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.

Uninformative non-reproduction: a pre-computable power frontier explains which driver–phenotype associations small cancer cohorts can and cannot test

Xueyi Lin
Yunnan Normal University
Xiayu Wu
3537@ynnu.edu.cn

Yunnan Normal University
SANjun Zhao
Yunnan Normal University

Research Article

Keywords: reproducibility, statistical power, sample size, detection limit, rare tumours, open science

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11401252/v1

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

Additional Declarations: No competing interests reported.

Uninformative non-reproduction: a pre-computable power frontier explains which driver–phenotype associations small cancer cohorts can and cannot test

Xueyi Lin[1], Xiayu Wu[1,2,3*], and Sanjun Zhao[1,2,3*]

[1] School of Life Sciences, Yunnan Normal University, Kunming 650500, Yunnan, China

[2] Yunnan Key Laboratory of Potato Biology, School of Life Sciences, Yunnan Normal University, Southwest United Graduate School, Kunming, Yunnan 650500, China

[3] Yunnan Key Laboratory of Potato Biology, School of Life Sciences, Yunnan Normal University, Southwest United Graduate School, Kunming, Yunnan 650500, China

[*] Correspondence: Xiayu Wu (xiayuwu3537@ynnu.edu.cn) and Sanjun Zhao (zhaosanjun4052@ynnu.edu.cn)

ORCID: Xiayu Wu: https://orcid.org/0000-0002-0612-3664

Abstract

Background

Small cancer cohorts are routinely used to test driver–phenotype associations, and negative results from them enter the literature as evidence that no association exists. Reproducibility frameworks classify such results only as reproduced or not reproduced, conflating a genuine null finding with a study that could never have detected the effect. We asked whether this third possibility can be identified before data are collected.

Methods

We derived a closed-form detection frontier, $d_{min}(n,p)$ — the smallest standardised effect detectable at 80% power, two-sided $\alpha = 0.05$, in a two-group comparison of total size n with alteration prevalence p — and calibrated it against empirical power by simulation across three data-generating mechanisms and four group-size configurations (60 points, 2 000 replicates each). We then applied the frontier to a complete scan of nine driver variables against eight expression signatures in TCGA malignant pleural mesothelioma ($n = 87$), and to fifteen rare-tumour TCGA cohorts.

Results

Under ideal-normal data the closed form was accurate to 0.009; under skew and dropout contamination it over-estimated the detection limit by up to 0.086, erring on the safe side in all eight non-normal configurations. Six of fifteen cohorts could detect a medium effect ($d = 0.5$) and none a small one ($d = 0.3$); the latter would require 416 samples. Of 72 comparisons in the case study, seven (9.7%) had an observed effect at or above their own detection limit and all seven were significant; the remaining 65 fell below it, and all 59 non-significant comparisons lay below their own limit, so none constituted an informative null. A published large-effect association reproduced closely (CDKN2A deletion 41.4% versus 41.5%), whereas a reported medium-effect association could not be addressed: that comparison's detection limit was 0.705

against an observed effect of 0.011.

Conclusions

A non-significant driver–phenotype result in a small cohort may carry no evidence either way, and this is determinable from cohort size and driver prevalence before any data are collected. Reporting the detection limit alongside negative results would distinguish an underpowered comparison from a genuine absence of association.

Keywords: reproducibility; statistical power; sample size; detection limit; rare tumours; open science

Background

The binary framing of reproducibility

Efforts to improve the reliability of biomedical research have largely been organised around a binary. A finding is either reproduced or it is not. When a study fails to recover a published association, that failure is counted — implicitly or explicitly — as evidence against the original claim, and it joins a growing tally of results that the literature treats as doubtful [1, 2].

That framing works well when the second study is large enough to have detected the effect in question. It breaks down when it is not. A study can fail to reproduce an association because the association does not exist, or because the study was never capable of detecting it. Under the binary framing these two situations are recorded identically, and the second is silently

attributed to the first.

The difficulty is not hypothetical. In brain-wide association studies, where the same question has been examined systematically, individual brain–phenotype correlations proved so small that samples of thousands of individuals were required before effect sizes stabilised [3].

The asymmetry is not merely philosophical. Low power biases the direction of reported evidence in a well-characterised way: among studies that reach significance, small ones overstate the effects they find, an effect known since the genetic-association literature named it the winner's curse [4, 5]. The corresponding problem on the negative side has received less formal treatment. Its classic statement — absence of evidence is not evidence of absence — dates to 1995 [6] and is universally endorsed, but it remains a caution rather than an operational rule. Nothing in routine practice requires an author, when reporting that no association was found, to state what association the study could have found.

Small cohorts for rare tumours

The problem is sharpest where cohorts are smallest. Rare tumours are systematically underrepresented in cancer genomics, and the resulting sample sizes — frequently below 100 cases, sometimes below 50 — are the only ones that exist for many diseases. The pan-cancer compendia contain cohorts such as cholangiocarcinoma (n = 36) and diffuse large B-cell lymphoma (n = 48) whose entire molecular characterisation rests on a few dozen specimens [7, 8]. Pleural mesothelioma, for which median survival is under two years [9], is represented in the same compendium by 87 primary tumours [10].

These cohorts are nonetheless used to test driver–phenotype associations, and the results enter the literature at face value. A negative finding in such a cohort is typically reported, and read, as evidence that the driver in question is unrelated to the phenotype. Whether the cohort was capable of detecting the relationship is rarely asked, and the quantities needed to answer it — the smallest detectable effect given the cohort size and the driver's prevalence — are never

reported.

This study

We ask whether the informativeness of a driver–phenotype test can be determined before the data are collected, and we argue that the answer is yes, from two numbers that are known at design time: the number of samples and the prevalence of the alteration.

We derive a closed-form *detection frontier*, $d_{min}(n,p)$ — the smallest standardised effect detectable at 80% power in a comparison of total size n with alteration prevalence p — calibrate it against empirical power under conditions that violate the normal approximation, and apply it to a complete driver-by-phenotype scan in a cohort of $n = 87$. The applied analysis is not a new biological claim; it is a worked demonstration, using a disease whose published associations include both a large effect that reproduces and a medium effect that the cohort cannot address.

The contribution we claim is conceptual and immediately usable. In addition to "reproduced" and "failed to reproduce" there is a third outcome, **uninformative non-reproduction**: a result that agrees or disagrees with a published claim but carries no evidence about it, because the comparison could not have detected an effect of the size claimed. Unlike the first two, this one is decidable in advance, and the table that decides it depends on nothing but the study design.

Methods

Overview and study design

We report a two-component study. The first component is methodological: we derive a closed-form *detection frontier* that specifies, for any two-group comparison of size n with driver prevalence p, the smallest standardised effect that can be detected with 80% power at $\alpha = 0.05$, and we validate this closed form by simulation under three non-ideal data-generating regimes.

The second component is empirical: we apply the frontier retrospectively to a complete driver-by-phenotype scan in a small real cohort — TCGA malignant pleural mesothelioma (MESO, $n = 87$) — and use a pair of published, already-reproduced associations as an internal benchmark for what the same cohort *can* and *cannot* resolve.

The two components are deliberately coupled. A frontier derived analytically is only useful if its verdicts match what a real, finite, messy cohort actually delivers; and a case study is only interpretable if each of its negative results can be located relative to that cohort's own detection limit. Neither component is sufficient alone, and our principal claim — that a third class of result, the **uninformative non-reproduction**, exists and is decidable before data are collected — rests on the agreement between them.

No human or animal subjects were involved; all data are previously published, de-identified public resources (see *Data sources* below and the Availability of data statement).

The detection frontier

Effect size definition

Throughout, the effect size for a comparison of group A (driver-altered, size n_1) and group B (driver-wild-type, size n_2) is the standardised mean difference

$$d = (\bar{x}A - \bar{x}B)/SDcohort, \quad SD_{cohort}^2 = 1/n\Sigma_{i=1}^{n}(x_i - x)^2, \quad n = n_1 + n_2,$$

that is, the group difference standardised by the standard deviation of the **whole cohort** rather than by a within-group pooled standard deviation. This is a deliberate choice and it is not the Cohen's *d* convention. The quantity SD_{cohort} is precisely the scale that must be estimated in an unselected cohort when asking whether a driver-stratified contrast is resolvable, and it is the quantity our empirical pipeline computes; the two conventions must not be mixed.

Two consequences of this choice are used later and are worth separating, because only the first

drives the frontier. **First**, power depends on the effect through the non-centrality parameter $\lambda = d\sqrt{(n1n2/(n1+n2))}$, which for $n_1 = np$ and $n_2 = n(1-p)$ is approximately $d\sqrt{(np(1-p))}$. An unbalanced split therefore loses efficiency, so for fixed n the frontier rises as p moves away from 0.5 roughly as $1/\sqrt{(p(1-p))}$; this is why rare drivers are hard to test, and it is not a small-cohort effect — it persists at any n. **Second**, because the between-group shift contributes to the denominator,

$$SD_{cohort} = \sqrt{(\sigma 2 + \delta 2\, p(1-p))},$$

the standardised effect is bounded above by $1/\sqrt{(p(1-p))}$ regardless of the raw shift. This ceiling is *loosest* for rare drivers, so it is not the reason they are hard to test; it matters instead for interpreting large standardised effects, since an effect near the ceiling implies a raw shift far exceeding the within-group spread.

