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.

A threshold-based classifier of essential versus auxiliary driver mutations across 31 genes and 16 cancer lineages

Chen Shen
shenchen1110@126.com

Beijing Children's Hospital
Qiuyuan Huang
Beijing Children's Hospital
Xiaotong Wang
Beijing Children's Hospital
Hao Chen
Beijing Children's Hospital
Deze Li
Beijing Children's Hospital
Xin Ni
Beijing Children's Hospital

Research Article

Keywords: Lineage enrichment, mutual exclusivity, CRISPR dependency, essential mutation, cancer cell lineage, chromatin accessibility, tissue-agnostic therapy, variant allele frequency

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11462018/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Single-agent BRAF inhibition produces objective responses in roughly 48% of BRAF-mutant melanomas but only about 5% of BRAF-mutant metastatic colorectal cancers—what a mutation *does* depends on the cell lineage that reads it. Yet no computable, threshold-based criteria distinguish disease-defining from progression-associated mutations across lineages. We scored a 31-gene × 16-lineage grid with three fixed-threshold metrics—lineage enrichment, sample-level mutual exclusivity, and mutant-selective CRISPR dependency, with variant-allele-frequency clonality as corroboration—on fully public, checksum-verified data, yielding a 496-cell classification stress-tested across seven axes including the independent 54,331-sample MSK-IMPACT cohort.

Results

The screen returned 15 dual-threshold essential (defining) calls, assembled into a three-class taxonomy that separates mutation frequency from essentiality: 14 events retain Lineage-defining status (ten on the enrichment axis alone, flagged provisional), with TP53 and PIK3CA classed auxiliary. Mutual exclusivity corroborated pathway logic—four designated pairs and two colorectal grid pairs crossed the decision line; both co-occurrence controls significantly positive—mutant-selective dependency confirmed essential oncogenes as addiction nodes. Under the identical rule in MSK-IMPACT, 11 of 15 calls replicate and enrichment ranks correlate at ρ = 0.839. Within MSK-IMPACT, 1,376 of 54,331 samples (2.53%) carry a Lineage-defining event for which an approved targeted agent exists in that lineage (3.57% including VHL@KIRC, druggable indirectly through HIF-2α inhibition). Lineage identity factors were selectively required in their lineages (all six pairs FDR < 0.05; SOX10 and PAX8 reach the dependency line), expressed in all six pairs across 7,303 primary tumors, with lineage-elevated motif accessibility for MITF, NKX2-1, SOX10, and SOX2. Caveats: clonality supported the dispersion but not the position prediction (n = 2,801; IQR 0.187 vs 0.372; 'more subclonal' held for TP53 only); the ATAD2 control was not a clean null; OLIG2 accessibility was null and PAX8 not evaluable; 8 of 15 calls are threshold-sensitive (directionally preserved); and three of nine robustness conclusions hold stably, six partially.

Conclusions

The grid is SNV-centric; within these bounds, the essential/auxiliary distinction becomes a computable, lineage-aware classification—a reporting template for molecular tumor boards and a quantitative stratification factor for basket-trial design. Prospective validation is the necessary next step.

Background

The same mutation can be a drug target in one cancer and nearly irrelevant in another. Single-agent BRAF inhibition produces objective responses in roughly 48% of BRAF-mutant melanomas[1] but only about 5% of BRAF-mutant metastatic colorectal cancers[2]: colorectal epithelium reactivates MAPK signaling through EGFR-mediated feedback[3][4]. Among 295,316 molecularly profiled tumors across 57 cancer types, responses to the same biomarker-directed therapy varied widely by histology —"treatment effects are not necessarily tissue-agnostic"[5]. Regulators concede it: the 2022 tissue-agnostic approval of dabrafenib plus trametinib for BRAF V600E solid tumors explicitly excludes colorectal cancer[6]. What a mutation *does* depends on the cell lineage that reads it. Yet precision oncology lacks a formal criterion for which mutations define a lineage's tumor and which merely accompany progression.

Developmental biology supplies the anchor: BRAF V600E transformed neural crest cells and melanoblasts, yet mature melanocytes tolerated it—even triple co-deletion of p53, RB1, and p16 yielded senescence, not melanoma[7]. This **oncogenic competence**—a cell state's capacity to read a mutation as transforming—makes lineage part of the mutation's functional definition. Three literatures reinforce it: initiation-stage competence (neural-crest identity re-emerges as melanoma initiates[8]; BRAF V600E alone yields nevi without cooperating lesions[9]; Sox10 loss prevents melanoma[10]); lineage-survival oncogenes

(MITF in melanoma[11], NKX2-1 in lung adenocarcinoma[12], SOX2 in lung and esophageal squamous cancers[13], PAX8 in ovarian cancer[14]); and chromatin gating (lineage factors pre-select the enhancers reading oncogenic signals[15]; the pan-cancer accessibility landscape is lineage-organized[16]; cell-of-origin state constrains tumor phenotype[17][18]).

Prior work quantified tissue specificity without a per-event decision rule. Earlier enrichment analyses span 21 tumour types[19] and relate single-cancer enrichment to pan-cancer significance by score[20]; and a threshold-based framework across 146,394 tumors and 265 histologic subtypes concludes tissue specificity is the norm[21] (a preprint at the time of writing) yet quantifies subtype enrichment alone—no exclusivity or dependency axis, no per-event decision classes. To our knowledge, however, no published framework jointly requires lineage enrichment, intra-pathway mutual exclusivity, and mutant-selective dependency within one computable, threshold-based criterion set with falsifiability criteria and independent-cohort validation. Viral oncoproteins provide a natural control: HPV E6 substitutes for TP53 loss, collapsing TP53 mutation frequency from 84% to 3% in head and neck squamous carcinoma[22].

Here we turn that distinction into a decision framework with a concrete deliverable: applied to a 31-gene × 16-lineage grid, three metrics with fixed decision thresholds—the lineage enrichment index (LEI), the mutual exclusivity index (MEI), and the gene dependency difference (GDD), plus variant-allele-frequency (VAF) clonality as corroboration—yield a 496-cell classification of driver events into a three-class taxonomy (Lineage-defining, Lineage-enriched, Non-lineage-restricted), returning 15 dual-threshold essential (defining) calls, 11 of which replicate under the identical rule in the independent 54,331-sample MSK-IMPACT cohort (Spearman ρ = 0.839). In established vocabulary, the **essential (defining)** class formalizes context-specific essentiality—the oncogene- and lineage-addiction events a tumor cannot do without—while the **auxiliary (progression-associated)** class collects the events that accompany progression rather than define an entity. This differs from gene-level context-essentiality frameworks such as PRODE[55], which score essential versus context-essential genes from loss-of-function screening neighborhoods: the present framework operates at the mutation level on (gene/pathway-node, lineage) pairs, applies fixed decision thresholds rather than continuous scores, and is externally replicated in a 54,331-sample independent cohort.

Methods

All analyses use only publicly available, checksum-fingerprinted datasets (md5 or GDC file UUID; Table S5); no imputed, simulated, or estimated values enter any computation (full pipeline, checksums, and inclusion audit: Supplementary Note).

Table S5
Data sources, versions, and integrity fingerprints.

Dataset	Source & version	Size	Use	Access date / md5 or UUID
TCGA PanCancer Atlas somatic mutations (16 studies × 31 driver genes)	cBioPortal public API, PanCancer Atlas 2018 studies (/api/molecular-profiles/{study}_mutations/fetch) [24][25][26]	16921 mutation records across 16 studies	Mutation-frequency matrix for LEI and sample-level 0/1 co-mutation counts for MEI	2026-08-25 / md5 9e989b5da2c699794ab3980b50d5fb98
DepMap Public 24Q4 CRISPR gene-effect scores (Chronos)	DepMap 24Q4 release, figshare article 27993248, file 51064667 [27][28]	1178 models × 17916 genes (428678699 bytes, checksum-verified)	GDD computation (Chronos = gene-fitness effect score)	2026-08-25 / md5 6edf7ade09b9b34199210b559d4745d3
DepMap Public 23Q4 (CRISPR + Model + Hotspot matrix)	DepMap 23Q4 release, figshare files 43346616/43346895/43347528 [27]	1100 models (CRISPR file 400960609 bytes, checksum-verified)	Release-version robustness re-computation of GDD	2026-08-25 / md5 8bfb35c18b7469020dbd5ae4b6274492 (integrity re-archiving 2026-09-27)
TCGA mc3 consensus MAF (allele counts)	NCI GDC data endpoint, mc3.v0.2.8.PUBLIC.maf.gz [29][30]	753 MB; 10295 samples; ~3.6 million variants	t_alt/t_ref allele counts for VAF clonality analysis	2026-08-25 / GDC UUID 1c8cfe5f-e52d-41ba-94da-f15ea1337efc
MSK-IMPACT 50K cohort (mutations + clinical)	cBioPortal public API, study msk_impact_50k_2026 [31][32]	54331 samples; 132618 mutation records	Independent-cohort replication of LEI Lineage-defining calls	2026-08-25
TCGA ABSOLUTE purity/ploidy mastercalls	PanCanAtlas ABSOLUTE mastercalls, GDC data endpoint [33]	10786 samples (9847 with call status = “called”)	Purity-controlled VAF subsets and regression adjustment	2026-08-25 / GDC UUID 4f277128-f793-4354-a13d-30cc7fe9f6b5
TCGA pan-cancer ATAC-seq atlas (Corces et al., GDC ATACseq-AWG)	GDC publication page (ATACseq-AWG, hg38) [16]	562709 peaks × 796 technical replicates raw counts (1.24 GB) + peak set + replicate lookup + donor QC table (404 QC-pass donors)	Reader-motif chromatin accessibility (chromVAR-isomorphic recompute; Table S17)	2026-09-27 / counts md5 f1270f619894cdab381e50afa1b791c6; hg38 md5 7ff134953dcca8c8997453bbb80b6b5e

Primary downloads: 2026-08-25; integrity re-archiving (24Q4 Chronos re-partition; full archival of the 23Q4 CRISPR and 24Q4 damaging-call matrices) completed with md5 verification on 2026-09-27; the ATAC-seq atlas inputs were fetched and integrity-verified (per-chunk md5 repair of the hg38 reference; 16/16 random-locus Ensembl agreement) on 2026-09-27.

TCGA PanCancer Atlas mutations (LEI/MEI inputs). We queried the cBioPortal public API for the 16 PanCancer Atlas 2018 studies[23][24][25], denominators from each study's _sequenced list. The panel comprises 31 driver genes selected from pan-cancer driver censuses[26][27]. The fetch returned 16,921 mutation records. The ovarian cohort is dominated by high-grade serous carcinoma; its TP53 frequency (71.3%) falls below the classic ~ 96% rate[28] through cohort composition and calling convention, not variant-class exclusion (audit: Supplementary Note).

DepMap 24Q4 CRISPR Chronos scores (GDD input). Dependency was measured with DepMap Public 24Q4 Chronos scores[29][30], a model-based gene-fitness estimate that corrects copy-number–driven cutting bias[30]; more negative values indicate stronger dependency. The primary analysis is restricted to cell-line models; mutant status came from the 24Q4 hotspot-mutation matrix and lineages from the Model file's OncotreeLineage field. Sensitivity analyses used the 24Q4 damaging-mutation calls and the complete 23Q4 release re-computed under identical rules.

TCGA mc3 consensus MAF (VAF input). Because the cBioPortal endpoint returns no raw allele counts, we used the mc3.v0.2.8.PUBLIC MAF (753 MB; 10,295 samples) from the NCI GDC[31][32], taking t_alt_count and t_ref_count directly, without purity or ploidy adjustment.

MSK-IMPACT 50K (external replication). We accessed the prospective MSK-IMPACT cohort (msk_impact_50k_2026; 54,331 samples)[33][34] via cBioPortal; the GENIE portal was inaccessible from our compute environment, so we used the designated contingency cohort[35]. All 31 genes are covered in all four IMPACT panel generations, so no gene-specific denominator correction was needed. Cancer-type mapping followed fixed rules (Table S2); LAML is absent from this solid-tumor cohort and left empty with cause noted.

ABSOLUTE purity/ploidy (VAF control). TCGA ABSOLUTE mastercalls (10,786 samples; 9,847 "called")[36] anchor purity-controlled subsets and regression adjustment; a parallel CPE-purity analysis[37] is in the Supplement.

ATAC-seq atlas (chromatin axis). Reader-motif accessibility was measured on the TCGA pan-cancer ATAC-seq atlas[16] with a chromVAR-isomorphic pipeline[38] rebuilt from the raw inputs (562,709 peaks × 796 technical replicates), scanning each 501-bp peak with position-weight matrices for the six reader factors plus CTCF/YY1 flat controls at right-tail $P \le 5 \times 10^{-5}$. Per-motif bias-corrected accessibility deviations (normalised motif accessibility deviations, NMAD) were computed against background peaks matched on GC content × mean accessibility (25 × 25 bins, 25 sampling rounds) and averaged within donor (404 QC-pass donors). Tests contrast in-lineage donors against all other lineages (Welch t and Mann–Whitney U; Cohen's d; BH-FDR across the six-reader family). NMAD is not benchmarked against official chromVAR, so inference uses effect sizes and rank tests, never the score's magnitude.

