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.

What is shared across 33 cancer types is the tissue, not the tumour: a non-malignant axis of the microenvironment

Shuaidong Gao

gaoshuaidong@stu.cqifs.edu.cn

Chongqing Institute of Foreign Studies

Research Article

Keywords: Pan-cancer analysis, Causal discovery, Tumour composition, CXCR3 chemokine axis, Tumour microenvironment, NOTEARS

Posted Date: November 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11281904/v1

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

Additional Declarations: No competing interests reported.

What is shared across 33 cancer types is the tissue, not the tumour: a non-malignant axis of the microenvironment

Shuaidong Gao[1*]

[1]Chongqing Institute of Foreign Studies, Qijiang Campus, Chongqing 401420, China.

E-mail: gaoshuaidong@stu.cqifs.edu.cn

ORCID: 0009-0004-5641-3581.

Abstract

What a pan-cancer causal network reports is the composition of the samples, not the regulation of the tumour. Comparing 33 cancer types from The Cancer Genome Atlas (TCGA) (10,459 samples), each fitted by NOTEARS on a common 100-gene panel, the reproducible structure sits on non-malignant axes: a co-induced, interferon-$\boldsymbol{\gamma}$-inducible CXCR3-ligand triplet—CXCL9, CXCL10 and CXCL11—the only three-node module whose edges keep one orientation, together with neutrophil/CXCR2 chemokines, blood content, plasma-cell/immunoglobulin, stroma/complement and myeloid/cytokine signalling. The dominant axis is non-malignant content (first principal component, PC1; $\boldsymbol{\lambda_1 = 16.04}$, against the Marchenko–Pastur edge; $\boldsymbol{r = 0.80}$ to markers outside the panel, 33/33 cohorts). Canonical drivers are absent structurally, not by chance: their dispersion lies at the 50.8th percentile whereas the panel selects the 0.487th, so drivers enter only at $\boldsymbol{k \gtrsim 1000}$. These findings are bounded. Only 14 of 9,900 directed pairs recur in $\geq \mathbf{10}$ cohorts—above the gene-wise permutation null but not against covariance-matched nulls in count, so the recurring set is identified by composition, not by its size—and the comparison is defined only after alignment: misaligned cohorts give a zero median adjacency matrix (**0.048**) against **0.731** when aligned. Shuffling sample labels abolishes the module ($\mathbf{14 \rightarrow 0}$ shared edges) while preserving the sample-size dependence, locating an identifiability floor: every cohort still returning edges on shuffled labels has $\boldsymbol{n \lesssim 1.7d}$, so the floor is empirical, not universal. What recurs is the tissue, not the tumour: edge orientations are estimator-selected and unidentified from Gaussian observational data; all results are bulk HiSeqV2 tumours.

Keywords: Pan-cancer analysis, Causal discovery, Tumour composition, CXCR3 chemokine axis, Tumour microenvironment, NOTEARS

1 Introduction

Cancer comprises hundreds of diseases distinguished by tissue of origin, molecular subtype and clinical behaviour (Cancer Genome Atlas Research Network 2012; Hanahan and Weinberg 2011). A causal regulatory edge inferred in one tumour type is routinely assumed to transfer to the others, and pan-cancer studies read the edges their models share as evidence of a common regulatory programme. Whether the same structure recurs edge by edge across an aligned panel has not been assessed across the full set of 33 TCGA cohorts. Multi-domain causal discovery addresses a related but distinct problem—estimating a causal model whose parameters shift between environments (Huang et al. 2020; Ghassami et al. 2018; Gu et al. 2025; Yang et al. 2025)—and we return to the difference in the Discussion.

The obstacle is not the estimator but the comparison. TCGA cohorts are analysed with tissue-specific gene panels, so two cohorts index different gene pairs at the same position of their adjacency matrices and an element-wise comparison between them is undefined. Existing work sidesteps the mismatch by analysing each cohort in isolation (Kumar and Vindal 2023), which precludes comparison; by pooling samples across types, which assumes a homogeneity that the tissue-specific rewiring of oncogenic pathways contradicts (Sanchez-Vega et al. 2018); or by restricting comparison to shared regulators and leaving the edges themselves unexamined (Emad and Sinha 2021). Where comparison across independent expression studies has been carried out, it has been at the level of preserved co-expression modules rather than of individual edges (Jha et al. 2016). Continuous-optimisation causal discovery makes per-cohort graph estimation feasible (Zheng et al. 2018; Pearl 2009), but it does not by itself make two per-cohort graphs comparable.

Two questions follow. First, if the panel is held fixed, what cross-cohort structure is recoverable, and what does it describe—tumour-cell regulation, or the composition of the samples themselves? Second, which genes can such a panel contain at all? The standard design selects on expression dispersion, and we ask whether that criterion admits the genes pan-cancer studies are usually built around. The questions are linked: what the graph can contain determines what it can report. Recent single-gene and single-organ studies ask where in a tumour a gene acts—pan-cancer positioning of DNA2 (Pandit et al. 2026), molecular classification of non-muscle-invasive bladder cancer (Liu et al. 2026), pan-cancer subtyping from pathway clusters (Xu et al. 2024), and integrative analyses of individual drivers (Sun et al. 2026)—and the question here is how far such a claim travels between tumours.

We estimated a NOTEARS graph for each of 33 TCGA cancer types, aligned every cohort on a common gene panel, and compared the graphs edge by edge. Within this design the comparison is defined only under alignment; the structure it recovers is a small co-induced chemokine module on non-malignant axes rather than a driver programme; and the number of edges a cohort yields tracks its sample size.

We state the alignment requirement, the composition attribution and the sample-size floor as findings of this design, together with the ranges over which each was tested and the observation that would overturn each finding, and we release the graphs, the comparison layer and the diagnostics as open-source resources.

2 Methodology

2.1 Data

We used TCGA HiSeqV2 RSEM-normalised gene expression data (Li and Dewey 2011) for 33 cancer types, obtained from the UCSC Xena browser (Goldman et al. 2020); Table 1 gives the sample sizes, which total $n = 10{,}459$. For each cancer we selected the top $d = 100$ genes by median absolute deviation (MAD) across samples, the standard preprocessing of transcriptomic network analysis (Langfelder and Horvath 2008), preferring it to variance because it is insensitive to outliers. Data were standardised to zero mean and unit variance per gene. Sample counts differ slightly between analysis layers because each layer requires a different data type: of the 1,218 breast carcinomas with expression profiles, 1,082 have usable overall-survival records and 1,078 are present in the GISTIC 2 thresholded copy-number matrix. Three rows of the source matrix carry no gene symbol and were annotated by their Entrez identifiers (*TMPRSS11E*, *EZHIP* and *RGPD7*); they account for 12 of the 2,445 edges, none of them shared between cohorts. Preprocessing and analysis are implemented in a version-controlled, open-source pipeline.

2.2 Terminology: SEM causal versus interventional causal

Throughout this paper we use "causal" in the sense established by the causal discovery literature (Pearl 2009; Spirtes et al. 2000; Zheng et al. 2018). Under the linear Structural Equation Model (SEM) with Gaussian noise, and assuming no latent confounders, W is the adjacency matrix of a directed graph whose edges encode directed conditional independence, so that $X_i \to X_j$ means X_j is a function of X_i in the generative SEM. Two qualifications warrant mention before the limitations. NOTEARS imposes acyclicity as a differentiable penalty and returns the least-violating iterate reached within its iteration budget rather than an exactly acyclic matrix; the achieved violation is reported per cohort in Section 2.8. Direction is likewise a property of the fitted SEM: where the observational distribution leaves several orientations consistent with the data, the reported direction is the one the estimator selected, and every orientation-dependent result below is reported next to its orientation-free counterpart. This **SEM-level causality** is distinct from **interventional causality**, which requires knockout or pharmacological perturbation experiments.

To check the inferred edges against independent resources we compare them with TRRUST v2 (Han et al. 2018), a literature-curated database of experimentally validated transcription factor–target interactions (8,324 distinct pairs when direction is disregarded), and with STRING v12 (Szklarczyk et al. 2025) filtered to high-confidence interactions (combined score > 700). Results are reported in Section 3.10.

Table 1 Cohort panel and NOTEARS edge counts for all 33 TCGA cancer types under the per-cohort top-100 design. **Panel A** lists the 18 cohorts with $n \geq 200$, which clear the identifiability floor of Section 3.8; **Panel B** lists the 15 smaller cohorts, which complete the panel but are reported separately because they fall below the identifiability floor of Section 3.8 ($n \lesssim 1.7d$; for the seven with $n \leq d = 100$ the fitted sample covariance matrix is rank-deficient as well). Panel A edge counts span 12–54 (mean 26.0) and are independent of n in this design (Pearson $r = -0.31$, $p = 0.21$; Spearman $\rho = -0.10$, $p = 0.69$); Panel B carries 1,977 of the 2,445 edges (80.9%) while holding 17.0% of the samples. Reuse = fraction of a cancer's edges also found in at least one other cancer, counted as directed pairs so that the inferred direction must agree ($N(p) \geq 2$), expressed as a percentage; mean reuse across all cohorts is 14.1% (20.0% when orientation is disregarded; Sections 3.1 and 3.8)

Cancer	Abbr.	n	Edges	Reuse
Panel A — cohorts with $n \geq 200$ (18 cohorts, 8,678 samples, 468 edges)				
Breast Invasive Carcinoma	BRCA	1,218	16	12.5%
Kidney Renal Clear Cell Carcinoma	KIRC	606	14	14.3%
Lung Adenocarcinoma	LUAD	576	14	28.6%
Thyroid Carcinoma	THCA	572	36	11.1%
Head-Neck Squamous Cell Carcinoma	HNSC	566	33	0.0%
Lung Squamous Cell Carcinoma	LUSC	553	35	34.3%
Prostate Adenocarcinoma	PRAD	550	18	5.6%
Brain Lower Grade Glioma	LGG	530	32	9.4%
Skin Cutaneous Melanoma	SKCM	474	22	63.6%
Stomach Adenocarcinoma	STAD	450	25	40.0%
Bladder Urothelial Carcinoma	BLCA	426	20	30.0%
Liver Hepatocellular Carcinoma	LIHC	423	16	0.0%
Colon Adenocarcinoma	COAD	329	48	41.7%
Kidney Renal Papillary Cell Carcinoma	KIRP	323	28	14.3%
Cervical Squamous Cell Carcinoma	CESC	308	32	25.0%
Ovarian Serous Cystadenocarcinoma	OV	308	12	8.3%
Sarcoma	SARC	265	13	30.8%
Uterine Corpus Endometrial Carcinoma	UCEC	201	54	9.3%
Panel B — cohorts below the identifiability floor (15 cohorts, 1,781 samples, 1,977 edges)				
Esophageal Carcinoma	ESCA	196	66	19.7%
Pheochromocytoma and Paraganglioma	PCPG	187	51	7.8%
Pancreatic Adenocarcinoma	PAAD	183	92	12.0%
Acute Myeloid Leukemia	LAML	173	65	12.3%
Glioblastoma Multiforme	GBM	172	88	3.4%
Testicular Germ Cell Tumors	TGCT	156	67	3.0%
Thymoma	THYM	122	118	5.1%
Rectum Adenocarcinoma	READ	105	205	9.8%
Kidney Chromophobe	KICH	91	154	3.9%
Mesothelioma	MESO	87	132	0.8%
Uveal Melanoma	UVM	80	167	2.4%
Adrenocortical Carcinoma	ACC	79	127	3.1%
Uterine Carcinosarcoma	UCS	57	229	1.3%
Diffuse Large B-Cell Lymphoma	DLBC	48	233	0.9%
Cholangiocarcinoma	CHOL	45	183	1.6%
Total		**10,459**	**2,445**	**14.1%**

2.3 Two-stage multi-batch causal decomposition framework

Let $\{X_k \in \mathbb{R}^{n_k \times d}\}_{k=1}^K$ be K datasets from K cancer types. We model each cancer's causal adjacency matrix as

$$W_k = W_0 + \Delta_k, \quad k = 1, \ldots, K \tag{1}$$

where $W_0 \in \mathbb{R}^{d \times d}$ is a **shared causal backbone** (a valid DAG satisfying $h(W_0) = 0$) and Δ_k captures **cancer-specific regulatory deviations**. One design choice separates this formulation from prior multi-domain causal discovery: the DAG constraint is enforced *only on* W_0, not on $W_0 + \Delta_k$. Biologically this encodes the assumption that the shared regulatory skeleton across tissues is acyclic, as expected for a core transcriptional program, while tissue-specific perturbations may introduce context-dependent feedback loops. Computationally it eliminates K conflicting augmented Lagrangian terms and improves numerical stability.

Stage 1: Per-batch causal discovery. For each cancer k we solve the standard NOTEARS problem,

$$\min_{W_k} \frac{1}{2n_k} \|X_k - X_k W_k\|_F^2 + \lambda_1 \|W_k\|_1 \quad \text{s.t.} \quad h(W_k) = 0, \tag{2}$$

where $h(W) = \operatorname{tr}(e^{W \odot W}) - d$, using L-BFGS-B (Byrd et al. 1995) within the NOTEARS augmented Lagrangian scheme (Zheng et al. 2018) ($\rho_{\text{init}} = 0.1$, factor $\gamma = 2.0$, $\rho_{\max} = 10^{10}$). Instead of escalating ρ on a fixed schedule we increase it only when $|h(W)|$ *increases* over the previous iteration, which prevents the premature escalation that causes numerical instability in standard SGD-based implementations. We use $\lambda_1 = 0.01$ for TCGA data and $\lambda_1 = 0$ for synthetic benchmarks; the edge detection threshold $\tau = 0.3$ is a reporting convention rather than a fitted parameter, because the estimator returns a continuous weighted matrix (Zheng et al. 2018).

