validated by null model analysis, single-cell characterization of 28,617 epithelial cells, and three independent bulk cohorts. Inter-patient transcriptomic variance defines a previously underutilized axis of antigen biology, and the four-mode framework enables prospective mapping of novel candidates to distinct therapeutic strategies based on heterogeneity structure alone. CLDN6, nominated entirely from variance structure and confirmed by convergent multi-dimensional evidence, exemplifies this predictive capacity - potentially expanding therapeutic options for patients in cancer types where effective targeted therapies remain limited.

Data Availability

TCGA RNA-seq: https://gdc-hub.s3.us-east-1.amazonaws.com/download/

GEO GSE26942: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26942

GEO GSE66229: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE66229

GEO GSE183904: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE183904

GEO GSE132465: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132465

GEO GSE131907: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE131907

DepMap Public 23Q4: https://depmap.org/portal/download/all/

TCGA PanCanAtlas purity: https://gdc.cancer.gov/about-data/publications/pancanatlas

Code: https://github.com/ohahouhui/AI-CAR-Loop-1.0

References

1. Braumuller H, et al. PRAME-specific TCR-T cell therapy (IMA203) in advanced solid tumors. Nature Medicine. 2025.

2. Shitara K, et al. Zolbetuximab plus mFOLFOX6 in CLDN18.2-positive gastric cancer (SPOTLIGHT). Lancet. 2023;401:1655-1668.

3. Mackensen A, et al. CLDN6-specific CAR-T cells plus amplifying RNA vaccine in relapsed or refractory solid tumors (BNT211-01). Nature Medicine. 2023;29:2844-2853.

4. Goldman MJ, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nature Biotechnology. 2020;38:675-678.

5. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020;369:1318-1330.

6. Liu J, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource. Cell. 2018;173:400-416.

7. Love MI, et al. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15:550.

8. Uhlen M, et al. Tissue-based map of the human proteome. Science. 2015;347:1260419.

9. Cristescu R, et al. Molecular analysis of gastric cancer identifies subtypes associated with distinct clinical outcomes. Nature Medicine. 2015;21:449-456.

10. Aran D, et al. Systematic pan-cancer analysis of tumour purity. Nature Communications. 2015;6:8971.

11. Kim N, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nature Communications. 2020;11:2285.

12. Lee HO, et al. Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer. Nature Genetics. 2020;52:594-603.

13. Zhang P, et al. Dissecting the single-cell transcriptome network underlying gastric premalignant lesions and early gastric cancer. Cell Reports. 2022;38:110232.

14. Dempke WCM, et al. Claudin-6: an oncofetal target for cancer immunotherapy. Cancers. 2023;15:1366.

15. DepMap. DepMap 23Q4 Public. Figshare. 2023. https://doi.org/10.6084/m9.figshare.24278053.v1

16. Hu X. Variance-based transcriptomic screening identifies CLDN18.2 as a top-ranked antigen candidate in gastric cancer. Research Square. 2026.

