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.

# OncoDigger Prioritizes Mutation-Associated Genes in Cancer-Specific PubMed Abstracts

**Jimmie E. Fata**

jimmie.fata@csi.cuny.edu

College of Staten Island https://orcid.org/0000-0001-5797-6619

**Research Article**

**Keywords:** literature mining, cancer genes, mutations, BM25, IntOGen

**Posted Date:** November 2nd, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11556965/v1

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

**Additional Declarations:** The authors declare potential competing interests as follows: The author is the creator and developer of OncoDigger, the software evaluated in this study, and declares no financial competing interests. OncoDigger is a freely accessible website with no commercial licensing, subscription or sales, and none is planned.

# OncoDigger Prioritizes Mutation-Associated Genes in Cancer-Specific PubMed Abstracts

**Jimmie E. Fata** (ORCID 0000-0001-5797-6619)

Department of Biology, College of Staten Island, City University of New York, Staten Island, New York 10314, USA

**Corresponding author.** jimmie.fata@csi.cuny.edu

**Keywords.** literature mining, cancer genes, mutations, BM25, IntOGen

## Abstract

OncoDigger is a biomedical text-mining tool designed to identify and rank genes associated with user-defined queries across cancer-specific PubMed literature. This study systematically evaluates OncoDigger's performance in prioritizing genes associated with mutation-focused literature and examines the factors underlying its results. Ranking involves no gene-specific mutation extraction, so concordance with IntOGen measures agreement with a driver catalogue, not mutation-reporting accuracy. Using the query *mutation mutations*, we analyzed PubMed abstract collections representing 19 cancer types and evaluated the top 10 ranked genes per cancer against the IntOGen cancer-driver catalogue.

Across 190 gene–cancer pairs, 93.7% (178/190) involved genes recognized as drivers in at least one cancer, while 78.9% (150/190) matched drivers of the corresponding cancer type. Exact cancer-type concordance exceeded that of a pooled, cancer-blind control (53.2%), which by construction returns one list for all cancers, by 25.8 percentage points and CancerMine (60.0%) by 18.9 percentage points. Controlled comparisons identified cancer-specific document retrieval and gene-name quality control as principal contributors to performance, whereas the full scoring formula provided only a small, statistically nonsignificant improvement over matched gene counting. Sensitivity analyses supported the stability of the principal findings.

These results provide evidence of OncoDigger's effectiveness for cancer-specific, mutation-focused gene prioritization from published literature. By providing traceable gene rankings and distinguishing literature associations from driver genes called in tumour-sequencing cohorts, OncoDigger offers an accessible and reproducible resource for exploring cancer-associated genes.

**Availability:** The web application is available at https://oncodigger.com. Code, data, and reproducibility materials are available at https://github.com/jfataphd/OncoDigger-gene-prioritization (release v1.3, archived at Zenodo, https://doi.org/10.5281/zenodo.23774746). **Contact.** jimmie.fata@csi.cuny.edu

**Supplementary Information.** Supplementary data are available online.

# 1 Introduction

Cancer research has generated an immense and rapidly expanding body of scientific literature. In this study alone, PubMed collections representing 19 cancer types contained approximately 1.6 million abstract records (about 1.5 million distinct publications) published since 2000. Extracting meaningful gene associations from this literature is challenging, particularly because extensively studied genes can dominate search results and obscure associations specific to individual cancer types. This challenge is especially important because the genetic alterations involved in cancer vary considerably among tumour types (Lawrence et al., 2013; Martínez-Jiménez et al., 2020). The growing volume and complexity of published research highlight the need for computational tools that can systematically identify and prioritize relevant cancer-associated genes.

Several biomedical text-mining tools are currently available for identifying gene associations in the scientific literature. CancerMine extracts published evidence describing genes as cancer drivers, oncogenes, or tumour suppressors (Lever et al., 2019a). DISEASES integrates text-mined gene–disease associations with other evidence (Pletscher-Frankild et al., 2015; Grissa et al., 2022), while PubTator 3.0 provides automated identification and normalization of biomedical entities and their relationships (Wei et al., 2024). Query-driven gene ranking from PubMed retrieval has also been implemented: GLAD4U accepts any PubMed query and prioritizes the genes of the retrieved publications by publication count or a hypergeometric test (Jourquin et al., 2012), and Génie ranks genes by the literature attached to a selected topic (Fontaine et al., 2011). OncoDigger instead ranks genes from BM25-ranked documents using rank-weighted mentions, within separate cancer-specific collections. Despite these resources, an important question remains: can searching literature within individual cancer contexts improve the prioritization of mutation-associated genes compared with cancer-blind approaches? Addressing this question also requires careful gene recognition, as gene symbols and aliases may overlap with ordinary words or unrelated abbreviations, even when approved nomenclature is used (Tuason et al., 2004; Chen et al., 2005; Morgan et al., 2008).