Stage 2: Joint decomposition. Given per-batch estimates $\{\hat{W}_k\}_{k=1}^K$ we compute the elementwise median $\bar{W}$ as the initial shared estimate, preferred to the mean because a single atypical cohort can pull the mean with it. We then project $\bar{W}$ onto the DAG manifold,

$$\min_{W_0} \frac{1}{2} \|W_0 - \bar{W}\|_F^2 + \lambda_{\text{proj}} \|W_0\|_1 \quad \text{s.t.} \quad h(W_0) = 0, \tag{3}$$

using the same augmented Lagrangian with $\rho_{\text{init}} = 0.05$, $\gamma = 2.0$, $\lambda_{\text{proj}} = 0.001$ and at most 100 outer iterations. Cancer-specific deviations follow as $\Delta_k = \hat{W}_k - W_0$, with soft-thresholding at $|\Delta_k[i,j]| < 0.01$ to suppress estimation noise. The synthetic benchmark of Section 2.6 exercises the complete pipeline; on the TCGA cohorts Stage 1 provides the per-cohort graphs and Stage 2's element-wise median provides the aligned-panel estimate, whose DAG projection and Δ_k are reported in Section 3.4.

2.4 Cross-cancer gene-pair comparison

Because tissue-specific biology produces different sets of highly variable genes in each cancer, the top-100 MAD gene sets differ across cancer types and matrix positions are therefore specific to each cancer by design. All edges are converted to gene-name representations and overlap is analysed at the gene-pair level. For each unique gene-pair (g_i, g_j) we record the number of cancers in which it appears, $N(p)$, and its edge weights across those cancers; a pair is classified as "shared" if $N(p) \geq 3$, and as "rewired" if its weights change sign between cancers. For each cancer c the *edge reuse rate* is $|\{p \in \mathcal{G}_c : N(p) \geq 2\}|/|\mathcal{G}_c|$. We reserve "shared" for pairs detected in at least three cohorts and "recurring" for those detected in at least ten, and the two are not interchangeable: the 19 shared pairs include the sex-chromosome class, none of which is among the 14 recurring pairs.

2.5 Aligned-panel comparison and its diagnostics

The gene-pair layer restores comparability across cohorts that select different genes, but it does so by discarding the positional information an adjacency matrix carries. To test whether that information is recoverable we fixed a single panel of 100 genes and refitted NOTEARS on each of the 33 cohorts against it, holding every other setting at the operating point of Section 2.8 ($\lambda_1 = 0.01$, $\tau = 0.3$). The panel is built without reference to any outcome: for every gene present in all 33 cohorts (20,530 genes), its within-cohort MAD percentile is averaged over the cohorts and the 100 highest-ranked genes are retained. It is deliberately not an intersection—the 33 per-cohort top-100 sets share no gene at all—so the panel is a fixed frame applied to every cohort rather than a core common to them. Two of its entries require a note: the highest-ranked gene is GSTM1 and the seventh is XIST, and neither appears in the recurring set. GSTM1's rank follows from its dispersion alone (mean within-cohort MAD percentile 0.997), part of which traces to a common deletion polymorphism rather than to expression biology; XIST belongs to the sex-chromosome class discussed in Section 2.8. The element-wise median of the 33 adjacency matrices was then computed in the aligned design and, for comparison, in the original per-cohort top-100 design, with the identical estimator, panel width and threshold in both. Recurrence is counted per directed pair over the 9,900 pairs the panel admits, a pair being treated as recurring when it is recovered in at least ten of the 33 cohorts. To test whether the conclusions depend on the particular genes chosen, the whole procedure was repeated on a second seeded panel drawn from median-MAD ranks 201–2,000, which shares no gene with the first.

Three diagnostics were computed on the aligned panel. Significance of recurrence was assessed by permuting sample labels independently for every gene within every cohort—preserving each gene's marginal distribution and removing all between-gene dependence—and repeating the pipeline 500 times; since the total edge count is an invariant of node permutation, the test statistics were the count of pairs reaching each recurrence threshold and the collision statistic $Q = \sum_p N(p)^2$. A second, sample-level control ran the same permutation once per cohort so that each shuffled fitted matrix could be inspected directly and the empirical boundary located without extrapolation. For the attribution, the dominant axis was read from the pooled 33-cohort matrix

against the Marchenko–Pastur upper edge $\lambda_+ = (1+\sqrt{d/n})^2$, PC1 was correlated with 22 stromal, immune and endothelial marker genes drawn from outside the panel and listed in the replication package, and entry of canonical driver genes was quantified for $k \in \{100, 200, 500, 1{,}000, 2{,}000\}$ against a fixed list of 67 canonical drivers.

2.6 Synthetic validation

The framework was validated end-to-end on a sparse chain-structured DAG ($d = 10$, $K = 3$) carrying a known shared backbone and four batch-specific edges, one of them sign-flipped relative to that backbone. Stage 1 recovered 13, 13 and 11 edges against ground truths of 13, 12 and 11; the element-wise median returned a shared estimate with $h(\bar{W}) = 1.1 \times 10^{-3}$, and the Stage-2 projection lowered that to 5.2×10^{-6} while preserving 10 of the 11 shared edges. Fourteen of the 15 ground-truth edges were recovered overall (recall = 0.93), including the sign-flipped rewired edge, which was correctly assigned to Δ_1. The projection also removed every spurious edge that Stage 1 introduced at this sample size. The benchmark establishes that the pipeline separates shared from batch-specific structure under controlled, low-noise conditions; it is not evidence that the same separation holds at $d = 100$ under real expression noise, which the TCGA analysis tests directly rather than by analogy. Ground-truth recovery metrics, the threshold grid and the per-batch breakdown are reported in Online Resource 5, Section S1.

2.7 Parameter sensitivity and robustness

Both hyperparameters were selected on a grid spanning the sample-size range of the panel (BRCA, 1,218 tumours; LUAD, 576; COAD, 329). Edge counts stabilised by $\lambda_1 = 0.01$ in all three cohorts while the recovered networks stayed compact, and the augmented Lagrangian behaved stably across the grid; a single under-regularised fit at $\lambda_1 = 0.001$ in COAD has the profile of estimation noise rather than of biology. The edge threshold was set at $\tau = 0.3$, where counts were balanced rather than inflated ($\tau = 0.1$, median 401 edges) or collapsed ($\tau = 0.5$, median 7). Because a continuous-optimisation solution is threshold-free, τ is a reporting choice stated explicitly so that the edge counts can be reproduced (Zheng et al. 2018). Hyperparameter choice is itself a recognised source of variability in structure learning (Machlanski et al. 2024). The full grids, the selected operating point and the sensitivity figure are in Online Resource 5 (Section S2 and Fig. S1).

2.8 Acyclicity at the operating point

The acyclicity NOTEARS imposes is a property of the optimisation rather than of the thresholded graph, so we state how far it holds in practice. The augmented Lagrangian stopped short of its tolerance in all 33 cohorts, and the iterate returned is the least-violating one seen, with $|h(W)|$ between 1.1×10^{-3} and 1.8×10^{-1} (median 4.4×10^{-2}). The fitted matrices are therefore near-acyclic rather than exactly acyclic, and the graphs analysed below are the entries that survive thresholding at $\tau = 0.3$, not the raw optimiser output. Thresholding removes most of what remains: sixteen of the 33 graphs are acyclic, and the 109 edges lying on a directed cycle in the other 17

amount to 4.5% of the 2,445 edges. None of the 14 recurring pairs of the aligned panel lies on a directed cycle in any cohort. The per-cohort violation and a flag on every cyclic edge are written to the replication package; the cohort-level distribution and the composition of the cyclic edges are in Online Resource 5, Section S3.

2.9 Baseline methods for comparison

To verify that the low cross-cancer sharing rate is not an artifact of NOTEARS' linearity or DAG constraint we compared against two baselines. **GENIE3** (Huynh-Thu et al. 2010) is a random-forest method that imposes no DAG constraint and captures non-linear dependencies by ranking candidate regulators on Out-of-Bag variable importance; we used 100 trees per gene with the top-K edges selected per cancer, where K is set to the per-cancer NOTEARS edge count so that the two networks are equally sparse for the sharing-rate comparison. **Pooled NOTEARS** concatenates all 33 cancers into a single matrix ($n = 10{,}459$) with the top-100 genes selected by pan-cancer median MAD and fits one model, simulating the standard pan-cancer pipeline. We restrict the comparison to estimators that return a directed graph over named genes, so that the sharing rate is computed on the same object in every case. Non-linear and neural extensions of the continuous DAG program (Zheng et al. 2020; Lachapelle et al. 2020; Yu et al. 2019) and additive-noise formulations (Bühlmann et al. 2014) change the estimand and are treated as complementary methods.

2.10 Statistical analysis

All reported associations between continuous variables are Spearman rank correlations, chosen because gene expression is right-skewed and the relation to survival or immune abundance is expected to be non-linear; the relation between a cohort's sample size and its edge count, and the correlations of the dominant principal component with the external marker score and with aneuploidy burden, are reported as Pearson correlations with the rank counterpart given alongside. Group comparisons use two-sided Mann–Whitney U tests for independent samples and Wilcoxon signed-rank tests for paired samples; departures from a specified null distribution are assessed with the Kolmogorov–Smirnov statistic, and sets of p-values are combined with Fisher's method. Where a family of tests is reported as a set, the false-discovery rate is controlled by the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995). Over-representation of network genes in curated gene sets is assessed with a one-tailed hypergeometric test against the cohort's own top-100 MAD gene pool as the universe, which keeps gene-selection inflation out of the test (Subramanian et al. 2005; Liberzon et al. 2015). Survival associations are estimated by Kaplan–Meier analysis with two-sided log-rank tests, and hazard ratios with their confidence intervals come from univariate Cox proportional-hazards models fitted to the harmonised follow-up data (Liu et al. 2018).

Several results are resampling statements rather than significance statements and are reported as retention fractions over draws, using two schemes on the aligned panel—drawing a cohort's observations with replacement and refitting, and refitting

the same data from six random initialisations—with the denominator for each recurring pair restricted to the cohorts whose original fit placed that directed edge. Null panels replace each panel gene with a gene drawn, without replacement, from the same decile of mean expression and—in the second family—of across-cohort median MAD, among genes outside the panel, and a rotation control applies an independent random orthogonal matrix to each cohort's panel. Multiple-testing control is applied where a family of tests is quoted as a family—Benjamini–Hochberg across the six nominated hubs—and nowhere else: the BRCA 98-gene scan, the 33-hub scan, the pathway over-representation scan and the null-panel rankings are reported with nominal or descriptive statistics and are labelled exploratory where they appear. The full protocol, with the permutation seeds and draw counts, is restated in Online Resource 5, Section S22.

2.11 Implementation

All code is implemented in Python 3.14 with NumPy 2.4 (Harris et al. 2020), SciPy 1.17 (Virtanen et al. 2020) (L-BFGS-B via `scipy.optimize.minimise`), pandas 2.3 (McKinney 2010) and scikit-learn 1.9 (Pedregosa et al. 2011); figures are produced with Matplotlib 3.10 (Hunter 2007), and the exact versions are pinned in the replication package. The Stage 2 DAG projection uses the same solver as Stage 1. Total runtime for the 33 per-cancer fits on a single CPU core is approximately 25 minutes, with the permutation layers and the projection additional and recorded per stage in the replication package. All results are deterministic: NOTEARS is run from a fixed initialisation with fixed seeds, so repeated runs reproduce the reported numbers exactly. The complete replication package is publicly available at https://github.com/sgao-academics/multibatch-causal under the MIT license.

2.12 Use of AI-assisted tools

Generative AI tools (WorkBuddy, an AI writing assistant) were used solely for language polishing of the manuscript. The author designed the study, selected the data and the analytical design, developed the methodology, wrote all code, conducted all analyses and prepared all figures, and takes full responsibility for the content of the manuscript. No AI tool was used to generate, alter, interpret or fabricate data, images or results, and no AI tool is credited as an author.

3 Results

3.1 Edge-level comparison across cancer types requires an aligned panel

The prevailing pan-cancer reading is that a regulatory edge inferred in one tumour type transfers to the others, but that reading can only be tested if the objects compared occupy the same position in the two models, and the condition fails by default in TCGA. When every cohort selects its own 100 most variable genes the two lists barely intersect—the mean pairwise overlap is 7.4 of 100 genes (mean Jaccard 0.040, range

0.000–0.504)—so two cohorts' adjacency matrices index different gene pairs at the same matrix position and an element-wise comparison between them is not defined. Holding the panel fixed restores the comparison. We re-ran NOTEARS on a single 100-gene panel applied to all 33 cohorts (10,459 samples) and recomputed the element-wise median adjacency matrix under both designs (Table 2). Across the misaligned cohorts the median matrix is identically zero: no position clears τ, and the largest median |coefficient| is 0.048 (Fig. 1a). On the aligned panel the same summary is not zero: six positions clear $\tau = 0.3$ and the largest reaches 0.731. The result is not the product of a fortunate subset—recomputed on the first four cohorts the median matrix returns nine edges (largest median coefficient 0.765), and extended to all 33 it returns six (largest 0.731)—and Section 3.7 tests it against nulls that retain each cohort's covariance structure. Alignment is therefore required for element-wise cross-cohort comparison, and every analysis below is computed on the aligned panel.

