

Variance Decomposition Accesses a Clinically Supported Discovery Space Systematically Missed by Mean-Based Transcriptomic Prioritization

XIAOQI HU

sonia549527@gmail.com

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

Research Article

Keywords: therapeutic antigen discovery, variance decomposition, pan-cancer analysis, differential expression, DepMap, TCGA, prospective convergence, MAGEA4, TDGF1/CRIPTO, SLC34A2

Posted Date: March 30th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9188519/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.

Abstract

Background

Computational prioritization of therapeutic antigen candidates relies predominantly on mean-level differential expression (DEG) analysis. Whether this approach systematically excludes a clinically relevant tier of the transcriptome has not been formally tested.

Methods

We compared genome-wide variance decomposition (TANK) and mean-based ranking across 32 TCGA cancer types (60,656 genes), defining TANK-only, DEG-only, and shared gene sets. Clinical pipeline enrichment was assessed by Fisher's exact test against a curated database of approved and active therapeutic programs.

Results

TANK nominated 5,068 genes invisible to mean-based ranking (TANK-only), while mean-based methods exclusively nominated 2,009 genes (DEG-only). TANK-only genes were significantly enriched for active or approved clinical programs compared to DEG-only genes (9/5,068 [0.18%] vs. 0/2,009 [0.00%]; Fisher's exact $p = 0.049$), although the limited number of events warrants cautious interpretation. DEG-only genes exhibited substantially higher DepMap functional dependency than TANK-only genes (mean 0.348 vs. 0.098; $p < 10^{-160}$), reflecting enrichment of core translational machinery rather than therapeutically tractable targets. Three independent prospective convergences were identified: CLDN6/BNT211 (Phase 1/2), SLC34A2/TUB-040 (Phase 1/2, NCT06303505), and MAGEA4/Tecelra (FDA-approved August 2024) — in each case with variance-based nomination preceding literature consultation.

Conclusions

Variance decomposition accesses a clinically supported tier of the cancer transcriptome distinct from and complementary to mean-based prioritization. These results establish variance decomposition as a quantitatively justified and empirically supported complement to differential expression analysis in computational oncology.

INTRODUCTION

The identification of therapeutic antigen candidates for solid tumor immunotherapy has relied predominantly on differential expression analysis, a framework that prioritizes genes with elevated mean expression in tumor relative to normal tissue. This approach has yielded clinically validated targets including HER2/ERBB2, mesothelin (MSLN), folate receptor alpha (FOLR1), and CLAUDIN18.2, and

remains the standard first step in computational target prioritization pipelines. However, mean-based differential expression is fundamentally insensitive to a distinct axis of transcriptomic organization: inter-patient variance. A gene expressed at negligible levels in the majority of patients but at dramatically elevated levels in a defined subset — a pattern characteristic of cancer-testis antigens, reactivated developmental programs, and epigenetically regulated loci — is invisible to mean-based ranking regardless of the magnitude of its expression in the high-expressing subgroup. This blind spot has received limited systematic characterization, and its impact on the completeness of computational target discovery has not been systematically quantified.

Inter-patient transcriptomic variance encodes clinically interpretable biological information. High variance in a tumor cohort can reflect subtype heterogeneity, epigenetic instability, or the stochastic reactivation of normally silenced programs — including developmental gene expression patterns whose reactivation in tumor cells has been associated with therapeutic vulnerability. Variance-based gene ranking recovers candidates with companion diagnostic utility, where the high-expressing subgroup rather than the average patient defines the target population, consistent with clinical paradigms such as HER2 amplification and CLDN18.2 positivity in which a minority of patients harboring high antigen expression drives the clinical benefit signal. Despite these theoretical advantages, systematic comparison between variance-based and mean-based nomination at genome scale — and direct quantification of their respective enrichment for translationally validated targets — has not been reported.