Power model and threshold solution

Power for a two-sided two-group comparison is computed by the normal approximation,

$$\pi(d, n_1, n_2) = 1 - \Phi(z_{1-\alpha/2} - d\sqrt{(n1\ n2/(n1+n2)))} + \Phi(-z_{1-\alpha/2} - d\sqrt{(n1\ n2/(n1+n2)))},$$

with $\alpha = 0.05$, $z_{1-\alpha/2} = 1.95996$ and Φ the standard normal CDF. Given a total cohort size n and a driver prevalence p, we set $n_1 = max(2, round(np))$ and $n_2 = n - n_1$ and define the **minimum detectable effect**

$$d_{min}(n, p) = argmin_d\{\pi(d, n_1, n_2) \geq 0.80\},$$

obtained by bisection on $d \in [0.01, 8]$ to a tolerance of 10^{-3}. The inverse problem — the sample size required to detect a nominated effect at 80% power,

$$n_{req}(d, p) = argmin_n\{\pi(d, max(2,round(np)), n - max(2,round(np))) \geq 0.80\},$$

is solved by the same bisection over n. The frontier $d_{min}(n, p)$ is the object we propose as a pre-

registration instrument: it depends only on the design (n, p, α, target power) and is therefore computable before any data exist.

For the applied scan we additionally report, for each realised comparison, the power at its *observed* effect, $\pi(|d|, \hat{n}_1, n_2)$, which locates that specific result relative to its own detection limit.

Simulation validation of the closed form

An analytic frontier built on a normal approximation is only credible if it survives conditions that violate the approximation. We therefore calibrated π and d_{min} against empirical power in a factorial simulation.

Generating mechanisms (three). (i) *Ideal normal*: $x \sim \mathcal{N}(\delta, 1)$ in group A and $\mathcal{N}(0, 1)$ in group B. (ii) *Gene-set score*: a signature score formed as the mean of $G = 40$ marker values, each drawn log-normal$(0, 1)$; because the mean of log-normal variates is right-skewed, this reproduces the actual distributional shape of a small-panel expression score, and the resulting score was rescaled to unit SD so that δ equals d directly. (iii) *Dropout-contaminated*: equal-variance normals ($\sigma = 0.9$ in both groups), with an independent 6% per-sample probability of a left-shifted outlier ($x \rightarrow x - \text{Exponential}(2)$); this injects the heavy left tail and the tied ranks that real assay data exhibit. The two groups share the same variance and the same dropout process by construction, so this regime is deliberately *not* heteroscedastic; the departure from normality is in the tail shape alone.

Group-size structures (four). Three configurations are the realised group sizes of driver strata in the case study itself, taken from the completed scan: 36 : 51 (CDKN2A homozygous deletion, $p = 0.414$), 48 : 39 (CDKN2A any-copy loss, $p = 0.552$) and 22 : 56 (BAP1 mutation, $p = 0.282$). The fourth is a balanced 180 : 180 design, which has no analogue in any of the cohorts analysed here and is included solely as a large-sample reference point; it is labelled synthetic wherever it is reported. Together: three mechanisms × four structures × five raw shifts

($\delta \in \{0, 0.3, 0.6, 0.9, 1.2\}$) = 60 calibration points, each with 2 000 independent replicates.

Realized effect size. At each calibration point δ is the *generating* shift, but reported comparisons are standardised by SD_{cohort}; the two are not equal, and by the identity above the realized d is a shrinking function of δ. We therefore computed the realized $\hat{d}$ from the generated samples at every replicate and compared empirical power **at that realized *d*** with the closed form. Comparing at the generating shift would have produced a spurious apparent failure of the approximation, and an earlier version of this calibration did exactly that before being corrected.

Tests. Empirical power was computed from the tie-corrected Mann–Whitney test — the same test used in the applied scan (formula in *Association testing* below) — with the Welch t-test carried alongside as a comparator. Rejection was at $\alpha = 0.05$, two-sided.

Empirical threshold. For each mechanism and group-size structure the empirical 80% threshold was obtained by bisecting on δ to a tolerance of 0.03 with 1 200 replicates per step, and the converged δ^* was then converted into the realized-d scale at 6 000 replicates before comparison with the closed-form d_{min}. The conversion step is required: δ and d differ by roughly 6% near the threshold of interest, so subtracting a δ-scale threshold from a d-scale threshold would have compared incommensurable quantities. An earlier version of the calibration omitted this conversion; the correction changed the apparent error at some configurations by a factor of four.

Seeds. Random draws are seeded from a CRC32 digest of the scenario label rather than from Python's built-in hash(). String hashing is salted per interpreter process, so a hash()-derived seed is not stable across runs and would have made the calibration irreproducible; the CRC32 derivation is stable and the entire grid regenerates bit-identically. This too was corrected after an initial version used hash().

Case study: TCGA malignant pleural mesothelioma

Data sources and processing

All case-study data are public and were retrieved from the Genomic Data Commons (GDC) [11] and the UCSC Xena TCGA pan-cancer compendium [12].

Layer	Source	Analysed size	Notes
mRNA expression	GDC / Xena per-cancer TCGA-MESO	20,530 genes × 87 samples	All samples primary tumour
Somatic mutation	GDC Masked Somatic Mutation (MAF)	2,709 variant records, 78 samples with evaluable calls	Merged from per-sample files
Copy number	GDC ASCAT3 gene-level	58,650 genes × 87 samples	See mapping note below
Protein (RPPA)	GDC RPPA	458 peptides × 62 samples	367 total-protein analytes after excluding 91 phospho-peptides
Clinical / histology	GDC clinical matrix	97 fields	Tumour site pleura in 100% of cases

Sample-type verification. Every expression sample carried primary-tumour barcodes (barcode positions 14–15 = 01). The cohort contains **no normal-tissue controls** (zero 11 samples), so no tumour-versus-normal contrast is available at any layer; all analyses are within-tumour driver-stratified contrasts, and no normal-adjacent interpretation is offered anywhere in this work.

Copy-number mapping. CNV files were mapped to samples through each file's associated_entities.entity_submitter_id rather than through the case-level cases.samples[0] field. The latter does not encode which aliquot a file belongs to and, in an initial pass, produced 41 primary-tumour and 45 blood-derived-normal assignments for a tumour-only cohort — an impossibility that the barcode check detected. The corrected mapping yielded 87/87 primary-tumour assignments. We report this because it is a reproducible failure mode of a widely used access pattern, and because an unnoticed version of it would have silently corrupted every copy-number result downstream.

Driver definitions

Nine driver variables were defined, spanning point mutation and copy-number loss (Table S1):

Mutation: BAP1, NF2, TP53, SETD2 — binary, from the merged MAF, evaluated in the 78 samples with mutation calls.

Homozygous deletion: CDKN2A, MTAP — gene-level copy number ≤ 0 (evaluated in 87 samples).

Any loss: CDKN2A, BAP1, NF2 — gene-level copy number ≤ 1.

Samples lacking an evaluable call at a given layer were excluded from that comparison only, and never counted as wild-type. For every driver, p was taken as the realized altered fraction in the analysed subset, not from an external estimate.

Phenotype (signature) scores

Eight phenotypes were scored as the mean expression of a fixed marker panel, computed on log_2-scale expression (Table S2):

Score	Markers
CD8_T	CD8A, CD8B, GZMA, GZMB, PRF1, CD3D, CD3E
CHECKPT	PDCD1, CD274, CTLA4, LAG3, HAVCR2, TIGIT
IFN_G	STAT1, IRF1, CXCL9, CXCL10, IFNG, IDO1, GBP1
MACROPHAGE	CD68, CD163, MRC1, CSF1R
FIBROBLAST	COL1A1, COL1A2, FAP, ACTA2, DCN
EMT	VIM, CDH2, FN1, SNAI1, SNAI2, TWIST1, ZEB1
PROLIF	MKI67, TOP2A, CCNB1, BUB1, AURKA
ANTIGEN	HLA-A, HLA-B, HLA-C, B2M, TAP1, TAP2, PSMB9

Panels were fixed from prior mesothelioma and pan-cancer immunogenomic literature [13, 14] and were not optimised against the outcome, which is a necessary condition for the per-test detection limits below to be interpretable.