Table 2 Cross-cohort causal comparison under two panel designs, 33 TCGA cohorts. Both panels are fitted with the same estimator, sparsity ($\lambda_1 = 0.01$) and edge threshold ($\tau = 0.3$); only the gene panel differs. The element-wise median adjacency matrix is the quantity that makes cohorts comparable, and it is zero across the whole misaligned design; an edge enters it when the median of the absolute coefficients across cohorts exceeds τ

	Aligned panel	Misaligned (per-cohort top-100)
Cohorts compared	33	33
Total inferred edges	2,239	2,445
Median-matrix edges ($> \tau$)	**6**	**0**
Largest median |coefficient|	**0.731**	**0.048**
Distinct genes spanned	100	1,775

3.2 Fourteen directed pairs recur across the panel, and the strongest form a co-induced chemokine module

On the aligned panel the 33 cohorts contribute 2,239 edges over the 9,900 directed gene pairs the panel admits. Recurrence is rare and unevenly distributed: fourteen pairs—0.14% of the space—are recovered in ten or more cohorts (Fig. 1c; Table 3). The count is significant against the gene-wise permutation null but not against covariance-matched nulls in count, so what singles this set out is its composition rather than its size (Section 3.7). The most prominent member is a single three-gene module rather than a scattered set: CXCL9–CXCL10 is recovered in 26 of 33 cohorts and CXCL10–CXCL11 in 28. We do not read that component as a directed cascade, because the known regulation is parallel rather than sequential. CXCL9, CXCL10 and CXCL11 are the three interferon-γ-inducible ligands of CXCR3, and reviews of the system conclude that they act in a non-redundant, ligand-specific manner rather than as a chain, recording claims of "mutual redundancy, ligand dominance, collaboration or even antagonism" that depend on context (Metzemaekers et al. 2017); the ligands

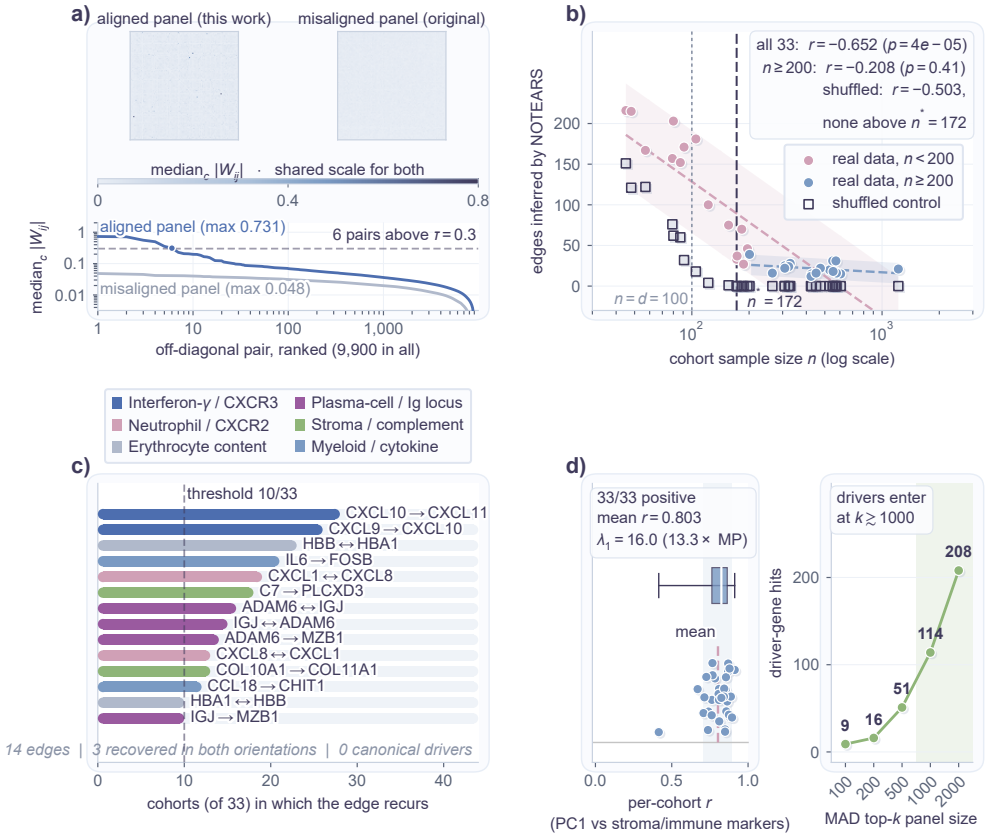


Fig. 1 Cross-cohort causal comparison across 33 TCGA cohorts on a common 100-gene panel. **a)** Element-wise median of the inferred adjacency under the two panel designs, read on one colour scale (top) and with the same off-diagonal entries ranked by magnitude below; the misaligned design clears the edge threshold $\tau = 0.3$ nowhere, the aligned panel retains shared structure. **b)** Edges per cohort against sample size on the aligned panel; coefficients are Pearson, and the gradient is confined to cohorts whose sample count falls below the number of variables fitted. **c)** Recurrence of directed pairs, coloured by the six non-malignant axes onto which the recurring set resolves; the strongest two edges link the CXCL9→CXCL10→CXCL11 co-induced module. **d)** Dominant axis of the aligned panel: PC1 of the pooled matrix against an external stromal/immune marker score, cohort by cohort, together with the dispersion threshold at which canonical drivers begin to enter the panel

are not even induced together, since CXCL10 also responds to innate type I interferon whereas CXCL9 induction is restricted to interferon-γ (Groom and Luster 2011). No evidence establishes that CXCL9 regulates CXCL10 or that CXCL10 regulates CXCL11, so the arrows on these two edges are the orientation the estimator selected rather than a direction the data identify (Section 3.3).

Every other recurring pair sits on the same kind of axis: interferon-γ/CXCR3 chemokines, neutrophil/CXCR2 chemokines (CXCL1, CXCL8), blood content (HBA1, HBB), the plasma-cell immunoglobulin locus (ADAM6, IGJ, MZB1), stromal abundance (COL10A1, COL11A1, C7, PLCXD3) and myeloid/macrophage abundance

(IL6, FOSB, CCL18, CHIT1). None of the 14 involves a canonical driver gene. The recurrence is far above what the panel geometry generates: under 500 permutations no pair reaches ten cohorts (observed 14), and the collision statistic drops from 8,037 to 2,715 $\pm$ 32 ($z = 166.8$, $p \leq 0.002$). Three qualifications are carried through the rest of the paper. Of the 14 edges, three are recovered in both orientations (ADAM6↔IGJ, CXCL1↔CXCL8, HBA1↔HBB), so only eight carry an unambiguous direction. The permutation test shows that recurrence is not expected under exchangeability, but it does not license a causal reading of any individual edge, so we describe the recurring set as conditional dependencies at the transcript level throughout. And correlated expression can be produced without any regulatory relation when two genes share a copy-number state, which Section 3.12 measures directly. The count moves with the reporting threshold—44 pairs at $\tau = 0.15$, 14 at 0.30 and 7 at 0.50—while the chemokine module stays above ten cohorts across the whole range (Online Resource 5, Table S6). The length-two path enumeration is in Online Resource 5, Section S9.

Table 3 The 14 directed gene pairs recovered in ten or more of the 33 TCGA cohorts on the aligned panel, grouped by the non-malignant axis they belong to. "Cohorts" gives the number of cohorts in which the directed pair is recovered; three pairs are recovered in both orientations and are marked ↔. "CN ρ" is the Spearman correlation between the thresholded copy-number states of the two genes over the 9,415 samples of the aligned panel (Section 3.12); the four pairs at $\rho \geq 0.98$ lie within one locus each and are therefore indistinguishable from co-amplification, which the reader should carry as a limit on their *set* and *abundance* entries alike. "Basis" records what can be said about the orientation: *set* means the two genes are known to be co-induced as a regulatory set, so the arrow is the estimator's ordering of a symmetric relation; *abundance* means both genes mark the same non-malignant cell type and their co-variation is that cell type's varying share of the sample, with no established direction between them; *both* means the pair is recovered in both orientations; *none* means neither relation is established for the pair, so its arrow is carried without a mechanistic reading. No pair involves a canonical driver gene, and none of the 14 lies on a directed cycle in any cohort.

Axis	Directed pair	Cohorts	CN ρ	Basis	Note
Interferon-γ / CXCR3	CXCL10→CXCL11	28	1.00	set	co-induced module
	CXCL9→CXCL10	26	1.00	set	
Neutrophil / CXCR2	CXCL1↔CXCL8	19/13	0.99	both	
Blood content	HBB↔HBA1	23/10	0.12	both	
Plasma-cell / Ig locus	ADAM6↔IGJ	16/15	0.07	both	
	ADAM6→MZB1	14	−0.03	abundance	
	IGJ→MZB1	10	0.13	abundance	
Stroma / complement	C7→PLCXD3	18	0.98	abundance	PLCXD3 poorly characterised
	COL10A1→COL11A1	13	0.05	abundance	
Myeloid / cytokine	IL6→FOSB	21	0.10	none	immediate-early gene
	CCL18→CHIT1	12	0.04	abundance	

3.3 How far the orientations survive resampling

Every recurring pair arrives with an orientation, and orientation is the part of a NOTEARS output that observational Gaussian data do not identify. Two resampling

schemes were run on the aligned panel. Bootstrap resampling of each cohort's samples reproduces the original directed edge in 0.77 of draws on average and never produces the reverse edge, so what resampling removes is the edge rather than its direction; refitting the 20 densest cohorts from six random initialisations, with the data held fixed, agrees with the original orientation in only 0.49 of restarts. The second figure is the relevant one for a reader asking whether an independent run would recover the direction. The consequence for Table 3 is that the "Cohorts" column states that an edge is present at a directed position, while the "Basis" column is where the direction is adjudicated: for the co-induced chemokine set the arrow is an ordering imposed on a relation the data make symmetric, for the pairs marked *both* nothing prefers either orientation, and for the remaining pairs the arrow is the estimator's ordering of a dependence it detects reliably. Per-pair retention under both schemes is in Online Resource 5, Section S4.

3.4 The two-stage decomposition on the aligned panel

Running the second stage on the 33 aligned cohorts separates what all cohorts share from what only some do. The element-wise median $\bar{W}$ is already almost acyclic, with $h(\bar{W}) = 4.5 \times 10^{-3}$, and every one of the 33 individual fits has a larger violation—so the DAG constraint is imported at a cost the shared estimate has nearly paid already. Projecting at $\lambda_{\mathrm{proj}} = 0.001$ leaves the thresholded edge set unchanged: six edges above τ at exactly the same positions, and a largest coefficient of 0.730 against 0.731 before. What the stage mainly establishes is how thin the shared backbone is relative to any single cohort: averaged over the 33 cohorts, 84.6% of a cohort's τ-level edges are absent from it, about 64 cohort-specific edges each, so the recurring set is the small part of each graph that survives across cohorts rather than a compressed description of it. The projection's convergence under both penalty schedules is in Online Resource 5, Section S5.

3.5 The reproducible structure is microenvironmental, not tumour-intrinsic

If the recurring edges traced tumour-cell regulatory wiring, they would concentrate on the axes that distinguish one cancer from another. They do not: every one of the 14 sits on an axis contributed by non-malignant cells or by whole-blood content—chemokines that recruit T cells and neutrophils, haemoglobin, immunoglobulin, complement and stromal collagen—which in bulk tumour RNA is the signature of how much tumour, stroma and immune infiltrate a sample contains, the cellular mixture that deconvolution methods are built to recover (White et al. 2024). We tested this attribution directly rather than by inspection. Pooling the 10,459 samples, the largest eigenvalue of the correlation matrix is $\lambda_1 = 16.04$ against a Marchenko–Pastur upper edge of 1.205, with fifteen eigenvalues above the edge where a null matrix leaves none. Scored against 22 stromal, immune and endothelial markers drawn deliberately from outside the panel, so that the panel cannot define its own validator, the leading component correlates at a mean Pearson $r = 0.803$ (range 0.415–0.912), with the sign positive in 33 of 33 cohorts (Fig. 1d). A cross-platform check on copy-number data from the same

tumours agrees in sign ($r = -0.320$, $p = 4.4 \times 10^{-27}$, $n = 1{,}078$ breast carcinomas), a modest magnitude we read as supporting rather than decisive evidence.

The same mechanism explains why no driver gene appears in the recurring set, and shows the exclusion to be structural rather than accidental. Canonical drivers sit at the middle of the transcriptome by dispersion—median percentile 50.8%—whereas each cohort's own top-100 cut sits at the 0.487th percentile of the 20,530 ranked genes. Letting the panel widen moves them in systematically: nine driver selections across the panel at $k = 100$, 16 at $k = 200$, 51 at $k = 500$, 114 at $k = 1{,}000$ and 208 at $k = 2{,}000$ (Table 4), so drivers enter reliably only at roughly ten times the standard panel size. A causal graph estimated on a variance-extreme gene panel therefore reports the composition of the tissue, not its driver architecture. Per-cohort eigenvalue factors are in Online Resource 5, Section S12.

Table 4 Entry of canonical driver genes into the gene panel as a function of panel width. Counts are driver-gene selections summed over the 33 cohorts; the number of distinct cohorts contributing at least one driver is given alongside. Drivers enter systematically only at $k \gtrsim 1000$

Panel width k	Driver selections (33 cohorts)	Cohorts contributing
100 (as used)	9	9
200	16	11
500	51	28
1,000	114	33
2,000	208	33

3.6 Controlling for tissue composition leaves the shared structure standing

If the recurring pairs were a restatement of "this sample contains more non-malignant tissue", then removing a measurement of non-malignant content from every panel gene should dissolve them. Each sample was scored on the same 22 immune, stromal and endothelial markers drawn from outside the panel, every one of the 100 panel genes was residualised on that score, and all 33 cohorts were refitted—so the composition axis is removed from the data before the estimator sees it. The shared structure survives: thirteen of the 14 recurring pairs still recur in ten or more cohorts, the shared median matrix has seven edges above τ against six before, and its largest coefficient is 0.732 against 0.731. What the adjustment does change is the orientation balance of the pairs already flagged as two-sided—the neutrophil and plasma-cell pairs become symmetric and the haemoglobin pair loses its leading orientation, whereas the chemokine pairs keep their one-sidedness—so the adjustment removes a direction, not a pair. It bounds the confound rather than eliminating it, since one composite axis built from 22 markers cannot separate composition from biology as cleanly as deconvolution with cell-type-specific expression would, but it does establish that the

recurrence is not mediated by the bulk-content axis the panel's own first component measures. Per-pair before-and-after correlations are in Online Resource 5, Section S8.