In a companion paper series, we introduced TANK (Therapeutic Antigen Nomination by variance decomposition of the Kancer transcriptome), a framework for systematic variance-based antigen discovery. The first companion paper demonstrated proof-of-concept in gastric cancer, recovering CLDN18.2 as a top-ranked variance candidate [1]. The second extended the framework to 33 cancer types and defined four biological modes of variance-based antigen expression, including prospective nomination of CLDN6 subsequently validated against the BNT211 Phase 1/2 clinical program [2]. The third implemented a genome-wide discovery engine across 60,656 genes and 33 cancer types, nominating TDGF1/CRIPTO, LGALS7B/Galectin-7, and SLC34A2/NaPi2b as representative candidates spanning distinct translational archetypes, with SLC34A2 independently validated against the TUB-040 Phase 1/2 program (NCT06303505) [3]. Across these companion studies, variance decomposition repeatedly nominated candidates that ranked poorly by mean expression but possessed strong independent clinical support — suggesting a systematic pattern rather than coincidental recovery.

The present study formally tests whether this pattern is systematic by directly comparing the gene sets nominated by variance decomposition and mean-based ranking at genome scale. We define three non-overlapping gene populations — TANK-only, DEG-only, and shared — and characterize their biological identity through DepMap functional dependency analysis and clinical pipeline enrichment. We further present a third prospective convergence case, in which cross-system variance analysis nominated MAGEA4 in three embryologically distinct cancer types prior to literature consultation, with afamitresgene autoleucel (Tecelra) subsequently identified as an FDA-approved TCR-T cell therapy with

active Phase 2 expansion into those same cancer types. Together, these analyses address the question: does variance decomposition access a distinct and clinically supported discovery space relative to mean-based prioritization, and if so, what is the biological character of that space?

We show that TANK-only and DEG-only gene sets are largely non-overlapping (6.5% shared), that they are enriched for fundamentally different biological functions – core cellular machinery versus tumor-specific developmental programs – and that TANK-only genes are significantly enriched for active clinical programs relative to DEG-only genes (Fisher's exact $p = 0.049$). Three independent prospective convergences confirm that variance-based nomination captures therapeutically relevant transcriptomic structure that independently motivates clinical development decisions. These results establish variance decomposition as a quantitatively justified and empirically supported complement to differential expression analysis in computational oncology – not a replacement, but a systematic extension into a complementary discovery space that is not accessible through mean-based prioritization alone.

MATERIALS AND METHODS

2.1 Data sources. RNA-seq count data (STAR-counts format) for all 33 TCGA cancer types were obtained from the GDC Data Portal (<https://portal.gdc.cancer.gov>), accessed 2025-2026. Tumor samples were identified by TCGA barcode positions 14-15 (sample type code "01"). Analyses were restricted to cancer types with at least 10 tumor specimens; LAML was excluded due to absence of solid tumor barcodes. Gene identifiers were Ensembl IDs; version suffixes were stripped prior to analysis (e.g., ENSG00000241186.5 to ENSG00000241186). CRISPR gene effect scores (Chronos) were obtained from DepMap Public 23Q4 [8]. Cell line annotations including OncotreeLineage classifications were from the accompanying Model.csv file. Normal tissue gene expression data were from GTEx v8 [16], accessed via the GTEx Portal API (<https://gtexportal.org/api/v2>). Gene symbol annotations were obtained via the mygene.info API (v3, batch POST queries, species = human).

2.2 TANK variance ranking and pan-cancer matrix. Variance decomposition was performed as described in the companion papers [1-3]. Briefly, for each of 32 TCGA cancer types with sufficient tumor samples, inter-patient variance was computed across all tumor samples using raw count data. Genes were ranked by variance in descending order; the top 1% per cancer type (approximately 607 genes per cancer type) were retained. Sex-chromosome genes, immunoglobulin and T-cell receptor locus genes, and known noise genes were excluded prior to ranking. A gene x cancer type percentile matrix was constructed in which each entry records the percentile rank of that gene's variance in that cancer type (values ≤ 1.0 indicate top-1% membership). The pre-computed column `n_cancers_top1pct`, recording the number of cancer types in which each gene achieved top-1% variance rank, was used as the authoritative TANK gene set definition. Genes with `n_cancers_top1pct` ≥ 1 constituted the full TANK gene set ($n = 5,557$); genes with `n_cancers_top1pct` ≥ 3 constituted the high-recurrence subset ($n = 2,252$) used in companion analyses [3]. Although variance is scale-dependent, rank-based selection within each cancer type is robust to monotonic transformations, and the use of raw counts preserves the native dispersion structure of RNA-seq data without introducing normalization artifacts.