Metric definitions and decision rules

Each hypothesis maps to one computable index with a fixed decision threshold, stress-tested in dedicated sensitivity sweeps rather than asserted as pre-registered; BH denotes Benjamini–Hochberg FDR control throughout[39].

Lineage Enrichment Index (LEI) — H1. For gene *g* in cancer type *c*, let *f*(*g*,*c*) be the mutation frequency in percent. Then

$$\mathrm{LEI}\,(g,c) = \frac{f\,(g,c)}{\overline{f}_{\,c' \neq c}\,(g) + 1},$$

where the denominator is the mean frequency across the other 15 cancer types plus a 1-percentage-point smoothing constant. We call an event lineage-defining ("essential") when LEI > 5.0 **and** *f*(*g*,*c*) > 15%, as posited in H1. Statistical support comes from a one-sided Fisher exact test against the pooled other 15 cancers, with BH-FDR < 0.05 within each gene. The LEI is a fold-enrichment ratio in the enrichment-ratio family of prior tissue-specificity analyses[19][20].

Mutual Exclusivity Index (MEI) — H2. For a gene pair (A, B) within a cancer type, we build the 2 × 2 sample-level co-occurrence table; all sequenced samples, including double-wild-type, serve as denominator, with no frequency-derived approximation:

$$\mathrm{MEI} = \log_2\left(\frac{n_{\mathrm{both}} \cdot n_{\mathrm{neither}}}{n_{A\,\mathrm{only}} \cdot n_{B\,\mathrm{only}}}\right).$$

The MEI is the log2 odds ratio of the co-occurrence table, the standard effect-size companion to DISCOVER-style independence tests[40]. The tested grid comprises 12 gene pairs across five cancer types, each eligible only when both genes are mutated in ≥ 5% of sequenced samples. MEI < − 1.5 with Fisher-exact BH-FDR significance is called significant exclusivity, as posited in H2. Two designated co-occurrence controls (PIK3CA–PTEN in UCEC, APC–KRAS in COADREAD; expected MEI > 0) validate the exclusivity/co-occurrence separation. Because cross-pathway pairs can appear mutually exclusive for compositional reasons, the functional-redundancy interpretation is restricted to same-pathway pairs, with cross-pathway pairs recomputed within expression subtypes (BLCA)[41] or MSI strata (COADREAD) as stratified controls.

Gene Dependency Difference (GDD) — H3. Within each OncotreeLineage, let *m* and *w* index mutant and wild-type cell lines for gene *g*. Then

$$\mathrm{GDD}\,(g) = \overline{\mathrm{Chronos}}_m - \overline{\mathrm{Chronos}}_w,$$

so that more negative values indicate stronger in vitro addiction of mutant lines to *g*[42]. We call GDD < − 0.3 an in vitro functional dependency and GDD < − 0.5 a strong in vitro dependency, as posited in H3; because CRISPR screens assay cell-autonomous fitness in culture, we do not equate this signal with in vivo essentiality. The tested grid comprises 19 gene–lineage cells: the 15 LEI dual-threshold events mapped to their DepMap lineages where computable, plus designated oncogene–lineage control cells. Formal testing (two-sided Mann–Whitney U with BH-FDR) requires ≥ 4 cell lines per group. Groups of 2–3 lines yield descriptive GDD values flagged † with no test. Groups below 2 lines are left empty with the cause stated. A two-line group is degenerate under bootstrap resampling and treated as structurally empty, neither supporting nor refuting dependency. The tumor-suppressor APC in Bowel-lineage lines serves as the designated negative control: an essential tumor suppressor should show no in vitro knockout dependency.

Hotspot-matrix blind spots (disclosed, not repaired). Of the 19 grid cells, 10 (52.6%) cannot be formally tested under the hotspot-only definition: five have fewer than two mutant lines (EGFR@Lung among them) and five have only two to three. The sparsity is structural, concentrating in indel-dominated genes; EGFR is distinct, because the hotspot matrix records its classic lung-adenocarcinoma exon-19 deletions as non-hotspot. A post-hoc sensitivity analysis on protein-altering calls from the full 24Q4 mutation file (six mutant versus 68 wild-type lung lines) shifts in the predicted dependency direction (Δmedian = − 0.70, Cohen's d = − 0.97; Mann–Whitney P = 0.029, Welch P = 0.135), underpowered with significance split between tests. Complete repair requires OncoKB oncogenicity annotation[43]; the sensitivity analysis stands as the interim bound.

Variant allele frequency (VAF) and clonality — H4. For each mc3 mutation passing a depth filter of *t*_alt + *t*_ref ≥ 20,

$$\mathrm{VAF} = \frac{t_{\mathrm{alt}}}{t_{\mathrm{alt}} + t_{\mathrm{ref}}}.$$

H4 predicts narrow VAF dispersion for essential events and broader, more subclonal distributions for auxiliary events, with no positional prediction. The essential group comprises the five LEI-defined Lineage-defining events; the auxiliary group comprises TP53 and PIK3CA mutations from the same five cancer types. Dispersion uses the median-centered Levene (Brown–Forsythe) test and location two-sided Mann–Whitney U, repeated within ABSOLUTE purity-controlled subsets (≥ 0.6, ≥ 0.8) and an OLS model (VAF ~ group + purity; n = 2,716). A symmetric GISTIC-stratified, purity-normalized (VAF/purity) correction is applied to both pools (Supplementary Note).

Resampling machinery. For formally tested GDD cells, 95% confidence intervals come from 1,000-fold bootstrap resampling of cell lines (seed 20260825). Because 57.4% of VAF-pool samples contribute more than one mutation, primary VAF inference uses patient-level cluster bootstrap (10,000 resamples, seed 20260927) and cluster permutation ($P \le 1 \times 10^{-4}$ resolution floor); cross-cohort LEI correlation uses gene-block permutation (2,000 nulls).

Cells that cannot be computed robustly are left empty with causes stated, and borderline results are reported as computed, without post-hoc threshold tuning.

Covariates. We assessed five patient-level covariates (tobacco, ultraviolet, HPV, HBV/HCV etiology, ancestry) and handled them outside the primary model, scoping conclusions to the populations assayed (per-covariate rationale: Supplementary Note: Extended discussion, S5).

Conceptual framework and hypotheses

An **essential (defining) mutation** is an event a lineage must acquire, or have functionally substituted, for the tumor to exist as that entity; an **auxiliary (progression-associated) mutation** modulates grade, stage, or fitness without being required. Malignancies must solve five universal bottleneck problems, but the molecular solutions are lineage-restricted, and viral oncoproteins can substitute functionally at the same node (HPV E6 degrades p53)[44]. The axiom: necessity attaches to a (pathway node, lineage) pair, with genes as interchangeable solutions—BRAF, NRAS, and NF1 serve the MAPK node in melanoma, their deep mutual exclusivity (MEI = − 4.32) the population-genetic signature of node-level necessity. Whether a mutation transforms (oncogenic competence) depends on the DNA mutation, lineage-specific transcription factors, and the developmentally set chromatin state[7][8][9]. Our units are TCGA organ cohorts—imperfect lineage surrogates validated by cell-of-origin dominance[23]; we flag known misregistrations (COADREAD, SKCM, the squamous pair, the glioma boundary) where they could matter. Table 1 assembles six representative lineages into the essential/auxiliary matrix.

Table 1
Representative lineages and their essential (defining) versus auxiliary (progression-associated) mutations

Lineage	Essential (defining)	Auxiliary (progression-associated)	Grid verdict (this study)
Squamous epithelium (skin, head and neck, esophagus)	p53 node (TP53 mutation or HPV E6); NOTCH1 loss	CDKN2A loss, PIK3CA, FAT1; TP53 3% (HPV+) versus 84% (HPV−) [44]	Discordant stress case: TP53@LUSC LEI = 2.23 and NOTCH1@LUSC LEI = 1.78 fall below every line—events shared across all squamous organs, or all lineages, cannot register on a lineage ratio (Discussion)
Colorectal adenocarcinoma (crypt stem cell)	WNT node (APC|CTNNB1) + MAPK node (KRAS)	TP53, SMAD4, PIK3CA (late)	Concordant: APC@COADREAD and KRAS@PAAD Lineage-defining
Clear cell renal cell carcinoma (proximal tubule)	VHL–HIF node (VHL inactivation with 3p loss, ~ 90%)[45]	PBRM1, SETD2, BAP1	Concordant: VHL@KIRC Lineage-defining
Germinal-center and post-GC B cell	GC: IG-enhancer hijack node (IGH::BCL2|MYC::IG|BCL6); post-GC: MYD88 node (L265P, 91%)[46]	Secondary events	Outside the computable grid (MYD88 not in the 31-gene panel; lymphoid lineages absent from the TCGA panel)
Melanocyte (neural crest)	MAPK node (BRAF|NRAS|NF1; requires SOX10/MITF × ATAD2 competence)[7][10]	TERT promoter mutation, CDKN2A loss	Concordant: BRAF@SKCM and NRAS@SKCM Lineage-defining
Mesenchymal (sarcoma)	Fusion-defined transcriptional node (EWSR1::FLI1, SS18::SSX, PAX3/PAX7::FOXO1); or p53/cell-cycle node (MDM2 + CDK4)	TP53, STAG2	Outside the computable grid (fusion-defined; SNV-centric design)
(extended all-16-lineage version in Table S18)			

Essential entries are stated at the (pathway node, lineage) level with interchangeable gene-level solutions; auxiliary entries remain gene-level. The right-hand column confronts each concept-level entry with this study's computable verdict. We retain the squamous row deliberately as a falsifiable stress case, rather than tuning the panel or thresholds to it.

The framework separates two dimensions gene-centric accounts conflate (Fig. 1): **lineage-restriction** and **necessity**. Three metrics operationalize them with fixed decision lines (Methods): the lineage enrichment index (LEI), fold-enrichment within a lineage against pan-cancer background; the mutual exclusivity index (MEI), sample-level exclusivity of same-pathway events, because selection eliminates functional redundancy[40]; and the gene dependency difference (GDD), the mutant-line versus wild-type-line Chronos score difference in genome-scale CRISPR screens[29][30]. An event is **Lineage-defining** when it passes LEI > 5 and within-lineage frequency > 15% plus, for an oncogene with an informative and computable knockout, dependency at the strong line (GDD < − 0.5); **Lineage-enriched** when it passes LEI > 3 but fails one Lineage-defining criterion; **Non-lineage-restricted** when frequent, even functionally important, but not lineage-defining by our criteria—not a clinical judgment. The classes describe statistical lineage economics and license no treatment decision by themselves; “essential” throughout qualifies mutations, not genes, distinct from the common-essential-gene usage of the CRISPR-screen literature.

H1 (enrichment): essential events are lineage-enriched—LEI > 5 together with within-lineage frequency > 15%. **H2 (exclusivity)**: essential events acting within the same pathway are mutually exclusive at the sample level (MEI < − 1.5); auxiliary events may co-occur freely. **H3 (dependency)**: cells carrying an essential mutation are selectively dependent on the mutated gene or its immediate node (GDD < − 0.3; strong dependency, GDD < − 0.5). **H4 (clonality)**: essential events show narrow VAF dispersion under purity and ploidy control, auxiliary events broader, subclonal-tolerant distributions; no positional prediction is made[36]. Each hypothesis carries an explicit falsifiability clause (Supplementary Note: Metric derivations and threshold rationale, S10).

Results

We tested the four framework hypotheses: lineage enrichment (LEI, H1), intra-pathway mutual exclusivity (MEI, H2), mutant-selective CRISPR dependency (GDD, H3), and variant-allele-frequency clonality (VAF, H4). All numbers derive from version-pinned datasets (TCGA PanCancer Atlas mutations[23][31], DepMap Public 24Q4 Chronos scores[29][30], TCGA mc3 consensus MAF[31]). Non-computable cells are left empty, with causes stated (per-cell narratives: Supplementary Note; full grid in Table S1).