Figure Legends

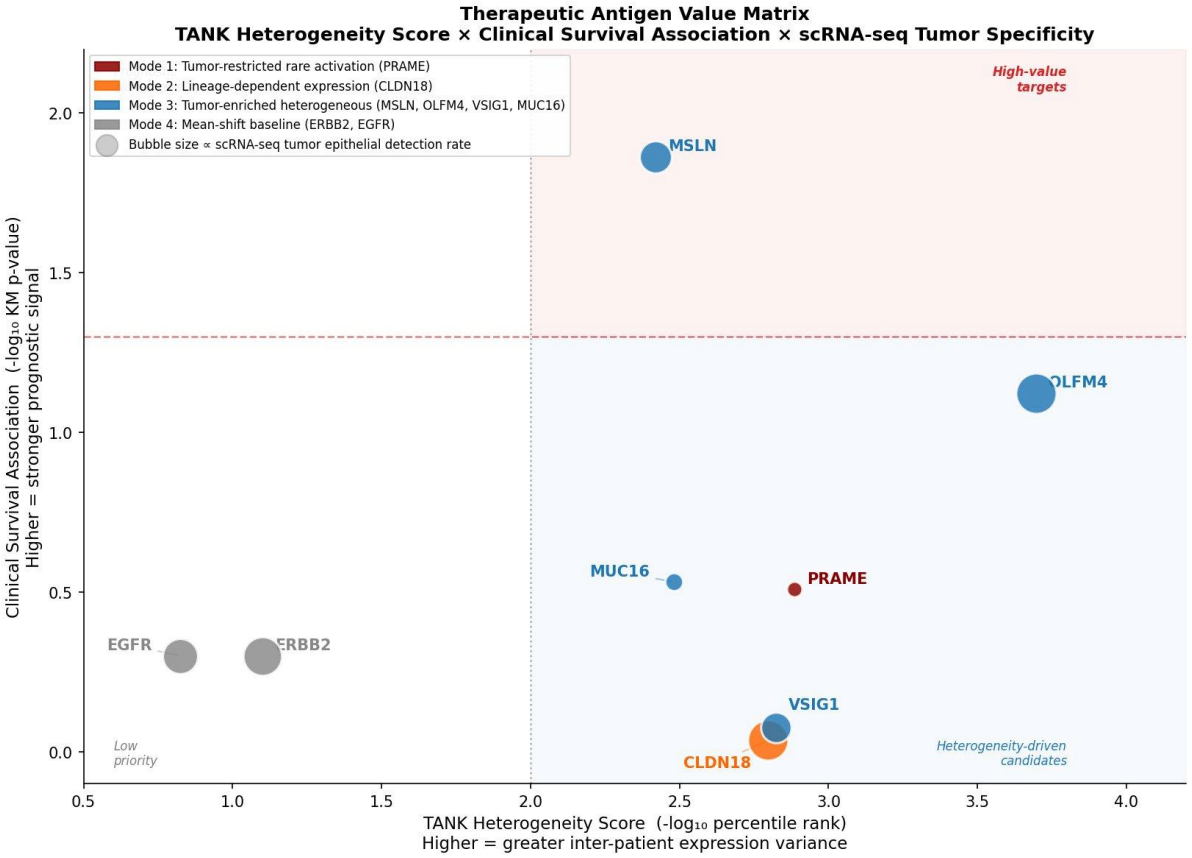

Therapeutic Antigen Value Matrix
TANK Heterogeneity Score × Clinical Survival Association × scRNA-seq Tumor Specificity
Mode 1: Tumor-restricted rare activation (PRAME)
Mode 2: Lineage-dependent expression (CLDN18)
Mode 3: Tumor-enriched heterogeneous (MSLN, OLFM4, VSIG1, MUC16)
Mode 4: Mean-shift baseline (ERBB2, EGFR)
Bubble size ∝ scRNA-seq tumor epithelial detection rate
High-value targets
MSLN
OLFM4
MUC16
PRAME
EGFR
ERBB2
VSIG1
CLDN18
Low priority
Heterogeneity-driven candidates
Clinical Survival Association (-log₁₀ KM p-value)
Higher = stronger prognostic signal
TANK Heterogeneity Score (-log₁₀ percentile rank)
Higher = greater inter-patient expression variance
0.0
0.5
1.0
1.5
2.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Figure 1. Therapeutic Antigen Value Matrix. TANK heterogeneity score vs clinical survival association, bubble size proportional to scRNA-seq tumor epithelial detection rate. Four modes color-coded. CLDN6 positioned alongside reference panel targets.

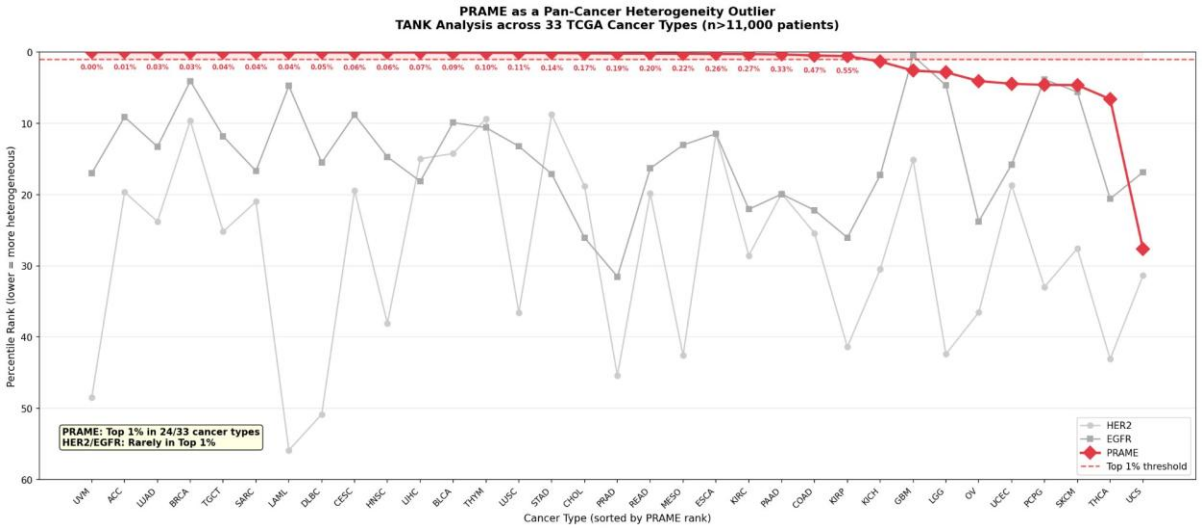


Figure 2. PRAME pan-cancer variance ranking across 33 TCGA cancer types.

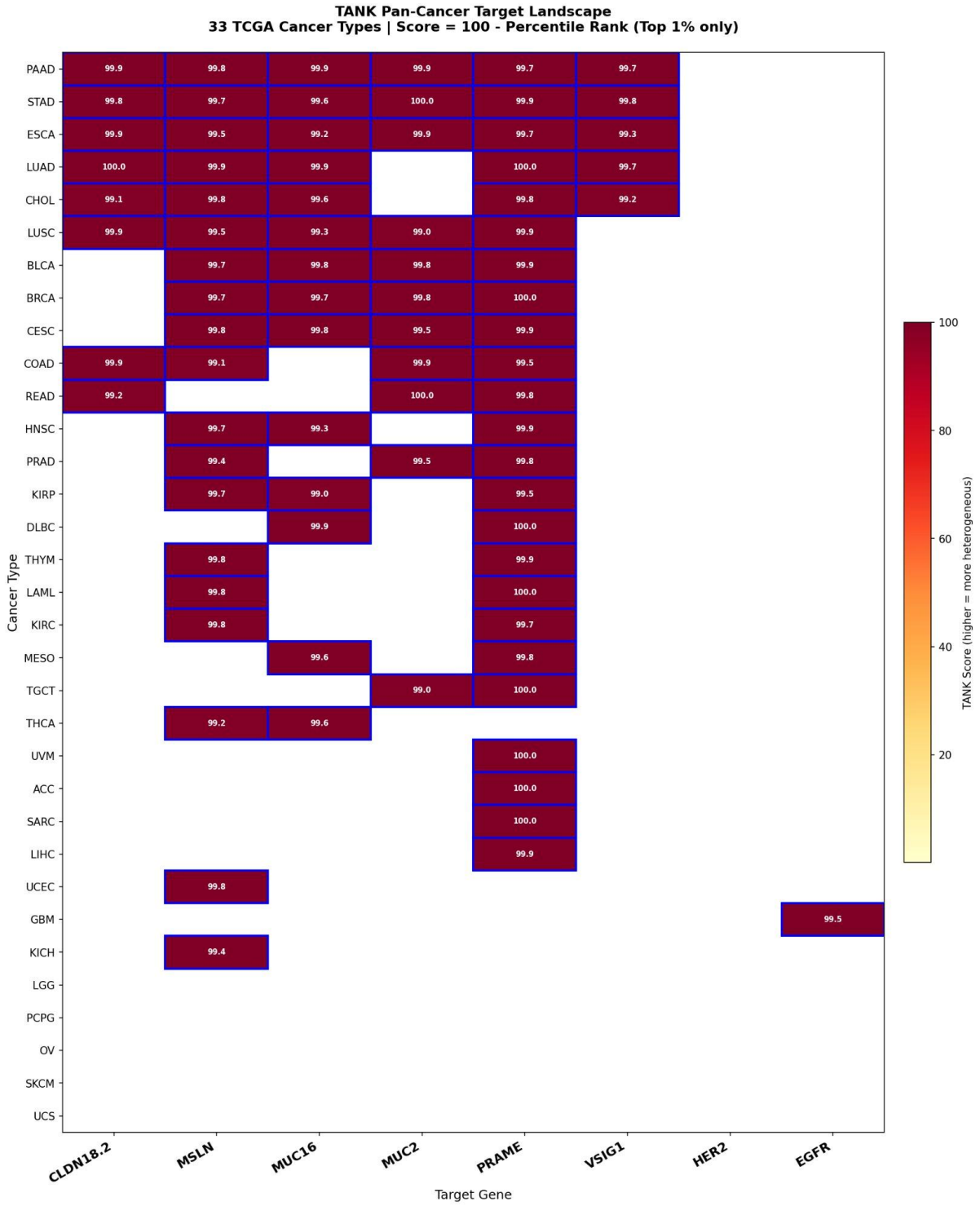


Figure 3. Pan-cancer TANK heatmap for 8 reference targets across 33 TCGA cancer types.