OncoDigger (oncodigger.com) takes a different approach to exploring the cancer literature. Rather than relying on a single collection of publications, it maintains separate PubMed abstract collections for 19 cancer types, allowing researchers to investigate associations within the context of individual cancers. Users can enter their own search queries, select publication years, and identify genes, pathways, drugs, phytochemicals, environmental exposures, and other biological or clinical features associated with their topic of interest. OncoDigger uses the established BM25 information-retrieval algorithm to identify publications most relevant to a user's query (Robertson and Zaragoza, 2009). It then ranks gene associations using evidence from these publications, including rank-weighted gene mentions, the number of supporting publications, and enrichment relative to the cancer-specific literature. Although OncoDigger supports a broad range of searches, this study focuses specifically on its ability to identify and prioritize mutation-associated genes. Importantly, OncoDigger identifies associations reported in published literature; it does not independently establish whether a gene functions as a somatic cancer driver.

Here, we systematically evaluate OncoDigger's performance across all 19 cancer types using a common mutation-focused query. We assess whether the highest-ranked genes correspond to established cancer drivers using IntOGen, a cancer-type-specific driver-gene compendium derived from tumour-sequencing analyses and literature-informed curation (Martínez-Jiménez et al., 2020; intogen-plus, 2024). We then compare OncoDigger with established biomedical text-mining resources (Table 1) and a pooled, cancer-blind search to determine whether organizing literature by cancer type provides a measurable advantage. Finally, we investigate which components of OncoDigger contribute most to its performance, including document retrieval, gene recognition, and ranking methods. Together, these analyses address a central question: can the organization and systematic mining of published cancer research improve the prioritization of genes that agree with an external driver

reference within individual cancer types?

### Table 1. Comparison of literature-mining resources

| Feature | OncoDigger | CancerMine | DISEASES 2.0 | PubTator 3.0 |
|---|---|---|---|---|
| Literature scope and text unit | 19 cancer-specific collections of PubMed titles and abstracts, 2000–2026 | PubMed abstracts + PMC full text across cancers | PubMed abstracts + PMC full text across diseases | PubMed abstracts + PMC full text across biomedical entities |
| Ranking/evidence used here | BM25 retrieval + gene enrichment | Cancer-role extraction + citation counts | Gene-disease z-score | Extracted Gene-Disease relations + per-disease publication counts |
| Separate cancer-specific source corpora | Yes | No | No | No |
| Primary output | Per-cancer ranked gene list | Driver/oncogene/TSG gene-cancer relations | Gene-disease association score | Normalized entities and relations |
| Paper-level traceability | Yes | Yes | Yes; underlying abstracts via viewer | Yes |
| Ranking explicitly conditioned on the mutation query in this study | Yes | No | No | No |
| Benchmarked against IntOGen here | Yes | Yes, audited | Yes, audited | Yes, audited |

---

# 2 System and methods

## 2.1 Cancer-specific literature collections

The study included 19 cancer types (Breast, Lung, Colorectal, Prostate, Melanoma, Bladder, Kidney, Pancreatic, Liver, Stomach, Esophageal, Ovarian, Endometrial, Cervical, Thyroid, Brain Cancer, Leukemia, Lymphoma, Myeloma). Each collection contained English-language PubMed abstracts published 2000 through September 2026, selected by a single cancer-specific MeSH term through the NCBI PubMed E-utilities API (Sayers et al., 2025). Collection sizes ranged from 22,079 to 250,061 abstracts (1,640,505 across all 19 before cross-collection deduplication, Supplementary Tables S1 and S2).

All primary results were computed on one frozen build, the 2026-09-27 release (19 SQLite databases, file hashes in `data/provenance/corpus_build_2026-09-27.json` in the companion repository). The manual vocabulary audits of the first correction round (Methods 3.3) were read on the earlier 2026-05-08 text of the same records and the corrections were applied unchanged, whereas the later v4 and v5 term readings were drawn from the 2026-09-27 collections (Supplementary Methods S1). A post hoc sensitivity analysis that excludes records indexed under the cancer term's /secondary MeSH subheading is described in Section 7.

Documents were scored by BM25 with title and MeSH matches, publication recency and OncoDigger's "Primary Evidence" discovery mode, which up-weights original-research and clinical-evidence terminology and

down-weights reviews by 0.65 without excluding them. The query `mutation mutations` selected up to 1,000 of the highest-ranked abstracts per cancer for gene scoring. The pool held 1,000 abstracts for 17 cancers, and 861 for Cervical Cancer and 817 for Myeloma, the only abstracts in those collections that contain either query word. Pool sizes from 250 to 2,000 abstracts gave reasonably robust results, with some change in top-10 membership (Supplementary Methods S1).

## 2.2 IntOGen calibration and internal controls

The query `mutation mutations` was applied to all 19 cancers and scored against IntOGen (release 2024.09.20, 633 driver genes; IntOGen, 2024; Martínez-Jiménez et al., 2020), used as an **external, cohort-computed calibration reference** and not ground truth. Its construction also incorporates prior cancer-gene knowledge (the Cancer Gene Census and CancerMine), so a post hoc sensitivity analysis restricts it to its statistical admission rule (Methods 2.6, Results 5.9). IntOGen's gene membership, cancer-type assignments and significance calls were never used to tune OncoDigger's ranking algorithm, scoring parameters or vocabulary rules, except in the post hoc proximity analysis (Supplementary Results S15). One audit (Methods 3.3) used IntOGen membership only to select genes for vocabulary-collision scrutiny, each correction was determined by reading PubMed text, and whether a correction raised or lowered concordance was not a criterion. The v4 and v5 corrections were decided by a rule fixed before the rebuild. A secondary `driver mutation driver mutations` query served as a wording sensitivity check.