LEI: lineage enrichment identifies 15 dual-threshold essential events

Across the 31-gene × 16-cancer-type matrix (Fig. 2), 15 events crossed both thresholds (LEI > 5, within-cancer frequency > 15%), each at FDR ≤ 1.2×10^{-20}. The strongest recapitulate canonical lineage drivers (VHL@KIRC, IDH1@LGG, APC@COADREAD, KRAS@PAAD, BRAF@SKCM; LEI 7.12–31.91); the remaining ten (LEI 5.24–9.28) match the lineage literature[45][47]. Two negative controls behaved as predicted (TP53 in ovarian cancer and KRAS@COADREAD; Table 2). Three near-line events are disclosed rather than forced: FGFR3@BLCA (LEI = 6.37; frequency 14.39%, just short of 15%), STK11@LUAD (LEI = 7.49; 13.25%, 75/566), and IDH2@LAML (LEI = 5.18; 10.50%, 21/200) are all classed Lineage-enriched. The first two cross both thresholds in MSK-IMPACT (robustness below); IDH2@LAML lacks an MSK-IMPACT counterpart (Table S2) and remains an unresolved boundary case. Sample-level aggregation of the MAPK node (BRAF, NRAS, NF1, KRAS) confirms the interchangeable-solution axiom—occupancy reaches 87.3% in SKCM (90.7% including KIT; Fisher $P = 1.1 \times 10^{-164}$), 67.3% in THCA, 67.0% in PAAD, and 57.9% in COADREAD—but aggregation compresses the enrichment ratio, so the LEI > 5 line is gene-level only (Table S11). Integrated with the dependency axis, the full 496-cell grid yields 14 Lineage-defining, 14 Lineage-enriched, and 46 Non-lineage-restricted events, with 422 cells below the screening line (Table S1); suspending the single descriptive-grade EGFR@GBM demotion would yield 15 Lineage-defining calls. We report the classification both ways where the count is load-bearing. Ten of the 14 Lineage-defining calls rest on LEI and frequency alone—tumor suppressors and chromatin regulators by design, plus structurally empty or degenerate GDD cells—and are flagged provisional (Table 2, ‡). Table 2 is a worked example of 15 representative events spanning all three tiers, not the 15 LEI calls themselves. Row-level caveats: FGFR3@BLCA, failing the 15% cut at 14.39% but crossing both thresholds in MSK-IMPACT (22.6%, LEI = 9.09)[48], is recorded as Lineage-enriched (borderline); TP53@OV is a pan-lineage defining event in high-grade serous ovarian carcinoma[28], its LEI failure a design boundary; PIK3CA@BRCA is an approved OncoKB Level 1 target[43]—Non-lineage-restricted denotes “not lineage-defining by the present criteria”, not “clinically unimportant”.

MEI: intra-pathway mutual exclusivity at sample level

Four designated pairs crossed the MEI < − 1.5 decision line with FDR significance (Fig. 3): BRAF–NRAS in SKCM, CTNNB1–AXIN1 in LIHC, and FGFR3–RB1 and FGFR3–TP53 in BLCA; two COADREAD grid pairs (KRAS–BRAF, APC–BRAF) also qualified. Both co-occurrence controls were significantly positive (PIK3CA–PTEN in UCEC, + 1.11; APC–KRAS in COADREAD, + 1.45), so the index separates the two regimes. The PIK3CA–PTEN co-occurrence is entirely attributable to hypermutated UCEC samples—within mutation-burden tertiles the pair inverts toward exclusivity (Cochran–Mantel–Haenszel log2 odds ratio = − 1.27, P = 0.0025)—and is reported as burden-driven, not biological, co-occurrence. Three of the six qualifying pairs carry 95% confidence intervals spanning the decision line (CTNNB1–AXIN1 [− 5.59, − 0.79], resting on a single co-mutant sample; FGFR3–RB1 [− 3.94, − 0.74]; FGFR3–TP53 [− 2.97, − 1.12]) and are read as consistent with exclusivity, effect size not resolved; the other three exclude the line by interval. Subtype- and MSI-stratified retests[41] leave exclusivity intact, with partial subtype contribution only (Table S22 row 8).

GDD: mutant-selective CRISPR dependency

Seven cells passed formal testing at the strong-dependency line (Fig. 4), spanning NRAS@Skin (− 1.44) to BRAF@Bowel (− 0.63); the tumor-suppressor control APC@Bowel is null (GDD = − 0.12, FDR = 0.20), as theory requires. Within the 19-cell dependency grid, descriptive cells (2–3 lines in one group, †) are concordant or null, five cells are empty (mutant n < 2), and this structural coverage loss—10 of the 19 cells, 52.6%, untestable under the hotspot-only definition—concentrates in indel-dominated genes and is disclosed rather than imputed. The EGFR@Lung gap is a hotspot-matrix annotation blind spot, not a biological null: a post-hoc sensitivity analysis on protein-altering calls shifts dependency in the predicted direction (Δmedian = − 0.70, Cohen's d = − 0.97; Mann–Whitney P = 0.029, Welch P = 0.135), directionally concordant but underpowered. Applied mechanically, the dependency decision rule changes exactly one of the 496 classifications: EGFR@GBM (+ 0.0046†, three mutant lines) is demoted to Lineage-enriched on its null-direction GDD—a single-cell reclassification. FGFR3@BLCA (+ 0.03†), the only other null-direction case, already fails the 15% frequency floor. Structurally empty GDD cells neither support nor refute; the call rests on LEI and frequency, the gap disclosed (IDH1@LGG, independently anchored by diagnostic classification[49]; CTNNB1@UCEC, outside the 19-cell grid; CTNNB1@LIHC, a two-line estimate degenerate under bootstrap resampling).

Reader dependency, expression, and chromatin accessibility measure the lineage machinery directly

Tested directly on DepMap 24Q4 (1,178 screened cell lines), all six lineage identity transcription-factor pairs are selectively required in their own lineages at FDR < 0.05 (SOX10 and MITF in Skin, PAX8 in Ovary/Fallopian Tube, NKX2-1 in Lung, SOX2 and OLIG2 in CNS/Brain; Table S12), but only SOX10 (Δmedian Chronos = − 1.12, Cohen's d = − 3.61) and PAX8 (Δmedian = − 0.53) reach the framework's − 0.30 dependency line; the other four are detectable but below it (median shifts − 0.21 to − 0.02). The seventh pair, ATAD2, is not a clean null: its shift is negligible and CI-crossing (Δmedian = + 0.02) yet nominally significant (FDR = 4.2×10^{-2}), reported as a partial control failure. Within Skin, BRAF mutation further deepens SOX10 dependency (Δ = −0.93, FDR = 1.3×10^{-3}); a formal cross-lineage interaction test is left as follow-up. Expression provides a positive-control check of the lineage assignments, expected by construction: all six identity pairs show lineage-restricted expression in primary tumors (7,303 samples; Cohen's d 1.56–6.87; Table S15), while ATAD2 shows none (d = − 0.03, P = 0.19). In the TCGA pan-cancer ATAC-seq atlas[16][38], motif accessibility is lineage-elevated for MITF in Skin (d = 2.36), NKX2-1 in Lung (d = 1.50), SOX10 in Skin (d = 1.45), and SOX2 in CNS tumors (d = 0.59); OLIG2 in CNS is null (d = 0.10, FDR = 0.70), reported as computed, and PAX8 in Ovary is not evaluable—the ATAC panel contains no ovarian cancer. Tumor-suppressor effector coupling is concordant where powered (KEAP1-mutant lung lines depend on NFE2L2, Δ = −0.65, FDR = 6.2×10^{-4}; PTEN-mutant lines on AKT1 pan-lineage, FDR = 6.2×10^{-4}, but the within-Uterus test is null; Table S13).

Table 2
Integrated three-metric classification of representative driver events.

Gene	Lineage	LEI	Freq (%)	GDD	Class	Targeted agent (this lineage)
VHL	KIRC (Kidney)	31.91	41.29	— (mut n = 1)	Lineage-defining‡	Indirect (downstream HIF-2α inhibitor approved)
IDH1	LGG (CNS/Brain)	22.58	76.85	— (mut n = 1)	Lineage-defining‡	Approved (mutant-IDH inhibitor in glioma)
APC	COADREAD (Bowel)	12.42	72.47	−0.12 (TSG, expected null)	Lineage-defining‡	None
PTEN	UCEC	9.28	65.18	not tested	Lineage-defining‡	None
NRAS	SKCM (Skin)	8.40	28.41	−1.44	Lineage-defining	None approved (MEK inhibition investigational)
BRAF	THCA (Thyroid)	8.30	58.57	−1.18†	Lineage-defining	Approved (BRAF + MEK; tissue-agnostic[6])
KRAS	PAAD (Pancreas)	7.46	65.36	−1.36†	Lineage-defining	Investigational (allele-specific KRAS inhibitors)
BRAF	SKCM (Skin)	7.12	52.95	−1.01	Lineage-defining	Approved (BRAF + MEK)
KEAP1	LUAD (Lung)	6.06	18.02	not tested	Lineage-defining‡	None
CTNNB1	LIHC (Liver)	5.39	25.96	−1.26† (degenerate bootstrap)	Lineage-defining‡	None
FGFR3	BLCA (Bladder)	6.37	14.39	+ 0.03†	Lineage-enriched (borderline)	Approved (erdafitinib)
EGFR	GBM (CNS/Brain)	5.24	23.68	+ 0.0046†	Lineage-enriched (GDD uninformative)	None approved in this lineage
KRAS	COADREAD (Bowel)	3.92	40.82	−1.09	Lineage-enriched	Approved (KRAS G12C + EGFR blockade)
PIK3CA	BRCA (Breast)	2.72	32.55	−0.80	Non-lineage-restricted (auxiliary)	Approved (alpelisib; OncoKB Level 1)
TP53	OV	1.86	71.32	not tested	Non-lineage-restricted (auxiliary)	None

LEI and frequency from the TCGA grid (Fig. 2); GDD from DepMap 24Q4 (Fig. 4). MEI is evaluated at the (pathway pair, cancer type) level (Fig. 3) and does not enter the per-cell rule. Targeted-agent status follows current OncoKB/ESCAT actionability grading and regulatory approvals[43][50] and is orthogonal to the lineage-definition axis. Eleven of the 15 dual-threshold calls appear here. The remaining four and the full 496-cell classification are in Table S1.

† descriptive GDD (2–3 lines in one group; no formal test); "—" empty cells (mutant n < 2). ‡ classified on LEI and frequency alone—tumor suppressors by design, IDH1@LGG because its GDD cell is structurally empty, and CTNNB1@LIHC because its two-line GDD is degenerate under resampling and therefore structurally empty; single-metric calls are provisional pending orthogonal evidence. Tumor suppressors are classified on LEI and frequency alone (validated by the APC@Bowel null, GDD = −0.12, FDR = 0.20). Non-lineage-restricted denotes "not lineage-defining by the present criteria", not "clinically unimportant": a Non-lineage-restricted cell can host an approved target (PIK3CA@BRCA), and a Lineage-defining cell can lack one (KEAP1@LUAD) (caveats in the Supplementary Note: Per-cell grid narratives).

VAF clonality corroboration

H4 predicts narrow VAF dispersion for essential events. This corroboration is partly circular: the five Lineage-defining events are the top of the LEI screen, and the comparator pool is the two genes the framework classes as auxiliary—the same panel-selection circularity disclosed for the clinical-outcome comparison (Table S7). Across 2,801 TCGA mc3 mutations[31], the five Lineage-defining events are markedly narrower than the Non-lineage-restricted TP53/PIK3CA pool (IQR 0.187 versus 0.372; Fig. 6A). Because 57.4% of analysis-pool samples contribute more than one mutation, cluster-corrected inference is primary (cluster-bootstrap ΔIQR 95% CI [0.153, 0.230]; cluster-permutation $P \le 1 \times 10^{-4}$; Table S22 row 5; Fig. 6B) and mutation-level values are anti-conservative. The dispersion prediction holds; the position prediction does not—the Lineage-defining pool's median VAF is lower than the auxiliary pool's (0.351 versus 0.414), and subclonal fractions are essentially equal (24.6% versus 24.3% below VAF 0.25)—so H4 is supported in its dispersion dimension only. The one per-event violation, KRAS@PAAD (median VAF 0.195, 64.8% below 0.25), resolves as a purity and copy-number artifact: purity ≥ 0.6 raises the median to 0.305 and halves the sub-0.25 fraction (65% → 31%), and GISTIC-stratified VAF/purity rises from 0.41 (diploid) to 0.84 (amplified), consistent with clonality pending cancer-cell-fraction modeling. Symmetric correction of the comparator pools shows the TP53 high-VAF tail is a clonal-plus-LOH group (44%, 368/836; Fig. 6C); excluding it, the pool remains significantly wider, whereas the PIK3CA pool normalizes to near-Lineage-defining width (IQR 0.213 versus 0.273)—the "more subclonal" reading holds for the TP53 pool only.