3.7 What the recurrence does and does not establish

The permutation check destroys all between-gene dependence within a cohort, so it answers only whether recurrence is an artefact of the panel's geometry. The harder question is whether a panel assembled by the same rule from unrelated genes would reproduce the same structure. We built four such nulls and refitted all 33 cohorts on each. The two that destroy gene identity return nothing: rotating each cohort's panel by an independent random orthogonal matrix preserves that cohort's covariance spectrum exactly while making every variable a mixture of the 100 genes, and no entry clears τ and no pair reaches ten cohorts. The two that preserve the selection regime do return structure, which is the informative outcome. Twelve replacement panels drawn from the same decile of mean expression are quiet—largest median coefficient at most 0.538 against 0.731, recurring pairs numbering 2–11—whereas twelve more matched on mean expression *and* across-cohort median MAD reach a largest median coefficient of 0.638 and 6–15 recurring pairs, bracketing the real panel's fourteen.

The families separate two quantities that the count alone conflates: density and coefficient strength discriminate, since no constructed panel reaches the real panel's six entries above τ and 0.731, while the recurring *count* does not (14 against a maximum of 15). What the matched families recover is again an axis rather than a pair, and a different axis, dominated by the AKR1C1–AKR1C2 tandem duplicate on 10p15 and by the interferon-inducible GBP5. Selection on dispersion therefore enlarges the recurring set, which is why we report the count with its null and do not use it as evidence, but it does not manufacture the module—that is identified instead by its convergence with STRING, its correlation with panel-exogenous markers, and its survival under composition adjustment. Per-panel results for all 24 nulls are in Online Resource 5, Section S10.

3.8 Recurrence is a property of the fit, and its sample-size dependence is too

The number of edges a cohort yields is strongly tied to its sample size, and the tie is a property of the estimator rather than of the biology. Across the 33 aligned cohorts the correlation between edge count and sample size is $r = -0.652$ ($p = 4.0 \times 10^{-5}$; Pearson, as for every edge-count correlation in this section), but restricted to the 18 cohorts with $n \geq 200$ it disappears ($r = -0.208$, $p = 0.407$; the rank correlations $\rho = -0.85$ and $\rho = -0.24$ change the values, not either reading). The 15 cohorts below $n = 200$ hold 82.6% of the aligned panel's 2,239 edges, so the gradient is carried by the cohorts whose sample count approaches or falls below the 100 variables fitted (Fig. 1b). Permuting the sample labels supports the estimator-property reading: the recurring set collapses from 14 pairs to zero and the largest median coefficient falls from 0.731 to 0.056, yet the sample-size dependence survives at a similar magnitude ($r = -0.503$, $p = 2.9 \times 10^{-3}$). A gradient of the same sign and similar size is generated from pure noise, so any pan-cancer statistic computed without holding sample size

fixed inherits it. The shuffled runs also give the boundary directly: cohorts that still return edges have a maximum of $n = 156$, cohorts that return none have a minimum of $n = 172$, and the regimes separate at $n^* \approx 1.7d$ for $d = 100$, which turns the identifiability floor into an operational rule—place the sample floor at the boundary the study measures on its own data, or reduce d for cohorts below it.

The conclusion does not depend on the particular 100 genes chosen. A second panel drawn from median-MAD ranks 201–2,000, sharing no gene with the first, repeats the qualitative result (Fig. 2): its recurring edges lie once more on lymphoid and immune axes, with no driver edge, and the sample-size dependence reproduces almost exactly ($r = -0.644$ over 33 cohorts; $r = +0.040$ over the 18 with $n \geq 200$). One difference separates two quantities that should not be conflated: the *identity* of the recovered axis and the sample-size dependence are panel-independent and reproduce, while the *number* of recurring edges is not, falling from 14 to six and the largest median coefficient from 0.731 to 0.440 in the same substitution. We therefore decline to treat the edge count as a biological quantity, since it moves with a choice we made rather than with the tumours. Panel-by-panel recurrences are in Online Resource 5, Section S11.

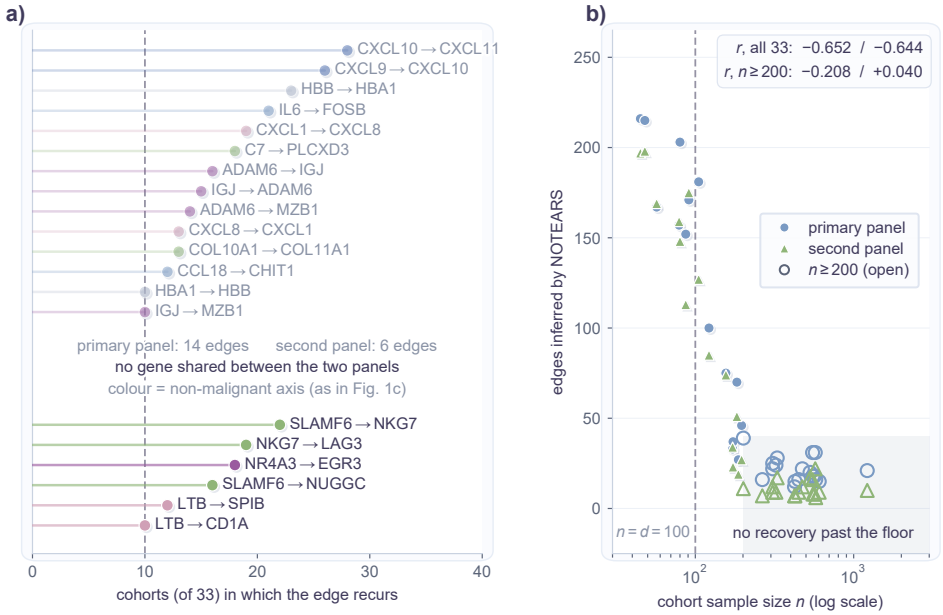


Fig. 2 Robustness of the recurring structure to the choice of gene panel. A second panel of 100 genes drawn from median-MAD ranks 201–2,000 shares no gene with the primary panel. **a)** Recurrence of directed pairs across the two panels; the six edges of the second panel are lymphoid and T-cell-activation genes, which fall outside the six classes of Table 3. **b)** Sample-size dependence of edge count for the two panels, over all 33 cohorts and restricted to $n \geq 200$; both vanish above the floor

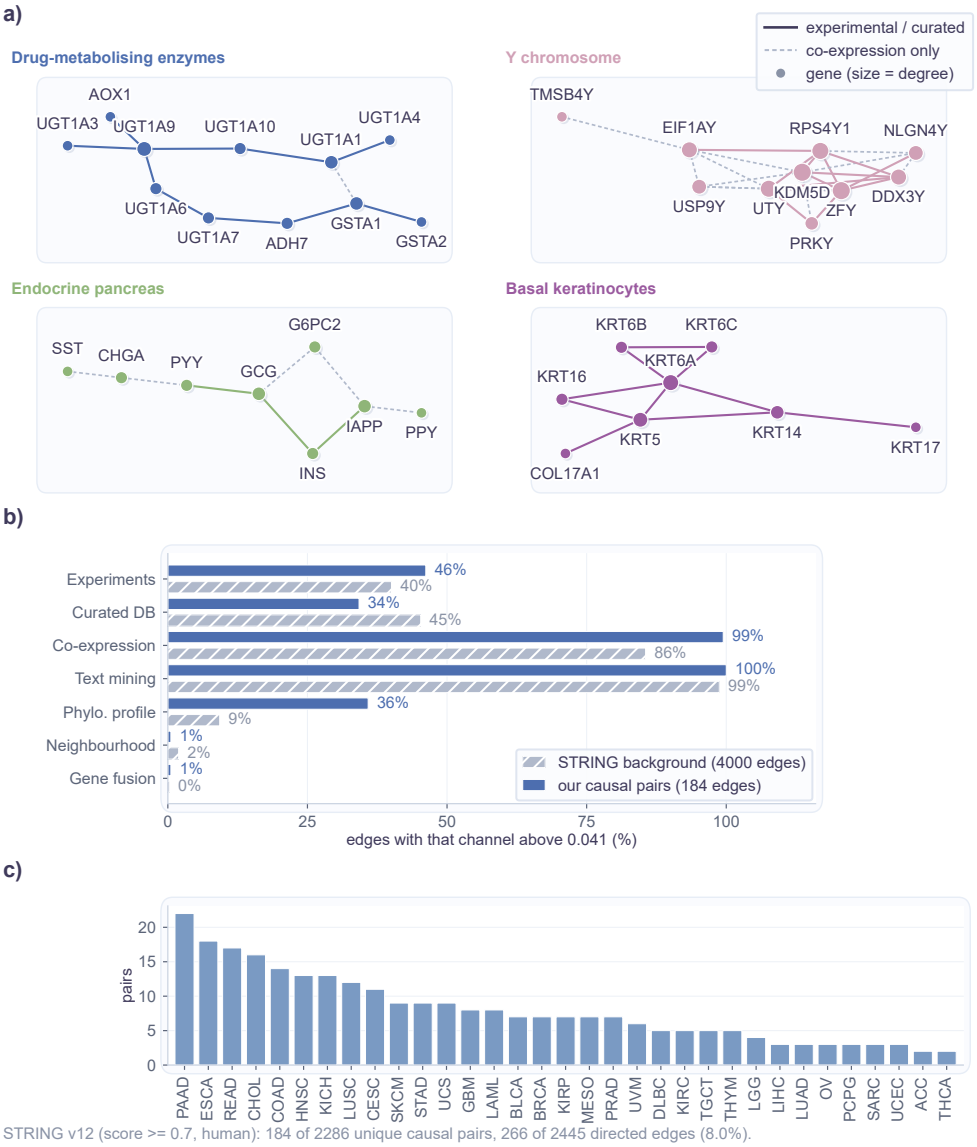
3.9 The boundary is measured at each width, not assumed to scale with it

If the floor were a property of the estimator rather than a measurement on this dataset, the boundary would keep a fixed ratio to the panel width. Repeating the shuffle control at four widths on the same top-d dispersion panels shows that it does not: the smallest cohort returning no edges on shuffled labels sits at $n^* = 80$ for $d = 50$ ($1.60d$), 156 for $d = 100$ ($1.56d$), 173 for $d = 150$ ($1.15d$) and 187 for $d = 200$ ($0.94d$), and at $d = 200$ the two regimes overlap. What is stable across the scan is that the largest cohort still generating edges on shuffled labels lies at or below $1.6d$ at every width. A margin stated as a multiple of d is therefore only as good as the width at which it was measured. The full width scan is in Online Resource 5, Section S6.

3.10 The recurring edges agree with independent resources

We checked the recurring structure against resources that were not used in fitting and that probe different biological layers. Against STRING v12 high-confidence physical interactions (Szklarczyk et al. 2025), 184 of the 2,286 distinct inferred gene pairs have a counterpart (8.0%) against 0.25% of random pairs drawn from the same 1,775-gene pool, a 31.7-fold enrichment (Fisher $p = 7.7 \times 10^{-206}$; Fig. 3); 122 of the 184 carry experimental or curated evidence rather than co-expression alone, which matters because the search is run on expression data. The overlap is partial by construction—STRING records physical binding and stable complexes, the search records transcript-level conditional dependence—so we read it as convergence with independently documented interactions rather than as validation of the edges or of their orientation, and it is exactly the chemokine, complement and immunoglobulin-adjacent modules that carry it. Against TRRUST v2 (Han et al. 2018) the overlap is near-null and we report it as such: 1 of the 2,286 distinct pairs is a documented TF–target pair (Fisher $p = 1$). That scarcity is the predicted consequence of the selection rule—the panel is cut on expression dispersion, transcription factors are transcribed at stable mid-range levels—so we treat it as a reported negative result and not as support.

Two baseline comparisons benchmark the design. Substituting a non-DAG ensemble estimator (GENIE3) on the same panels returns the same qualitative answer—recurrence remains rare, 1.9% of directed pairs in three or more cohorts against 0.8% for NOTEARS (Fig. 4b, upper)—so the low sharing rate is not an idiosyncrasy of one estimator. Pooling all 33 cohorts into a single fit returns 37 edges against 2,445 from the per-cancer fits, a 66-fold reduction: a single sparse DAG cannot represent a mixture of 33 heterogeneous covariance structures. A pathway over-representation scan against MSigDB C2 (one-tailed hypergeometric test with each cohort's own panel as the universe; Fig. 4c) is reported as exploratory and uncorrected, and is read as a consistency check rather than as independent evidence. The DepMap cross-platform check (Tsherniak et al. 2017) covers only eight edges, six of them paralogous family members whose co-regulated expression cannot be independent evidence, so it is reported as a technical check on the expression matrix rather than as confirmation (Fig. 4a). Details of all three checks are in Online Resource 5, Section S13.



STRING v12 (score >= 0.7, human): 184 of 2286 unique causal pairs, 266 of 2445 directed edges (8.0%).

Random pairs from the same 1775 genes reach that score in 0.25% of cases, a 31.7-fold enrichment (Fisher p = 8e-206).

Solid edges carry experimental or curated evidence, dashed edges co-expression only (61 of 184).

Four largest of 87 connected components; the rest are 131 edges including 60 two-gene pairs.