The ranking is a retrieval-and-enrichment procedure and performs no gene-specific relation extraction, so a gene can rank because it is discussed alongside mutations in other genes (for example as a drug target, biomarker or methylation marker, Supplementary Table S4). Precision@10 against IntOGen therefore measures agreement of the top-ranked genes with somatically mutated driver genes and not the proportion of supporting abstracts that report a mutation in the gene.

Precision@10 is the percentage of a cancer's top 10 genes listed as drivers anywhere in IntOGen's 633-gene compendium. Confidence intervals use a case bootstrap (19 cancers resampled with replacement, 2,000 times), because the cancer type and not the gene is the experimental unit (a Wilson interval is in Supplementary Results S1). Paired t-tests and Wilcoxon tests, where reported, use the 19 per-cancer values and are secondary. No multiplicity correction is applied across the comparators apart from a Bonferroni correction of the per-cancer Spearman correlations, so the weaker comparisons are descriptive. The two exact-concordance differences, against CancerMine ($p=1.9\times10^{-6}$) and against the pooled control ($p=4.4\times10^{-7}$), would survive a Bonferroni correction for 20 tests. The 19 cancers share genes, biology and some abstracts, so the intervals describe variation across these categories and not independent biological replicates. Random selection would hit an IntOGen driver about 3.3% of the time (633/19,250 protein-coding genes; HGNC, Seal et al., 2026). Rank AUC and a Spearman correlation summarised quality across each top 50.

Two controls test which part of the pipeline drives results. The **pooled-literature control** merges all 19 collections into one pool (1,514,148 unique papers) and runs the same procedure once, which evaluates the combined effect of cancer-aware organization and retrieval and not index separation alone. The **full-pipeline-versus-counting control** keeps the collections separate but replaces BM25, rank weighting and enrichment with a raw mention count, so it compares complete pipelines and is not an isolated formula test (Results 5.8 gives the isolated decomposition). A stricter variant counted only HGNC primary symbols (`scripts/analysis/intogen_ablation_panel.py`).

## 2.3 Exact cancer-type driver concordance

Beyond the any-cancer benchmark, **exact cancer-type driver concordance** asks whether IntOGen lists a gene as a driver specifically in the cancer it was ranked in (release 2024.09.20, cancer-type-resolved, Martínez-Jiménez et al., 2020). IntOGen's cancer-type codes were mapped to the 19 study categories (Supplementary Methods S5), and the same rule was applied to CancerMine, the pooled control and ChatGPT. This measure is not cancer-specific biological truth. IntOGen detects somatic point-mutation signal only, so germline- or fusion-driven genes can be real drivers without reaching significance there (Results 5.3).

## 2.4 External resources and benchmark auditing

OncoDigger was compared with CancerMine (Lever et al., 2019a), DISEASES 2.0 (Pletscher-Frankild et al., 2015; Grissa et al., 2022) and PubTator 3.0 (Wei et al., 2024). DisGeNET and CIViCmine were excluded as scope-mismatched (Piñero et al., 2017; Lever et al., 2019b). These resources differ from OncoDigger in gene recognition, source data and scoring, so a score difference reflects the whole system and not corpus separation. They were downloaded between 2026-09-26 and 2026-09-30 (identities in `data/external/PROVENANCE.md`) and audited and corrected before any number was final (Supplementary Table S3, Repository Table RT10, Supplementary Methods S2 to S4). The audit found a real entity-recognition error in CancerMine, where all 1,329 sentences credited to PTGDR were unrelated lncRNA mentions, so PTGDR was removed from the benchmark and the author notified. PubTator genes were ranked by the sum of PubTator's per-disease publication counts across each cancer's matched subtypes (the API returns counts and not PMID lists, so a paper tagged under two matched subtypes can be counted more than once). A single ChatGPT free-tier response (27 September 2026) listing each cancer's top 10 most frequently mutated genes served as an informal practical baseline and not a competing system; it is one response without repeated sampling, and it stated that it used live web searches grounded in TCGA/cBioPortal data (Cerami et al., 2012; Gao et al., 2013).

## 2.5 Literature weight of unrecovered annotations

All 1,740 IntOGen gene-cancer pairs were scored for rank at any depth and for supporting-paper count by rerunning the search engine at an extended entity limit, to test whether pairs absent from the top 10 or 50 reflect weak literature coverage and not a ranking failure (Results 5.5). The extended run reproduced the main run's top 50 exactly for all 458 IntOGen-driver pairs present in both lists. Support count is not independent of rank, and genes that never appeared were also checked against the vocabulary (Supplementary Results S10).

## 2.6 Sensitivity to IntOGen's literature-informed admission rules

IntOGen's construction incorporates prior cancer-gene knowledge through the Cancer Gene Census (CGC) and CancerMine (intogen-plus, 2024). All systems' top-10 lists were therefore rescored, unchanged, against the 1,780 of 4,478 catalogue rows that satisfy its stricter non-CGC statistical rule (combined $q<=0.05$ and at least 2 individually significant methods), using the same bootstrap. This post hoc analysis is not an independent reference, because CGC-calibrated method weights shape every q-value (Supplementary Results S13).