Robustness across cohorts and parameters

Robustness summary statistics are provided in Table S22. Seven axes stress-tested every directional conclusion (Fig. 7; Figures S1–S3); three of the nine conclusions hold stably, and six are partially stable or refined, each resolving in the direction that strengthens or sharpens the original interpretation. In the prospective MSK-IMPACT 50K cohort (54,331 samples)[33][34]—an external-cohort comparison under differing ascertainment, not an exact replication—LEI across the 360 shared gene–lineage cells correlates strongly with TCGA (Spearman $\rho = 0.839$; gene-block permutation $P \le 5 \times 10^{-4}$ over 2,000 block-shuffled nulls, maximum null $\rho = 0.276$, effective $n \approx 24$ genes; bootstrap 95% CI [0.747, 0.896]; the cell-level asymptotic $P = 1.6 \times 10^{-96}$ treats cells as independent units—the wrong inferential unit—and is reported only for completeness). Restricted checks among non-empty cells give $\rho = 0.892$ ($n = 70$, LEI > 1 in both cohorts; further restrictions in Fig. 7); the restriction choices were made after seeing the full-grid correlation and are acknowledged as post-hoc. Under the identical dual-threshold rule applied within each cohort, 11 of the 15 calls replicate, and all 15 preserve enrichment direction and significance (MSK-side FDR < 0.05; Table S1); shared-grid overlap gives Jaccard = 0.611 (11/18). The four non-replicating calls are decision-line borderline cases: ARID1A@UCEC (43.9% → 42.4%, LEI = 3.73), EGFR@GBM (23.7% → 21.8%, LEI = 3.84), CTNNB1@UCEC (25.7% → 18.7%, LEI = 3.19), and KEAP1@LUAD (18.0% → 13.8%, LEI = 4.41, now failing the 15% frequency line). Conversely, FGFR3@BLCA crosses both thresholds in MSK (22.6%, LEI = 9.09)[48], as does STK11@LUAD (15.16%, LEI = 6.6); both remain Lineage-enriched—the primary calls were conservative. On DepMap 23Q4 (release-version stability, not independent replication), the GDD grid is near-identical (Pearson $r = 0.989$), and purity-controlled VAF dispersion remains significant at every ABSOLUTE tier[36] (Table S22 rows 4–5). On an LEI-independent grid, strong GDD does not by itself discriminate the classes (odds ratio 1.39, $P = 1.0$); we therefore present the GDD axis as corroborating, not confirming, evidence (Supplementary Note: Metric derivations and threshold rationale, S13).

Clinical actionability of the grid

Six of the 15 representative events classified in Table 2 carry an approved targeted agent in their lineage; three are Lineage-defining cells with a directly approved agent (IDH1@LGG, mutant-IDH inhibitor in glioma; BRAF@THCA and BRAF@SKCM, BRAF + MEK blockade), and VHL@KIRC is druggable indirectly through the approved downstream HIF-2α inhibitor. Quantified in the independent 54,331-sample MSK-IMPACT cohort, 1,376 samples (2.53%) carry such a Lineage-defining, in-lineage-actionable event—598 IDH1@LGG, 407 BRAF@SKCM, and 371 BRAF@THCA, non-overlapping at the sample level—rising to 1,942 samples (3.57%) when the indirectly druggable VHL@KIRC is included. Context controls bound the interpretation: TP53-mutant (45.2%) and PIK3CA-mutant (14.1%) samples are an order of magnitude more frequent yet are classed auxiliary by design, and most Lineage-defining cells (APC, PTEN, KEAP1, CTNNB1, SMAD4) have no approved agent at all. The 2.53% figure is therefore a lower bound on the classification's clinical footprint, computed in a tertiary-care, metastasis-enriched cohort, and is not a population incidence.

Pediatric out-of-sample test

The grid is trained on adult tumor cohorts; pediatric malignancies provide an out-of-sample test of whether Lineage-defining status tracks shared developmental lineage or universal necessity. Eight of the 15 calls have a mappable pediatric counterpart—Wilms tumor whole-exome MAFs from the NCI TARGET program[58] and curated variant prevalence from the St. Jude PeCan portal[57] (Table S23, Figure S8); the remaining seven calls (in UCEC, COADREAD, PAAD, and LUAD) have no pediatric counterpart and are untestable. Four adult Lineage-defining events are significantly depleted in their pediatric counterparts: VHL (41.3% in TCGA KIRC) is absent from Wilms tumor (0/38 in TARGET, FDR = 1.5×10^{-8}; 0/245 in PeCan, FDR = 2.7×10^{-41}), as is PBRM1 (36.1% → 0/38 and 0/245; FDR < 10^{-6} in both cohorts); IDH1 (76.9% in adult LGG) is absent from pediatric-type diffuse low-grade glioma (0/18, FDR = 3.4×10^{-11}); and EGFR (23.7% in adult GBM) reaches at most 5.0% (≤ 9/179) in pediatric-type diffuse high-grade glioma (FDR = 1.5×10^{-8}). Three comparisons are directionally concordant—CTNNB1 in hepatoblastoma (30.8%, 4/13, versus 26.0% in adult LIHC), BRAF in pediatric papillary thyroid carcinoma (52.6%, 10/19), and BRAF in pediatric melanoma (42.9%, 9/21)—while NRAS in pediatric melanoma runs lower (9.5%, 2/21) without reaching significance (P = 0.078; small cohorts, limited power, so non-significance is not equivalence). The adult grid therefore does not transfer wholesale to pediatric disease: the depletions align with the distinct developmental origins of the pediatric entities and demonstrate that the grid measures lineage definition rather than universal necessity (Discussion).

Exploratory survival association

As an exploratory outcome check, we stratified TCGA patients within each informative lineage by framework class—defining-event carriers, the auxiliary TP53/PIK3CA pool, and wild-type—and tested overall and progression-free survival using TCGA-CDR endpoints[56] (three-group log-rank; Cox models adjusted for age and, where ≥ 95% complete, AJCC stage; BH-FDR across the 30 tests of each family; Table S24, Figure S9). No consistent prognostic gradient separates the classes. The only associations surviving FDR control in both the log-rank and adjusted Cox families are IDH1@LGG (overall survival HR = 0.32; log-rank q = 2.7×10^{-14}, Cox q = 5.8×10^{-7}; progression-free survival also significant) and ARID1A@UCEC (overall survival HR = 0.33; log-rank q = 6.8×10^{-5}, Cox q = 0.006), and both couple to known molecular-subtype biology (IDH-mutant glioma; ARID1A-mutant endometrial cancer) rather than to the classification itself. KRAS@PAAD illustrates the fragility of the remaining signal: significant by log-rank (q = 0.004) but not after Cox adjustment (q = 0.069). The auxiliary pool is uniformly null (0/30 Cox tests pass FDR). We therefore make no prognostic claim for the classification: the gradient is sparse and, where present, confounded by known molecular subtypes.

Discussion

Essential (defining) mutations are events a lineage must acquire or functionally substitute for the tumor to exist as that entity; **auxiliary (progression-associated)** mutations modulate grade, stage, or fitness without being required. Essential status attaches to a (gene/pathway-node, lineage) pair, not a gene. Across the 16-lineage panel, H1–H3 were supported and H4 was partially supported, with a pool-specific mechanistic refinement. The LEI screen returned 15 dual-threshold essential calls; 14 retain Lineage-defining status in the final three-metric grid (Table S1). The result is a computable essential (defining) versus auxiliary (progression-associated) distinction, expressed as a three-class taxonomy—Lineage-defining, Lineage-enriched, Non-lineage-restricted—that separates frequency from essentiality. The negative controls make the point: TP53 in ovarian cancer and KRAS in colorectal cancer are frequent yet Non-lineage-restricted (Table S1). High frequency without lineage restriction is not essentiality. The 15 calls collapse onto fewer than ten pathway bottlenecks, as H1 and the bottleneck model require (Fig. 2). H2 exclusivity concentrates in these pathways, and H3 confirms essential oncogenes as addiction nodes. The framework's explanatory claim is now measured, not inferred. An essential mutation is read by lineage identity-factor machinery in a permissive chromatin landscape, tested on independent data layers (Fig. 5). The two layers use different assays, cohorts, and units, so their concordance cannot be a single-platform artifact; post-transcriptional or cell-state mechanisms are not excluded. We stress-tested borderline outcomes on seven robustness axes spanning nine directional conclusions. Only one axis, the MSK-IMPACT cohort, adds new data. Three conclusions hold stably; six are partially stable or refined (Table S22).

The framework builds on established concepts: oncogene addiction[42], lineage dependency created by differentiation wiring[11], and formal mutual-exclusivity testing[40]. Viral oncoproteins independently support the same nodes. HPV E6/E7 substitute for TP53 and RB1 loss, and TP53 mutation collapses in HPV-positive HNSCC[44], a contrast our own HPV-stratified 31-gene analysis reproduces (83.5% versus 11.6%; Table S3). HBV and HCV substitute at shallower depths, and at every depth the corresponding somatic route is depleted[47][51][52]. HPV, HBV, and HCV thus trace a depth-of-substitution hierarchy: exact node occupancy, locus-specific cis-activation, and population-level mutation supply. Our increment is the computable dual-thresholding plus the lineage × node matrix, not the concepts themselves. Catalog cross-checking makes the increment precise: all 15 events involve catalogued driver genes in previously annotated (gene, cancer-type) pairs (15/15 in both the Bailey et al. 2018 driver consensus[60] and the IntOGen compendium[59]; Table S25); the contribution of this framework is the essential/auxiliary tiering at (gene, lineage) resolution, which no existing catalog provides.

A single oncogene's essentiality varies across lineages in a graded way. KRAS is Lineage-defining in pancreas, Lineage-enriched in bowel, and strongly depended upon in lung (Table S1). The KRAS clonality signal was re-adjudicated as pre-analytic: stroma-rich pancreatic tumors are systematically low-purity, and focal amplification distorts allele fractions. Purity- and copy-number-controlled re-adjudication fits a clonal heterozygous mutation, pending cancer-cell-fraction-level modeling[36]. The TP53 paradox resolves the same way. TP53 mutation is clonal, frequent, and positively selected even in normal squamous epithelium. Yet it scores Non-lineage-restricted, because its niche fitness at the clonal-ecology level diverges from entity definition at the established-tumor level[26][53]. Essentiality attaches to the (node, lineage) pair, not to the gene. This case yields two boundary classes the framework must own. Pan-lineage defining events fail LEI by construction—a false-negative direction. Progression-associated enrichments, such as KEAP1 in lung and SMAD4 in pancreas, enter the Lineage-defining list as a known false-positive direction[54]. A further boundary is concept–computation discordance: events shared across all squamous organs cannot register on a lineage ratio. We report the discordance rather than tune the panel or thresholds to the concept table. H4 is thus partially supported, with pool-specific mechanisms rather than a uniform auxiliary signature.

LEI, MEI, and GDD quantify lineage dependence. They do not, by themselves, predict single-agent efficacy. Across the KRAS cells, the dependency gradient runs opposite to the single-agent response gradient (Table S7). A response rate additionally reflects drug exposure, adaptive resistance, and receptor-feedback topology. Dependency strength is therefore necessary but not sufficient for single-agent efficacy. Regulators have already encoded lineage effects into labels: the tissue-agnostic dabrafenib–trametinib approval excludes colorectal cancer[6]. The metrics instead generate trial-stratification hypotheses. Where a Lineage-defining event pairs strong dependency with clonal VAF and lacks dominant upstream receptor feedback, cross-organ single-agent activity is a reasonable trial hypothesis. Where such feedback dominates, front-line combination blockade is more plausible. The trio complements rather than replaces OncoKB and ESCAT actionability grading[43][50], pricing the histology axis before the data arrive. The actionable fraction is now quantified rather than asserted: 1,376 of 54,331 MSK-IMPACT samples (2.53%; 3.57% including the indirectly druggable VHL@KIRC) carry a Lineage-defining event with an approved agent in that lineage—a deliberately narrow, approved-therapy-anchored denominator that excludes investigational and out-of-lineage uses.

The pediatric out-of-sample test disciplines the concept itself. The depletions are not mere absences of evidence: Wilms tumor arises from embryonal nephrogenic rests rather than the proximal-tubule epithelium that gives rise to clear-cell renal cell carcinoma, so the complete absence of VHL and PBRM1 mutations in Wilms tumor is simultaneously a negative transfer result and a biological fact—the renal lineage definition the grid learned is specific to the adult cell of origin. IDH1's absence from pediatric-type diffuse low-grade glioma reads the same way: the adult IDH1-defined entity has no pediatric counterpart in these data. Where developmental lineage is genuinely shared, the grid does transfer—CTNNB1 in hepatoblastoma, BRAF in pediatric papillary thyroid carcinoma and melanoma. Lineage-defining status is thus a statement about a (gene, lineage) pair in its developmental context, not a claim of universal necessity, which is what the framework's formal definition asserts.

Several limitations bound these claims. First, the grid is SNV-centric. Fusion-, copy-number-, and promoter-driven tumors are under-represented, and pediatric malignancies sit outside it by construction. Low TERT, CDKN2A, and RB1 values are coverage statements, not biological verdicts. Second, the hotspot matrix leaves blind spots that we disclose rather than repair. The EGFR@Lung cell is empty because canonical exon-19-deletion lines score as non-hotspot; a post-hoc sensitivity analysis shifts in the predicted direction but is underpowered. DepMap also over-represents aggressive, high-proliferation tumors, and the

KRAS@Pancreas comparison rests on few wild-type lines and is descriptive. Third, the ATAD2 intended negative control does not behave as a clean null. We report this partial control failure, not designed validation of specificity. Fourth, the APC@Bowel control returning the expected null is the design working, not a defect: knockout cannot validate essentiality by loss. Fifth, CTNNB1@LIHC rests on LEI alone; its two-line GDD is degenerate under resampling and structurally empty, and other two-to-three-line cells are reported only descriptively. Sixth, all decision lines are author-chosen heuristics, and the sensitivity sweeps and restricted external-cohort checks are acknowledged as post-hoc. Directions hold, but borderline calls shift. The 15-call set is a property of this 31-gene panel and 16-lineage partition, not an invariant; the panel and partition were fixed before any metric was computed and were not revised thereafter (leave-one-lineage-out jackknife: Supplementary Note: Metric derivations and threshold rationale, S15). Seventh, the Lineage-defining suppressors VHL, PTEN, and KEAP1 rest on LEI and frequency alone—a single-metric classification we flag as provisional. Eighth, the framework is trained on untreated, primary-disease genomics: acquired resistance mutations plausibly represent an auxiliary→essential conversion under therapy-imposed selection, and the classifier has no standing on ctDNA or minimal-residual-disease contexts; the metrics make no single-agent-efficacy claims.