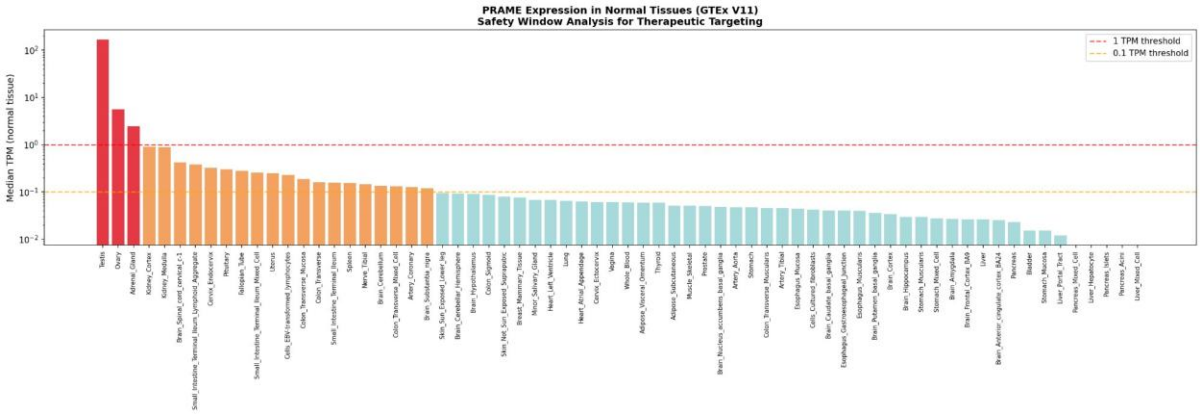


Figure 4. PRAME normal tissue safety (GTEx V11, n=68 tissues).

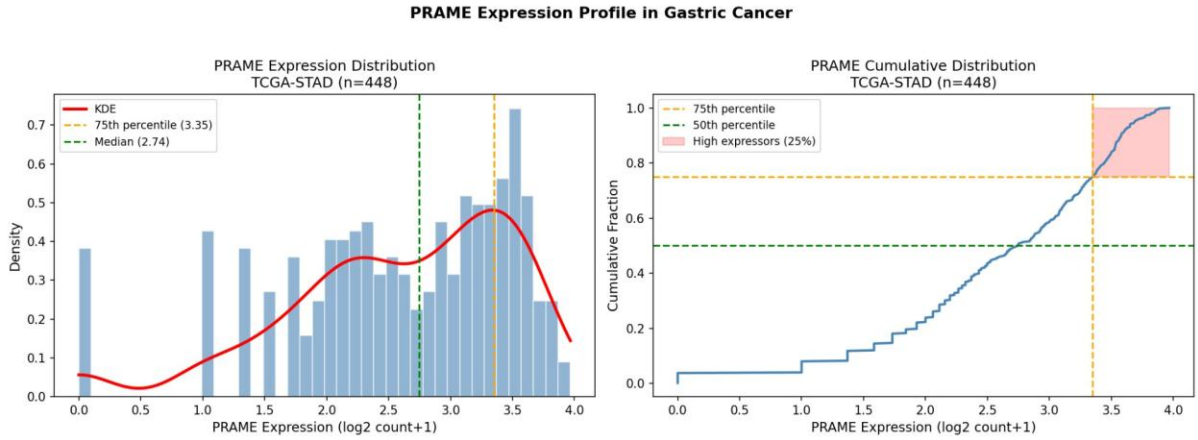


Figure 5. PRAME expression distribution in TCGA-STAD (n=448).

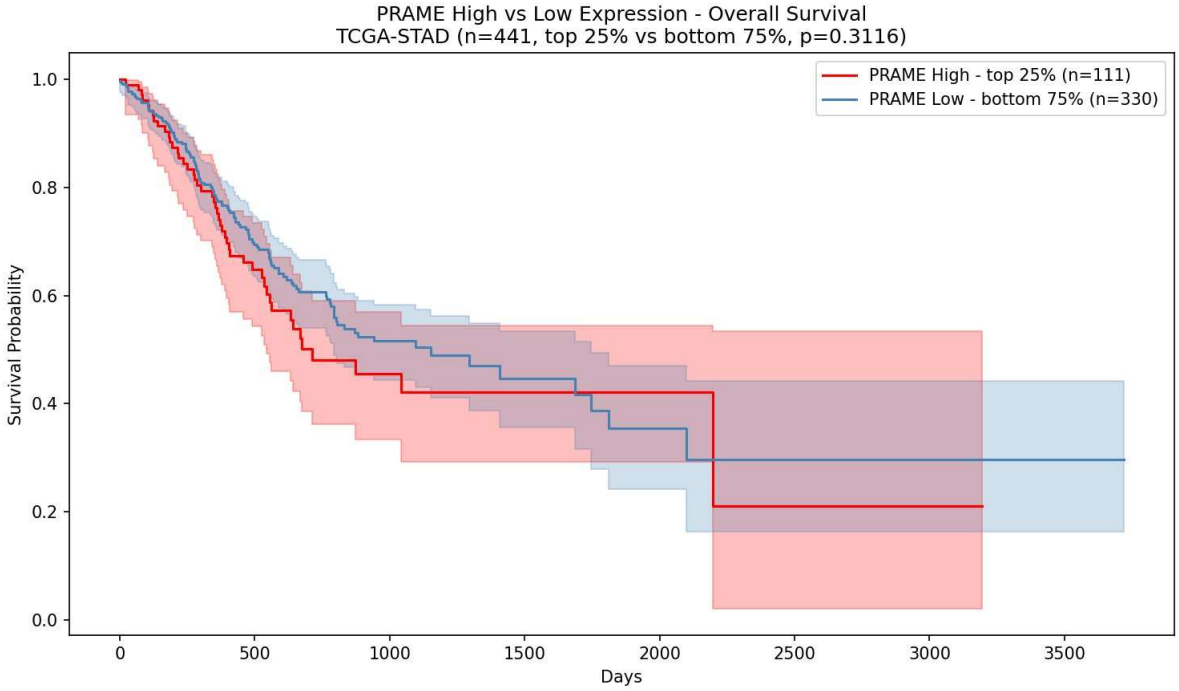


Figure 6. PRAME survival independence (log-rank p=0.31).