### 2.7 Literature-cutoff-matched comparison (post hoc)

To bound the effect of newer literature (CancerMine's snapshot ends on 2023-03-01, OncoDigger's collections run to September 2026), the unchanged pipeline was rerun on collections truncated at publication year 2022 and 2023, with collection statistics recomputed on the truncated set, scored against the same IntOGen reference and compared with CancerMine using the same bootstrap. The MET, REST and FH redaction of Methods 3.3 was applied to the truncated collections and changed no top-10 membership (Supplementary Results S14). This is a literature-age sensitivity analysis and not a reconstruction of PubMed on 2023-03-01, because later revisions of earlier records are included.

---

## 3 Algorithm

### 3.1 Document ranking

Each abstract was ranked with Okapi BM25 (Robertson and Zaragoza, 2009), with k1=1.5 and b=0.75 held fixed across all 19 cancers (Supplementary Results S8 shows Precision@10 is not sharply sensitive to these values). Inverse document frequency and average document length were computed within each cancer collection. The BM25 score was combined with fixed multipliers, namely 1.2 for a title match, 1.15 for a matching MeSH term, a recency multiplier (1.1 for 2020 onward, 0.95 before 2010), 0.65 for reviews and the "Primary Evidence" content-type multiplier, as defined in OncoDigger's source code (private live-application repository, commit `4e39fc068102`, 2026-05-31) and vendored in the companion repository (Supplementary Methods S1). The highest-scoring papers, up to 1,000, formed the gene-ranking pool.

### 3.2 Gene enrichment

Genes were recognized with an HGNC-derived vocabulary of approved symbols and curated aliases (Seal et al., 2026). A mention's weight decayed with the inverse square root of its paper's BM25 rank. An enrichment score compared a gene's frequency in the query-relevant pool with its frequency in the background collection.

```
enrichment = (supporting_papers / pool_size) / (background_papers / corpus_size)
```

```
Relevance = weighted_score × log(1 + enrichment) × log(1 + supporting_papers)
```

`background_papers` is the number of distinct papers in the whole cancer collection in which the same recognition procedure that credits the gene in the pool finds it, so `supporting_papers` ≤ `background_papers` ≤ `corpus_size` for every reported gene. It was obtained by running the engine's own per-paper matching code over every paper of the collection, once per query (Supplementary Methods S1). The engine's built-in background count is an approximation that understates this number, and results computed with it are reported only as a comparison (Supplementary Methods S1).

A gene needed at least 2 distinct supporting publications to be reported, which rewards genes preferential to the most query-relevant papers and suppresses genes common throughout the background literature. The score is an engineered heuristic. It was introduced in OncoDigger's May 2026 redesign (application commit `eb4da47`), before this benchmark, and was not derived from a statistical model, tuned on IntOGen or ablated term by term when it was introduced. Because every supporting paper adds to the weighted score, the number of supporting papers enters the product twice, once with rank sensitivity and once directly. Results 5.8 tests the formula against flat counting.

**Computational requirements.** The analyses ran on one Windows workstation with 16 logical processors and 63 GB of RAM. Computing the exact background counts and the 19-cancer ranking for the first time took 86 minutes (6 worker processes), and the pooled-corpus control over 1,514,148 papers took 75 minutes (Supplementary Methods S1).

## 3.3 Gene-name quality control

The HGNC vocabulary was cleaned before benchmarking. 143 aliases colliding with another gene's own approved symbol were removed, aliases doubling as ordinary English words were filtered out, a longest-match rule was applied to 8,203 nested-name cases (for example EGF is not counted inside `epidermal growth factor receptor`), 7 genes were excluded entirely, and 31 bare symbols and 51 single-word aliases were suppressed at term level, so that only the ambiguous term is removed and the gene's other terms stay matchable (frozen policy 2026-10-16-v5; Supplementary Methods S1). The result is an audited, ambiguity-filtered vocabulary and not an ambiguity-free one (Section 7). A symbol-presence eligibility filter also excludes a gene whose own approved symbol has no posting anywhere in a cancer's collection, even when the engine recognised it through a full name or an alias (Supplementary Methods S1).

On the 2026-05-08 text, every gene in the top 30 of any cancer and all 190 top-10 pairs of the May ranking were checked by hand against a representative PubMed sample, which could miss a gene contaminated in only a minority of postings. A blinded human check of the top-10 pairs has not been done. Extending the check through rank 50 and sweeping every short symbol found nine false aliases in all (Supplementary Methods S1). All were corrected and the pipeline was rerun from scratch.

A further audit applied a two-step verification (a standalone capitalized match, then a pre-identified alternate meaning) to every confirmed IntOGen driver gene in the top-50 rankings (195 genes on the 2026-09-27 build). Two required correction: MET (Lung Cancer, rank 9; 1,733 of 4,105 raw postings, 42.2%, passed under the final rule) and REST (Thyroid Cancer, rank 41; all 136 postings excluded). The false-positive mentions were redacted and the engine rerun. MET stayed at rank 9 with both headline metrics unchanged, and REST left Thyroid Cancer's ranking. REST was later also suppressed globally in policy v5, so in the final pipeline the redaction changes no ranking list (Supplementary Methods S1).