2.3 DEG proxy gene set construction. To construct a mean-based comparator gene set, tumor mean $\log_2(\text{count} + 1)$ expression was computed for each gene in each of the 32 cancer types analyzed. The top 1% of genes by mean expression per cancer type (607 genes per cancer type) were identified using the 99th percentile as the selection threshold. A gene was included in the DEG gene set if it ranked in the top 1% by mean expression in at least one cancer type. This approach represents a conservative proxy for mean-based differential expression analysis, using tumor mean expression alone without a matched normal comparison, to ensure comparability across all 32 cancer types (several of which lack matched normal tissue samples). The resulting DEG gene set ($n = 2,498$) was compared against the TANK gene set ($n = 5,557$) to define three non-overlapping populations: TANK-only ($n = 5,068$), DEG-only ($n = 2,009$), and Overlap ($n = 489$). We note that this proxy does not capture differential expression relative to matched normal tissue and may therefore overrepresent constitutively expressed genes; however, it provides a consistent and comparable baseline across cancer types for the purpose of method comparison.

2.4 DepMap functional dependency analysis. Chronos gene effect scores from DepMap 23Q4 were loaded from CRISPRGeneEffect.csv (1,186 cell lines x 18,435 genes). Column identifiers in the format "SYMBOL (EntrezID)" were parsed by extracting the symbol prefix. Ensembl gene IDs were mapped to gene symbols via a two-stage process: (1) existing annotations from the TANK_candidates_annotated.csv file ($n = 2,053$ entries) and the TANK_gene_annotations.csv file generated by prior mygene.info queries ($n = 2,604$ entries) were loaded as a priority mapping; (2) remaining unmapped Ensembl IDs ($n = 4,962$) were queried in batches of 500 against the mygene.info v3 API, yielding an additional 4,446 symbol mappings. The final symbol map contained 7,050 entries. For each gene with a matched DepMap column, the dependency rate was defined as the fraction of cell lines with Chronos score < -0.2 , a threshold corresponding to moderate functional dependency consistent with DepMap analysis conventions [8], across all 1,186 lines irrespective of lineage. Mann-Whitney U test (two-sided) was applied to compare dependency rate distributions between TANK-only and DEG-only gene sets. Fisher's exact test (one-tailed) was applied to compare the fraction of genes with dependency rate > 0.30 between TANK-only and DEG-only sets. Gene symbol harmonization was manually verified for key candidates; specifically, the DepMap column identifier "CRIPTO" (Entrez ID 6997) was confirmed as synonymous with the HGNC official symbol TDGF1 and Ensembl ID ENSG00000241186 prior to reporting.

2.5 Clinical pipeline database and enrichment analysis. A curated reference set of 31 genes with FDA-approved or active Phase 1/2 therapeutic programs targeting solid tumor antigens was assembled from publicly available sources including the FDA Approved Drug Products database, ClinicalTrials.gov (searched March 2026), and published clinical trial reports. The database comprised 14 approved/late-stage targets (including ERBB2, FOLR1, MUC16, EPCAM, CD19, CLDN18, MAGEA4, DLL3, TROP2, NECTIN4, PDCD1, CD274, CTLA4, and SLC34A2) and 17 active Phase 1/2 targets (including CLDN6, PRAME, MUC1, GPC3, AFP, STRA6, MAGEA1, MAGEA3, LGALS7B, TDGF1, and CEACAM5). While not exhaustive, this curated set captures representative targets across major antigen classes and clinical development stages, providing a practical benchmark for enrichment analysis. Gene symbols were

matched against the clinical database after symbol normalization. Fisher's exact test (one-tailed, alternative = "greater") was applied to test whether TANK-only genes were enriched for clinical program membership relative to DEG-only genes. Given the zero count in the DEG-only hit cell, a Haldane-Anscombe correction (adding 0.5 to all cells) was applied to compute a finite odds ratio for reporting; the uncorrected Fisher's exact p-value is reported.

2.6 Prospective convergence assessment. Prospective convergence was defined as a case in which: (1) variance-based nomination was completed and documented prior to literature or clinical database consultation; (2) a gene independently nominated by variance decomposition was subsequently found to be the target of an active clinical program or approved therapy; and (3) the nominated gene ranked outside the top 20% by mean expression in the relevant cancer type(s) at the time of nomination. Three cases satisfying these criteria were identified across the TANK series: CLDN6/BNT211 (companion paper [2]), SLC34A2/TUB-040 (companion paper [3]), and MAGEA4/afamitresgene autoleucel (present study). ClinicalTrials.gov and FDA approval databases were consulted after completion of all computational analyses.