The necessary next step is prospective validation of this lineage-aware readout, together with fusion-grid extension and cancer-cell-fraction-level clonality modeling.

Conclusions

We set out to replace an expert intuition—which mutations a tumor cannot exist without—with a computable criterion set, bounded to an SNV-centric grid. Essential status is a property of a (pathway node, lineage) pair, not a gene. Three metrics with fixed, author-chosen dual thresholds, plus partial, pool-specific clonality corroboration, yield a three-class taxonomy that survived stress-testing on seven robustness axes. The clinical corollary is deliberately restrained: a lineage-aware, hypothesis-generating readout for biomarker interpretation, anchored to a quantified actionable fraction (2.53% of an unselected clinical-sequencing cohort carrying a Lineage-defining event with an in-lineage approved agent; 3.57% counting indirect HIF-2α druggability), whose prospective validation is the necessary next step.

Abbreviations

ATAC-seq
assay for transposase-accessible chromatin with sequencing
BH
Benjamini–Hochberg
BLCA
bladder urothelial carcinoma
ccRCC
clear-cell renal cell carcinoma
CI
confidence interval
COADREAD
colorectal adenocarcinoma
CPE
consensus purity estimation
CRISPR
clustered regularly interspaced short palindromic repeats
FDR
false discovery rate
GBM
glioblastoma
GDD
gene dependency difference

GISTIC
genomic identification of significant targets in cancer
HBV
hepatitis B virus
HCV
hepatitis C virus
HNSCC
head and neck squamous cell carcinoma
HPV
human papillomavirus
IQR
interquartile range
KIRC
kidney renal clear-cell carcinoma
KIRP
kidney renal papillary cell carcinoma
LAML
acute myeloid leukemia
LEI
lineage enrichment index
LGG
lower-grade glioma
LIHC
liver hepatocellular carcinoma
LOH
loss of heterozygosity
LUAD
lung adenocarcinoma
LUSC
lung squamous cell carcinoma
MEI
mutual exclusivity index
MSK-IMPACT
Memorial Sloan Kettering Integrated Mutation Profiling of Actionable Cancer Targets
MWU
Mann–Whitney U
ORR
objective response rate
OV
ovarian serous cystadenocarcinoma
PAAD
pancreatic adenocarcinoma
PFM/PWM
position frequency/weight matrix
SKCM
skin cutaneous melanoma
SNV
single-nucleotide variant
TARGET

Therapeutically Applicable Research to Generate Effective Treatments
TCGA
The Cancer Genome Atlas
THCA
thyroid carcinoma
UCEC
uterine corpus endometrial carcinoma
VAF
variant allele frequency
WT
wild-type.

Declarations

Clinical trial number: not applicable.

Ethics approval and consent to participate

This study analyzes only publicly available, de-identified datasets; no new human-participant data were collected, and no ethics approval was required.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Use of artificial intelligence tools

Open artificial intelligence (AI) tools contributed to language editing and drafting assistance during the preparation of this manuscript. The authors reviewed and edited all content and take full responsibility for the integrity and accuracy of the work.

Funding

This work was supported by the Beijing Science and Technology Nova Program Interdisciplinary Project (grant no. 20230484439, to SC). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contribution

SC: Conceptualization, Formal analysis, Software, Validation, Funding acquisition, Writing–original draft. QH: Conceptualization, Formal analysis, Software, Writing–original draft. XW: Data curation, Methodology, Investigation, Validation. HC: Data curation, Methodology, Visualization. DL: Data curation, Investigation, Project administration. XN: Supervision, Writing–review & editing. SC and QH contributed equally as co-first authors. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Availability of data and materials

All datasets analyzed in this study are publicly available. TCGA PanCancer Atlas somatic mutations were obtained through the cBioPortal public API (PanCancer Atlas 2018 studies; molecular profiles {study}_mutations). DepMap Public 24Q4 CRISPR Chronos gene-effect scores are available from figshare (article 27993248, file 51064667; DOI 10.25452/figshare.plus.27993248.v1), and DepMap Public 23Q4 files from figshare (files 43346616, 43346895, and 43347528). The TCGA mc3 consensus MAF (mc3.v0.2.8.PUBLIC.maf.gz) is available from the NCI Genomic Data Commons (file UUID 1c8cfe5f-e52d-41ba-94da-f15ea1337efc). TCGA PanCanAtlas ABSOLUTE purity/ploidy mastercalls are available from the NCI GDC (file UUID 4f277128-f793-4354-a13d-30cc7fe9f6b5). The MSK-IMPACT 50K cohort was accessed through the cBioPortal public API (study msk_impact_50k_2026). TARGET Wilms tumor whole-exome ensemble masked MAF files (open access) were obtained from the NCI GDC data endpoint, pediatric variant prevalence from the St. Jude PeCan public API (https://pecan.stjude.cloud), and TCGA clinical outcome endpoints from the TCGA Pan-Cancer Clinical Data Resource (TCGA-CDR). Access dates and integrity fingerprints (md5 checksums or GDC file UUIDs) are listed in Table S5.

Code availability

The analysis scripts used in this study (fetch_tcga.py, fetch_msk50k.py, analysis1_enrichment.py, analysis2_mutex.py, analysis3_depmap.py, analysis4_vaf.py, analysis_final.py, figs_final.py, Fig. 4_vaf.py, together with the round37_*.py sensitivity-analysis scripts, the analysis/mechanism/atac_chromatin/ ATAC pipeline scripts and logs, the round39_fig_*.py figure scripts, and the round47_*.py clinical-validation scripts), maintained under analysis/, are provided to editors and reviewers as a review-only archive accompanying this submission and will be deposited in a public repository with a persistent identifier upon acceptance; until then they are available from the corresponding author upon reasonable request.

References

1. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, BRIM-3 Study Group, et al. Improved survival with vemurafenib in melanoma with BRAF V600E mutation. N Engl J Med. 2011;364(26):2507–16. 10.1056/NEJMoa1103782.
2. Kopetz S, Desai J, Chan E, Hecht JR, O'Dwyer PJ, Maru D, et al. Phase II pilot study of vemurafenib in patients with metastatic BRAF-mutated colorectal cancer. J Clin Oncol. 2015;33(34):4032–8. 10.1200/JCO.2015.63.2497.
3. Prahallad A, Sun C, Huang S, Di Nicolantonio F, Salazar R, Zecchin D, et al. Unresponsiveness of colon cancer to BRAF(V600E) inhibition through feedback activation of EGFR. Nature. 2012;483(7387):100–3. 10.1038/nature10868.
4. Corcoran RB, Ebi H, Turke AB, Coffee EM, Nishino M, Cogdill AP, et al. EGFR-mediated reactivation of MAPK signaling contributes to insensitivity of BRAF mutant colorectal cancers to RAF inhibition with vemurafenib. Cancer Discov. 2012;2(3):227–35. 10.1158/2159-8290.CD-11-0341.
5. Sledge GW Jr, Yoshino T, Xiu J, Helmstetter A, Ribeiro JR, Klimov S, et al. Real-world evidence provides clinical insights into tissue-agnostic therapeutic approvals. Nat Commun. 2025;16:2646. 10.1038/s41467-025-57941-0.
6. Subbiah V, Kreitman RJ, Wainberg ZA, Gazzah A, Lassen U, Stein A, et al. Dabrafenib plus trametinib in BRAFV600E-mutated rare cancers: the phase 2 ROAR trial. Nat Med. 2023;29(5):1103–12. 10.1038/s41591-023-02321-8.
7. Baggiolini A, Callahan SJ, Montal E, Weiss JM, Trieu T, Tagore MM, et al. Developmental chromatin programs determine oncogenic competence in melanoma. Science. 2021;373(6557):eabc1048. 10.1126/science.abc1048.
8. Kaufman CK, Mosimann C, Fan ZP, Yang S, Thomas AJ, Ablain J, et al. A zebrafish melanoma model reveals emergence of neural crest identity during melanoma initiation. Science. 2016;351(6272):aad2197. 10.1126/science.aad2197.
9. Patton EE, Widlund HR, Kutok JL, Kopani KR, Amatruda JF, Murphey RD, et al. BRAF mutations are sufficient to promote nevi formation and cooperate with p53 in the genesis of melanoma. Curr Biol. 2005;15(3):249–54. 10.1016/j.cub.2005.01.031.

10. Shakhova O, Zingg D, Schaefer SM, Hari L, Civenni G, Blunschi J, et al. Sox10 promotes the formation and maintenance of giant congenital naevi and melanoma. Nat Cell Biol. 2012;14(8):882–90. 10.1038/ncb2535.
11. Garraway LA, Sellers WR. Lineage dependency and lineage-survival oncogenes in human cancer. Nat Rev Cancer. 2006;6(8):593–602. 10.1038/nrc1947.
12. Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007;450(7171):893–8. 10.1038/nature06358.
13. Bass AJ, Watanabe H, Mermel CH, Yu S, Perner S, Verhaak RG, et al. SOX2 is an amplified lineage-survival oncogene in lung and esophageal squamous cell carcinomas. Nat Genet. 2009;41(11):1238–42. 10.1038/ng.465.
14. Cheung HW, Cowley GS, Weir BA, Boehm JS, Rusin S, Scott JA, et al. Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. Proc Natl Acad Sci U S A. 2011;108(30):12372–7. 10.1073/pnas.1109363108.
15. Vierbuchen T, Ling E, Cowley CJ, Couch CH, Wang X, Harmin DA, et al. AP-1 transcription factors and the BAF complex mediate signal-dependent enhancer selection. Mol Cell. 2017;68(6):1067–e108212. 10.1016/j.molcel.2017.11.026.
16. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. Science. 2018;362(6413):eaav1898. 10.1126/science.aav1898.
17. Alcantara Llaguno S, Chen J, Kwon CH, Jackson EL, Li Y, Burns DK, et al. Malignant astrocytomas originate from neural stem/progenitor cells in a somatic tumor suppressor mouse model. Cancer Cell. 2009;15(1):45–56. 10.1016/j.ccr.2008.12.006.
18. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019;178(4):835–e84921. 10.1016/j.cell.2019.06.024.
19. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature. 2014;505(7484):495–501. 10.1038/nature12912.
20. Schaefer MH, Serrano L. Cell type-specific properties and environment shape tissue specificity of cancer genes. Sci Rep. 2016;6:20707. 10.1038/srep20707.
21. Chen J, Collins RL, Haigis KM et al. The tissue specificity of cancer genes. bioRxiv 2025. 10.1101/2025.09.21.677175. [Preprint].
22. Seiwert TY, Zuo Z, Keck MK, Khattri A, Pedamallu CS, Stricker T, et al. Integrative and comparative genomic analysis of HPV-positive and HPV-negative head and neck squamous cell carcinomas. Clin Cancer Res. 2015;21(3):632–41. 10.1158/1078-0432.CCR-13-3310.
23. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer (TCGA PanCancer Atlas). Cell. 2018;173(2):291–e3046. 10.1016/j.cell.2018.03.022.
24. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4. 10.1158/2159-8290.CD-12-0095.
25. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal. 2013;6(269):pl1. 10.1126/scisignal.2004088.
26. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA Jr, Kinzler KW. Cancer genome landscapes Sci. 2013;339(6127):1546–58. 10.1126/science.1235122.
27. Kandoth C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013;502(7471):333–9. 10.1038/nature12634.
28. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15. 10.1038/nature10166.
29. Tsherniak A, Vazquez F, Montgomery PG, Weir BA, Kryukov G, Cowley GS, et al. Defining a cancer dependency map. Cell. 2017;170(3):564–e57616. 10.1016/j.cell.2017.06.010.
30. Dempster JM, Boyle I, Vazquez F, Root DE, Boehm JS, Hahn WC, Tsherniak A, McFarland JM. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol. 2021;22:343. 10.1186/s13059-021-02540-7.

31. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandoth C, Stewart C, MC3 Working Group; Cancer Genome Atlas Research Network, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. Cell Syst. 2018;6(3):271–e2817. 10.1016/j.cels.2018.03.002.
32. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. N Engl J Med. 2016;375(12):1109–12. 10.1056/NEJMp1607591. Toward a shared vision for cancer genomic data (NCI Genomic Data Commons).
33. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR, et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients (MSK-IMPACT). Nat Med. 2017;23(6):703–13. 10.1038/nm.4333.
34. Bandlamudi C, Muldoon D, de Bruijn I, Zhang M, Gormally MV, Harrold EC, et al. Cancer type-specific variation in patterns of driver alterations across 50,000 tumors. Cancer Cell. 2026;44(5):1045–e10626. 10.1016/j.ccell.2026.03.003.
35. AACR Project GENIE Consortium. AACR Project GENIE: powering precision medicine through an international consortium. Cancer Discov. 2017;7(8):818–31. 10.1158/2159-8290.CD-17-0151.
36. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30(5):413–21. 10.1038/nbt.2203.
37. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. Nat Commun. 2015;6:8971. 10.1038/ncomms9971.
38. Schep AN, Wu B, Buenrostro JD, Greenleaf WJ. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. Nat Methods. 2017;14(10):975–8. 10.1038/nmeth.4401.
39. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57(1):289–300. 10.1111/j.2517-6161.1995.tb02031.x.
40. Canisius S, Martens JWM, Wessels LFA. A novel independence test for somatic alterations in cancer shows that biology drives mutual exclusivity but chance explains most co-occurrence. Genome Biol. 2016;17(1):261. 10.1186/s13059-016-1114-x.
41. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, The Cancer Genome Atlas Research Network, et al. Comprehensive molecular characterization of muscle-invasive bladder cancer. Cell. 2017;171(3):540–e55625. 10.1016/j.cell.2017.09.007.
42. Weinstein IB. Addiction to oncogenes—the Achilles heal of cancer. Science. 2002;297(5578):63–4. 10.1126/science.1073096.
43. Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J et al. OncoKB: a precision oncology knowledge base. JCO Precis Oncol 2017;2017:PO.17.00011. 10.1200/PO.17.00011
44. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature. 2015;517(7536):576–82. 10.1038/nature14129.
45. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. Nature. 2013;499(7456):43–9. 10.1038/nature12222.
46. Treon SP, Xu L, Yang G, Zhou Y, Liu X, Cao Y, et al. MYD88 L265P somatic mutation in Waldenström's macroglobulinemia. N Engl J Med. 2012;367(9):826–33. 10.1056/NEJMoa1200710.
47. Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. Cell. 2017;169(7):1327–e134123. 10.1016/j.cell.2017.05.046.
48. Al-Obaidy KI, Cheng L. Fibroblast growth factor receptor (FGFR) gene: pathogenesis and treatment implications in urothelial carcinoma of the bladder. J Clin Pathol. 2021;74(8):491–5. 10.1136/jclinpath-2020-207115.
49. Louis DN, Perry A, Wesseling P, Brat DJ, Cree IA, Figarella-Branger D, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. Neuro Oncol. 2021;23(8):1231–51. 10.1093/neuonc/noab106.
50. Mateo J, Chakravarty D, Dienstmann R, Jezdic S, Gonzalez-Perez A, Lopez-Bigas N, et al. A framework to rank genomic alterations as targets for cancer precision medicine: the ESMO Scale for Clinical Actionability of molecular Targets (ESCAT). Ann Oncol. 2018;29(9):1895–902. 10.1093/annonc/mdy263.
51. Sung WK, Zheng H, Li S, Chen R, Liu X, Li Y, et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. Nat Genet. 2012;44(7):765–9. 10.1038/ng.2295.

52. Bartosch B, Thimme R, Blum HE, Zoulim F. Hepatitis C virus-induced hepatocarcinogenesis. J Hepatol. 2009;51(4):810–20. 10.1016/j.jhep.2009.05.008.
53. Calonaci N, Krasniqi E, Colic D, Scalera S, Gandolfi G, Milite S, et al. Gene mutant dosage is associated with prognosis and metastatic tropism in 60,000 clinical cancer samples. Nat Genet. 2026;58(8):1906–17. 10.1038/s41588-026-02666-z.
54. Skoulidis F, Goldberg ME, Greenawalt DM, Hellmann MD, Awad MM, Gainor JF, et al. STK11/LKB1 mutations and PD-1 inhibitor resistance in KRAS-mutant lung adenocarcinoma. Cancer Discov. 2018;8(7):822–35. 10.1158/2159-8290.CD-18-0099.
55. Cantore T, Gasperini P, Bevilacqua R, Ciani Y, Sinha S, Ruppin E, et al. PRODE recovers essential and context-essential genes through neighborhood-informed scores. Genome Biol. 2025;26(1):42. 10.1186/s13059-025-03501-0.
56. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. Analytics Cell. 2018;173(2):400–e41611. 10.1016/j.cell.2018.02.052. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome.
57. McLeod C, Gout AM, Zhou X, Thrasher A, Rahbarinia D, Brady SW, et al. St. Jude Cloud: a pediatric cancer genomic data-sharing ecosystem. Cancer Discov. 2021;11(5):1082–99. 10.1158/2159-8290.CD-20-1230.
58. NCI Genomic Data Commons (GDC) data endpoint. TARGET whole-exome sequencing ensemble masked somatic mutation MAF files (open access). https://api.gdc.cancer.gov/data/{file-uuid}. Accessed 2026-10-21.
59. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nat Rev Cancer. 2020;20(10):555–72. 10.1038/s41568-020-0290-x.
60. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–e38518. 10.1016/j.cell.2018.02.060.

Figures

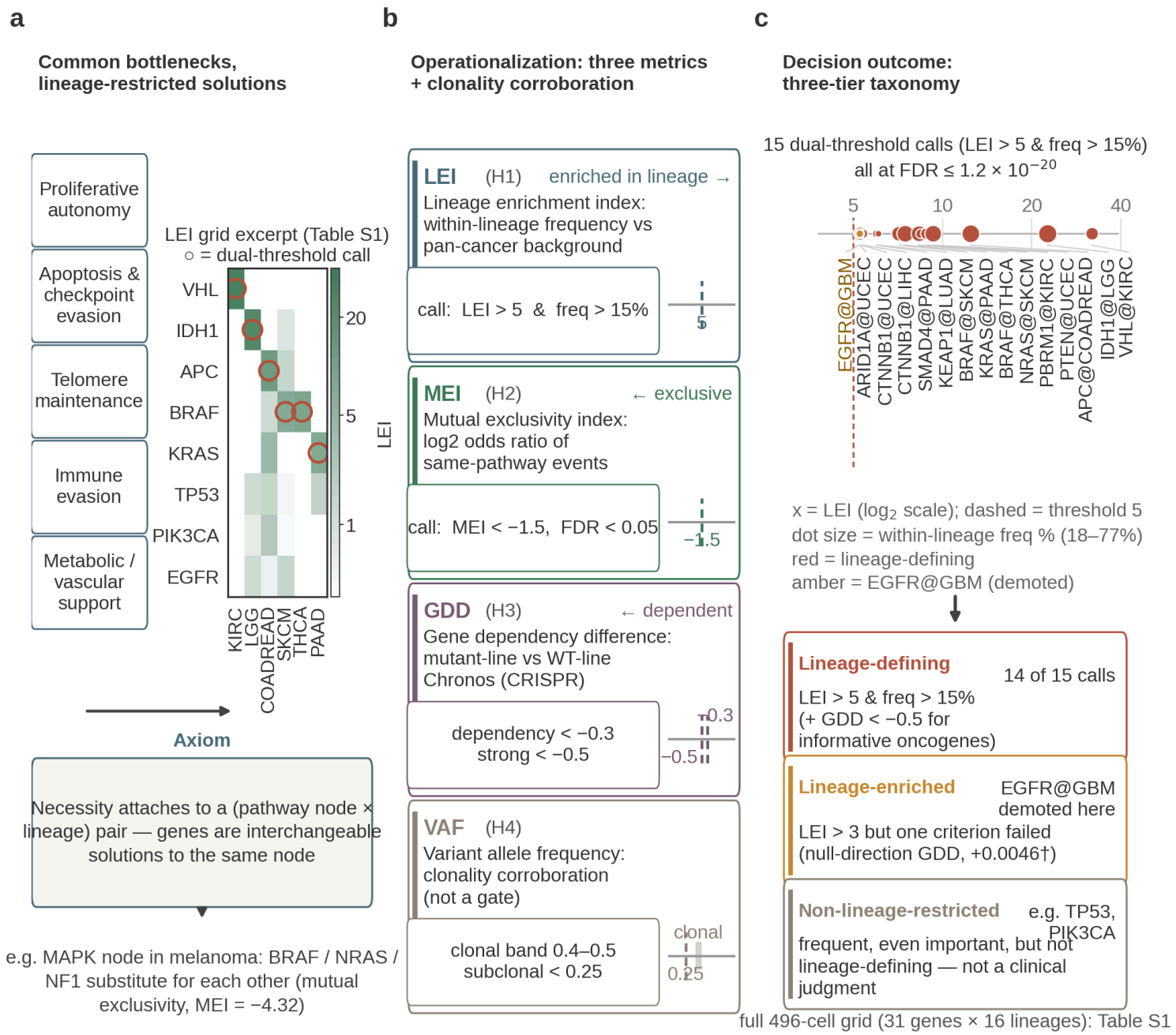


Figure 1

Conceptual framework schematic. (A) Common bottlenecks in tumor evolution are solved by lineage-restricted solutions; necessity attaches to a (pathway node, lineage) pair rather than to a gene. (B) Two orthogonal dimensions—lineage-restriction and necessity—operationalized as three framework metrics (LEI, MEI, GDD) with their decision thresholds, plus VAF clonality as corroboration. (C) Screen outcome: 15 dual-threshold essential calls at FDR ≤ 1.2×10^{-20}, positioned by LEI ($\log_2$ scale) and sized by within-lineage frequency; 14 retain Lineage-defining status in the final three-metric grid (red), with EGFR@GBM demoted at the dependency step (amber) (full 496-cell grid in Table S1).

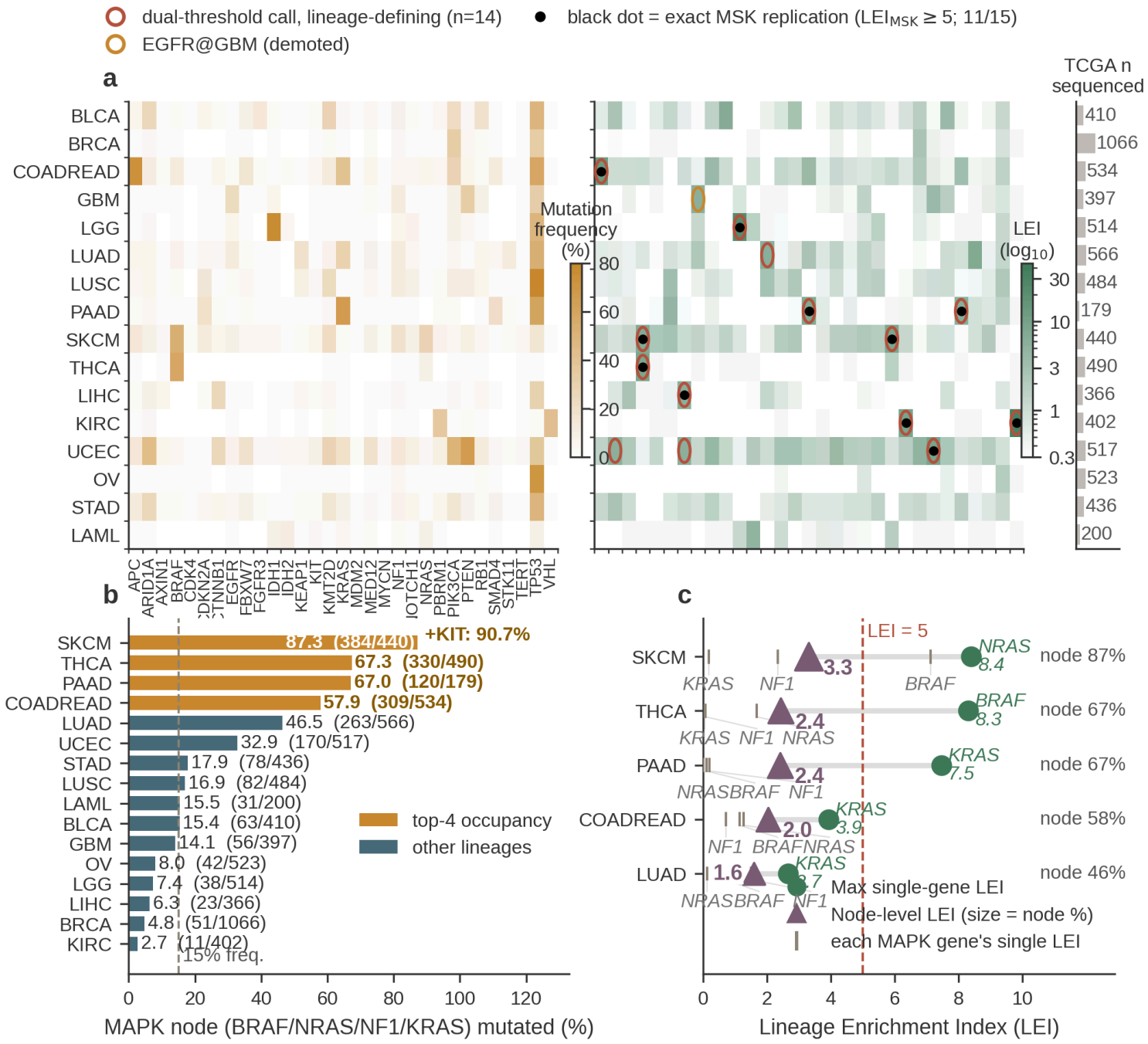


Figure 2

Mutation frequency, lineage enrichment, and MAPK node occupancy across 16 TCGA PanCancer Atlas cancer types. (A) Mutation frequency (%) of 31 driver genes (left); LEI=within-cancer frequency/(mean of other 15 types + 1), log10 (right). Circles, the 15 dual-threshold calls by final tier (red, Lineage-defining; amber, EGFR@GBM demoted); black dots, the 11 calls replicating in MSK-IMPACT 50K. (B) Sample-level MAPK node occupancy (BRAF/NRAS/NF1/KRAS union), mutated/sequenced counts at bar ends; amber bars, four highest-occupancy lineages; dashed line, 15% reference; SKCM includes the KIT-extended node. (C) Gene-level (circles) versus node-level (triangles) LEI, five highest-occupancy lineages: aggregation raises occupancy but compresses the enrichment ratio; the LEI=5 line (dashed) is gene-level only.