Fig. 3 External support for the inferred edges. **a)** The four largest connected components of the subnetwork formed by the causal gene pairs that STRING v12 also reports; solid edges carry experimental or curated evidence, dashed edges co-expression only. **b)** Share of edges whose STRING evidence channel exceeds 0.041, for the supported pairs and for the STRING background. **c)** Supported pairs per cancer type. STRING does not distinguish orientation, so the undirected unit is the one used here and in Section 3.10: the 2,445 directed edges of the per-cohort fits collapse to 2,286 undirected pairs, 159 of them recovered in both orientations, in different cohorts

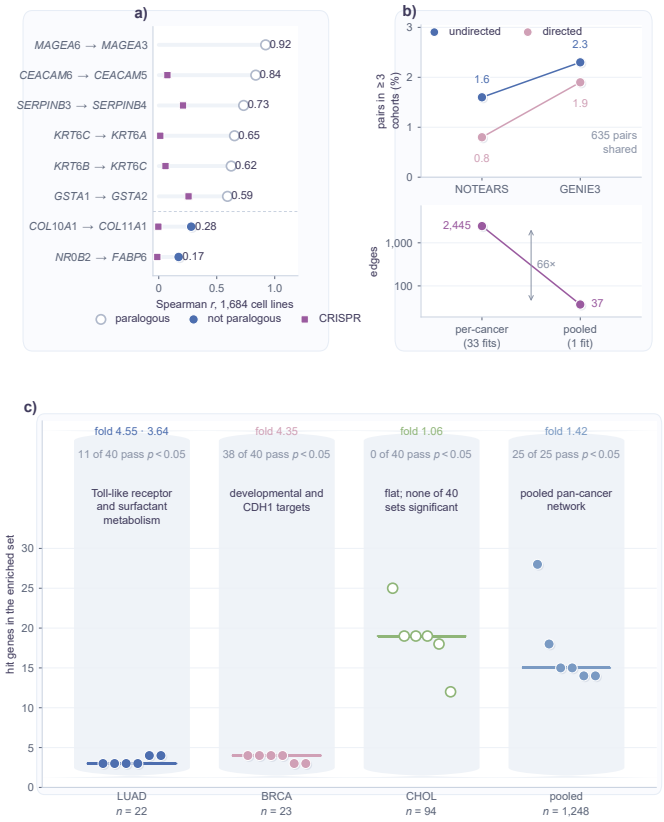


Fig. 4 Three independent checks on the inferred structure, each swapping one component of the pipeline. **a)** Cross-platform check on the DepMap cell-line panel ($n = 1{,}684$ lines), which covers eight of the network edges: the Spearman correlation between the expression levels of the two genes of each edge (circles, one edge per row). Six of the eight are paralogous family members (open), whose co-regulated expression cannot be independent evidence for a regulatory edge; the two that are not paralogous (filled) carry the two lowest correlations. The squares are the CRISPR co-dependency of the same pairs, near zero throughout; across the seven pairs carrying both readouts the two are rank-correlated at $\rho = 0.57$. Only one of the eight, COL10A1–COL11A1, is among the 14 recurring pairs of Table 3, so the resource carries nothing about the axes where the reproducible structure lies and the panel is read as a technical check on the expression matrix, not as validation (Section 3.10). **b)** Estimator and aggregation. Upper: substituting a non-DAG ensemble estimator (GENIE3) for NOTEARS on the same panels, with the edge budget matched, returns the same qualitative answer, the two estimators sharing 635 undirected pairs (of 2,286 and 1,732 respectively). Lower: pooling all 33 cohorts into a single fit and forcing one common panel. **c)** Pathway over-representation of the inferred per-cancer networks (MSigDB C2, one-tailed hypergeometric test, each cohort's own panel used as the universe); the cohorts are ordered by the number of genes their panel can test at all, and each column draws that cohort's six highest-fold sets. A dot sits at the number of hit genes the set draws, filled dots reach $p < 0.05$, and the bar is the column median. The member sets of each column, with their fold, hit count and p, are named in the replication package; the p-values are nominal and uncorrected across the 7,411 C2 sets tested, so the scan is an exploratory consistency check on the axes rather than a significance claim

3.11 Prognostic signal concentrates in the reproducible hubs

The genes the networks nominate are associated with outcome in their own cohort, and the association is confined to the cohorts whose graphs are above the identifiability floor. In each cohort we tested the hub the network had already nominated—a gene fixed by the structure-learning step, not chosen from the outcome—by log-rank at the median split. Across all 33 cohorts six hubs are prognostic at the 5% level against 1.65 expected (3.64-fold; binomial $p = 5.4 \times 10^{-3}$), while the 33 p-values taken as a set do not depart from uniformity, so the excess sits in the tail rather than in the body of the distribution. Restricting to the 18 cohorts that clear the floor raises the ratio to 5.56-fold (5 of 18 against 0.90 expected), while the 15 cohorts below it contribute a single hit against 0.75 expected—indistinguishable from noise. The floor was located from edge counts alone, before any survival data were read, so the split is not outcome-circular, though it is post-hoc and would need confirmation in an independent cohort set. The six hubs are six different genes in six different tissues (FCRL5 in SKCM, SLC22A6 in KIRC, SHOC1 in ACC, CHGB in HNSC, NAPSA in LUSC, SFRP2 in BLCA; Fig. 5), so prognostic value is specific to a gene's position in its own network. Two caveats qualify the ratio, both in the same direction: the 33 tests are not independent, and hubs are the highest-degree genes of panels selected on dispersion, so a hub is also a gene for which a median split carries more power. The excess is therefore an upper bound rather than a calibrated rate.

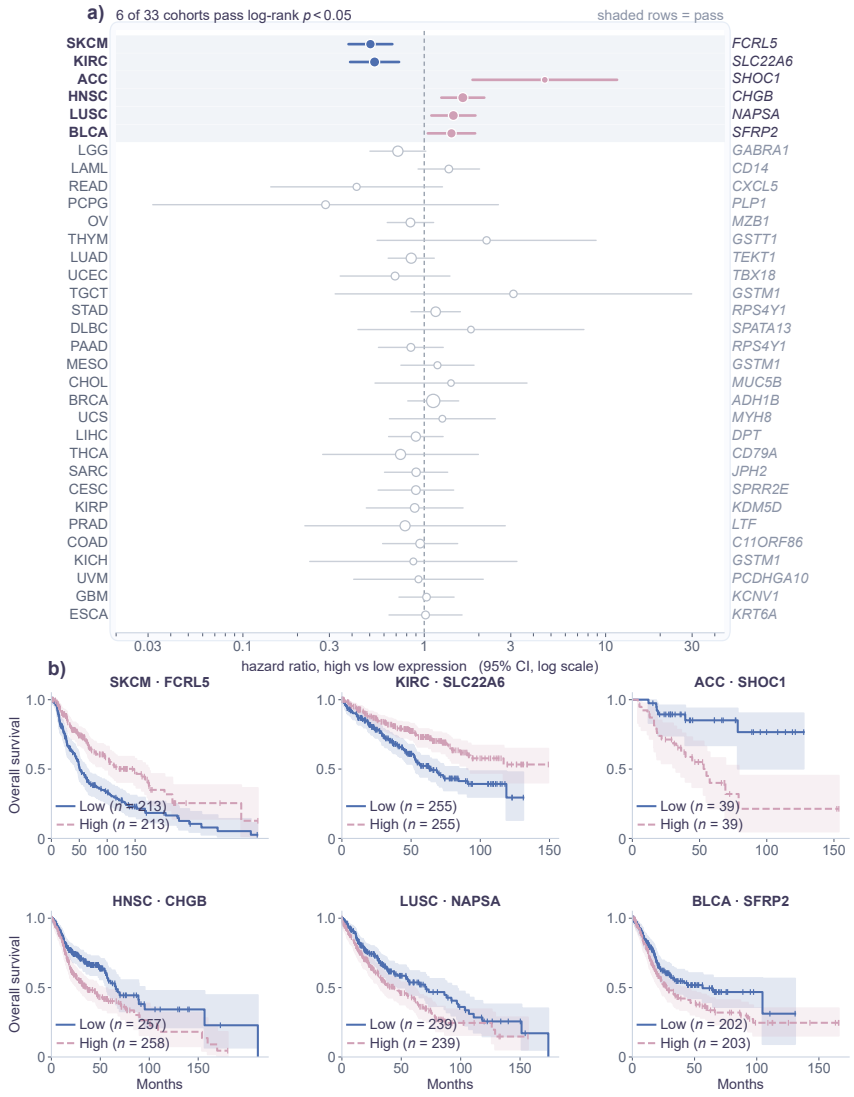


Fig. 5 Overall-survival association of the per-cohort network hubs across all 33 TCGA cohorts. In each cohort the hub gene (highest-degree gene in that cohort's inferred causal network) was split at its median expression and tested by log-rank, so the gene tested was fixed by the network and not selected on the outcome. **a)** Every cohort in the scan, ordered by log-rank p: hazard ratio with 95% confidence interval, dot area the number of patients. The six intervals that exclude 1 are the six significant at the 5% level, filled and coloured by direction (blue, better survival at high expression; pink, worse); open intervals do not reach significance. **b)** Kaplan–Meier curves for those six hubs, with the same hazard ratio (95% CI) and log-rank p: FCRL5 0.50 (0.38–0.67), $p = 9.3 \times 10^{-7}$; SLC22A6 0.53 (0.39–0.73), $p = 5.3 \times 10^{-5}$; SHOC1 4.62 (1.85–11.57), $p = 3.5 \times 10^{-4}$; CHGB 1.63 (1.24–2.14), $p = 3.8 \times 10^{-4}$; NAPSA 1.45 (1.09–1.92), $p = 9.1 \times 10^{-3}$; SFRP2 1.41 (1.05–1.91), $p = 0.023$. Solid curves, low expression; dashed curves, high expression; shaded bands are 95% confidence intervals. The 33 tests are not independent and the ratio is an upper bound (Section 3.11)

Retesting the six hubs with continuous per-SD Cox models on the same harmonised follow-up leaves all six significant after Benjamini–Hochberg correction, and none violates proportional hazards (Table 5); the floor arbitrates the list as well, since SHOC1 is nominated by a cohort of 78 patients and restricting the cohorts to $n \geq 172$ removes it while leaving the other five unchanged. We therefore present the six as hypothesis-generating associations between a hub's expression and outcome in its own cohort: none is adjusted for stage, age, tumour purity or treatment era, none has been validated in an independent cohort, and no clinical utility is claimed. The hubs are positioned as tissue markers rather than as broadly acting genes, with a median tissue-specificity index of 0.63 against 0.73 for the non-hub genes of the same networks (Mann–Whitney $p = 0.047$; Fig. 7a); a hub's own cohort also tends to be one in which that gene sits relatively high (Fig. 7b), and the seven hub genes whose cohort medians are shown in Fig. 7c span the full observed rank range rather than the best cases only. The hub genes are distal to somatic selection: their median non-synonymous mutation frequency is 0.34% against 31.23% for the highest canonical driver in the same cohort (paired Wilcoxon $p = 1.3 \times 10^{-6}$; Fig. 6a,b). The BRCA 98-gene scan, the per-hub Cox models, the immune-abundance comparison and the TISIDB layers are in Online Resource 5, Section S14.

Table 5 The six nominated hubs, under the median split used for nomination and under a continuous per-SD Cox model on the same harmonised overall-survival follow-up. q is Benjamini–Hochberg across the six tests; PH is the proportional-hazards test on scaled Schoenfeld residuals (rank transform). All six remain significant after the correction and none violates proportional hazards.

Cohort	Gene	n	Events	HR (split)	HR per SD	95% CI	q (BH)	PH p
SKCM	FCRL5	425	210	0.50	0.72	0.62–0.82	8.9×10^{-6}	0.26
KIRC	SLC22A6	508	168	0.53	0.70	0.61–0.81	8.9×10^{-6}	0.79
ACC	SHOC1	78	27	4.62	2.39	1.55–3.69	1.6×10^{-4}	0.95
HNSC	CHGB	515	216	1.63	1.27	1.11–1.45	5.5×10^{-4}	0.65
LUSC	NAPSA	477	203	1.45	1.20	1.04–1.40	1.4×10^{-2}	0.21
BLCA	SFRP2	404	177	1.41	1.28	1.09–1.51	2.8×10^{-3}	0.28

3.12 The recurring pairs carry no copy-number coupling beyond physical proximity

The recurrence is read on the expression axis, so the one genomic layer that can produce correlated expression between two genes with no regulatory relation between them has to be separated from it: somatic copy number. Correlating the thresholded GISTIC 2 states of each recurring pair over the 9,415 analysed samples separates the set in two at the genomic scale. The five cis pairs lie 13–462 kb apart and correlate at $\rho = 0.98$–1.00; the nine trans pairs join genes on different chromosomes and correlate at $\rho = -0.03$ to 0.13, indistinguishable from the cross-chromosome background ($p =$