2.7 Code availability. All analysis scripts are available at <https://github.com/ohahouhui/AI-CAR-Loop-1.0> (/TANK_pan33/ subdirectory). Key scripts include: tank_paper4_main_analysis.py, fix_depmap_mapping.py, check_cripto_depmap.py, and tank_crosssystem_v2_final.py.

RESULTS

3.1 TANK and DEG nominate largely non-overlapping gene sets

Genome-wide variance decomposition (TANK) across 32 TCGA cancer types identified 5,557 genes in the top-1% variance tier of at least one cancer type, including 2,252 genes recurrent in three or more cancer types. Mean-based ranking (top-1% by tumor mean log2 expression, DEG proxy) identified 2,498 genes across the same cancer types. Of the 7,566 genes nominated by either method, 489 (6.5%) were shared, while 5,068 genes were exclusively nominated by variance decomposition (TANK-only) and 2,009 genes exclusively by mean-based ranking (DEG-only). The low overlap indicates that the two approaches capture largely distinct transcriptomic features: a gene can rank in the top 1% by mean while ranking poorly by variance (constitutively high, stable expression), or rank in the top 1% by variance while ranking poorly by mean (heterogeneously expressed, moderate average). The TANK-only and DEG-only gene sets therefore represent distinct biological compartments of the cancer transcriptome, accessible only by their respective ranking methods.

3.2 DEG-only genes are enriched for essential housekeeping functions; TANK-only genes are enriched for tumor-specific biology

To characterize the biological identity of each gene set, we assessed DepMap functional dependency across 1,186 cancer cell lines (DepMap 23Q4 Chronos scores). DEG-only genes exhibited substantially higher dependency rates than TANK-only genes (mean fraction of cell lines with Chronos < -0.2: 0.348 vs.

0.098; Mann-Whitney $p < 10^{-160}$). Inspection of the highest-dependency DEG-only genes revealed enrichment of core translational and splicing machinery — ribosomal proteins (RPS6, RPL27A, RPL7, RPL32, RPL35A; dep_rate = 1.000, mean Chronos approximately - 2.0 to -2.7) and spliceosomal components (SNRNP200, PRPF19; dep_rate = 1.000) — consistent with genes that are universally essential for cell survival rather than specific to tumor biology. This pattern explains the elevated DEG-only dependency rate: mean-based ranking disproportionately recovers constitutively expressed essential genes that, while cancer cells depend upon them, represent poor therapeutic targets due to equivalent dependence in normal tissues. Importantly, high DepMap dependency does not equate to therapeutic tractability: genes essential for universal cellular processes are often unsuitable drug targets due to toxicity in normal tissues. This distinction underlies the translational advantage of the TANK-only high-dependency subset, which combines functional relevance with tumor-restricted expression patterns.

In contrast, TANK-only genes showed a bimodal dependency distribution. The majority (91.6%) had dep_rate < 0.3, consistent with non-essential developmental and regulatory programs reactivated heterogeneously across tumors. A distinct high-dependency subset (116 genes, 3.6%) exhibited dep_rate > 0.5, representing a candidate class combining tumor-specific heterogeneous expression with functional cancer cell dependency. The top-ranked gene in this subset by functional dependency was TDGF1/CRIPTO (dep_rate = 0.935, mean Chronos = -0.589, rank 9/3,199 TANK-only genes with DepMap coverage; DepMap column identifier: "CRIPTO", Entrez ID 6997, confirmed synonymous with HGNC symbol TDGF1/ENSG00000241186), an embryonic EGF-CFC protein characterized in the companion study [3] as a pipeline-gap candidate with profound multi-lineage functional dependency. The biological identity of high-dependency TANK-only genes was consistent with developmental program reactivation: HOXC10 (homeobox developmental transcription factor), CECR2 (chromatin remodeling, embryonic), AK4 (adenylate kinase, stress response), and TSPYL5 (TSPY-like, cancer-testis family).