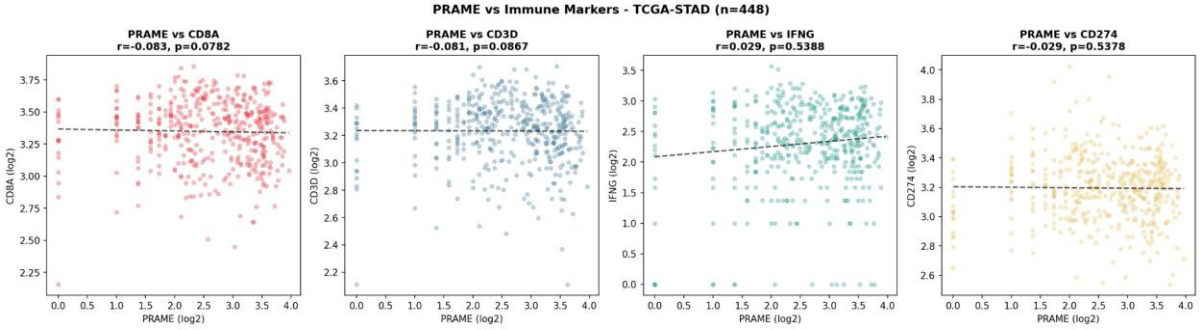


Figure 7. PRAME immune orthogonality.

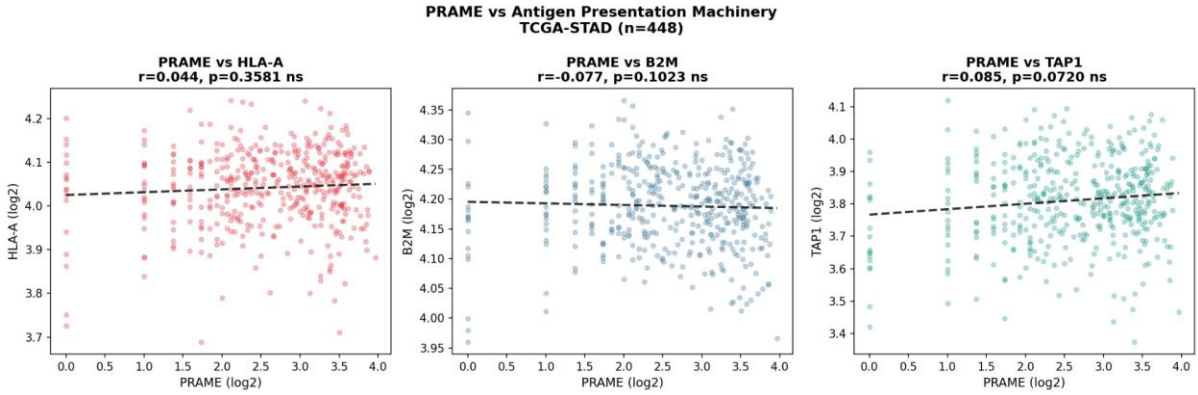


Figure 8. PRAME antigen presentation compatibility.

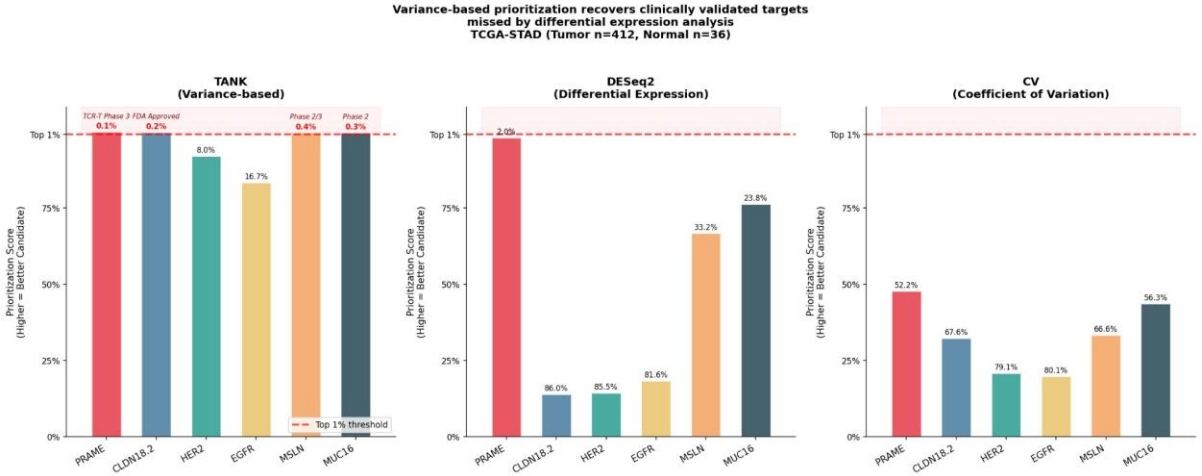


Figure 9. TANK vs DESeq2 vs CV comparison for six reference targets.