A 40-posting sample of MET postings that passed the rule without a positive gene-context term found 33/40 (82.5%) genuine, which is diagnostic and not independent validation. Rerunning the pipeline with only the 1,268 postings (30.9% of 4,105) that had both a capitalized match and a positive context term moved MET to rank 11, replaced by the Lung Cancer driver NRAS, with both headline metrics unchanged (178/190 and 150/190), so the headline is not sensitive to MET's retention rule (Supplementary Methods S1).

The audits above were done on rankings computed with the engine's approximate background count (Methods 3.2). Under the exact count 36 genes entered a top-50 list for the first time, and the collision screen of Supplementary Methods S1, rerun on the corrected lists (347 genes), flagged SHE, RHO and CCS. A whole-vocabulary screen of single-word aliases and ordinary-word symbols then found 165 candidate terms, 119 of which belong to the 602 genes appearing in any reported ranking or control. Each candidate was classified once, under a rule written before the rebuild and independent of IntOGen: a term was suppressed if it referred to the gene in at most 10% of the sampled abstracts in every collection examined, and retained if it is a genuine biological name. The abstracts were read by a language-model agent that had seen the manuscript, so the readings are not independent validation. The result is the frozen policy v5 (Supplementary Methods S1): 69 terms suppressed (51 aliases and 18 symbols, including SHE, RHO, CCS and NODAL), 45 retained as genuine names, and 8 real gene names with mixed, collection-dependent meanings retained globally and disclosed (Section 7). Every cleaned-vocabulary analysis was rebuilt under v5. The top 10 was identical in all 19 cancers and 30 of 950

top-50 positions changed. Whether a correction raised or lowered concordance was not a criterion for any decision, the frozen table was not revised after the results were seen, and the true-raw control arms (Methods 3.4) remain unfiltered.

### 3.4 Post hoc decomposition of pipeline components

A post hoc analysis under a frozen written protocol (`FROZEN_ANALYSIS_SPEC_retrieval_vocab_decomposition.md`, companion repository) examined how much of the IntOGen agreement depends on retrieval, vocabulary cleaning and the relevance formula. Two experiments were run across all 19 cancers using production code. The first crossed document retrieval (BM25 pool versus complete corpus) with vocabulary state (cleaned versus true-raw) using flat mention-count ranking, which requires retrieval to be unaffected by vocabulary (confirmed byte-identical for all 19 cancers). The second held retrieval at the BM25 pool and compared the full relevance formula with flat counting under both vocabulary states. A three-way factorial was not run, because the enrichment term equals 1 for every gene when the pool is the full corpus. Both used the case bootstrap of Methods 2.2.

---

## 4 Implementation

The environment was fixed for reproducibility (`PYTHONHASHSEED=0`, per-connection caches cleared between cancers), and repeated runs gave byte-identical top-50 lists and rank order. OncoDigger rankings were produced with the engine source pinned at application commit 4e39fc0 (2026-05-31; vendored in `reference_engine_code/`), run on the 2026-09-27 corpus build with the exact background count (Methods 3.2) and vocabulary policy v5 (Methods 3.3). The MET, REST and FH rules, the v4 and v5 corrections and the decomposition (Methods 3.4) were added to the analysis scripts after the engine was frozen and are part of the release. The companion repository (Data and code availability) carries the code, the derived outputs and a build-provenance record that ties every output to the corpus build and engine version that produced it. The public `oncodigger.com` interface defaults to "Overview" mode and not this paper's "Primary Evidence" mode, and its corpora update monthly. It also differs from the frozen pipeline in its vocabulary policy (it does not yet include the v4 and v5 additions) and in eligibility filtering (it does not apply the symbol-presence filter of Methods 3.3; Supplementary Methods S1), so it can return different rankings from those reported here, and exact reproduction requires the frozen release. Figure 1 summarises the pipeline.