Association testing and per-test detection limits

The complete scan is 9 drivers × 8 phenotypes = **72 comparisons**. Each comparison was tested by the two-sided Mann–Whitney *U* test with tie correction at $\alpha = 0.05$:

$$z = (U1 - n1\ n2/2)/(\sqrt{(n1\ n2/12[(N+1) - (\Sigma t\ (t3 - t))/(N(N-1))]}), \quad N = n_1 + n_2,$$

with the sum over tied groups. Effect size was computed as defined above, standardised by SD_{cohort} over all samples with evaluable data for that driver. Each comparison was then annotated with three derived quantities: its own d_{min} at n_1, n_2; the power at its own observed $|d|$; and a flag for whether $|d| \geq \hat{d}_{min}$. Multiplicity was assessed both by a Bonferroni threshold ($\alpha/72 = 6.94 \times 10^{-4}$) and by comparing the observed count of $p < 0.05$ to its chance expectation under a global null [15].

Specificity and confound controls

Two controls were applied to the associations that did reach significance.

Expression-matched null model. For a candidate association we asked whether its statistic exceeded what the same test produces for random gene sets of matched expression. Genes were ranked by mean expression across samples and partitioned into 20 expression strata; 1 000 random gene sets were drawn with the *same stratum composition* as the candidate panel, scored and tested identically, yielding a null distribution of the test statistic. A candidate was accepted only if it exceeded the null at the 95th percentile or above. This control addresses the fact that co-expressed, high-abundance immune panels will associate with one another regardless of any driver effect.

Histological confounding. Because MESO comprises epithelioid, biphasic and sarcomatoid histologies that differ in both driver spectrum and transcriptional programme, every retained association was re-estimated within the epithelioid subset alone, and within strata of the histological variable. An association that weakened on stratification was treated as confounded rather than reported.

Direction-of-effect check. For each negative result we verified that the observed effect fell below that comparison's own d_{min}, rather than merely failing to reach significance. This distinction is the applied counterpart of the methodological argument, and it is what permits a negative result here to be classified as *uninformative* rather than as evidence of absence.

Reproduction assessment

Reproduction was assessed against two published results for the same disease and, where possible, the same data source [16, 17, 18]. The criteria were fixed before the comparison was run: a reproduction was counted as **successful** if the direction of effect agreed, the association was significant at $\alpha = 0.05$ in our cohort, and the estimated effect fell within the 95% confidence interval reported by the original study; as **uninformative non-reproduction** if the direction of effect agreed but the comparison's own d_{min} exceeded the effect size claimed by the original study, so that the reported effect was not resolvable at this cohort size; and as **discordant** if the direction of effect was opposite under adequate power. Reporting the middle category requires the per-comparison d_{min} developed above; without it, all three cases collapse into the single label "failed to reproduce".

Software and reproducibility

All analyses were performed in Python 3.14 (NumPy 2.5, SciPy 1.18, pandas 3.0). The power surface, the simulation calibration and the applied scan are separate scripts, each with fixed seeds, and are provided as Additional file 1 together with the derived result tables. No proprietary or license-restricted software was used. Per the journal's Collection on reproducibility, transparency and open science, the analysis code is deposited at https://github.com/xiayu98wu/detection-frontier and archived on Zenodo (DOI to be inserted at acceptance) and the derived tables required to regenerate every figure and table are provided as additional files.

Results

1. The detection frontier and its calibration

1.1 The frontier

For a two-group comparison of total size n with the driver altered in a proportion p of the cohort, the smallest standardised effect detectable at 80% power with a two-sided $\alpha = 0.05$ spans a wide range across the sizes and prevalences that occur in practice (Fig. 1, Additional file 2). At prevalence $p = 0.30$ — close to the alteration frequencies of the principal mesothelioma drivers — the frontier falls from $d_{min} = 0.97$ at $n = 40$ to 0.66 at $n = 87$, 0.50 at $n = 150$, 0.35 at $n = 300$ and 0.25 at $n = 600$.

Prevalence matters as much as cohort size. At a fixed cohort size the frontier is governed by the non-centrality parameter of the comparison, $\lambda = d\sqrt{(n1n2/(n1+n2))}$, which for $n_1 = np$ and $n_2 = n(1-p)$ is approximately $d\sqrt{(np(1-p))}$. Because $p(1-p)$ is maximised at an even split and falls toward zero at both extremes, an unbalanced comparison wastes samples: at $n = 87$ the frontier at $p = 0.40$ is 0.61, while at $p = 0.10$ it rises to 0.99, close to the ratio $1/\sqrt{(p(1-p))}$ that the approximation predicts. A driver altered in one tenth of the cohort therefore requires an effect about 60% larger than one altered in two fifths of it, for the same number of samples.

A second, separate consequence of the standardisation bounds how large an effect can be expressed at all. Because d is scaled by the standard deviation of the whole cohort and the between-group shift contributes to that standard deviation, d cannot exceed $1/\sqrt{(p(1-p))}$ whatever the underlying raw difference. This ceiling is loosest for rare drivers and is not what makes them hard to test — at $p = 0.10$ it lies at 3.33, far above any frontier value — but it does mean that a very large standardised effect and a rare driver are mutually exclusive, which matters when interpreting extreme effect sizes in small strata.

Inverting the frontier gives the sample size a study would need. A large effect ($d = 0.90$)

requires 46 samples at $p = 0.30$; a medium effect ($d = 0.50$) requires 150; a small effect ($d = 0.30$) requires 416 (Table 1).

1.2 Simulation calibration

The closed form rests on a normal approximation, so we calibrated it against empirical power under three data-generating mechanisms and four group-size configurations (60 calibration points, 2 000 replicates each).

Under ideal-normal data the approximation is essentially exact: the empirical and closed-form 80% thresholds differ by at most 0.009 across all four configurations (Fig. 5a, Table 2). Under the two non-normal mechanisms the closed form errs consistently in one direction. For gene-set scores built from log-normal marker values — the shape an expression signature actually has — the closed form over-estimates the threshold by 0.024 to 0.060. For dropout-contaminated data, which carries the heavy left tail and tied ranks of real assay measurements, it over-estimates by 0.037 to 0.086. Nine of the twelve configurations lie below the identity line (Fig. 5b).

The direction is the important part. An over-estimated threshold means the frontier never reports an association as detectable when the data could not have resolved it; the cost is that it may occasionally demand slightly more samples than strictly necessary. At the level of power rather than threshold, the discrepancy is larger — up to +0.119 in the dropout-contaminated regime, where the rank-based test outperforms the normal approximation in heavy tails — but the quantity the frontier is used for, the 80% crossing point, moves by at most 0.086.

2. Reachability across existing cohorts

Applying the frontier at $p = 0.30$ to the fifteen rare-tumour TCGA cohorts large enough to be analysed at all, fourteen can detect a large effect ($d = 0.9$). The exception is CHOL, at $n = 36$, whose detection limit of 1.01 exceeds even that.

Six of the fifteen can detect a medium effect ($d = 0.5$): TGCT (150), PCPG (179), ESCA (184), KIRP (290), CESC (303) and BLCA (407). The remaining nine cannot, and malignant pleural mesothelioma (MESO, $n = 87$) sits among them, with a detection limit of 0.66.

No cohort can detect a small effect ($d = 0.3$) at $p = 0.30$ (Fig. 2b, Table 3). The smallest cohort that would is not represented in TCGA at all; the required sample size of 416 exceeds every rare-tumour cohort in the compendium.

3. Case study: a complete driver–phenotype scan

3.1 The scan

We crossed nine driver variables with eight expression signatures in TCGA-MESO, giving 72 comparisons (Table 4). Thirteen reached nominal significance at $\alpha = 0.05$ against 3.6 expected by chance if no driver were associated with any signature. Five survived a Bonferroni threshold of 6.94×10^{-4}, all of them involving CDKN2A or its neighbour MTAP.

Not one of the 72 comparisons could detect an effect as small as $d = 0.40$ with 80% power. The minimum detection limit across the whole scan was 0.60.

3.2 Reading each result against its own limit

Comparing every observed effect with the detection limit specific to that comparison gives the picture in Fig. 3. Seven comparisons (9.7%) had an observed effect at or above their own limit, and all seven were statistically significant. Sixty-five comparisons (90.3%) had an observed effect below their own limit.

The 59 comparisons that were not significant are the informative group. Every one of them lies below its own detection limit; not a single non-significant comparison exceeded the effect size its group sizes could have resolved. In other words the scan produced no instance of a genuine null — no case where a comparison was adequately powered and still found nothing. The negative results in this cohort are uniformly uninformative, and the reason is visible before any

data are examined.

The spread within a single cohort is large. CDKN2A any-copy loss against the proliferation signature reached $d = 1.171$ with 99.97% power, while BAP1 mutation against the $CD8^+$ T-cell score reached $d = 0.011$ with 5.0% power — barely above the 5% that a test of no effect would produce. The same 87 samples support a decisive positive result in one comparison and no information whatever in another.

3.3 Reproducible and unreproducible effects in the same cohort

Two published associations for this disease allow the frontier's verdicts to be checked against an external standard (Fig. 4).