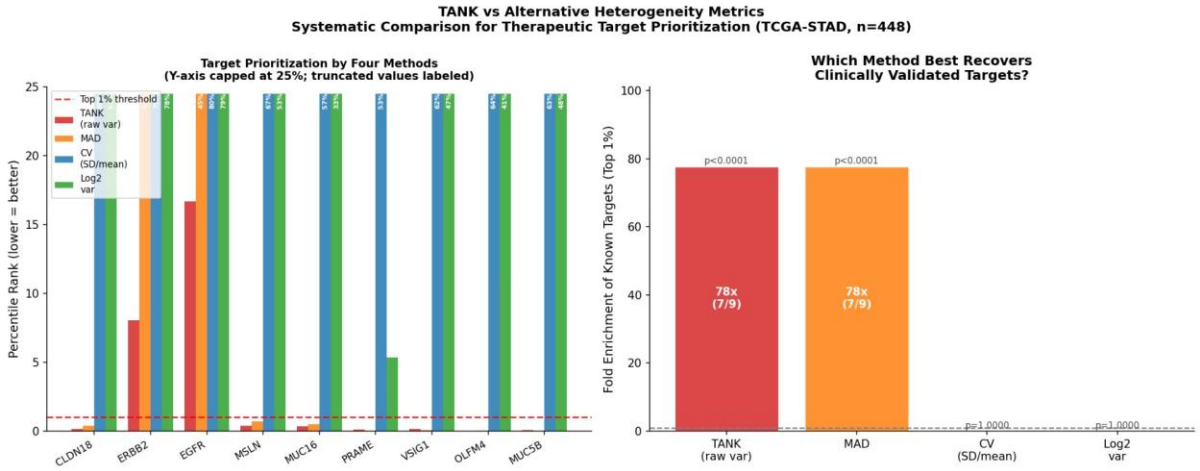


Figure 10. Four-metric comparison (TANK, MAD, CV, log2 variance). TANK and MAD equivalent (78x).

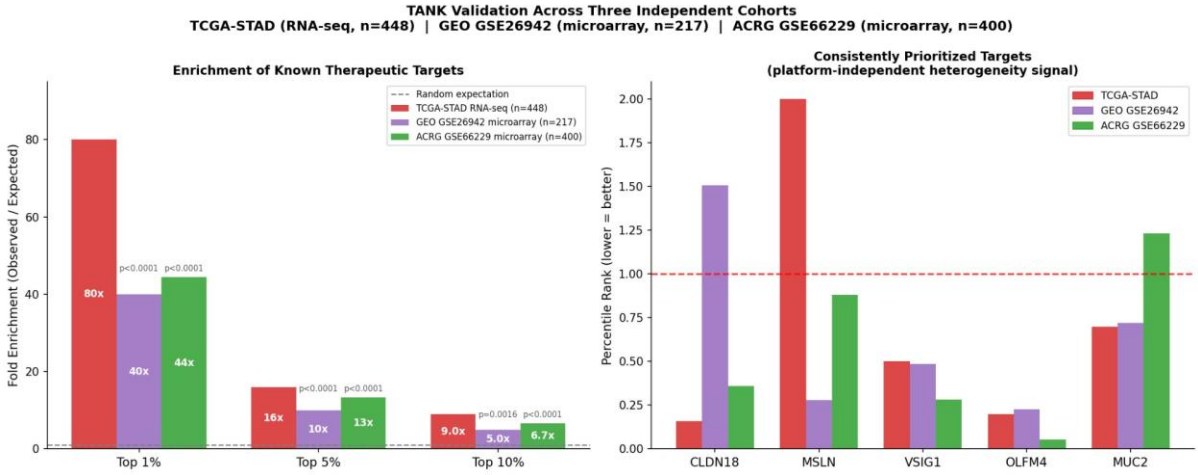


Figure 11. Three-cohort external validation (TCGA 80x, GEO 40x, ACRG 44x, all p<0.0001).

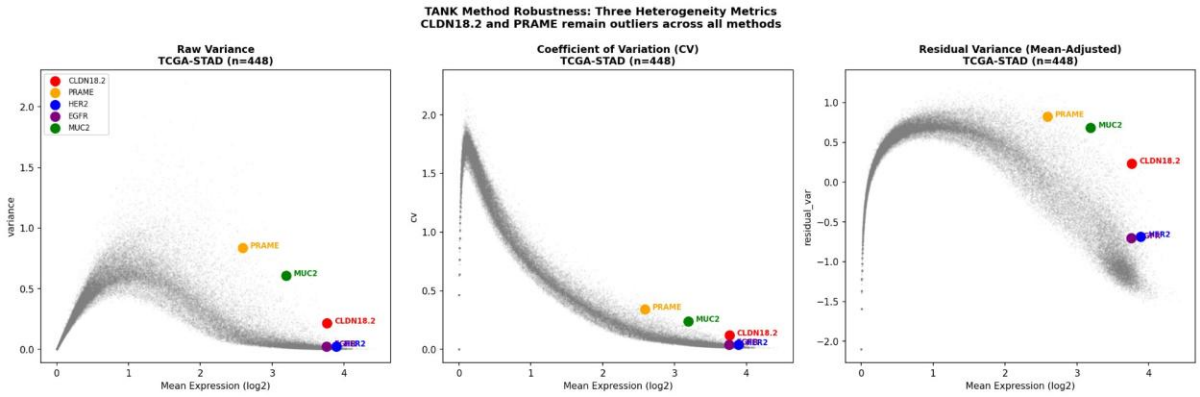


Figure 12. TANK method robustness scatter plots.

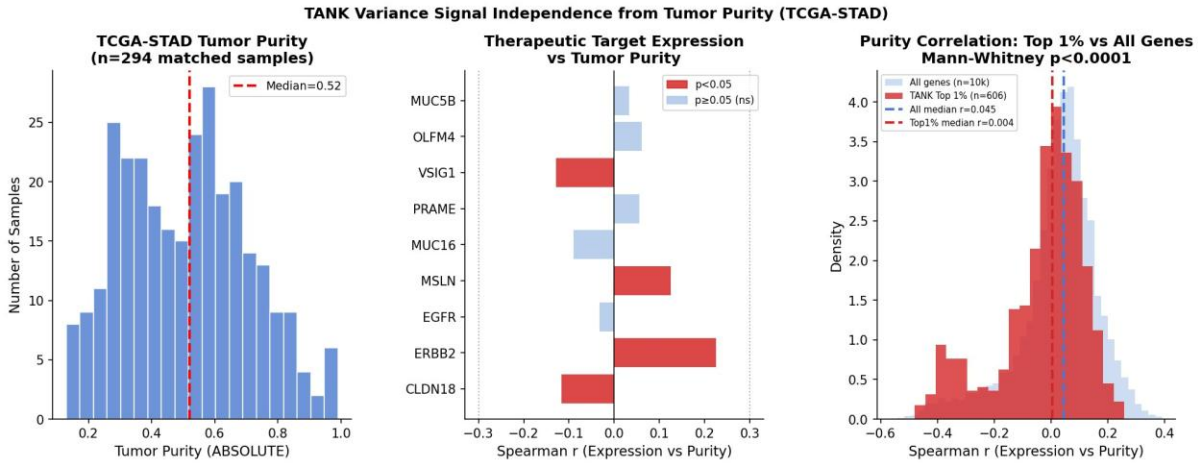


Figure 13. Tumor purity independence (TANK Top 1% r=0.004 vs background r=0.045, p<0.0001).