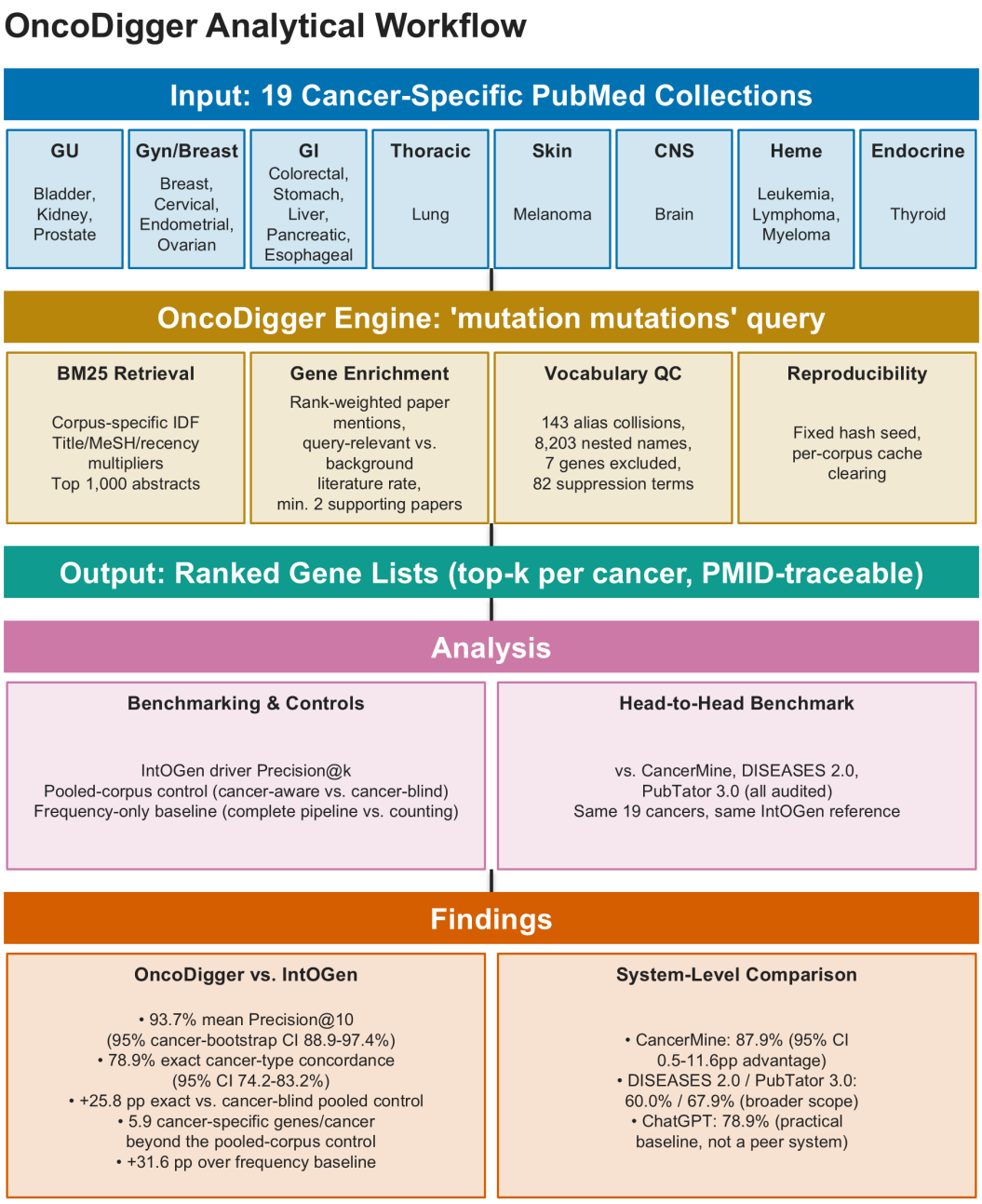



**Figure 1. OncoDigger analysis workflow.** Nineteen separate cancer-specific PubMed collections are searched with the same mutation query. BM25 ranks papers, enrichment ranks genes, vocabulary quality control reduces ambiguous entity matches (82 suppression terms are declared in the policy and 78 change the matcher, Supplementary Methods S1), and supporting PMIDs preserve traceability. The resulting rankings are evaluated against IntOGen, internal pooled/frequency controls and external literature-mining systems.

# 5 Results

## 5.1 Cancer-specific collections produce different rankings

The common query surfaced pan-cancer drivers (TP53 in the top 10 for 18/19 cancers, ranking 12th in Melanoma, and KRAS, PIK3CA and BRAF in 11 to 13) and drivers confined to one cancer type (Figure 2, Supplementary Results S1). Mean pairwise top-10 Jaccard similarity across 171 cancer pairs was 0.165 (1.0 = identical lists), comparable to CancerMine (0.207), DISEASES 2.0 (0.160) and PubTator 3.0 (0.184), so the lists are cancer-differentiated but not markedly more so than the comparators.

A pooled control merging all 19 collections returned the identical top-10 list for every cancer, with 100% any-cancer Precision@10 but no cancer-type differentiation. Separating the literature recovered a mean of 5.9 additional genes per cancer (5.3 listed IntOGen drivers, Supplementary Figure S1), 74.1% of which were drivers of that specific cancer, including VHL, PBRM1 and BAP1 for Kidney Cancer (Cancer Genome Atlas Research Network, 2013). Scored against cancer-type-specific calls, the pooled control's exact concordance was 53.2% (101/190), 25.8 points below the separated pipeline's 78.9% (95% CI 20.0 to 32.1, paired t-test $p=4.4x10^{-7}$, separated pipeline higher in all 19 cancers), because its any-cancer score reflects well-known pan-cancer genes (Supplementary Results S1).