The first is reproducible. CDKN2A homozygous deletion was present in 36 of 87 samples (41.4%), against 41.5% reported in an independent series of 89 cases [16]; the difference is 0.12 percentage points, and the association between deletion and the proliferation signature was among the largest effects in the scan ($d = 0.898$, $p = 1.5 \times 10^{-5}$). Both the prevalence and the downstream association agree with the published account, and the effect size concerned is the largest class the cohort can address.

The second is not reproducible, and the frontier says why. BAP1 mutation has been reported to reshape the immune microenvironment of mesothelioma [17, 18]. In this cohort, BAP1 mutation against the $CD8^+$ T-cell score gave $d = 0.011$, $p = 0.67$ — and the comparison's own detection limit is 0.705, some sixty times the effect observed. The reported literature effects are of medium size, well below 0.705. This cohort could not have detected them. The result is not a failed reproduction; it is a comparison that cannot reproduce anything, and that fact was determinable from the group sizes alone.

The distinction matters for how the literature reads. Reported as "BAP1 mutation was not associated with immune infiltration in TCGA-MESO", this result would count against the published association. Reported against its detection limit, it carries no evidence either way.

3.4 Three molecular layers agree on the null

Because the negative result above is central to the argument, we asked whether it reflects the mutation calls rather than a genuine absence of association, by estimating the same relationship independently at three layers (Fig. 4b).

At the DNA layer, BAP1 mutation against the $CD8^+$ T-cell score gave $d = +0.011$ ($p = 0.67$). At the RNA layer, BAP1 mRNA against the same score gave Spearman $\rho = -0.044$ ($p = 0.69$). At the protein layer, BAP1 measured by reverse-phase protein array against the same score gave $\rho = +0.077$ ($p = 0.55$). All three estimates lie within 0.08 of zero, and none approaches significance.

A control argues against the mutation calls being simply uninformative. The same pipeline applied to the same MAF recovers the expected loss of transcript for the other drivers: NF2 mutation gives −0.49 ($p = 0.010$), TP53 −0.63 ($p = 0.0016$) and SETD2 −0.96 ($p = 4.5 \times 10^{-4}$). The assay detects nonsense-mediated and loss-of-function effects where they exist; it does not detect one for BAP1, and neither does the protein layer.

Extending the protein-layer test to all eight signatures, one of eight reaches nominal significance (fibroblast signature, $\rho = +0.328$, $p = 0.007$), which does not survive correction for eight tests (Bonferroni threshold 0.0063). CDKN2A and MTAP are absent from the RPPA panel, so the deletion association that did reproduce could not be examined at the protein layer.

4. Summary of the empirical picture

In one small cohort, scanned exhaustively and read against its own detection limits:

One association reproduced, and it is a large effect.

One association failed to reproduce, and it is a medium effect whose detection limit exceeds it roughly sixtyfold.

Ninety per cent of the comparisons carry no information about their driver.

No comparison anywhere in the scan was powered to detect an effect of the size the literature most commonly reports.

The pattern is not a property of this cohort's quality. It is the arithmetic of n and p, and it can be computed before the data are collected.

Discussion

A third class of result, and why it needs a name

The reproducibility literature has a rich vocabulary for results that fail to replicate, and a well-developed statistical apparatus for predicting, in aggregate, which findings will survive. What it lacks is a per-result criterion separating a failure that constitutes evidence from one that constitutes none.

The distinction has a precise form. For a comparison with realised group sizes n_1 and n_2, let d_{min} be the smallest standardised effect detectable at the study's intended power, and let d_{lit} be the effect size the published claim implies. If $d_{lit} < d_{min}$, a non-significant result is uninformative: the study could not have found the effect even if it were present and exactly as reported. This condition is computable from n_1, n_2 and d_{lit} alone, before any data exist.

Our scan illustrates how far the two classes can diverge within one cohort. Of 72 comparisons, 59 were non-significant, and **all 59 fell below their own detection limit**. Not one was adequately powered and still null. The cohort produced no evidence of absence anywhere; it produced 59 results that look like evidence of absence and are not. Reporting each with its detection limit converts a body of apparent negative findings into an accurate statement of what the cohort could address — which in this case was very little.

What the frontier says that sample-size conventions do not

Conventional effect-size categories — small, medium, large — are not tied to any particular

study design, and that is the source of the difficulty. Our results show how far a real cohort can be from resolving the category the literature most often reports. Across the fifteen rare-tumour TCGA cohorts, six can detect a medium effect ($d = 0.5$) and none can detect a small one ($d = 0.3$); the 416 samples needed for a small effect at 30% prevalence exceed every cohort in the compendium. In the case study, not one of 72 comparisons could detect even $d = 0.40$.

The frontier also makes visible a constraint that is not about sample size at all. Because power depends on $d\sqrt{(np(1-p))}$, an unbalanced split wastes samples, and a driver altered in 10% of a cohort requires an effect about 60% larger than one altered in 40% of it. Enlarging a cohort does not remove this penalty; it applies at any n. For rare drivers, prevalence and sample size are separate obstacles, and only one of them is addressed by recruiting more patients.

A reporting recommendation

The practical upshot is a small change to how negative driver–phenotype results are reported. We suggest that any study reporting a non-significant association between a molecular alteration and a phenotype state, alongside the effect estimate and its confidence interval, the smallest effect the comparison could have detected at the study's intended power. The computational cost is negligible — the frontier is a closed-form expression in n and p — and the value is that it distinguishes "we found no association" from "we could not have found one".

We do not propose this as a substitute for power calculation at the design stage, but as a reporting requirement for results already in hand, since the information needed is available in every study that reports group sizes and prevalence. Where the detected limit exceeds the effect size the literature reports, the appropriate description is that the comparison was underpowered to address the question, not that the association was absent.

Implications for the mesothelioma literature specifically

BAP1 mutation has been reported to reshape the immune microenvironment of malignant pleural mesothelioma, on the basis of immunohistochemical and cohort evidence [17, 18]. In TCGA-MESO the corresponding comparison has a detection limit of 0.705, while the observed effect is 0.011 — the test carries no information about the claim. The protein layer is stricter still, not merely comparable: RPPA covers 62 of the 87 samples, and at the BAP1 prevalence of this cohort the protein-layer detection limit is 0.80, above the DNA layer's 0.705. That layer is therefore even further from being able to address the reported association.

This is not a refutation, and we are careful not to present it as one. The published reports may well be correct; our cohort simply cannot speak to them. What follows is a caution about how such a result should be read if it appears in a meta-analysis or a review: a negative TCGA-MESO finding on BAP1 and immune phenotype should be excluded from any evidence synthesis for this question, because it contributes no information.

By contrast, the association between CDKN2A deletion and proliferative programmes reproduced closely: a deletion frequency of 41.4% against 41.5% in an independent series [16], with one of the largest effect sizes in the scan. The contrast between these two cases, in the same cohort, is the paper's empirical core. Which published association a small cohort can address is determined by the effect size, not by the cohort's quality.

What to do when the cohort cannot be enlarged

The frontier is diagnostic, not therapeutic. For most rare tumours, recruiting 416 patients is impossible, and a recommendation to do so is not useful. Several alternatives remain, and the frontier helps choose among them.

Pool across cohorts. The power penalty for a given driver falls with total n, so consortium-level pooling is the most direct remedy, at the cost of between-cohort heterogeneity that must itself be modelled.

Choose questions the cohort can answer. The same cohort that cannot test a medium effect

can test a large one, and our scan shows this concretely: CDKN2A deletion against proliferation reached 99.97% power. Designing the analysis around effects the cohort can resolve is more productive than testing every available driver–phenotype pair.

Pre-register the detection limit. Because d_{min} depends only on n and p, it can be stated before the analysis, together with the effect size the study is powered to detect rather than merely the alpha it will use. This converts an after-the-fact excuse into an ex ante commitment, and is the natural place for the frontier to enter the reporting standard.

Report sensitivity to the assumption. Where a plausible effect size can be bounded from prior literature, reporting the power at that effect — as we do for each of the 72 comparisons — lets readers judge informativeness without recomputing anything.

Limitations

The frontier rests on a normal approximation, and we have calibrated rather than exactly solved it. The simulation shows the approximation to be close under ideal normality and, under the deviations we tested — skew and dropout contamination — to over-estimate the limit in every non-normal configuration, with the empirical threshold below the closed-form value in nine of twelve configurations overall and the largest discrepancy 0.086. Conservatism is the safe direction for this purpose, since it never labels a comparison resolvable when it is not, but it does mean the required sample sizes we report may be slightly generous. The three configurations that err the other way are all ideal-normal, and all by at most 0.009.

The case study is a single tumour type. It is, however, a complete scan of one cohort rather than a selection, and the pattern it shows — negatives that are uniformly uninformative — follows from group sizes and prevalences that are not specific to mesothelioma. The cohort itself is a TCGA cohort, and shares the ascertainment biases of that programme; in particular it contains no normal-tissue controls and no peritoneal cases, so the analyses here are within-tumour contrasts only and cannot address histology-specific effects.