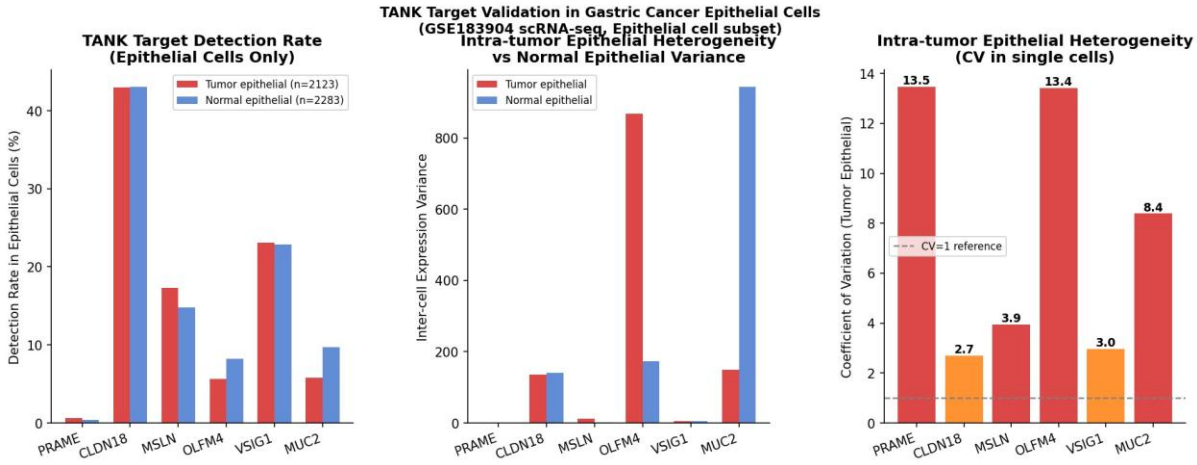


Figure 14. Multi-cancer scRNA-seq detection rates across gastric, colorectal, and lung cancer.

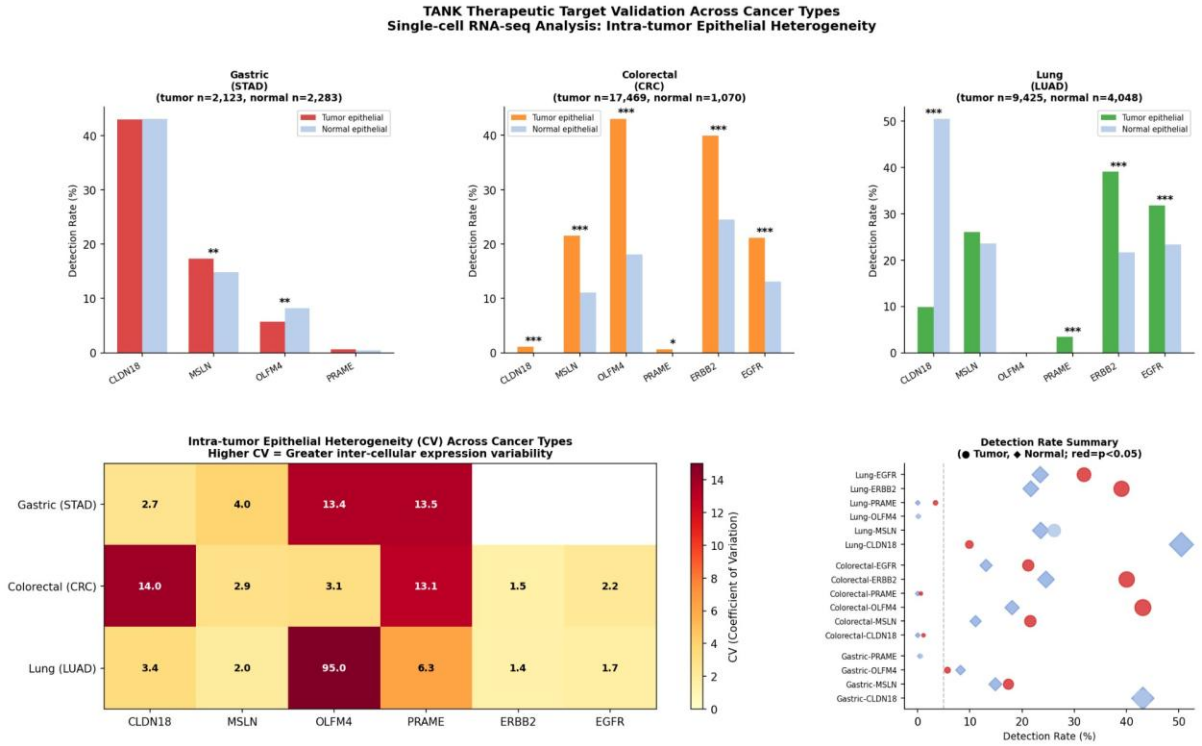


Figure 15. Multi-cancer scRNA-seq CV heatmap and detection rate summary.

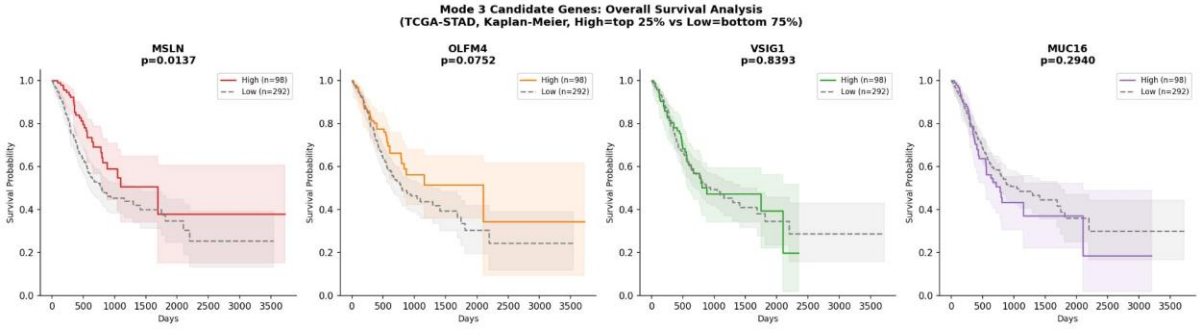


Figure 16. Mode 3 survival analysis. KM curves for MSLN (p=0.014), OLFM4 (p=0.075), VSIG1 (p=0.839), MUC16 (p=0.294), TCGA-STAD (n=390).