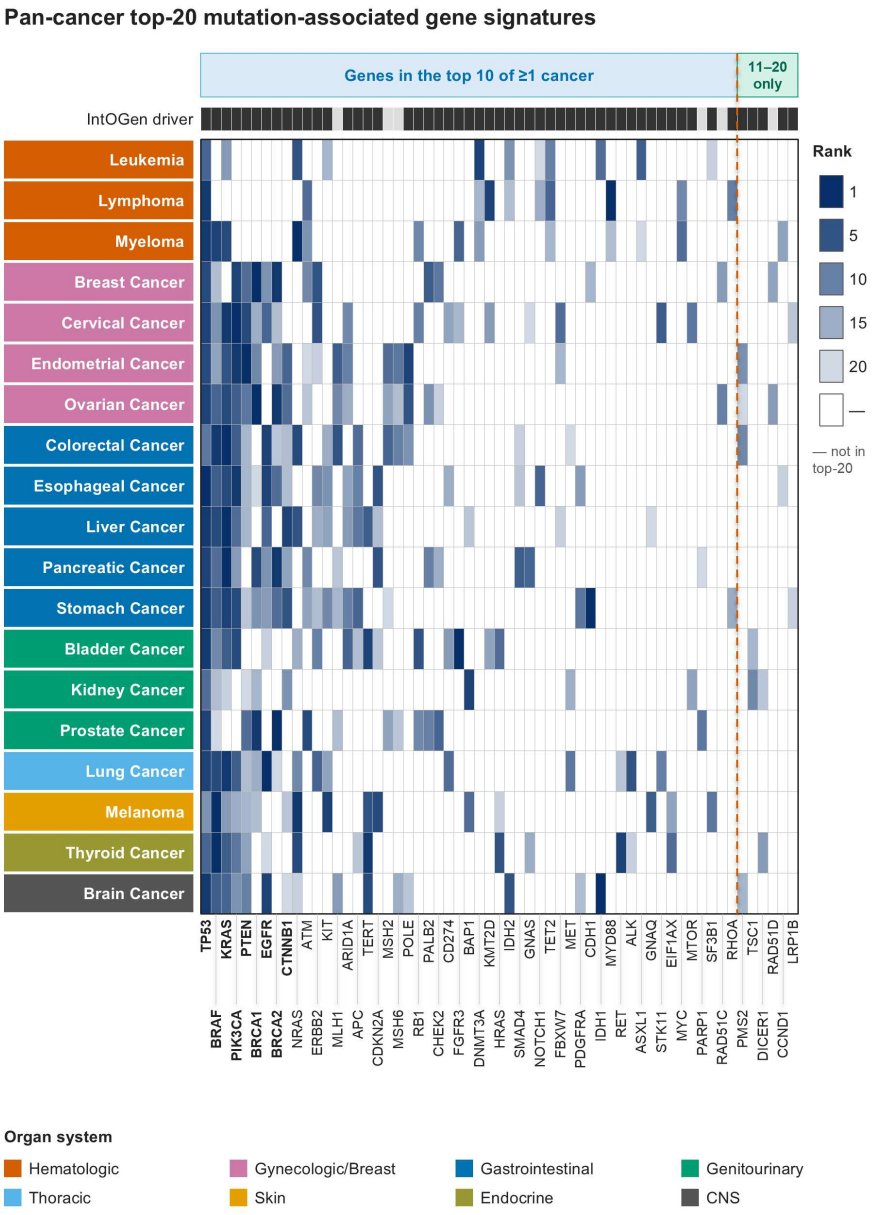



**Figure 2. Shared and cancer-differentiated gene-ranking patterns.** Heatmap of genes reaching the top 20 in at least two cancers. Genes are grouped in two column blocks: genes reaching the top 10 in at least one cancer, and genes reaching only ranks 11-20. The blocks classify genes and do not give the rank shown in a given cell, and a gene in the first block can occupy ranks 11-20 in other cancers. Rows are cancers and columns are genes, and shading is a continuous gradient from darkest (rank 1) to lightest (rank 20), white indicating the gene is outside that cancer's top 20. The strip above the heatmap marks genes listed by IntOGen as drivers in any cancer type (dark) or not listed (light grey), and bold gene symbols are in the top 20 of at least half of the 19 cancers. The key's 1/5/10/15/20 ticks sample this scale, not discrete bins. The figure illustrates a shared pan-cancer backbone together with cancer-differentiated literature signals.

## 5.2 IntOGen calibration and the scoring controls

Here Precision@10 means the share of a cancer's top-10 genes that IntOGen lists as drivers in any cancer, that is agreement with a driver compendium and not the precision of mutation relations extracted from abstracts. Of 190 top-10 slots, 178 were listed IntOGen drivers, a Precision@10 of 93.7% (95% cancer-bootstrap CI 88.9 to 97.4%). Eleven of 19 cancers scored 10/10 and the rest 70 to 90%, with no significant association between collection size and Precision@10 across an 11-fold range (Spearman rho=0.05, permutation p=0.84, Supplementary Results S1). Below the top 10, listed drivers ranked above non-drivers through the top 50 (mean rank AUC 0.706, above 0.5 in every cancer, range 0.530 to 0.886), and the rank-driver Spearman correlation was negative in all 19 cancers and significant after Bonferroni correction in 2 (Supplementary Results S1).