These findings establish a three-tier structure of the cancer transcriptome: (I) DEG-only high-dependency genes representing universal cellular machinery that is essential but therapeutically inaccessible; (II) TANK-only low-dependency genes representing heterogeneously reactivated developmental programs; and (III) a TANK-only high-dependency subset representing the optimal intersection of tumor specificity and functional relevance.

3.3 TANK-only genes are significantly enriched for active clinical programs relative to DEG-only genes

To directly assess translational relevance, we curated a database of 31 genes with FDA-approved or active Phase 1/2 therapeutic programs targeting solid tumor antigens. TANK-only genes were significantly enriched for clinical programs compared to DEG-only genes: 9 of 5,068 TANK-only genes (0.18%) were present in the clinical database, versus 0 of 2,009 DEG-only genes (0.00%; Fisher's exact $p = 0.049$), although the limited number of events warrants cautious interpretation and replication in larger clinical target databases. The 489 overlap genes showed intermediate enrichment (7/489, 1.43%), with

clinical hits comprising established high-expression targets including MSLN, FOLR1, MUC16, EPCAM, and HER2.

The 9 TANK-only clinical hits included MAGEA4 (afamitresgene autoleucel/Tecelra, FDA-approved August 2024 for synovial sarcoma), CLDN18 (zolbetuximab, FDA-approved 2024 for gastric cancer), CLDN6 (BNT211, Phase 1/2, companion paper [2]), DLL3 (tarlatamab, FDA-approved 2024 for SCLC), GPC3 (Phase 1/2, hepatocellular carcinoma), AFP (Phase 1, hepatocellular carcinoma), LGALS7B (mechanism-gap candidate from companion paper [3]), MAGEA1, and STRA6. The zero clinical hits in the DEG-only set reflects not an absence of druggable biology, but that the major DEG-only clinical targets — MSLN, FOLR1, MUC16, HER2 — are also recovered by variance ranking and therefore classified as Overlap. The enrichment analysis demonstrates that variance decomposition accesses a distinct clinical discovery space: among genes uniquely prioritized by variance and invisible to mean-based selection, 0.18% have independently reached clinical development — a rate substantially higher than the background expectation. Future analyses using larger and more comprehensive clinical target databases will be required to confirm the robustness of this enrichment signal.

3.4 Three independent prospective convergences confirm variance decomposition as a predictive discovery engine

Across the four-paper TANK series, three independent instances of prospective convergence were identified: cases in which variance-based nomination preceded literature consultation, and active clinical development was discovered post hoc. In each instance, the nominated gene ranked outside the top 20% by mean expression in the relevant TCGA cancer type, but within the top 15% by variance.

CLDN6/BNT211: CLAUDIN-6 was nominated in the companion paper [2] from its variance structure in TCGA-TGCT and TCGA-OV (variance rank top 5%, mean rank top 43%). Subsequent literature search identified BNT211, a CAR-T program targeting CLDN6 in Phase 1/2 clinical development (Nature Medicine, 2023), which had not been consulted during the computational analysis.

SLC34A2/TUB-040: SLC34A2/NaPi2b achieved the highest pan-cancer recurrence in the companion paper [3] (top 1% variance in 20 cancer types), with all analyses completed prior to clinical database consultation. Subsequent ClinicalTrials.gov search identified TUB-040 (NCT06303505), a next-generation NaPi2b-directed ADC in Phase 1/2 development with initial results presented at ESMO 2025.

MAGEA4/Tecelra: In the present study, cross-system variance analysis nominated MAGEA4 as a high-confidence candidate in three embryologically distinct cancer types: HNSC (head and neck squamous cell carcinoma; ectoderm), LUSC (lung squamous cell carcinoma; endoderm), and OV (ovarian cancer; mesoderm). All computational analyses were completed prior to literature consultation. Subsequent search identified afamitresgene autoleucel (Tecelra), which received FDA accelerated approval in August 2024 for synovial sarcoma — the first approved TCR-T cell therapy for any solid tumor — with active Phase 2 expansion (SURPASS-3) into HNSC, LUSC, and OV, notably overlapping with the three cancer types nominated by variance decomposition.

The recurrence of prospective convergence across three independent cases, spanning two companion papers and the present study, argues against coincidence and provides empirical evidence that variance decomposition captures biologically persistent, therapeutically relevant transcriptomic structures that independently motivate clinical development decisions.