Finally, the frontier addresses statistical detectability given a chosen test and alpha. It does not address whether the phenotype signature is a valid measure of the biological process, whether the driver variable is correctly defined, or whether the comparison is confounded. These remain capable of producing wrong answers at any sample size, and a comparison that clears its detection limit is resolvable, not correct.

Conclusions

In small cancer cohorts, a failure to detect a driver–phenotype association is not the same as evidence that the association is absent. This distinction is ordinarily invisible in the literature, but it is decidable before a study begins, from the cohort size and the driver's prevalence alone — and in the cohort we examined, it reclassified every negative result as uninformative.

List of abbreviations

ASCAT3 Allele-Specific Copy number Analysis of Tumours, version 3

CNV Copy-number variation

CRC32 Cyclic redundancy check, 32-bit (used here to derive stable random seeds)

GDC Genomic Data Commons

GWAS Genome-wide association study

MAF Mutation Annotation Format

MESO Malignant pleural mesothelioma (TCGA cohort code)

mRNA Messenger RNA

RPPA Reverse-phase protein array

SD Standard deviation

TCGA The Cancer Genome Atlas

Xena UCSC Xena visual exploration and analysis platform

Declarations

Ethics approval and consent to participate

Not applicable. This study analysed only previously published, de-identified public data resources. No human participants or animals were involved.

Consent for publication

Not applicable.

Availability of data and materials

All datasets analysed during this study are publicly available. Expression, mutation, copy-number, protein and clinical data for TCGA malignant pleural mesothelioma were obtained from the Genomic Data Commons [11] and the UCSC Xena TCGA pan-cancer compendium [12]. The analysis code supporting the conclusions of this article is available in the GitHub repository, https://github.com/xiayu98wu/detection-frontier, and archived on Zenodo (DOI to be inserted at acceptance). Derived result tables and figure source data are included in that repository.

Competing interests

The authors declare no competing interests.

Clinical trial number

Not applicable.

Funding

This work received no specific funding.

Authors' contributions

Conceptualization: Xueyi Lin and Xiayu Wu; investigation (data acquisition, formal analysis, evidence integration and visualization of figures): Xueyi Lin; writing - original draft: Xiayu

Wu; writing - review and editing: Xiayu Wu and Sanjun Zhao; supervision: Xiayu Wu and Sanjun Zhao. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

Not applicable.

Authors' information

Not applicable.

References

1. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14(5):365–76.
2. Stanley TD, Carter EC, Doucouliagos H. What meta-analyses reveal about the replicability of psychological research. Psychol Bull. 2018;144(12):1325–46.
3. Marek S, Tervo-Clemmens B, Calabro FJ, Montez DF, Kay BP, Hatoum AS, et al. Reproducible brain-wide association studies require thousands of individuals. Nature. 2022;603(7902):654–60.
4. Kraft P. Curses—winner's and otherwise—in genetic epidemiology. Epidemiology. 2008;19(5):649–51.
5. Xiao R, Boehnke M. Quantifying and correcting for the winner's curse in genetic association studies. Genet Epidemiol. 2009;33(5):453–62.
6. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ. 1995;311(7003):485.
7. Nakamura H, Arai Y, Totoki Y, Shirota T, Elzawahry A, Kato M, et al. Genomic spectra

of biliary tract cancer. Nat Genet. 2015;47(9):1003–10.

8. Schmitz R, Wright GW, Huang DW, Johnson CA, Phelan JD, Wang JQ, et al. Genetics and pathogenesis of diffuse large B-cell lymphoma. N Engl J Med. 2018;378(15):1396–407.

9. Fennell DA, Sekido Y, Baas P, Husain AN, Curioni-Fontecedro A, Lim E, et al. Pleural mesothelioma. Nat Rev Dis Primers. 2025;11(1):56.

10. Hmeljak J, Sanchez-Vega F, Hoadley KA, Shih J, Stewart C, Heiman D, et al. Integrative molecular characterization of malignant pleural mesothelioma. Cancer Discov. 2018;8(12):1548–65.

11. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12.

12. Goldman MJ, Craft B, Hastie M, Repecka K, McDade F, Kamath A, et al. Visualizing and interpreting cancer genomics data via the Xena platform. Nat Biotechnol. 2020;38(6):675–8.

13. Thorsson V, Gibbs DL, Brown SD, Wolf D, Bortone DS, Ou Yang TH, et al. The immune landscape of cancer. Immunity. 2018;48(4):812–830.e14.

14. Hiltbrunner S, Mannarino L, Kirschner MB, Opitz I, Rigutto A, Laure A, et al. Tumor immune microenvironment and genetic alterations in mesothelioma. Front Oncol. 2021;11:660039.

15. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. Stat Med. 2014;33(11):1946–78.

16. Torricelli F, Donati B, Manicardi V, Gugnoni M, Reggiani F, Manzotti G, et al. CDKN2A deletion is associated with immune desertification in diffuse pleural mesothelioma. J Exp Clin Cancer Res. 2025;44(1):256.

17. Xu D, Gao Y, Yang H, Spils M, Marti TM, Losmanova T, et al. BAP1 deficiency

inflames the tumor immune microenvironment and is a candidate biomarker for immunotherapy response in malignant pleural mesothelioma. JTO Clin Res Rep. 2024;5(5):100672.

18. Dagogo-Jack I, Mitchell O, Codd E, Li A, Mitchell D, Flynn SE, et al. Immune composition and immunotherapy outcomes of mesothelioma with BAP1, CDKN2A, MTAP, and NF2 alterations. J Thorac Oncol. 2025;20(11):1615–25.

Figure legends

Figure 1

Minimum detectable standardised effect by cohort size and driver prevalence

The detection frontier $d_{min}(n,p)$. Each cell gives the smallest standardised effect — the group mean difference divided by the standard deviation of the whole cohort — that a two-group comparison of total size n, with the driver altered in a proportion p of the cohort, can detect with 80% power at a two-sided $\alpha = 0.05$. Group sizes are $n_1 = \text{round}(np)$ and $n_2 = n - n_1$; d_{min} is obtained by bisection on the effect size under the normal approximation to the two-sample comparison. Cells shaded green mark comparisons whose $d_{min} \leq 0.50$, that is, those able to resolve an effect of the size conventionally described as medium. The red box marks the grid cell nearest to the cohort analysed in this study (TCGA malignant pleural mesothelioma); that cohort is $n = 87$ with a CDKN2A deletion prevalence of 0.41, which falls between grid cells, and its exact d_{min} is 0.61. Values rise steeply as prevalence falls: at any fixed n, a driver altered in 10% of the cohort requires an effect about 1.7 times as large as one altered in 50%. Rows are cohort sizes from 40 to 1200; columns are prevalences from 0.50 to 0.10. No value is interpolated — every cell is an independent solution of the power equation.

Figure 2

Most TCGA rare-tumour cohorts cannot detect the effects they are used to test

Reachability of a driver–phenotype association as a function of cohort size. (a) The detection limit d_{min} at prevalence $p = 0.30$ for each of fifteen TCGA rare-tumour cohorts, plotted against cohort size on a logarithmic axis. Shaded bands mark the effect-size ranges conventionally described as small ($d = 0.3$), medium ($d = 0.5$) and large ($d = 0.9$); dashed horizontal lines mark the two boundaries. Each point is one cohort; the case-study cohort (MESO) and the largest cohort (BLCA) are labelled. (b) For each cohort, whether an association of each effect size can be detected at 80% power with $\alpha = 0.05$ at $p = 0.30$. Green cells (Y) indicate that detection is possible, grey cells (N) that it is not. Fourteen of the fifteen cohorts can detect a large effect; the exception is CHOL ($n = 36$), whose detection limit of 1.01 exceeds $d = 0.9$. Six of fifteen can detect a medium effect. None can detect a small effect. Cohort sizes are the numbers of primary tumours with expression data in the TCGA pan-cancer compendium, and detection limits are computed from the same closed-form model as in Figure 1.

Figure 3

Seventy-two driver–phenotype comparisons against their own detection limits

Every driver–phenotype comparison in one cohort, plotted against the detection limit specific to that comparison. Points are the 72 comparisons formed by crossing nine driver variables (point mutations in BAP1, NF2, TP53 and SETD2; homozygous deletion of CDKN2A and MTAP; any-copy loss of CDKN2A, BAP1 and NF2) with eight expression signatures ($CD8^+$ T cells, checkpoint transcripts, interferon-γ response, macrophages, fibroblasts, epithelial–mesenchymal transition, proliferation and antigen presentation) in TCGA malignant pleural mesothelioma ($n = 87$). The horizontal axis is the absolute observed standardised effect; the vertical axis is the smallest effect that comparison could detect at 80% power given its own group sizes. The dashed diagonal is $|d| = d_{min}$. Points below the diagonal carry no information about the driver: the test could not have detected an effect of the size actually present, so a non-

significant result is uninformative rather than evidence of absence. Filled green points are the seven comparisons (9.7%) whose observed effect met or exceeded their own detection limit; each of these is statistically significant. All 59 non-significant comparisons lie below the diagonal.