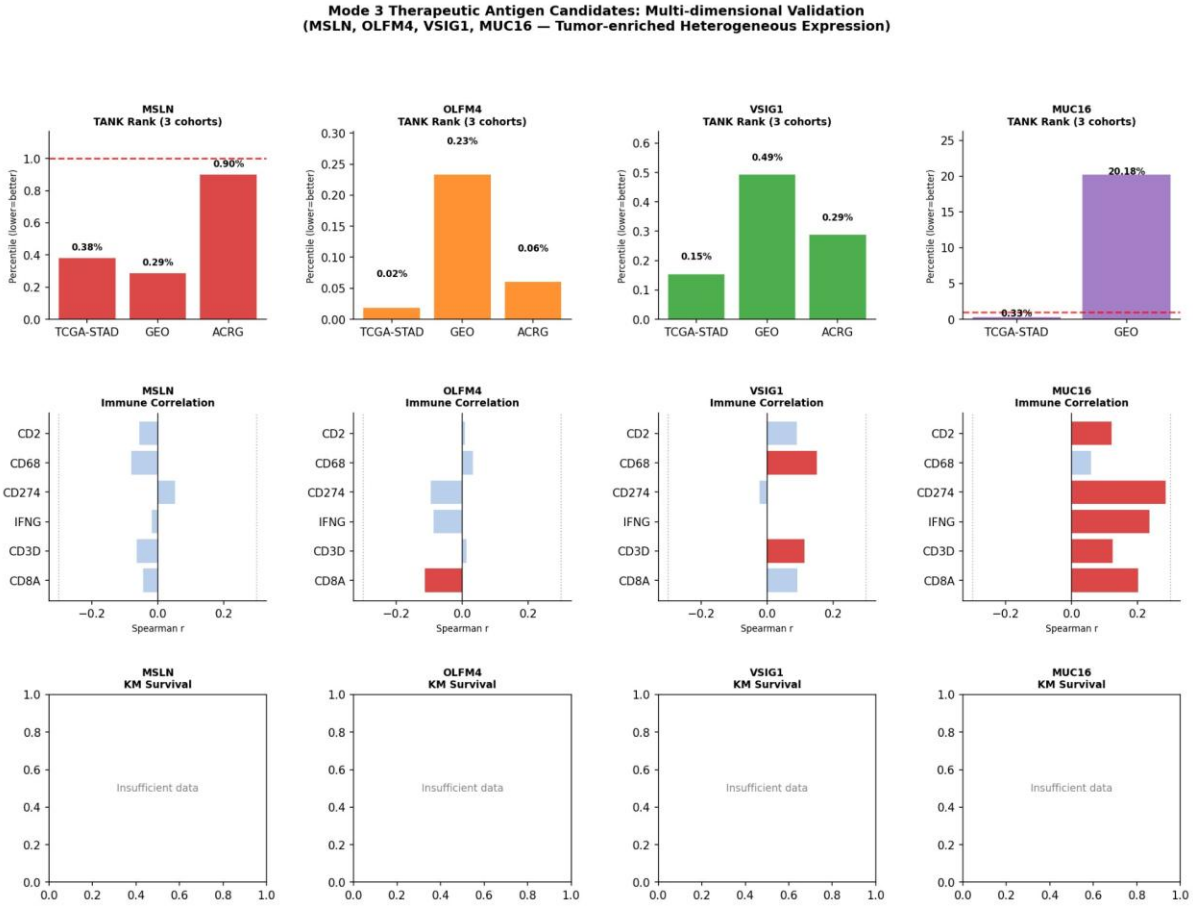


Figure 17. Mode 3 immune correlation and cross-cohort characterization.