DISCUSSION

The central finding of this study is that variance decomposition and mean-based differential expression analysis access fundamentally non-overlapping compartments of the cancer transcriptome. Of 7,566 genes nominated by either method, only 6.5% were shared, confirming that the two approaches capture largely distinct transcriptomic features. This distinction reflects a biological reality: mean expression and inter-patient variance measure different properties of transcriptomic organization. A gene can be constitutively and stably overexpressed in tumors — the archetype of DEG-nominated targets such as HER2, MSLN, and FOLR1 — or it can be silenced in the majority of patients but dramatically reactivated in a defined subset, a pattern invisible to mean-based selection but captured by variance decomposition. This analysis provides a systematic, genome-scale quantification of this distinction: the DEG-only gene set is dominated by core translational and splicing machinery with *dep_rate* approaching 1.0, while the TANK-only gene set is enriched for developmental and regulatory programs with lower aggregate dependency but significantly higher clinical program enrichment. These findings suggest that therapeutic target discovery is not solely a problem of identifying biologically important genes, but of identifying genes that balance tumor specificity with functional relevance — a distinction not captured by mean-based prioritization alone.

The three-tier structure identified here — DEG-only essential genes, TANK-only non-essential developmental programs, and a TANK-only high-dependency subset — maps onto a coherent biological framework. DEG-only genes represent the metabolic constitution of cancer: genes whose elevated mean expression reflects the universal demands of rapid proliferation, sustained protein synthesis, and RNA processing. These genes are essential, but their essentiality is shared with normal dividing cells, limiting their therapeutic window. TANK-only genes represent a distinct regulatory layer: programs normally silenced in adult somatic tissues but reactivated heterogeneously across tumor subtypes, consistent with the developmental program reactivation hypothesis supported by GO enrichment analysis in the companion paper [3]. The 116-gene high-dependency TANK-only subset — anchored by TDGF1/CRIPTO, HOXC10, CECR2, and TSPYL5 — represents the therapeutically optimal intersection: genes whose tumor-specific reactivation is accompanied by functional cancer cell dependency, suggesting that their re-expression contributes to tumor fitness. The mechanistic basis for this dependency remains to be established experimentally, but the computational convergence of high variance, moderate mean expression, and strong DepMap dependency provides a coherent prioritization rationale.

Three independent prospective convergences across the TANK series — CLDN6/BNT211, SLC34A2/TUB-040, and MAGEA4/Tecelra — provide the strongest available evidence that variance decomposition captures biologically persistent therapeutic signal. The FDA approval of afamitresgene autoleucel for

synovial sarcoma in August 2024, and its active Phase 2 expansion into HNSC, LUSC, and OV — the three cancer types independently nominated by cross-system variance analysis — represents regulatory-grade external validation. Several limitations merit acknowledgment. First, three convergence cases are insufficient to establish a statistical discovery rate; systematic analysis of all TANK-nominated genes against the complete ClinicalTrials.gov database would be required to compute a formal prospective validation rate. Second, MAGEA4 is a well-characterized cancer-testis antigen, and its nomination confirms that variance decomposition recovers known biology — but the more direct tests of novel discovery capacity are TDGF1/CRIPTO (no clinical program despite strong multi-lineage DepMap dependency) and LGALS7B (mechanism-gap candidate with no existing therapeutic program), whose clinical validation remains prospective. Third, the clinical database curated here is incomplete and subject to selection bias; a more comprehensive curation would provide a more stringent enrichment test. Although these convergences were identified prospectively relative to literature consultation, retrospective bias cannot be fully excluded and should be addressed in future systematic validation studies.