Figure 4

Reproduction succeeds for a large effect and is uninformative for a medium one

Reproduction and three-layer evidence in TCGA mesothelioma. (a) CDKN2A homozygous deletion frequency in this study (36 of 87 samples, 41.4%) and in the published series against which it was compared (41.5%, n = 89); the difference is 0.12 percentage points. (b) Association between BAP1 and the $CD8^+$ T-cell expression score, estimated independently at three molecular layers: mutation status (standardised mean difference d), BAP1 mRNA (Spearman ρ) and BAP1 protein measured by reverse-phase protein array (Spearman ρ). All three estimates lie close to zero and none is significant at $\alpha = 0.05$; the shaded band marks $|effect| < 0.20$. (c) Of the 72 comparisons in Figure 3, seven have an observed effect at or above their own detection limit and 65 fall below it. (d) BAP1 protein against each of the eight signatures; the shaded band marks $|\rho| < 0.20$. One of the eight tests reaches nominal significance (fibroblast signature, $\rho = +0.328$, $p = 0.007$), which does not survive correction for eight tests (Bonferroni threshold 0.0063). CDKN2A and MTAP are absent from the protein panel, so the deletion association that did reproduce could not be examined at the protein layer.

Figure 5

Closed-form detection limits are calibrated and err conservatively

Calibration of the closed-form detection limit by simulation. (a) Empirical 80%-power thresholds against the closed-form values, for three data-generating mechanisms and four group-size configurations. The dashed line is the identity line and the grey band is ±0.03. Points

below the line indicate that the closed form is conservative. The group-size configurations are the realised driver strata of the case study (36:51, 48:39 and 22:56) together with a synthetic balanced 180:180 design that has no counterpart among the cohorts analysed. (b) Signed error, empirical minus closed form, at the threshold, by mechanism. Each point is one configuration; boxes show the median and interquartile range. Ideal-normal data deviate by at most 0.009. Under skew and under dropout contamination the closed form over-estimates the limit by between 0.02 and 0.09, and it does so in all eight non-normal configurations: in that regime it never reports an association as detectable when the data could not have resolved it, at the cost of occasionally requiring slightly more samples than strictly necessary. Nine of the twelve configurations lie below zero, and the largest absolute discrepancy is 0.086.

Tables

Table 1

Minimum cohort size required to detect each effect class at 80% power

Effect class	d	p = 0.50	p = 0.30	p = 0.20	p = 0.10
Large	0.90	40	46	64	106
Medium	0.50	126	150	198	346
Small	0.30	350	416	546	966

Group sizes are n1 = round(np) and n2 = n - n1; two-sided alpha = 0.05.

Table 2

Simulation calibration of the closed-form detection limit

Mechanism	Group sizes	Empirical d_min	Closed-form d_min	Error
Ideal normal	CDKN2A homozygous deletion 36:51	0.614	0.610	+0.004
Ideal normal	CDKN2A any-copy loss 48:39	0.602	0.604	-0.002
Ideal normal	BAP1 mutation 22:56	0.710	0.705	+0.005

Mechanism	Group sizes	Empirical d_min	Closed-form d_min	Error
Ideal normal	Synthetic balanced 180:180	0.304	0.295	+0.009
Skewed gene-set score	CDKN2A homozygous deletion 36:51	0.568	0.610	-0.041
Skewed gene-set score	CDKN2A any-copy loss 48:39	0.554	0.604	-0.050
Skewed gene-set score	BAP1 mutation 22:56	0.645	0.705	-0.060
Skewed gene-set score	Synthetic balanced 180:180	0.271	0.295	-0.024
Dropout-contaminated	CDKN2A homozygous deletion 36:51	0.537	0.610	-0.072
Dropout-contaminated	CDKN2A any-copy loss 48:39	0.535	0.604	-0.069
Dropout-contaminated	BAP1 mutation 22:56	0.619	0.705	-0.086
Dropout-contaminated	Synthetic balanced 180:180	0.258	0.295	-0.037

2000 replicates per calibration point; 1200 per bisection step, 6000 at the converged threshold.

Empirical and closed-form values are both in units of the whole-cohort standardised effect.

Negative error indicates the closed form over-estimates the limit, which is the safe direction.

Table 3

Reachability of each effect class in fifteen rare-tumour TCGA cohorts

Cohort	n	d_min at p = 0.30	Large (d = 0.9)	Medium (d = 0.5)	Small (d = 0.3)
CHOL	36	1.014	No	No	No
DLBC	48	0.890	Yes	No	No
UCS	57	0.811	Yes	No	No
KICH	66	0.750	Yes	No	No
ACC	79	0.685	Yes	No	No
UVM	80	0.684	Yes	No	No
MESO	87	0.656	Yes	No	No
READ	94	0.632	Yes	No	No
THYM	120	0.558	Yes	No	No
TGCT	150	0.499	Yes	Yes	No
PCPG	179	0.456	Yes	Yes	No
ESCA	184	0.451	Yes	Yes	No
KIRP	290	0.359	Yes	Yes	No

Cohort	n	d_min at p = 0.30	Large (d = 0.9)	Medium (d = 0.5)	Small (d = 0.3)
CESC	303	0.351	Yes	Yes	No
BLCA	407	0.303	Yes	Yes	No

Cohort sizes are primary tumours with expression data. Reachability is evaluated at a representative prevalence of 0.30 with 80% power and two-sided alpha = 0.05. Six of fifteen cohorts can detect a medium effect; none can detect a small effect.

Table 4

All 72 driver x phenotype comparisons in TCGA malignant pleural mesothelioma

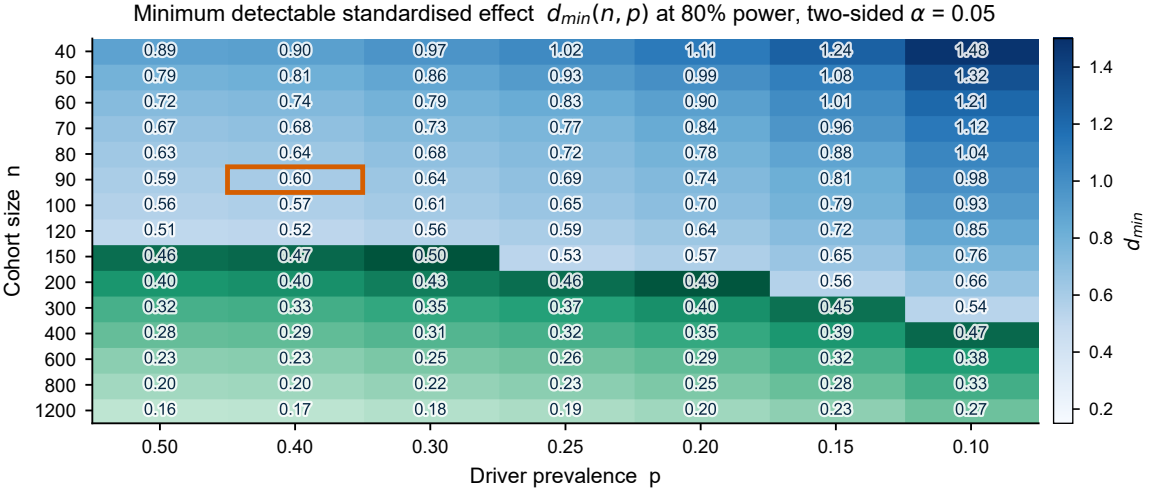
Driver	Signature	n altered	n wild type	Prevalence	d_min	Observed d	Power	p	Detectable
BAP1_mut	CD8_T	22	56	0.282	0.705	+0.011	5.02%	0.665	No
BAP1_mut	CHECKPT	22	56	0.282	0.705	+0.067	5.82%	0.505	No
BAP1_mut	IFN_G	22	56	0.282	0.705	+0.037	5.24%	0.747	No
BAP1_mut	MACROPHAGE	22	56	0.282	0.705	-0.258	17.61%	0.262	No
BAP1_mut	FIBROBLAST	22	56	0.282	0.705	-0.266	18.47%	0.226	No
BAP1_mut	EMT	22	56	0.282	0.705	-0.280	19.95%	0.307	No
BAP1_mut	PROLIF	22	56	0.282	0.705	-0.308	23.21%	0.380	No
BAP1_mut	ANTIGEN	22	56	0.282	0.705	-0.024	5.11%	0.885	No
NF2_mut	CD8_T	19	59	0.244	0.739	+0.124	7.55%	0.418	No
NF2_mut	CHECKPT	19	59	0.244	0.739	+0.310	21.76%	0.314	No
NF2_mut	IFN_G	19	59	0.244	0.739	+0.341	25.26%	0.115	No
NF2_mut	MACROPHAGE	19	59	0.244	0.739	+0.380	30.26%	0.057	No
NF2_mut	FIBROBLAST	19	59	0.244	0.739	-0.060	5.60%	0.958	No
NF2_mut	EMT	19	59	0.244	0.739	+0.211	12.56%	0.386	No
NF2_mut	PROLIF	19	59	0.244	0.739	+0.240	14.87%	0.276	No
NF2_mut	ANTIGEN	19	59	0.244	0.739	+0.182	10.60%	0.418	No