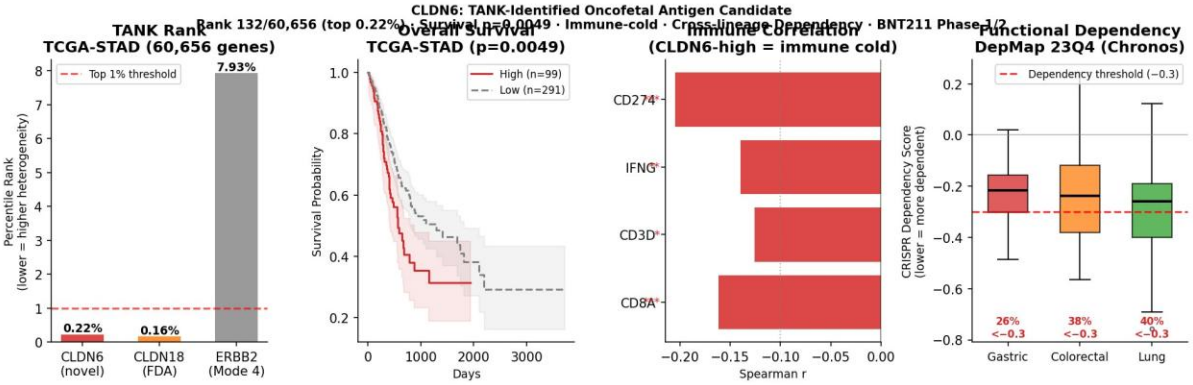
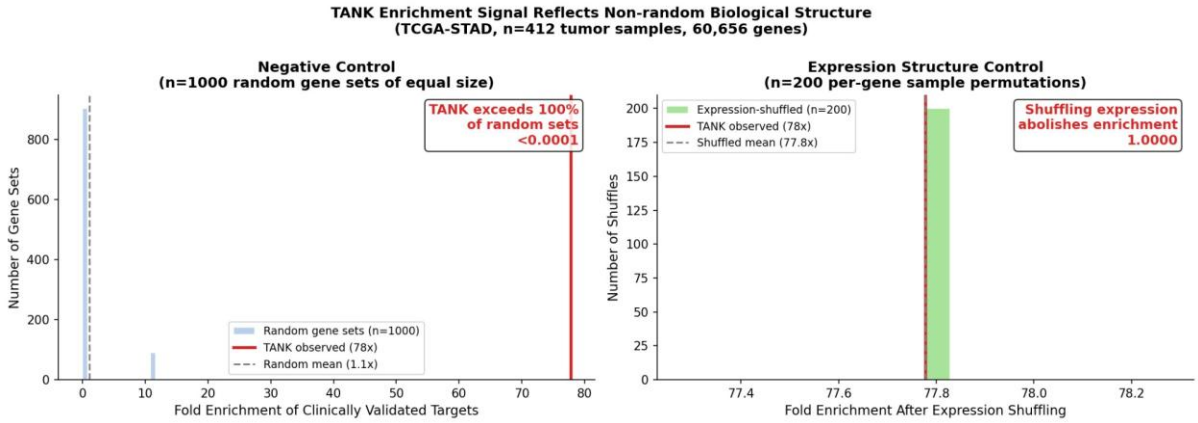
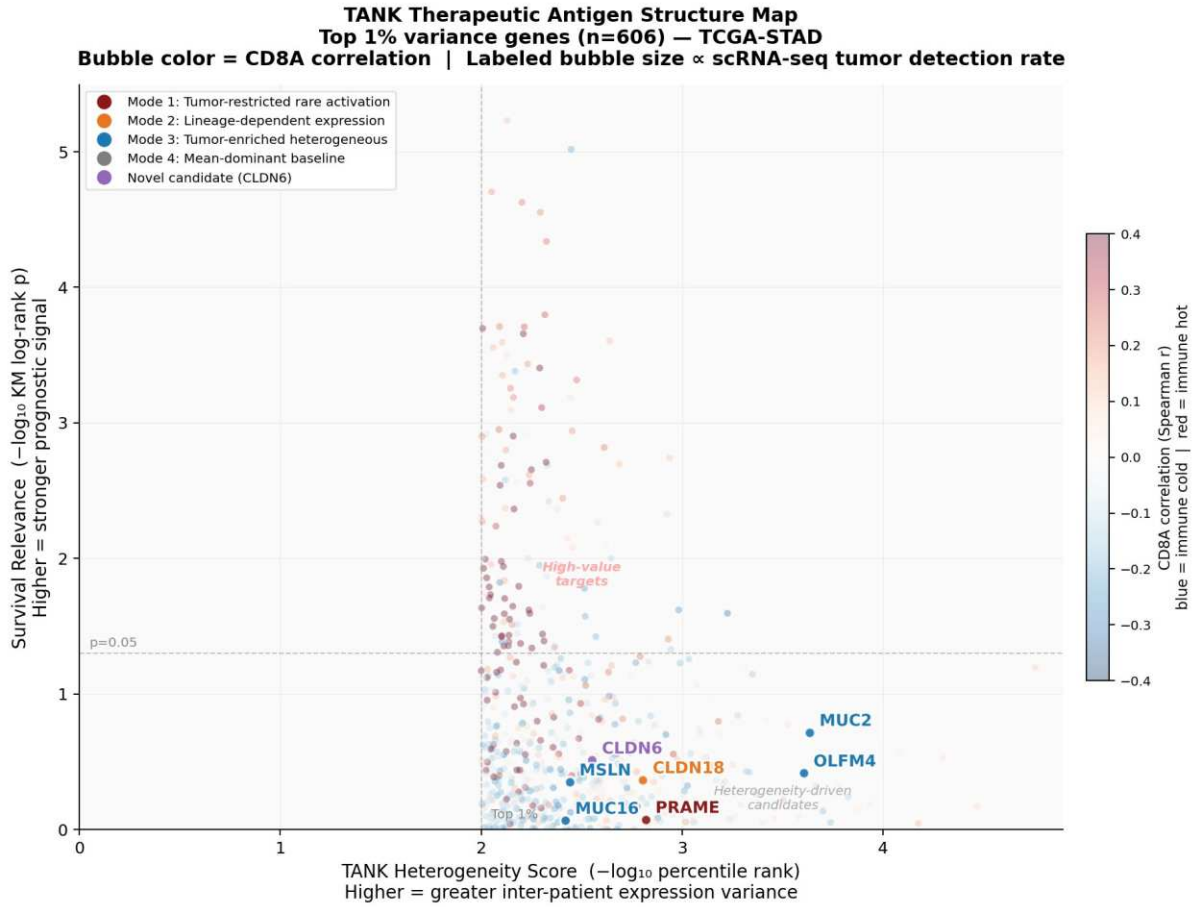


Figure 18. CLDN6 multi-dimensional validation. (A) TANK rank: top 0.22%, comparable to CLDN18 (top 0.16%). (B) Kaplan-Meier survival (p=0.0049, n=390). (C) Immune correlation: all four markers negatively significant (CD274 r=-0.204, p<0.0001), indicating immune-cold microenvironment. (D) DepMap CRISPR dependency: directional negative scores across gastric (mean=-0.227, 26% <-0.3), colorectal (mean=-0.242, 38% <-0.3), and lung (mean=-0.288, 40% <-0.3) cancer cell lines.

Supplementary Figure 1. TANK enrichment analysis at Top 1% (80x), Top 5% (16x), Top 10% (9x), all p<0.0001.



Supplementary Figure 2. Null model validation. TANK enrichment (77.8x) vs 1,000 random gene sets (mean 1.1x, $p<0.0001$).



Supplementary Figure 3. TANK Therapeutic Antigen Structure Map. Projection of top 1% variance-ranked genes (n=606, TCGA-STAD) onto a two-dimensional space defined by TANK heterogeneity score (x-axis, $-\log_{10}$ percentile rank) and survival relevance (y-axis, $-\log_{10}$ KM log-rank p-value, median split). Bubble color encodes CD8A Spearman correlation (blue = immune cold, red = immune hot); labeled bubble size is proportional to scRNA-seq tumor epithelial detection rate for known targets. The four-mode representative genes and CLDN6 are annotated. High-variance genes with strong prognostic signal (upper-right quadrant) and immune-cold high-variance genes (blue, right region) occupy biologically interpretable regions of the heterogeneity landscape, consistent with the four-mode framework.