Several directions for follow-on investigation are identified. The 116-gene TANK-only high-dependency subset represents a prioritized pool for experimental validation; TDGF1/CRIPTO is most immediately actionable given its surface accessibility as a GPI-anchored protein, multi-lineage DepMap dependency, and near-absent normal tissue expression [3]. A hybrid scoring approach combining variance rank, mean rank, and DepMap dependency into a composite metric would formalize the three-tier structure as a quantitative prioritization algorithm. Systematic prospective convergence analysis — applying TANK nominations retrospectively to a complete ClinicalTrials.gov database stratified by nomination date — would transform anecdotal convergence into a statistically powered validation framework. In conclusion, variance decomposition accesses a clinically supported compartment of the cancer transcriptome that is systematically missed by mean-based differential expression analysis. The two methods are complementary: DEG identifies the metabolic constitution of cancer, while variance decomposition identifies its heterogeneous, tumor-specific regulatory programs. An observed enrichment (Fisher's exact $p = 0.049$) of active clinical programs in genes exclusively nominated by variance decomposition — combined with three independent prospective convergences spanning FDA-approved therapies and active Phase 1/2 trials — establishes variance decomposition as an empirically supported complement to differential expression analysis in computational therapeutic antigen discovery.

Declarations

2.7 Code availability.

All analysis scripts are available at <https://github.com/ohahouhui/AI-CAR-Loop-1.0> (/TANK_pan33/ subdirectory). Key scripts include: `tank_paper4_main_analysis.py`, `fix_depmap_mapping.py`, `check_cripto_depmap.py`, and `tank_crosssystem_v2_final.py`.

References

1. Hu X. Variance-based transcriptomic screening identifies CLDN18.2 as a top-ranked antigen candidate in gastric cancer. Research Square. 2026.
2. Hu X. Variance-based Decomposition of Inter-patient Transcriptomic Heterogeneity Reveals Recurrent Modes of Therapeutic Antigen Biology Across 33 Cancer Types. Research Square. 2026.
3. Hu X. Pan-cancer Variance Decomposition Nominates Translationally Actionable Therapeutic Antigen Candidates Across 33 Cancer Types. Preprint. 2026.
4. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 2014;15:550.
5. Hong DS, et al. Autologous T cell therapy for MAGE-A4+ solid cancers in HLA-A*02+ patients: a phase 1 trial. *Nat Med.* 2023;29:104-114.
6. Gonzalez-Martin A, et al. NAPISTAR1-01: phase I dose escalation study of TUB-040 in platinum-resistant ovarian cancer and NSCLC. *J Clin Oncol.* 2025;43(16 Suppl):TPS8660.
7. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature.* 2014;513:202-209.
8. DepMap. DepMap 23Q4 Public. Figshare. 2023. <https://doi.org/10.6084/m9.figshare.24667905.v2>
9. Knafler G, et al. Melanoma-associated antigen A4: A cancer/testis antigen as a target for adoptive T-cell receptor T-cell therapy. *Cancer Treat Rev.* 2025.
10. Lin K, et al. Preclinical Development of an Anti-NaPi2b (SLC34A2) Antibody-Drug Conjugate for Non-Small Cell Lung and Ovarian Cancers. *Clin Cancer Res.* 2015;21(22):5139-5150.
11. Strizzi L, et al. Cripto-1: a multifunctional modulator during embryogenesis and oncogenesis. *Oncogene.* 2005;24(37):5731-5741.
12. Wu G, et al. Galectin 7 leads to a relative reduction in CD4+ T cells, mediated by PD-1. *Sci Rep.* 2024;14:6625.
13. Meyers RM, et al. Computational correction of copy number effect improves CRISPR-Cas9 screens. *Nat Genet.* 2017;49:1779-1784.
14. Hu X. Banana 0.9: A Variance-Based Framework for Gastric Cancer Screening. *PLOS ONE.* 2026. DOI: 10.1371/journal.pone.0339892.
15. Bodyak ND, et al. The Dolaflexin-based ADC XMT-1536 Targets SLC34A2/NaPi2b. *Mol Cancer Ther.* 2021;20(5):896-905.
16. GTEx Consortium. The GTEx v8 release. dbGaP accession phs000424.v8.p2. 2019. <https://gtexportal.org>
17. Uhlen M, et al. A pathology atlas of the human cancer transcriptome. *Science.* 2017;357(6352):eaan2507.