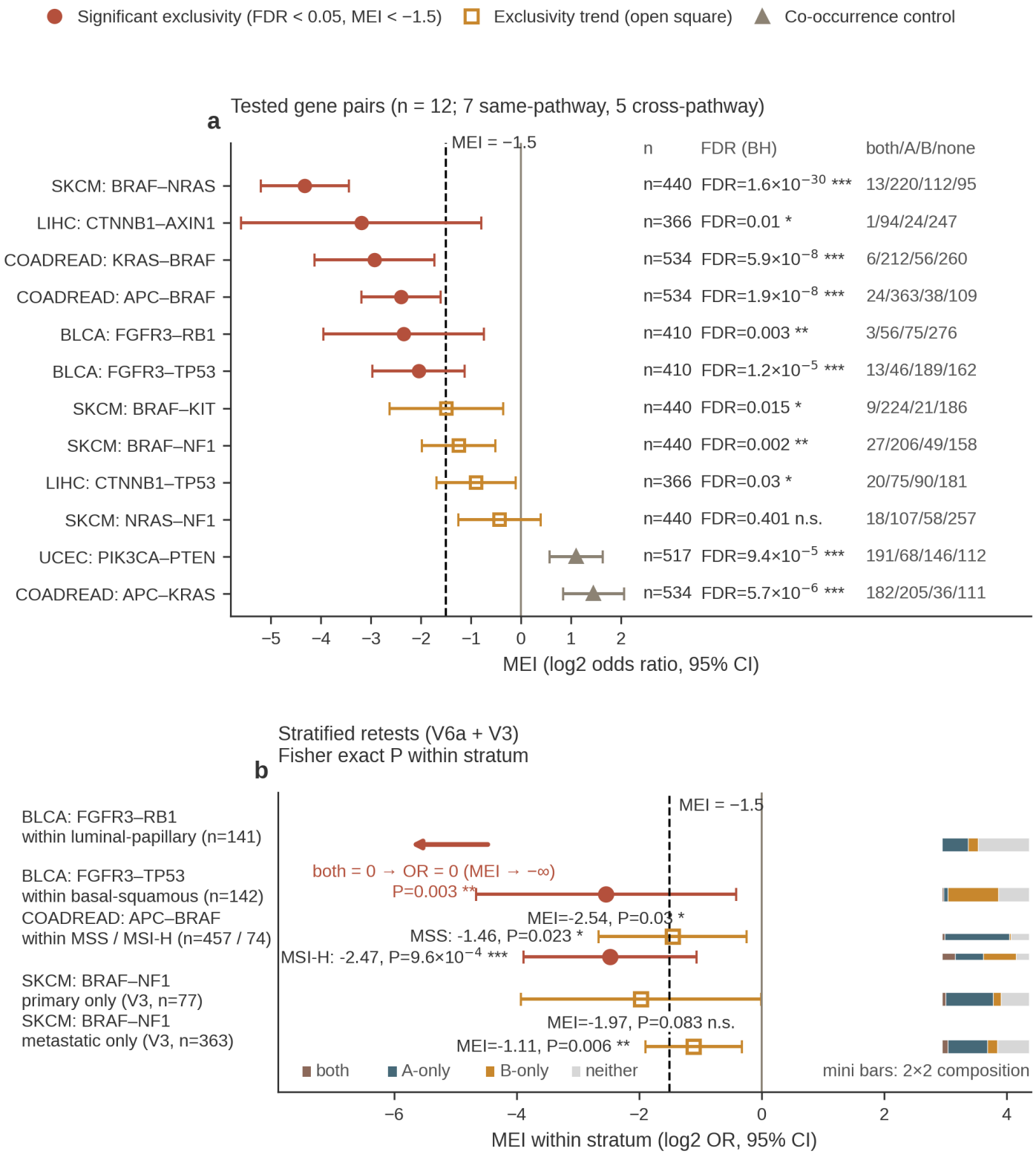
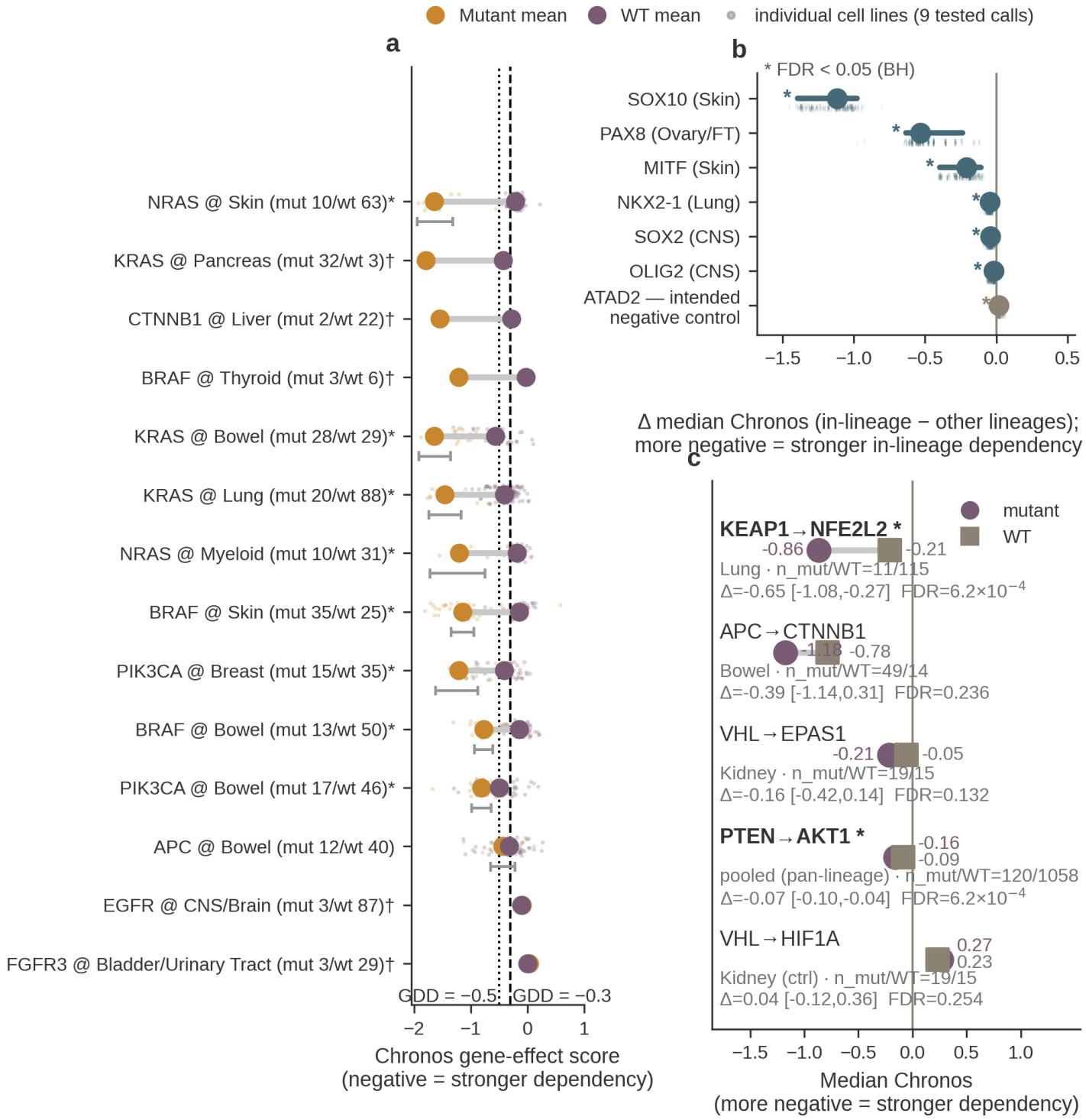


Figure 3

Sample-level MEI across the designated 12-pair grid (TCGA PanCancer Atlas). (A) MEI with 95% CIs; dashed line, exclusivity threshold (MEI=−1.5). Red, significant exclusivity (FDR<0.05, MEI<−1.5); amber, exclusivity trend (FDR<0.05, above the line); blue, the two co-occurrence controls; per-pair n, FDR, 2 × 2 counts annotated. (B) Subtype- and MSI-stratified retests (axis V6a, nominal P; Table S22 row 8): FGFR3–RB1 zero co-occurrences in luminal-papillary; FGFR3–TP53 exclusive in basal-squamous; APC–BRAF holds in MSS and MSI-H (−2.47, P=9.6×10⁻⁴); BRAF–NF1 more negative in primary (−1.97, n=77, P=0.083) than metastatic (−1.11, n=363, P=6.4×10⁻³), strata not formally compared.



* FDR < 0.05; † descriptive only (<4 lines in one group; KRAS @ Pancreas: 32 mutant vs 3 WT lines)
grey whisker = bootstrap 95% CI of GDD, anchored at WT mean (v2_gdd_bootstrap_ci.csv); (C) [..] = bootstrap 95% CI of Δ

Figure 4

Functional dependency evidence (DepMap 24Q4). (A) Mutant-selective GDD: dumbbells link mean Chronos of mutant (amber) and wild-type (purple) lines, bootstrap 95% CIs for the nine tested cells (five descriptive † summary-only); decision lines −0.3 (dashed), −0.5 (dotted); * FDR<0.05; Table S6. (B) Reader dependency: Δmedian Chronos (in-lineage minus all others), bootstrap 95% CI, seven reader–lineage pairs; all six identity-factor pairs reach BH-FDR<0.05, two reach the −0.30 dependency line (SOX10, PAX8); ATAD2 (grey), intended negative control, negligible CI-crossing yet nominally FDR<0.05—a partial control failure (Table S12). (C) Effector coupling: median Chronos of downstream effectors, mutant versus wild-type lines; n, Δ [95% CI], BH-FDR annotated (Table S13).