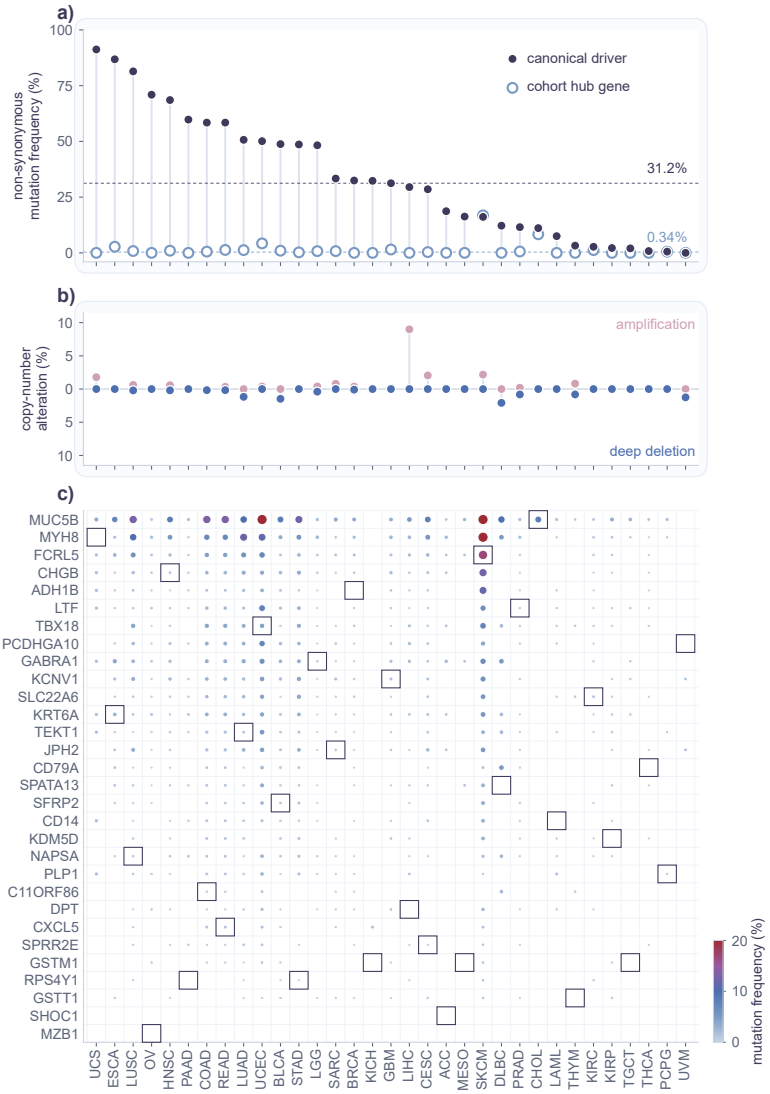


Fig. 6 Somatic alteration burden of the hub genes, queried in the TCGA PanCancer Atlas through cBioPortal (Cerami et al. 2012; Gao et al. 2013). Cohorts are ordered by the driver value in all three panels and their names are printed once, beneath panel **c)**. **a)** Non-synonymous mutation frequency of each cohort's own hub (open lower head) against the highest of the canonical drivers TP53, PIK3CA and KRAS in the same cohort (filled upper head), the stem spanning the gap; 15 of the 33 hub values are exactly zero, the pairing is reversed in one cohort (SKCM, 16.8% against 16.1%) and tied in two. **b)** Amplification (upward) and homozygous deletion (downward) frequencies of the same genes, drawn as points so that a measured zero stays visible; 19 and 21 of the 33 cohorts are zero in the two directions and the largest value anywhere is 9.0%. **c)** Mutation frequency of every hub gene in every cohort, scaled in both size and tone to 20%; a cell is drawn only when it is non-zero, so an empty cell is a measured zero rather than a missing value—463 of the 990 cells (47%) are zero and 14 exceed 10%. The outline marks a gene's own cohort

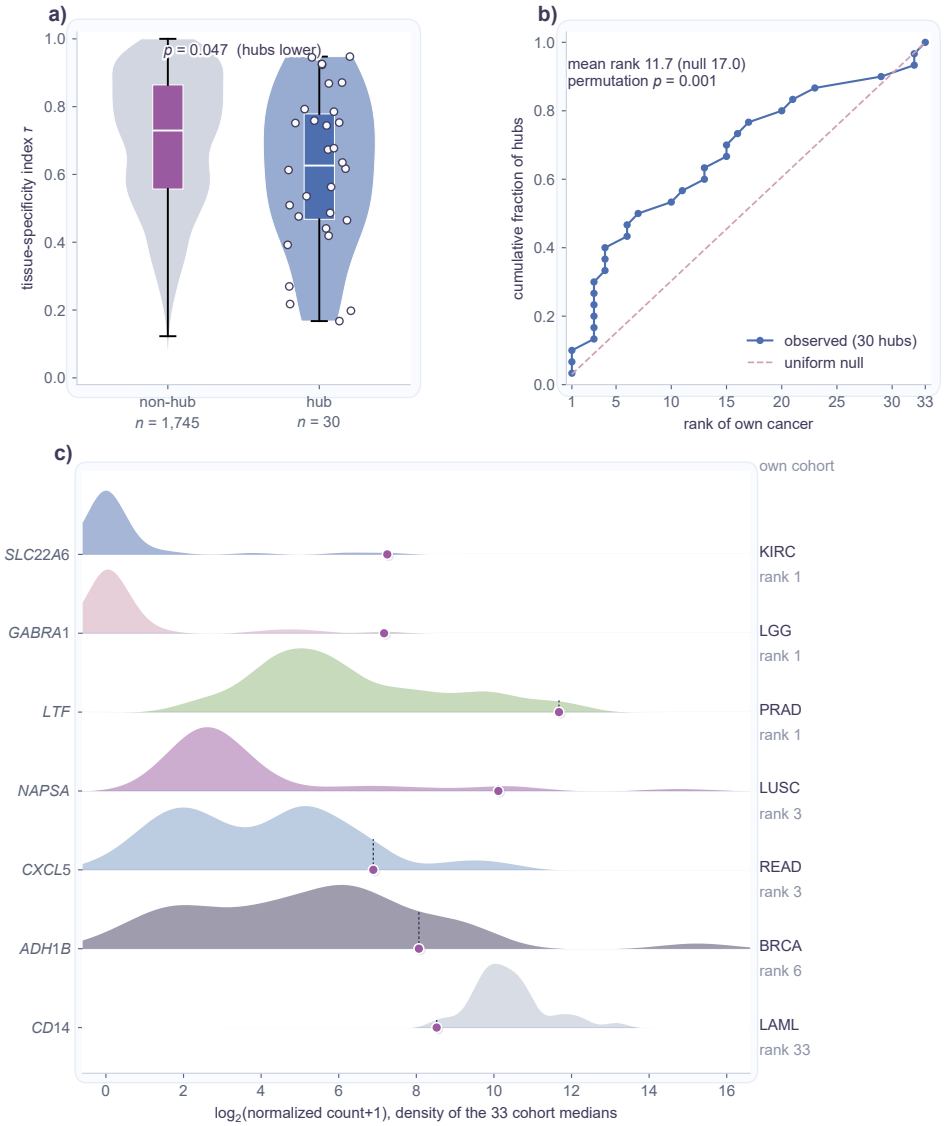


Fig. 7 Tissue-specificity of the per-cohort network hubs. **a)** Yanai's tissue-specificity index τ (Yanai et al. 2005) for the 30 hub genes with sufficient coverage against the 1,745 non-hub genes of the same per-cohort networks (two-sided Mann–Whitney); the non-hubs are shown as a distribution and the hubs as individual values. **b)** Cumulative distribution of the rank of each hub gene's own cohort among the 33 cohorts (rank 1 = highest median expression), against a uniform null. **c)** Expression of seven hub genes across the 33 cohorts, one density per gene over that gene's 33 cohort medians, ordered by how sharply each gene is confined: SLC22A6 has a median of zero in 28 of the 33 cohorts and GABRA1 in 24, whereas CD14 sits in a single high band everywhere. The marker on each row is the cohort in which that gene is the hub, with its rank among the 33; the seven span the full observed rank range rather than the best cases only

0.105), while the two classes differ from each other ($p = 0.0016$). Copy-number co-variation falls monotonically with distance (median ρ from 0.980 within 1 Mb to 0.042 across chromosomes), so a correlation near unity is the expected behaviour of neighbouring genes rather than evidence of anything, and all 14 pairs sit below the 95th percentile of their own distance stratum. Physical proximity alone cannot exclude co-amplification for the cis pairs, including the two chemokine edges in the 4q21.1 cluster, so it was tested rather than assumed: residualising each gene's expression on its own thresholded copy-number state leaves the correlation essentially untouched (CXCL10–CXCL11 at 0.911 both before and after; CXCL9–CXCL10 at 0.853 and 0.853), the 4q21.1 locus is gained in only 8.6% of the analysed samples, and in DLBC—where the locus is never gained—the pair still co-varies at $\rho = 0.941$. The cis pairs therefore enter Table 3 as a statement of what distance alone cannot exclude, not as a mechanism the data support, and the recurrence of the nine trans pairs is left to the abundance reading that Section 3.6 reaches independently. Distance strata and per-pair correlations are in Online Resource 5, Section S7 and Fig. 8.

4 Discussion

4.1 Cross-cancer comparison is defined only after alignment, and the structure it recovers is small and coherent

Pan-cancer causal networks are conventionally read as evidence of a shared regulatory programme, and cross-cohort studies treat recurrence as the quantity of interest (Huang et al. 2020; Ghassami et al. 2018; Gu et al. 2025; Yang et al. 2025; Jha et al. 2016). Our first result is that the comparison this reading presupposes is not defined until the panel is fixed: across the original per-cohort top-100 design the element-wise median adjacency matrix is identically zero, because the cohorts index different gene pairs at the same matrix position. This is a stronger statement than a low sharing rate—at the matrix level there is nothing to share until alignment is imposed. Once aligned, a small and biologically coherent structure is recoverable, and it is not diffuse. Fourteen of the 9,900 directed pairs the panel admits recur in ten or more cohorts—far more than the permutation null generates—and the strongest two edges share a node, giving a three-node module CXCL9–CXCL10–CXCL11 whose ties are named by the estimator rather than resolved by the data. CXCL9, CXCL10 and CXCL11 are the CXCR3-binding chemokines that recruit activated T cells into tissue; all three are interferon-γ-inducible and act as a parallel, co-regulated set rather than a chain (Groom and Luster 2011; Metzemaekers et al. 2017; Schroder et al. 2004), and their coordinate presence is the canonical signature of a T-cell-inflamed tumour (Thorsson et al. 2018; Wu et al. 2024; Liu et al. 2025); we do not test that association here. The remaining recurring edges fall on the same class of axis. What the analysis contributes is therefore not an oncogenic cascade but the reproducible content of a pan-cancer causal comparison: an axis that reports the composition of the samples it is computed on. The result also bounds how much transfer is available: because only 14 of 9,900 pairs recur and the two panels we tested share none of their recurring edges, a study that reports a specific conserved edge should expect it to be one instance of

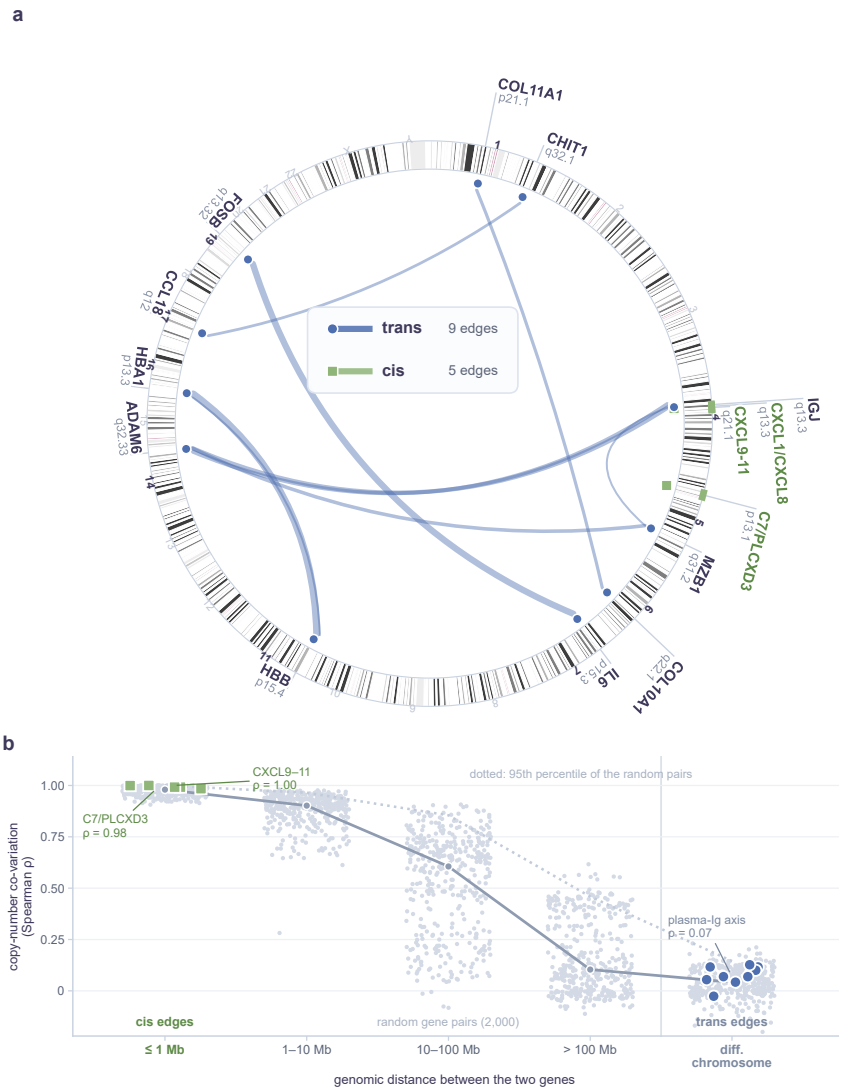


Fig. 8 The recurring pairs carry no copy-number coupling beyond physical proximity. **a)** The 14 recurrent loci on the hg19 ideogram, the chromosomes carrying one lettered in bold; filled squares mark the three loci whose edges are all cis, filled circles the other eleven, and the nine trans edges are drawn as ribbons whose width is proportional to the number of cohorts in which the edge recurs. **b)** Spearman correlation ρ between the thresholded copy-number vectors of the two genes of each recurring directed edge, against the genomic distance separating them; the 14 edges are plotted one by one, cis in green squares and trans in blue circles. Grey points are 2,000 random gene pairs in five distance strata (400 each), the solid line joining the stratum medians and the dotted line marking each stratum's 95th percentile; every recurring pair falls below the 95th percentile of its own stratum, the cis pairs (13–462 kb apart) inside the same-band cloud and the trans pairs inside the cross-chromosome cloud. Gene spans are UCSC hg19 refGene, the copy-number states are the pan-cancer GISTIC 2 calls restricted to the 9,415 samples of the aligned panel, and cis is defined as ≤ 1 Mb on the same chromosome

an axis, and a study that reports a conserved network should expect its content to be microenvironmental.