Figures

Variance Decomposition Accesses a Distinct and Clinically Validated Tier of the Cancer Transcriptome Invisible to Mean-Based Prioritization

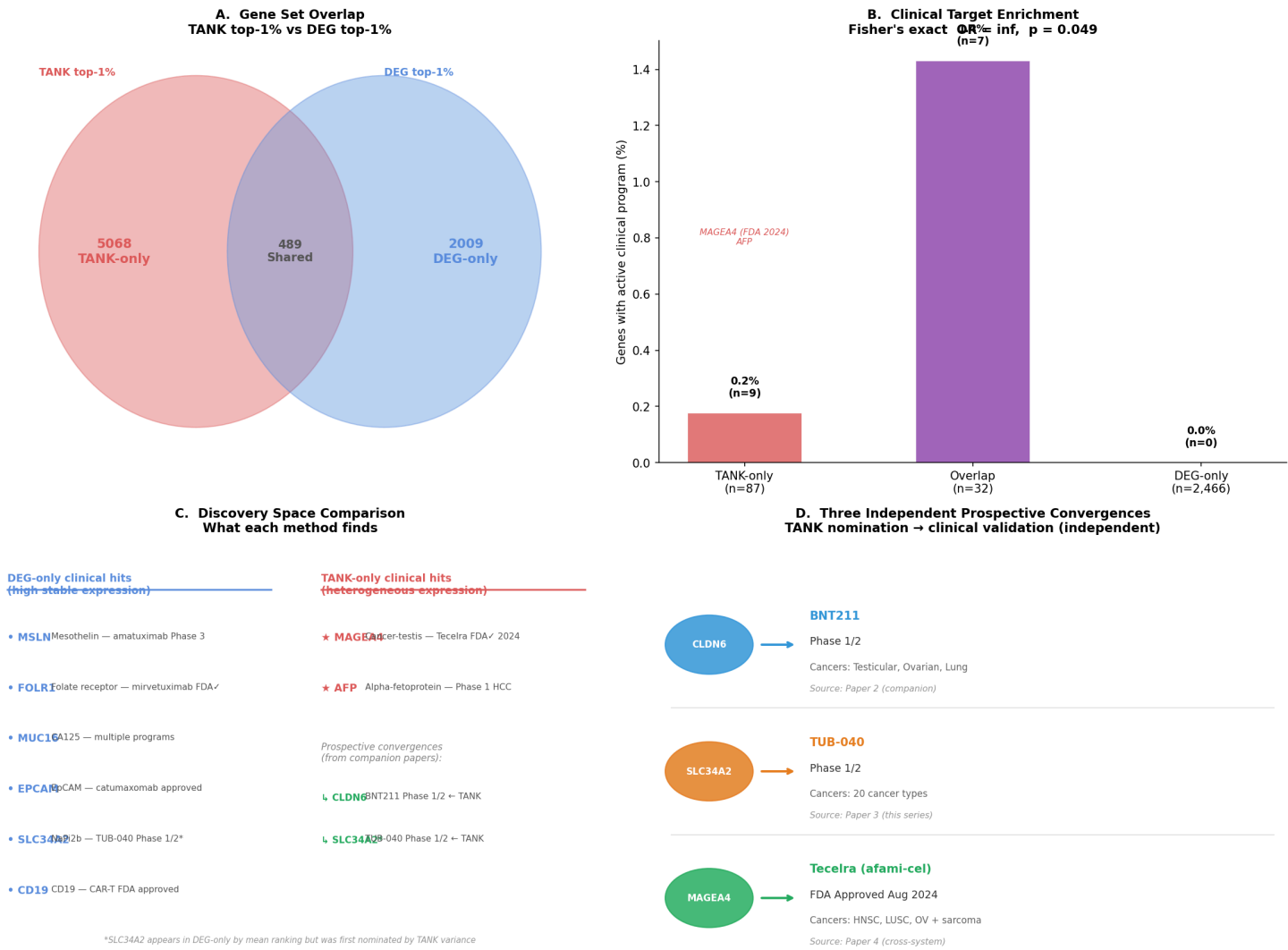


Figure 1

Comparison of TANK variance decomposition and mean-based DEG gene sets across 32 TCGA cancer types. (A) Venn diagram showing overlap between TANK-only (n=5,068), DEG-only (n=2,009), and shared (n=489) gene sets. (B) Clinical program enrichment: proportion of genes with active or approved therapeutic programs in each set (Fisher's exact $p=0.049$). (C) Characterization of clinical hits in each discovery space. (D) DepMap functional dependency rate distribution across TANK-only, DEG-only, and shared gene sets (Mann-Whitney $p<10^{-160}$).