Driver	Signature	n altered	n wild type	Prevalence	d_min	Observed d	Power	p	Detectable
TP53_mut	CD8_T	11	67	0.141	0.911	-0.259	12.49 %	0.417	No
TP53_mut	CHECKPT	11	67	0.141	0.911	-0.243	11.59 %	0.542	No
TP53_mut	IFN_G	11	67	0.141	0.911	-0.207	9.77%	0.768	No
TP53_mut	MACROPHAGE	11	67	0.141	0.911	+0.039	5.17%	0.736	No
TP53_mut	FIBROBLAST	11	67	0.141	0.911	+0.308	15.75 %	0.340	No
TP53_mut	EMT	11	67	0.141	0.911	+0.184	8.74%	0.581	No
TP53_mut	PROLIF	11	67	0.141	0.911	+0.254	12.20 %	0.451	No
TP53_mut	ANTIGEN	11	67	0.141	0.911	+0.153	7.57%	0.551	No
SETD2_mut	CD8_T	8	70	0.103	1.046	-0.154	6.96%	0.779	No
SETD2_mut	CHECKPT	8	70	0.103	1.046	+0.126	6.31%	0.668	No
SETD2_mut	IFN_G	8	70	0.103	1.046	+0.222	9.14%	0.499	No
SETD2_mut	MACROPHAGE	8	70	0.103	1.046	-0.128	6.37%	0.843	No
SETD2_mut	FIBROBLAST	8	70	0.103	1.046	-0.533	29.74 %	0.130	No
SETD2_mut	EMT	8	70	0.103	1.046	-0.714	48.15 %	0.0500	No
SETD2_mut	PROLIF	8	70	0.103	1.046	-0.367	16.59 %	0.236	No
SETD2_mut	ANTIGEN	8	70	0.103	1.046	+0.043	5.15%	0.961	No
CDKN2A_del	CD8_T	36	51	0.414	0.610	-0.440	52.53 %	0.043	No
CDKN2A_del	CHECKPT	36	51	0.414	0.610	-0.391	43.43 %	0.125	No
CDKN2A_del	IFN_G	36	51	0.414	0.610	-0.331	33.01 %	0.107	No
CDKN2A_del	MACROPHAGE	36	51	0.414	0.610	-0.307	29.21 %	0.134	No
CDKN2A_del	FIBROBLAST	36	51	0.414	0.610	+0.641	83.80 %	8.52 x 10^-4	Yes
CDKN2A_del	EMT	36	51	0.414	0.610	+0.839	97.09 %	5.50 x 10^-5	Yes
CDKN2A_del	PROLIF	36	51	0.414	0.610	+0.898	98.48 %	1.46 x 10^-5	Yes

Driver	Signature	n altered	n wild type	Prevalence	d_min	Observed d	Power	p	Detectable
CDKN2A_del	ANTIGEN	36	51	0.414	0.610	-0.573	74.93 %	0.010	No
MTAP_del	CD8_T	32	55	0.368	0.623	-0.427	48.47 %	0.058	No
MTAP_del	CHECKPT	32	55	0.368	0.623	-0.245	19.67 %	0.369	No
MTAP_del	IFN_G	32	55	0.368	0.623	-0.203	14.99 %	0.291	No
MTAP_del	MACROPHAGE	32	55	0.368	0.623	-0.208	15.48 %	0.379	No
MTAP_del	FIBROBLAST	32	55	0.368	0.623	+0.416	46.42 %	0.030	No
MTAP_del	EMT	32	55	0.368	0.623	+0.660	84.39 %	0.002	Yes
MTAP_del	PROLIF	32	55	0.368	0.623	+0.930	98.68 %	7.15 x 10^-6	Yes
MTAP_del	ANTIGEN	32	55	0.368	0.623	-0.496	60.74 %	0.028	No
CDKN2A_loss	CD8_T	48	39	0.552	0.604	-0.293	27.45 %	0.242	No
CDKN2A_loss	CHECKPT	48	39	0.552	0.604	-0.253	21.61 %	0.339	No
CDKN2A_loss	IFN_G	48	39	0.552	0.604	-0.234	19.26 %	0.242	No
CDKN2A_loss	MACROPHAGE	48	39	0.552	0.604	-0.324	32.45 %	0.177	No
CDKN2A_loss	FIBROBLAST	48	39	0.552	0.604	+0.509	65.58 %	0.008	No
CDKN2A_loss	EMT	48	39	0.552	0.604	+0.814	96.52 %	1.18 x 10^-4	Yes
CDKN2A_loss	PROLIF	48	39	0.552	0.604	+1.171	99.97 %	9.73 x 10^-9	Yes
CDKN2A_loss	ANTIGEN	48	39	0.552	0.604	-0.468	58.43 %	0.038	No
BAP1_cnvloss	CD8_T	33	54	0.379	0.619	+0.132	9.19%	0.637	No
BAP1_cnvloss	CHECKPT	33	54	0.379	0.619	+0.342	34.11 %	0.085	No
BAP1_cnvloss	IFN_G	33	54	0.379	0.619	+0.251	20.55 %	0.426	No
BAP1_cnvlo	MACROPHA	33	54	0.379	0.619	-0.088	6.86%	0.733	No

Driver	Signature	n altered	n wild type	Prevalence	d_min	Observed d	Power	p	Detectable
ss	GE								
BAP1_cnvloss	FIBROBLAST	33	54	0.379	0.619	-0.311	29.05 %	0.159	No
BAP1_cnvloss	EMT	33	54	0.379	0.619	-0.161	11.27 %	0.558	No
BAP1_cnvloss	PROLIF	33	54	0.379	0.619	+0.069	6.11%	0.861	No
BAP1_cnvloss	ANTIGEN	33	54	0.379	0.619	+0.405	45.02 %	0.100	No
NF2_cnvloss	CD8_T	44	43	0.506	0.601	-0.245	20.80 %	0.333	No
NF2_cnvloss	CHECKPT	44	43	0.506	0.601	-0.306	29.77 %	0.238	No
NF2_cnvloss	IFN_G	44	43	0.506	0.601	-0.244	20.62 %	0.222	No
NF2_cnvloss	MACROPHAGE	44	43	0.506	0.601	-0.394	45.18 %	0.086	No
NF2_cnvloss	FIBROBLAST	44	43	0.506	0.601	+0.154	11.08 %	0.581	No
NF2_cnvloss	EMT	44	43	0.506	0.601	+0.268	23.97 %	0.166	No
NF2_cnvloss	PROLIF	44	43	0.506	0.601	+0.075	6.40%	0.966	No
NF2_cnvloss	ANTIGEN	44	43	0.506	0.601	-0.174	12.80 %	0.368	No

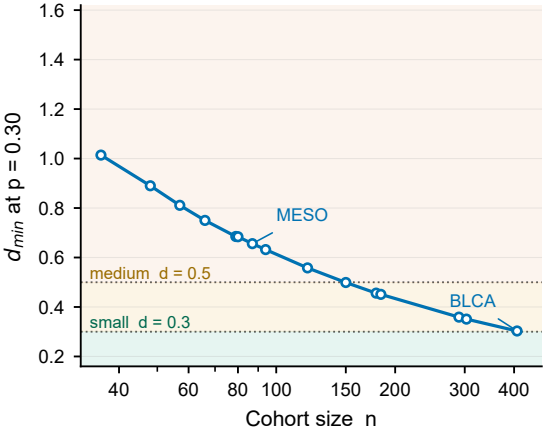
Comparisons are ordered by driver, then by signature. Effect size is the group mean difference divided by the standard deviation of the whole cohort. d_min is that comparison's own detection limit at 80% power; power is evaluated at the observed effect. Detectable marks comparisons whose observed effect meets or exceeds their own detection limit. p is given to three decimals, except where three decimals would render a value just above the 0.05 threshold as "0.050" (SETD2 mutation against the EMT score, p = 0.05001), which is shown to four decimals because it is not significant. Because the table exceeds one page it is placed at the end of the manuscript file, as required by the journal.