4.2 The reproducible structure is microenvironmental, and the standard panel cannot see drivers

The recurring edges could in principle have been tumour-intrinsic, in which case the panel would have recovered cancer-cell wiring. Two independent checks say otherwise. The dominant component of the aligned panel sits far outside the noise band ($\lambda_1 = 16.04$ against a Marchenko–Pastur edge of 1.205, with fifteen eigenvalues above the edge where a null matrix leaves none), and it correlates at $r = 0.803$ with stromal, immune and endothelial markers drawn deliberately from outside the panel, positively in all 33 cohorts; a cross-platform check on copy-number data agrees in sign. The first principal axis of a variance-extreme pan-cancer gene panel is therefore the composition of the sample—how much tumour, stroma and immune infiltrate it contains—and the recurring edges inherit that composition. This stands in tension with work showing that sample-specific transcriptional networks can be estimated at all (Ellington et al. 2025), but per-sample structure and cross-cohort recurrence are different quantities, and neither supports the other.

The second check explains why no driver appears in the recurring set, and shows the exclusion to be structural. Canonical drivers sit at the middle of the transcriptome by dispersion—their median percentile is 50.8%—whereas each cohort's own top-100 cut is at the 0.487th percentile, two orders of magnitude away, and widening the panel moves them in systematically (nine driver selections at $k = 100$, 208 at $k = 2{,}000$). The point is not specific to NOTEARS: pan-cancer gene networks have been built for two decades from the same dispersion-selected panels, and those edges carry the same composition signal whether or not their authors read them that way. We state this as the paper's second contribution, because it converts a recurring complaint about pan-cancer networks—that they lack driver signal—into a quantified and predictable property of the gene-selection step. A study that wants driver-level causal structure should widen the panel only where its own sample size supports the width, or condition on purity at the width the samples can carry; one that wants reproducible cross-cancer structure should read its edges as microenvironment-conditional dependencies (Thorsson et al. 2018).

The attribution is compositional in kind, and its individual instances are not portable. Removing an external non-malignant marker score from every panel gene and refitting leaves thirteen of the 14 recurring pairs unchanged, so the recurrence is not mediated by the bulk-content axis that the panel's own first component measures. The pairs we can check for a mechanism fall into two classes: those co-induced as a set, such as the CXCR3 ligands, and those that both mark the abundance of one non-malignant cell type and for which the arrow is the estimator's ordering of a symmetric relation. Reporting an edge without the dispersion percentile of its genes and without a purity control leaves a reader unable to tell a microenvironmental relation from a tumour-intrinsic one, which is why we report both here.

4.3 Recurrence tracks the fitting regime, which fixes an operational sample-size rule

Edge count is a finite-sample property of the estimator rather than a biological signal. The correlation between edge count and sample size is $r = -0.652$ across the 33 aligned cohorts but vanishes among the 18 with $n \geq 200$ ($r = -0.208$), and the 15 cohorts below that line hold 82.6% of the aligned panel's edges. Permuting the sample labels supports the attribution: the recurring set collapses from 14 pairs to zero, yet the dependence on sample size survives at much the same strength ($r = -0.503$). A gradient of the same sign and similar size is generated from pure noise, so no cross-cohort statistic computed without holding sample size fixed can be read as biology.

The same experiment yields the boundary directly. Cohorts that still return edges on shuffled data have a maximum of $n = 156$, cohorts that return none have a minimum of $n = 172$, and the regimes separate at $n^* \approx 1.7d$ for $d = 100$. This recovers the classical requirement that the sample covariance matrix be full rank—$n > d$—with a safety margin, and turns it into an operational rule: place the sample floor at the boundary the study measures on its own data, or reduce d for cohorts that fall below it. The rule comes with two limits. The constant is a measurement rather than a law: repeating the shuffle at $d = 50$, 100, 150 and 200 finds the ratio n^*/d falling from 1.60 to 0.94, so the boundary has to be located at the width in use. And two independent realisations at $d = 100$ place it between $1.2d$ and $1.7d$, so even at one width it is a range. It is not specific to NOTEARS only in the weaker sense that the mechanism—mass on noise directions under an under-determined covariance—is shared by every sparse continuous optimiser, while the position of the floor will shift from one to another (Bello et al. 2022; Zheng et al. 2020; Lachapelle et al. 2020; Yu et al. 2019; Olko et al. 2025). Sample-size requirements for learning from biomedical data have themselves been approached by fitting a learning curve to classifier performance under limited labels (Richter and Khoshgoftaar 2020); the floor reported here is a different quantity—where a sparse structure-learning fit stops returning structure on labels that carry none—and it is located on the studyś own data rather than extrapolated from a performance curve.

4.4 Implications for pan-cancer causal discovery

Three recommendations follow directly from the results above, each stated with the evidence that supports it. **Align the panel before comparing, or compare only named pairs.** Matrix-level cross-cohort comparison is undefined when feature sets differ; the element-wise median is zero by construction in that design. Where a common panel is used, alignment recovers the shared structure; where it is not, comparison must fall back on named gene pairs with the positional information surrendered explicitly. **State the dispersion percentile of every gene of interest, and condition on composition.** Because the top-100 panel sits at the 0.487th dispersion percentile it cannot contain canonical drivers, and the structure it recovers is the sample's composition; a study that selects genes on variance and then interprets edges biologically is reading tissue composition as regulation. **Locate the sample floor by shuffling,**

and do not import a constant. The floor is measurable directly—at $d = 100$ no cohort with $n \geq 172$ returned an edge on shuffled labels while every one with $n \leq 122$ did—so it can be verified rather than assumed, provided its position is read off at the width the study uses.

The comparison layer itself is estimator-agnostic: NOTEARS (Zheng et al. 2018) can be replaced by DAGMA (Bello et al. 2022) or by neural structure learning (Zheng et al. 2020; Lachapelle et al. 2020; Yu et al. 2019), and the alignment and shuffle diagnostics carry over unchanged because they operate on the fitted graph rather than on the objective. The same logic extends beyond cancer transcriptomics—to single-cell data, where the question is how far one cell state's skeleton transfers to another (Tejada-Lapuerta et al. 2025; Yuan and Duren 2025; Funk et al. 2025; Wu et al. 2025; Jiang et al. 2024; Wang et al. 2026; Dimitrov et al. 2026); to pharmacogenomics, where it is whether an edge survives drug perturbation; and to evolutionary comparisons, where it is whether it survives phylogenetic distance. In every case the two diagnostics are the same—align, then shuffle—and the units to report are the axis, the dispersion percentile and the sample floor. Normal-tissue atlases (GTEx Consortium 2020) and cross-species expression compendia (Stuart et al. 2003) supply the reference panels those comparisons will need.

4.5 Limitations

The results are bounded by the following scope limitations, each paired with the corresponding design choice; Online Resource 5, Section S24 carries the same list with the observation that would falsify each claim.

Panel width. The analysis uses $d = 100$ genes, 0.49% of the transcriptome, selected on dispersion. That width is what makes the driver exclusion structural, and the conclusions are stated for it. A $d = 500$ run would test whether the axis content is stable at larger width but is computationally out of reach here; we substitute a second, gene-disjoint 100-gene panel and a driver-entry curve, and state the limitation rather than implying wider coverage.

Composition control. Removing an external non-malignant marker score from every panel gene and refitting leaves the recurring set essentially intact (13 of the 14 directed pairs). Official ABSOLUTE purity estimates are not distributed with uniform processing across all 33 cohorts, so the adjustment rests on one composite axis built from 22 markers drawn from outside the panel rather than on a full multi-dimensional deconvolution; a residual compositional confound cannot be excluded in principle at this level of control, and conditioning on measured purity with an established deconvolution method is the natural next step.

External validity. All evidence comes from bulk human TCGA HiSeqV2 tumours. The indexing argument—that element-wise comparison is defined only when the compared cohorts share a gene panel—is a property of matrix comparison and holds generally; the empirical magnitudes do not, and we make no claim that they transfer to normal tissue, to single-cell or spatial data, to other expression platforms or to other species. The sample-size floor is calibrated on this dataset at this panel width.

Blood content. The erythrocyte axis (HBA1–HBB) is read as whole-blood content. In bulk tumour RNA that signal can come from tissue-resident erythroid cells, from blood carried over in the sample, or from haemolysis during processing, and we cannot separate the three, since sample-level RNA-quality and haemolysis metrics are not distributed with the matrices analysed here. The axis is reported as a composition axis, which it is under any of the three readings; only its cellular attribution is affected.

Axis coverage and model class. The six axes are those whose marker genes carry enough pan-cancer dispersion to enter a 100-gene panel; cell types whose markers are less dispersed—endothelium, adipocytes, the tumour cells themselves—do not appear, and their absence is a property of the selection rule rather than evidence that they contribute no cross-cohort structure. NOTEARS assumes a linear Gaussian SEM, the standard operating point of continuous-optimisation causal discovery; non-linear and neural extensions change the estimand while leaving the alignment and shuffle diagnostics intact, and added flexibility does not by itself improve structure recovery (Olko et al. 2025).

Evidential level, modality and alteration source. Each cohort's graph is a conditional-independence structure estimated under that model, so its edges are model-based hypotheses at SEM level, a step short of interventional effects, which would require PerturbSeq, CRISPR interference or an explicit instrumental-variable design (Pearl 2009; Sachs et al. 2005). The permutation test demonstrates specificity—recurrence is not expected under exchangeability—and not causality, which is why the recurring set is described throughout as conditional dependence. Three of the 14 pairs are recovered in both orientations, leaving eight with a definite direction, and the returned matrices are near-acyclic, with 4.5% of thresholded edges on a directed cycle, each flagged in the replication package. Transcriptomic data report mRNA abundance, so post-translational and metabolite-level regulation lie outside the window of this analysis, and external databases enter only as an orthogonal validation layer, never as discovery priors. The somatic-alteration check uses cohort-level frequencies from the PanCancer Atlas, in which colon and rectal adenocarcinoma share a single study; it establishes the expression-layer position of the hub genes, and the two cohorts with elevated background burden are reported directly in Fig. 6c.

5 Conclusion

This study estimated a graph for each of 33 TCGA cancer types and asked which edges survive across them, and what those edges describe. Within the data regime studied it returns a data-design requirement, a compositional structure and a sample-size heuristic.

The requirement is alignment. Cross-cohort comparison is not defined until the compared cohorts index the same genes: across the original per-cohort top-100 design the element-wise median adjacency matrix is zero, whereas holding the panel fixed recovers shared structure (largest median coefficient 0.731 against 0.048). The structure is small, reproducible and microenvironmental. Fourteen of the 9,900 directed pairs recur in ten or more cohorts, against none under permutation of sample labels ($z = 166.8$); the most prominent is a co-induced CXCR3-ligand module (26 and 28 of

33 cohorts) whose orientations are the estimator's selection, not a direction the data identify. Every recurring pair lies on one of six non-malignant axes, and the dominant component is a microenvironmental content score ($r = 0.803$ against markers from outside the panel, 33 of 33 cohorts). No recurrence involves a canonical driver: driver dispersion lies at the 50.8th percentile while each cohort's own panel is cut at the 0.487th, and drivers enter systematically only at $k \gtrsim 1{,}000$. A variance-extreme pan-cancer analysis reports the composition of the tissue, not its driver architecture.

The heuristic is a sample floor. Edge count tracks sample size ($r = -0.652$) until the cohort list is restricted to $n \geq 200$, where the association disappears ($r = -0.208$); shuffling sample labels abolishes the recurring set ($14 \to 0$) while leaving the dependence intact ($r = -0.503$), and separates cohorts that generate edges from noise between $1.2d$ and $1.7d$ at $d = 100$. A second, gene-disjoint panel reproduces the axis content and the dependence closely ($r = -0.644$), so the reproducible unit is the axis rather than the edge. Within the cohorts above the floor the survival reading is bounded by the caveats stated in Section 3.11: the tests are not independent, and the six continuous models are unadjusted for stage, age, tumour purity and treatment. Alignment, the shuffle diagnostic and the gene-pair database are released as open-source resources for reproducible pan-cancer analysis.

Supplementary Information

Five online resources accompany this paper: four spreadsheets and one file that carries the four supplementary figures together with the supplementary text. Supplementary material is submitted apart from the manuscript, so the spreadsheet tables are separate files. Every item is generated from the analysis outputs by a script in the replication package, so every entry can be regenerated rather than transcribed.

- **Online Resource 1** (`ESM_1.xlsx`) — Table S1, per-cancer network summary: sample size, edge count, edge-weight statistics, the highest-degree hub gene with its degree, its tissue-specificity index and own-cancer expression rank where these could be assessed, and the hub's share of the network's total degree.
- **Online Resource 2** (`ESM_2.xlsx`) — Table S2, all 2,445 inferred causal edges with weight, sign, the number of cancers in which each gene-pair is detected, and the identity of those cancers.
- **Online Resource 3** (`ESM_3.xlsx`) — Table S3, four sheets: the sex-chromosome and same-gene-family pairs detected in three or more cohorts; the DepMap concordance for all eight tested edges; the per-cohort hub genes; and the 22 external non-malignant markers.
- **Online Resource 4** (`ESM_4.xlsx`) — Table S4, two sheets: MSigDB C2 overrepresentation results for each individual cancer and for the pan-cancer pooled network, including every non-significant set so that the full search space is visible.
- **Online Resource 5** (`ESM_5.pdf`) — the four supplementary figures and the supplementary text. Figures: **Fig. S1**, parameter sensitivity across λ_1, τ and the per-cancer edge-count distribution; **Fig. S2**, two-stage decomposition validation on the synthetic benchmark; **Fig. S3**, co-expression structure of the STRING-supported sub-network; and **Fig. S4**, association of each cohort's own hub gene

with the abundance of the 28 TISIDB immune cell types. Text: sections S1–S24 covering the synthetic validation, the hyperparameter grids and diagnostics, the robustness analyses underlying every Result, the full statistical protocol and the extended limitations, with Tables S5–S6.