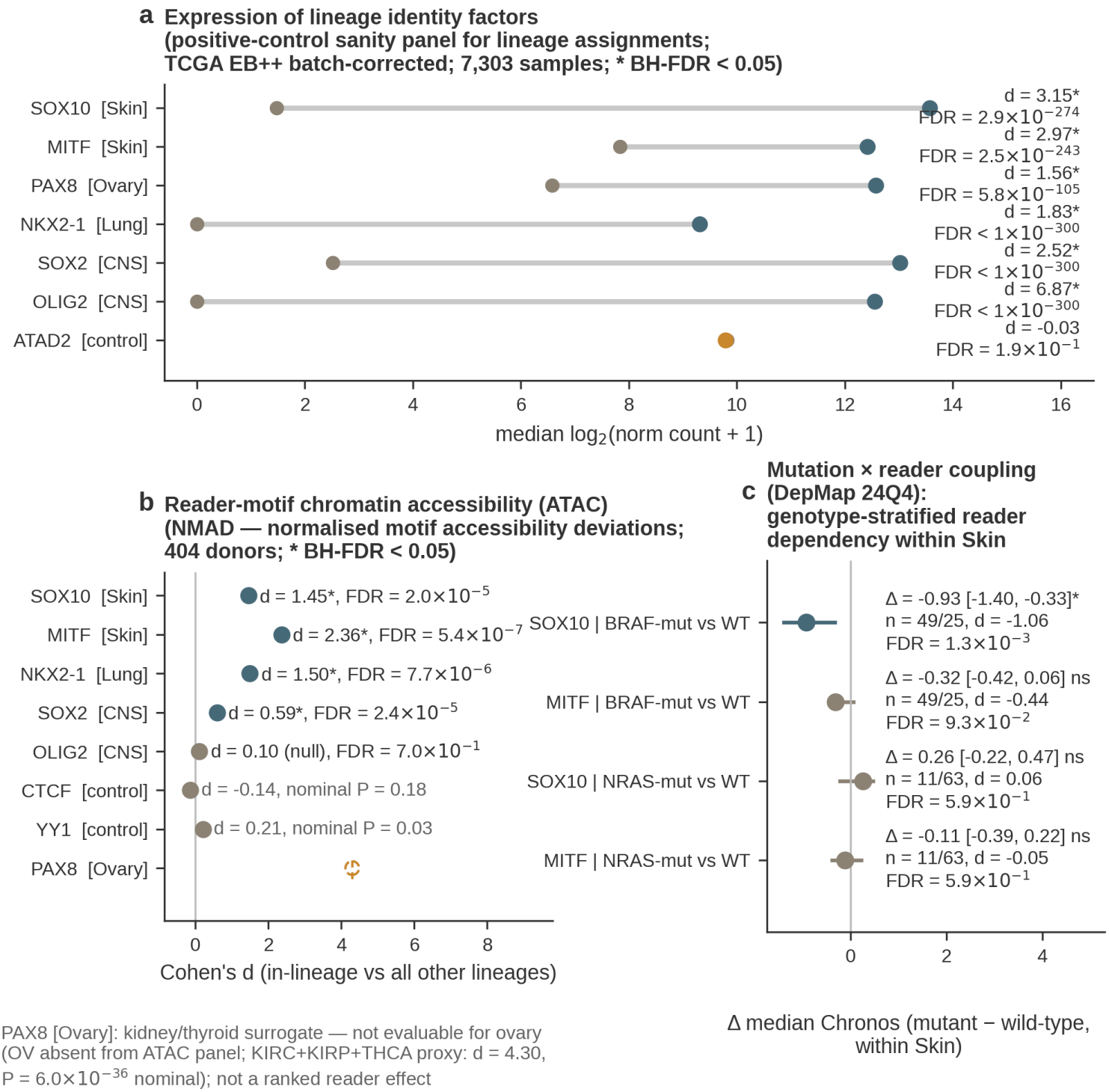
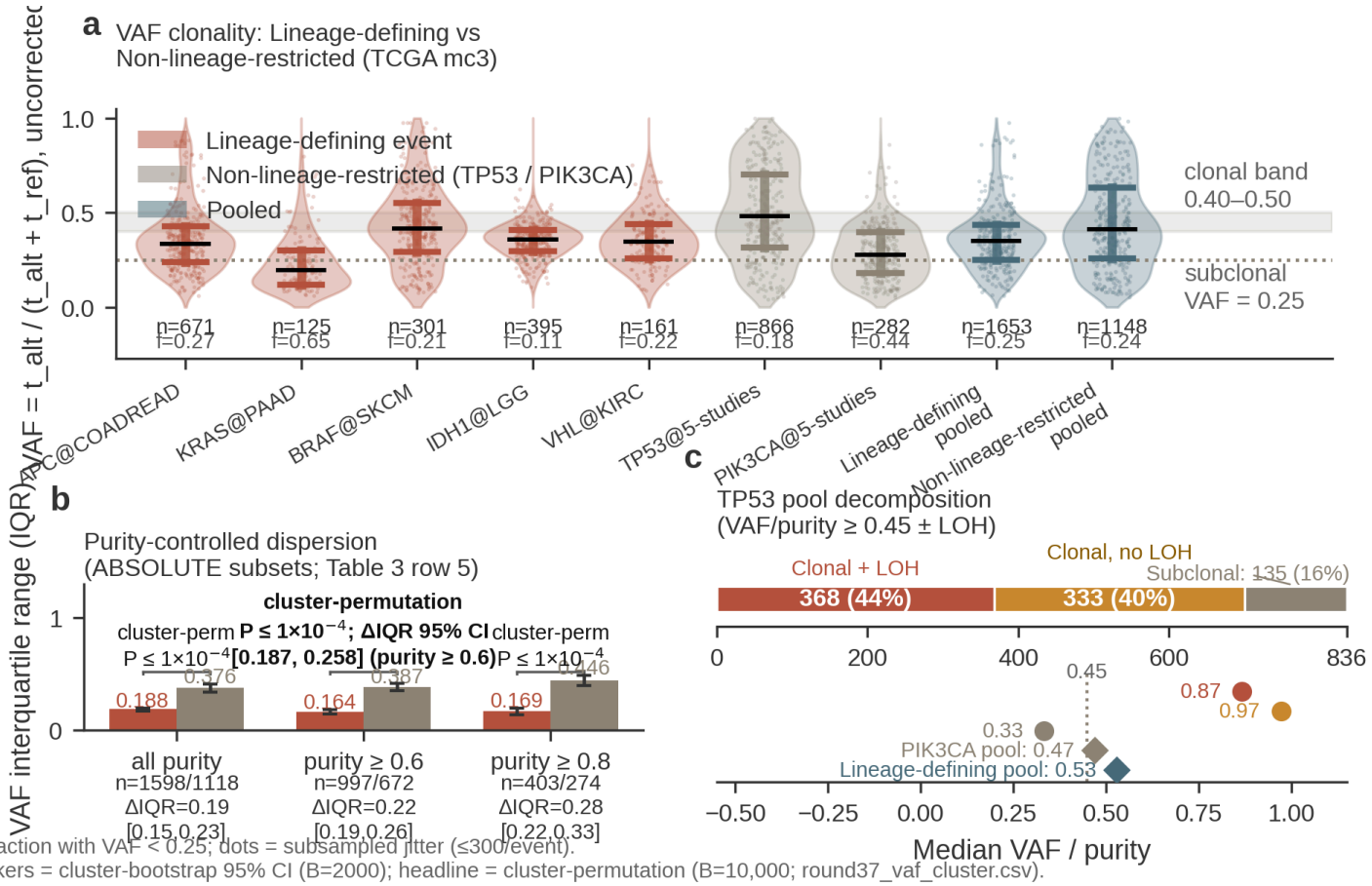


Figure 5

The lineage-reader factor measured on three orthogonal axes. (A) Reader-TF expression in primary tumors (TCGA EB++ matrix, 7,303 samples): a positive-control panel, not a test of the framework—in-lineage versus all-other-lineage medians, six identity-TF pairs plus ATAD2 merged control; Cohen's d, BH-FDR (Table S15). (B) Reader-motif chromatin accessibility (chromVAR-isomorphic reimplementation, 404 QC-pass donors): Cohen's d, BH-FDR for five evaluable pairs, including the OLIG2@CNS null; PAX8@Ovary not evaluable (open marker); dashed amber marker, kidney/thyroid surrogate, excluded from reader-effect ranking; CTCF and YY1 architectural controls (Table S17). (C) Genotype-stratified reader dependency within Skin (DepMap 24Q4): Δmedian Chronos with bootstrap 95% CIs across BRAF and NRAS strata (Table S12).



a: f = fraction with VAF < 0.25; dots = subsampled jitter (≤300/event).
b: whiskers = cluster-bootstrap 95% CI (B=2000); headline = cluster-permutation (B=10,000; round37_vaf_cluster.csv).
b, mutation-level tests (wrong inferential unit, anti-conservative): Levene P = 2.2×10^{-50} / 1.4×10^{-39} / 9.1×10^{-18};
MWU P = 5.3×10^{-18} / 6.2×10^{-26} / 1.1×10^{-10} (tiers: all purity / ≥0.6 / ≥0.8).

Figure 6

VAF clonality of Lineage-defining versus Non-lineage-restricted mutations (TCGA mc3; n=2,801). (A) Uncorrected VAF: five Lineage-defining events (red), Non-lineage-restricted TP53/PIK3CA pool (blue); f, fraction VAF<0.25; grey band, clonal heterozygous expectation (0.40–0.50); dotted line, subclonal reference (0.25). (B) Purity-controlled dispersion (ABSOLUTE: n=2,716; ≥0.6, n=1,669; ≥0.8, n=677): IQR gap 0.16 versus 0.39 at ≥0.6, significant at every tier (Levene P=2.2×10^{-50}/1.4×10^{-39}/9.1×10^{-18}; Mann–Whitney P=5.3×10^{-18}/6.2×10^{-26}/1.1×10^{-10}); error bars, cluster-bootstrap 95% CIs (Table S22 row 5). (C) TP53-pool decomposition (n=836): clonal with LOH 368 (44%), clonal without LOH 333 (40%), subclonal 135 (16%); group-median VAF/purity 0.87/0.97/0.33 versus PIK3CA 0.47, Lineage-defining 0.53; dotted line, threshold 0.45.

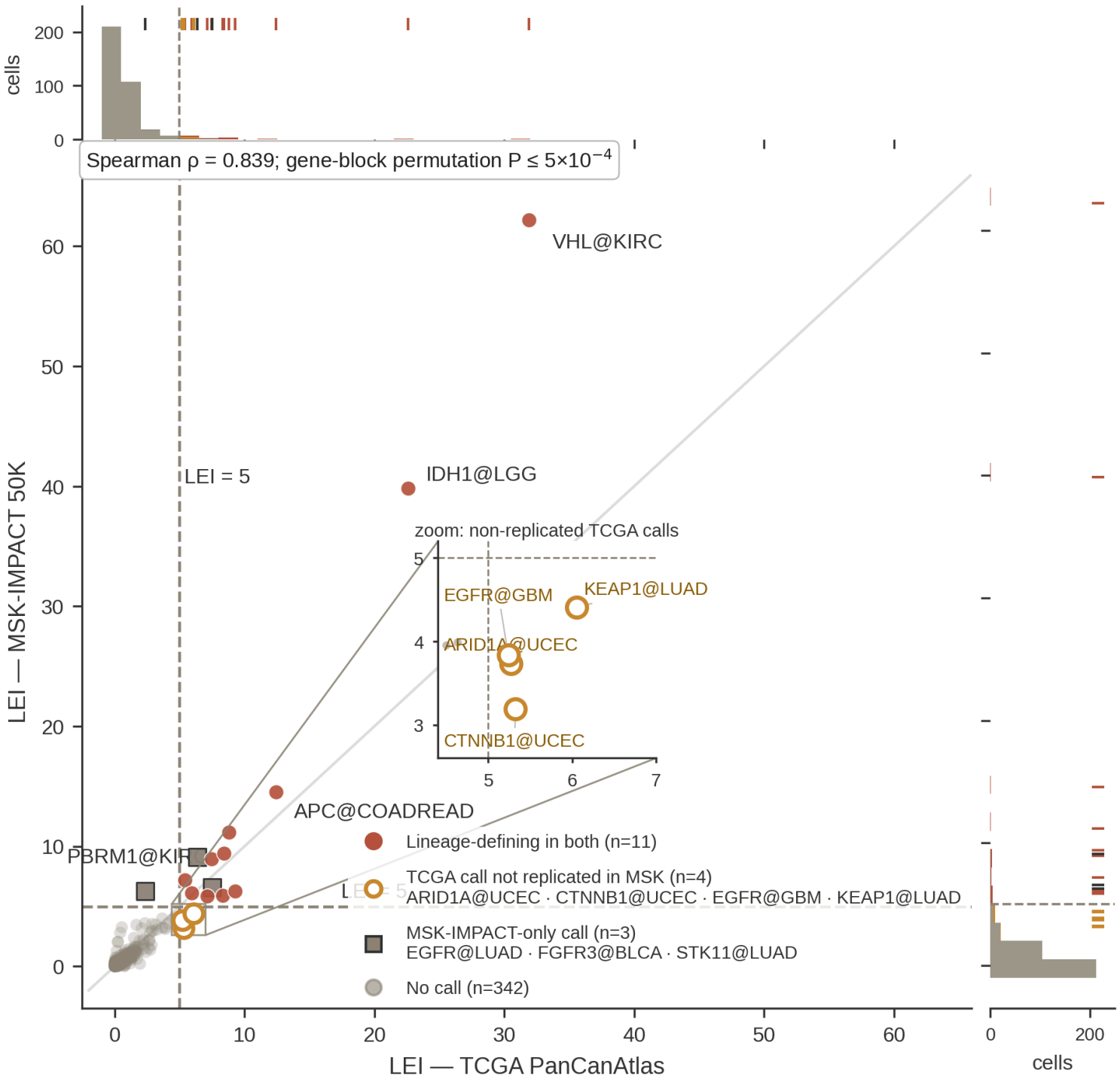


Figure 7

LEI replication: TCGA PanCanAtlas versus MSK-IMPACT 50K. LEI across the 360 shared gene–lineage cells (24 genes × 15 lineages): Spearman ρ=0.839 (gene-block permutation P≤5×10⁻⁴ (2,000 nulls, maximum null ρ=0.276), effective n≈24 genes; cell-level asymptotic P=1.6×10⁻⁹⁶ uses the wrong inferential unit; bootstrap 95% CI [0.747, 0.896]); restricted checks ρ=0.892/0.881/0.817/0.775 (LEI>1 in both/TCGA/MSK/either cohort, n=70/92/78/100); raw frequencies ρ=0.821. Filled red points, the 11 dual-threshold calls replicating in both cohorts; amber-edged open circles, the four TCGA calls scoring LEI<5 in MSK (zoom inset); steel-blue squares, three MSK-only calls (three further off-panel); marginal histograms, cell-count distributions; dashed lines, LEI=5 in each cohort.

Supplementary Files

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

- TableS1.csv
- TableS3.csv
- TableS4.csv
- SupplementaryMaterials.docx