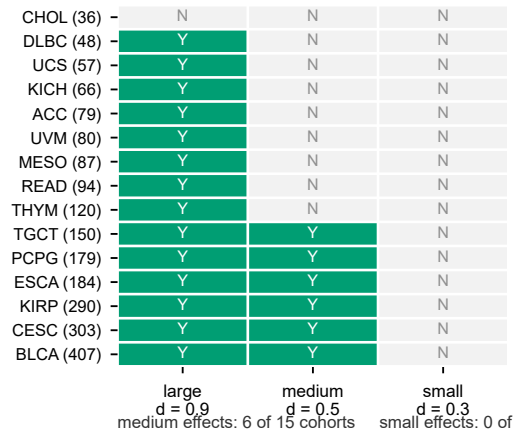


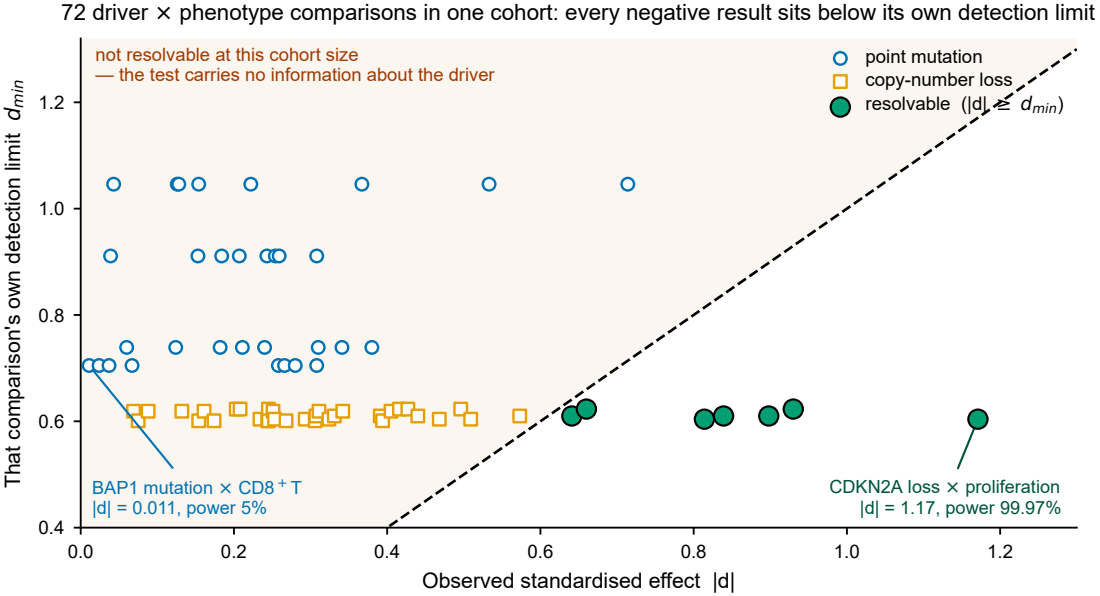
green: $d_{min} \leq 0.50$, i.e. a medium-sized effect is resolvable

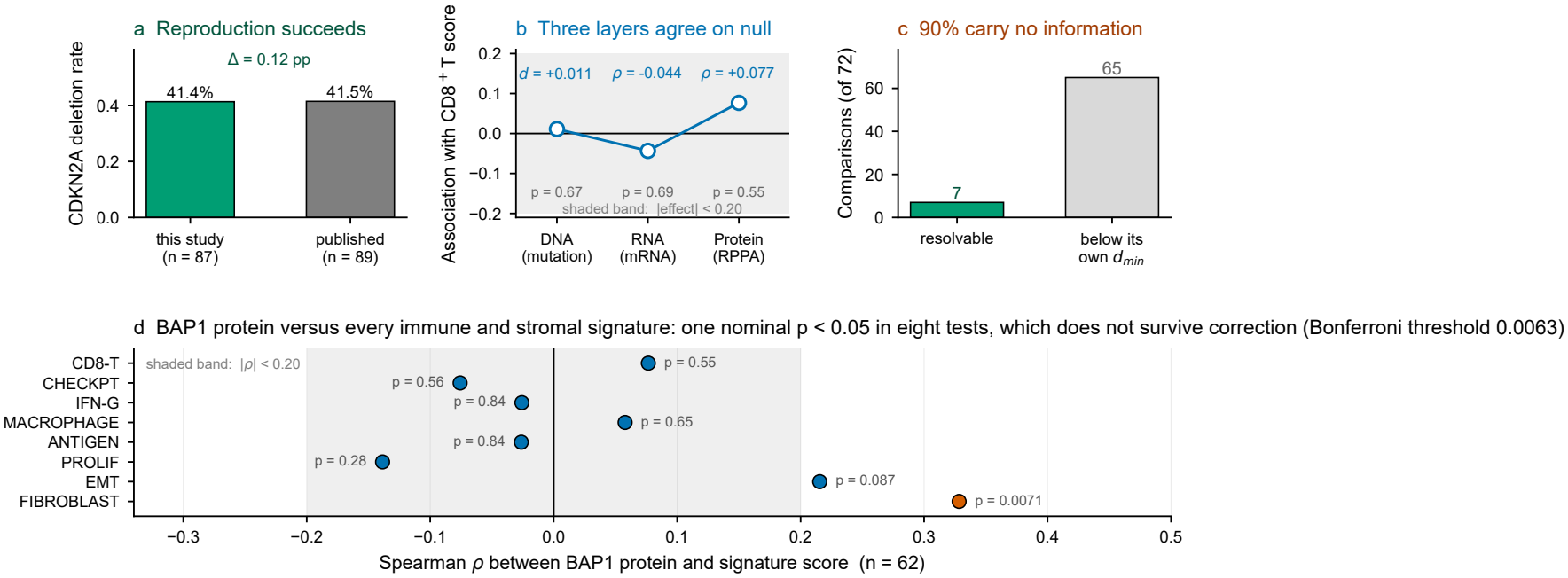
red box: grid cell nearest to the case study — TCGA-MESO is n = 87, p = 0.41, where $d_{min} = 0.61$

Most TCGA rare-tumour cohorts cannot test the effects they are used to test

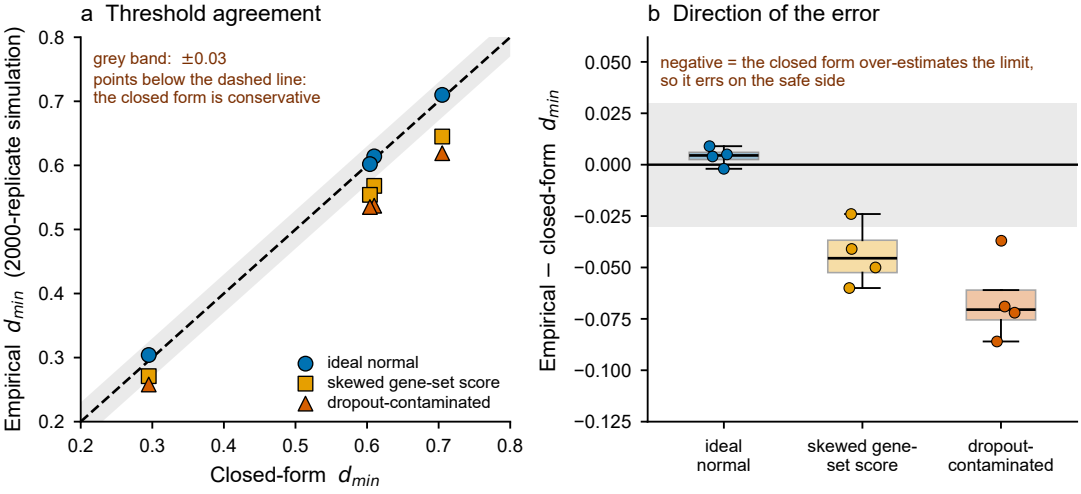





72 driver × phenotype comparisons in one cohort: every negative result sits below its own detection limit
not resolvable at this cohort size
— the test carries no information about the driver
point mutation
copy-number loss
resolvable (|d| ≥ d_{min})
That comparison's own detection limit d_{min}
Observed standardised effect |d|
BAP1 mutation × CD8⁺ T
|d| = 0.011, power 5%
CDKN2A loss × proliferation
|d| = 1.17, power 99.97%
0.0
0.2
0.4
0.6
0.8
1.0
1.2


a Reproduction succeeds
Δ = 0.12 pp
41.4%
41.5%
CDKN2A deletion rate
0.0
0.2
0.4
this study
(n = 87)
published
(n = 89)
b Three layers agree on null
Association with CD8+ T score
0.2
0.1
0.0
−0.1
−0.2
d = +0.011
ρ = -0.044
ρ = +0.077
p = 0.67
p = 0.69
p = 0.55
shaded band: |effect| < 0.20
DNA
(mutation)
RNA
(mRNA)
Protein
(RPPA)
c 90% carry no information
Comparisons (of 72)
0
20
40
60
7
65
resolvable
below its
own d_min
d BAP1 protein versus every immune and stromal signature: one nominal p < 0.05 in eight tests, which does not survive correction (Bonferroni threshold 0.0063)
shaded band: |ρ| < 0.20
CD8-T
CHECKPT
IFN-G
MACROPHAGE
ANTIGEN
PROLIF
EMT
FIBROBLAST
p = 0.55
p = 0.56
p = 0.84
p = 0.65
p = 0.84
p = 0.28
p = 0.087
p = 0.0071
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
Spearman ρ between BAP1 protein and signature score (n = 62)

The closed-form frontier is calibrated, and errs on the safe side under non-normal data



Supplementary Files

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

- Supplement.docx