Statements and Declarations

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Competing Interests

The author has no relevant financial or non-financial interests to disclose.

Author Contributions

S.G. conceived the study, developed the methodology, wrote all code, conducted all experiments, prepared all figures and tables, and wrote the manuscript. S.G. is the sole author and takes responsibility for the integrity of the work.

Data Availability

All code, results and reproduction material for this manuscript are available at https://github.com/sgao-academics/multibatch-causal under the MIT license; the archived version corresponding to this manuscript is tag `v1.0.3`. The TCGA gene expression and copy-number data analysed here are publicly available from the UCSC Xena browser (https://xenabrowser.net/) and are not redistributed with the repository.

Ethics Approval

Not applicable. This study used publicly available, de-identified data and did not involve human participants, animals, or biological specimens requiring ethics committee review.

Clinical Trial Number

Clinical trial number: not applicable.

Compliance with Ethical Standards

All procedures performed in this study complied with the current laws of the People's Republic of China, the country in which the research was conducted.

References

Bello K, Aragam B, Ravikumar P (2022) DAGMA: Learning DAGs via M-matrices and a Log-Determinant Acyclicity Characterization.

In: Advances in Neural Information Processing Systems 35, pp 8226–8239, URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/36e2967f87c3362e37cf988781a887ad-Abstract-Conference.html

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B 57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bühlmann P, Peters J, Ernest J (2014) CAM: causal additive models, high-dimensional order search and penalized regression. The Annals of Statistics 42(6):2526–2556. https://doi.org/10.1214/14-AOS1260

Byrd RH, Lu P, Nocedal J, et al (1995) A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing 16(5):1190–1208. https://doi.org/10.1137/0916069

Cancer Genome Atlas Research Network (2012) Comprehensive genomic characterization of squamous cell lung cancers. Nature 489(7417):519–525. https://doi.org/10.1038/nature11404

Cerami E, Gao J, Dogrusoz U, et al (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discovery 2(5):401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Dimitrov D, Schrod S, Rohbeck M, et al (2026) Interpretation, extrapolation and perturbation of single cells. Nature Reviews Genetics 27(5):349–370. https://doi.org/10.1038/s41576-025-00920-4

Ellington CN, Lengerich BJ, Watkins TBK, et al (2025) Learning to estimate sample-specific transcriptional networks for 7,000 tumors. Proceedings of the National Academy of Sciences 122(21):e2411930122. https://doi.org/10.1073/pnas.2411930122

Emad A, Sinha S (2021) Inference of phenotype-relevant transcriptional regulatory networks elucidates cancer type-specific regulatory mechanisms in a pan-cancer study. npj Systems Biology and Applications 7(1):9. https://doi.org/10.1038/s41540-021-00169-7

Funk MW, Wang Y, Wang L (2025) Airqtl dissects cell state-specific causal gene regulatory networks with efficient single-cell eQTL mapping. Nature Communications 16(1):11403. https://doi.org/10.1038/s41467-025-66214-9

Gao J, Aksoy BA, Dogrusoz U, et al (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling 6(269):pl1. https://doi.org/10.1126/scisignal.2004088

Ghassami A, Kiyavash N, Huang B, et al (2018) Multi-domain causal structure learning in linear systems. In: Advances in Neural Information Processing Systems 31, pp 6266–6276, URL https://papers.neurips.cc/paper/2018/hash/6ad4174eba19ecb5fed17411a34ff5e6-Abstract.html

Goldman MJ, Craft B, Hastie M, et al (2020) Visualizing and interpreting cancer genomics data via the Xena platform. Nature Biotechnology 38(6):675–678. https://doi.org/10.1038/s41587-020-0546-8

Groom JR, Luster AD (2011) CXCR3 ligands: redundant, collaborative and antagonistic functions. Immunology and Cell Biology 89(2):207–215. https://doi.org/10.1038/icb.2010.158

GTEx Consortium (2020) The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369(6509):1318–1330. https://doi.org/10.1126/science.aaz1776

Gu Y, Fang C, Bühlmann P, et al (2025) Causality pursuit from heterogeneous environments via neural adversarial invariance learning. The Annals of Statistics 53(5):2230–2257. https://doi.org/10.1214/25-AOS2541

Han H, Cho JW, Lee S, et al (2018) TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research 46(D1):D380–D386. https://doi.org/10.1093/nar/gkx1013

Hanahan D, Weinberg RA (2011) Hallmarks of cancer: the next generation. Cell 144(5):646–674. https://doi.org/10.1016/j.cell.2011.02.013

Harris CR, Millman KJ, van der Walt SJ, et al (2020) Array programming with NumPy. Nature 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2

Huang B, Zhang K, Zhang J, et al (2020) Causal discovery from heterogeneous/nonstationary data. Journal of Machine Learning Research 21(89):1–53. URL https://jmlr.org/papers/v21/19-232.html

Hunter JD (2007) Matplotlib: a 2D graphics environment. Computing in Science & Engineering 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

Huynh-Thu VA, Irrthum A, Wehenkel L, et al (2010) Inferring regulatory networks from expression data using tree-based methods. PLOS ONE 5(9):e12776. https://doi.org/10.1371/journal.pone.0012776

Jha M, Malhotra AG, Singh S, et al (2016) Gene co-expression network analysis reveals common system-level properties of genes involved in tuberculosis across independent gene expression studies. Network Modeling Analysis in Health Informatics and Bioinformatics 5(1):24. https://doi.org/10.1007/s13721-016-0131-8

Jiang H, Wang Y, Yin C, et al (2024) SLIVER: unveiling large scale gene regulatory networks of single-cell transcriptomic data through causal structure learning and modules aggregation. Computers in Biology and Medicine 178:108690. https://doi.org/10.1016/j.compbiomed.2024.108690

Kumar S, Vindal V (2023) Architecture and topologies of gene regulatory networks associated with breast cancer, adjacent normal, and normal tissues. Functional & Integrative Genomics 23(4):324. https://doi.org/10.1007/s10142-023-01251-5

Lachapelle S, Brouillard P, Deleu T, et al (2020) Gradient-based neural DAG learning. In: International Conference on Learning Representations (ICLR)

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12(1):323. https://doi.org/10.1186/1471-2105-12-323

Liberzon A, Birger C, Thorvaldsdóttir H, et al (2015) The Molecular Signatures Database hallmark gene set collection. Cell Systems 1(6):417–425. https://doi.org/10.1016/j.cels.2015.12.004

Liu J, Lichtenberg T, Hoadley KA, et al (2018) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 173(2):400–416.e11. https://doi.org/10.1016/j.cell.2018.02.052

Liu K, Chen Z, Jin X, et al (2026) An integrated molecular classification system identifies distinct prognostic clusters of non-muscle-invasive bladder cancer. Functional & Integrative Genomics 26(1):103. https://doi.org/10.1007/s10142-026-01884-2

Liu YT, Wang YL, Wang S, et al (2025) Turning cold tumors into hot tumors to ignite immunotherapy. Molecular Cancer 24(1):254. https://doi.org/10.1186/s12943-025-02477-6

Machlanski D, Samothrakis S, Clarke PS (2024) Robustness of algorithms for causal structure learning to hyperparameter choice. In: Proceedings of the Third Conference on Causal Learning and Reasoning, pp 703–739

McKinney W (2010) Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference, pp 56–61, https://doi.org/10.25080/Majora-92bf1922-00a

Metzemaekers M, Vanheule V, Janssens R, et al (2017) Overview of the Mechanisms that May Contribute to the Non-Redundant Activities of Interferon-Inducible CXC Chemokine Receptor 3 Ligands. Frontiers in Immunology 8:1970. https://doi.org/10.3389/fimmu.2017.01970

Olko M, Gajewski M, Wojciechowska J, et al (2025) Since faithfulness fails: the performance limits of neural causal discovery. In: Proceedings of the 42nd International Conference on Machine Learning (ICML), pp 47155–47175

Pandit D, Dhillon A, Chakrabarty S, et al (2026) Multi-omics analysis positions DNA2 at the interface of genome integrity programs and tumor behavior in pan-cancer. Functional & Integrative Genomics 26(1):147. https://doi.org/10.1007/s10142-026-01941-w

Pearl J (2009) Causality: Models, Reasoning, and Inference, 2nd edn. Cambridge University Press, New York, https://doi.org/10.1017/CBO9780511803161

Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: machine learning in Python. Journal of Machine Learning Research 12:2825–2830

Richter AN, Khoshgoftaar TM (2020) Sample size determination for biomedical big data with limited labels. Network Modeling Analysis in Health Informatics and Bioinformatics 9(1):12. https://doi.org/10.1007/s13721-020-0218-0

Sachs K, Perez O, Pe'er D, et al (2005) Causal protein-signaling networks derived from multiparameter single-cell data. Science 308(5721):523–529. https://doi.org/10.1126/science.1105809

Sanchez-Vega F, Mina M, Armenia J, et al (2018) Oncogenic signaling pathways in The Cancer Genome Atlas. Cell 173(2):321–337.e10. https://doi.org/10.1016/j.cell.2018.03.035

Schroder K, Hertzog PJ, Ravasi T, et al (2004) Interferon-γ: an overview of signals, mechanisms and functions. Journal of Leukocyte Biology 75(2):163–189. https://doi.org/10.1189/jlb.0603252

Spirtes P, Glymour C, Scheines R (2000) Causation, Prediction, and Search, 2nd edn. MIT Press, Cambridge, https://doi.org/10.7551/mitpress/1754.001.0001

Stuart JM, Segal E, Koller D, et al (2003) A gene-coexpression network for global discovery of conserved genetic modules. Science 302(5643):249–255. https://doi.org/10.1126/science.1087447

Subramanian A, Tamayo P, Mootha VK, et al (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102(43):15545–15550. https://doi.org/10.1073/pnas.0506580102

Sun K, He P, Chen L, et al (2026) A pan-cancer analysis reveals the multifaceted role of PABPC1L in immune regulation and therapy response. Functional & Integrative Genomics 26(1):221. https://doi.org/10.1007/s10142-026-02004-w

Szklarczyk D, Nastou K, Koutrouli M, et al (2025) The STRING database in 2025: protein networks with directionality of regulation. Nucleic Acids Research 53(D1):D730–D737. https://doi.org/10.1093/nar/gkae1113

Tejada-Lapuerta A, Bertin P, Bauer S, et al (2025) Causal machine learning for single-cell genomics. Nature Genetics 57(4):797–808. https://doi.org/10.1038/s41588-025-02124-2

Thorsson V, Gibbs DL, Brown SD, et al (2018) The immune landscape of cancer. Immunity 48(4):812–830.e14. https://doi.org/10.1016/j.immuni.2018.03.023

Tsherniak A, Vazquez F, Montgomery PG, et al (2017) Defining a cancer dependency map. Cell 170(3):564–576.e16. https://doi.org/10.1016/j.cell.2017.06.010

Virtanen P, Gommers R, Oliphant TE, et al (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 17(3):261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang M, Yang J, Lyu L, et al (2026) CSCN: inference of cell-specific causal networks using single-cell RNA-seq data. Bioinformatics 42(7):btag480. https://doi.org/10.1093/bioinformatics/btag480

White BS, de Reyniès A, Newman AM, et al (2024) Community assessment of methods to deconvolve cellular composition from bulk gene expression. Nature Communications 15(1):7362. https://doi.org/10.1038/s41467-024-50618-0

Wu AP, Singh R, Walsh CA, et al (2025) Unveiling causal regulatory mechanisms through cell-state parallax. Nature Communications 16(1):8096. https://doi.org/10.1038/s41467-025-61337-5

Wu B, Zhang B, Li B, et al (2024) Cold and hot tumors: from molecular mechanisms to targeted therapy. Signal Transduction and Targeted Therapy 9(1):274. https://doi.org/10.1038/s41392-024-01979-x

Xu M, Ai H, Wang D, et al (2024) Gene clusters-based pathway enrichment analysis identifies four pan-cancer subtypes with distinct molecular and clinical features. Functional & Integrative Genomics 24(6):224. https://doi.org/10.1007/s10142-024-01501-0

Yanai I, Benjamin H, Shmoish M, et al (2005) Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21(5):650–659. https://doi.org/10.1093/bioinformatics/bti042

Yang D, Yu G, Wang J, et al (2025) Causal discovery from shifted multiple environments. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp 1751–1762, https://doi.org/10.1145/3690624.3709247

Yu Y, Chen J, Gao T, et al (2019) DAG-GNN: DAG structure learning with graph neural networks. In: Proceedings of the 36th International Conference on Machine Learning (ICML), pp 7154–7163

Yuan Q, Duren Z (2025) Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. Nature Biotechnology 43(2):247–257. https://doi.org/10.1038/s41587-024-02182-7

Zheng X, Aragam B, Ravikumar P, et al (2018) DAGs with NO TEARS: Continuous Optimization for Structure Learning. In: Advances in Neural Information Processing Systems 31, pp 9472–9483, URL https://proceedings.neurips.cc/paper/2018/hash/e347c51419ffb23ca3fd5050202f9c3d-Abstract.html

Zheng X, Dan C, Aragam B, et al (2020) Learning sparse nonparametric DAGs. In: Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), pp 3414–3425

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- supplementarymaterial.pdf
- ESM1.xlsx
- ESM2.xlsx
- ESM3.xlsx
- ESM4.xlsx