Three controls address why the complete pipeline works. They compare complete pipelines and do not show that the scoring formula is better than counting. A frequency-only control, which counted the papers containing any single-word term of a gene in the same vocabulary after the term-level suppressions of Methods 3.3, with no BM25 retrieval, multi-word names or enrichment, reached 62.1% Precision@10, 31.6 points lower and lower in all 19 cancers ($p=2.9\text{x}10^{-7}$). A genuinely unfiltered vocabulary through the full pipeline gave a 57.4-point drop (Supplementary Results S7). These comparisons change several pipeline components at once and do not isolate the contribution of aliases within the complete pipeline (Supplementary Results S1).

Together with the pooled control, these results show that neither cancer-specific separation nor accurate document selection and vocabulary alone is sufficient (Figure 3). With retrieval and vocabulary matched, flat counting reached 92.1% Precision@10 and 76.3% exact concordance against 93.7% and 78.9% for the full formula, so the formula's further contribution is small and not statistically significant (+1.6 and +2.6 points, Results 5.8).

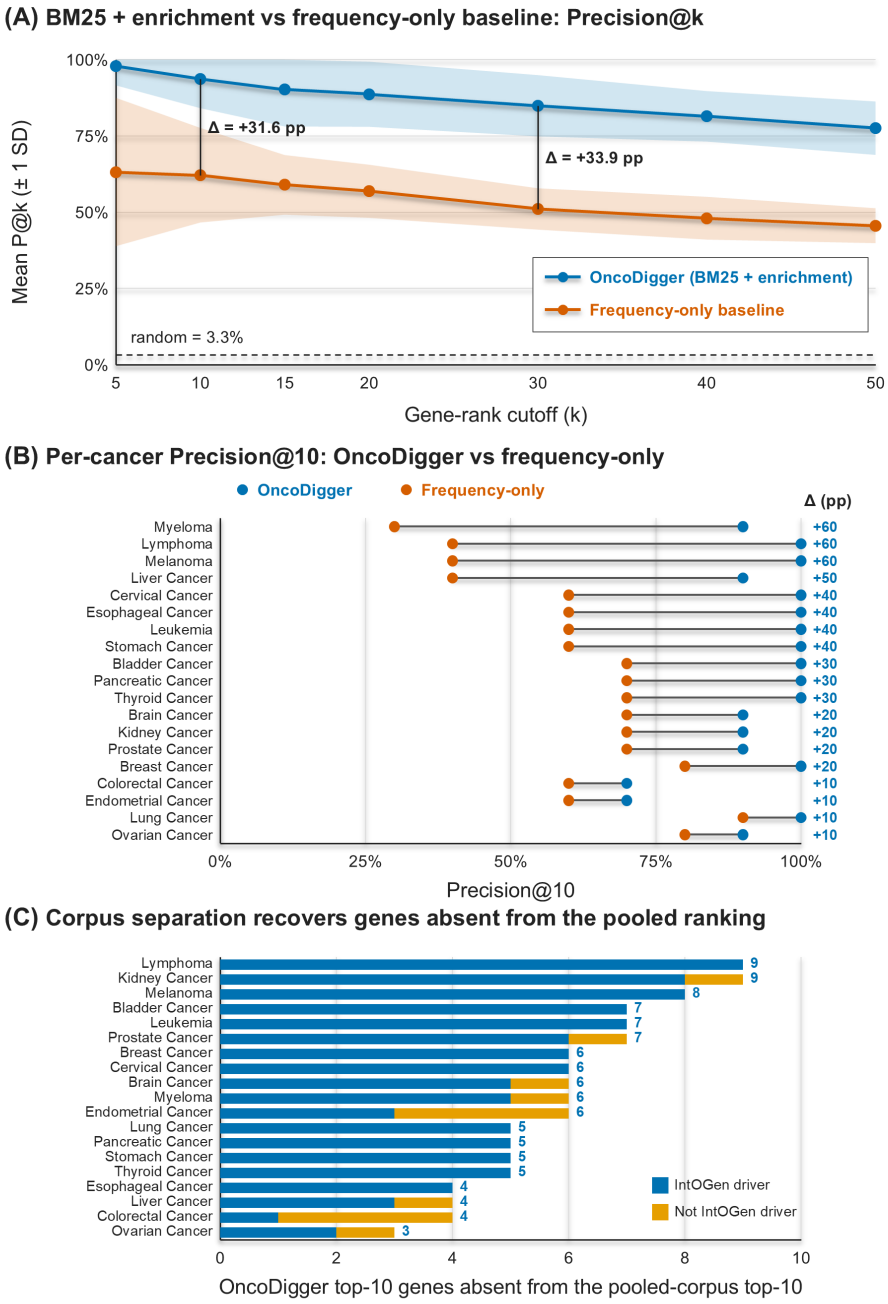



**Figure 3. Cancer separation and the complete pipeline contribute different properties.** (A) Mean Precision@k for OncoDigger and the frequency-only baseline. Shaded bands show ±1 standard deviation across the 19 cancer types (clipped to 0-100%), not confidence intervals. (B) Per-cancer Precision@10 for the same methods. (C) Number of OncoDigger top-10 genes absent from the pooled global top 10. Panels A/B assess the combined effect of the complete pipeline relative to simple publication counting, which changes document retrieval, vocabulary behaviour and scoring together rather than isolating any one of them (Results 5.8 reports the properly isolated decomposition). Panel C assesses cancer differentiation.

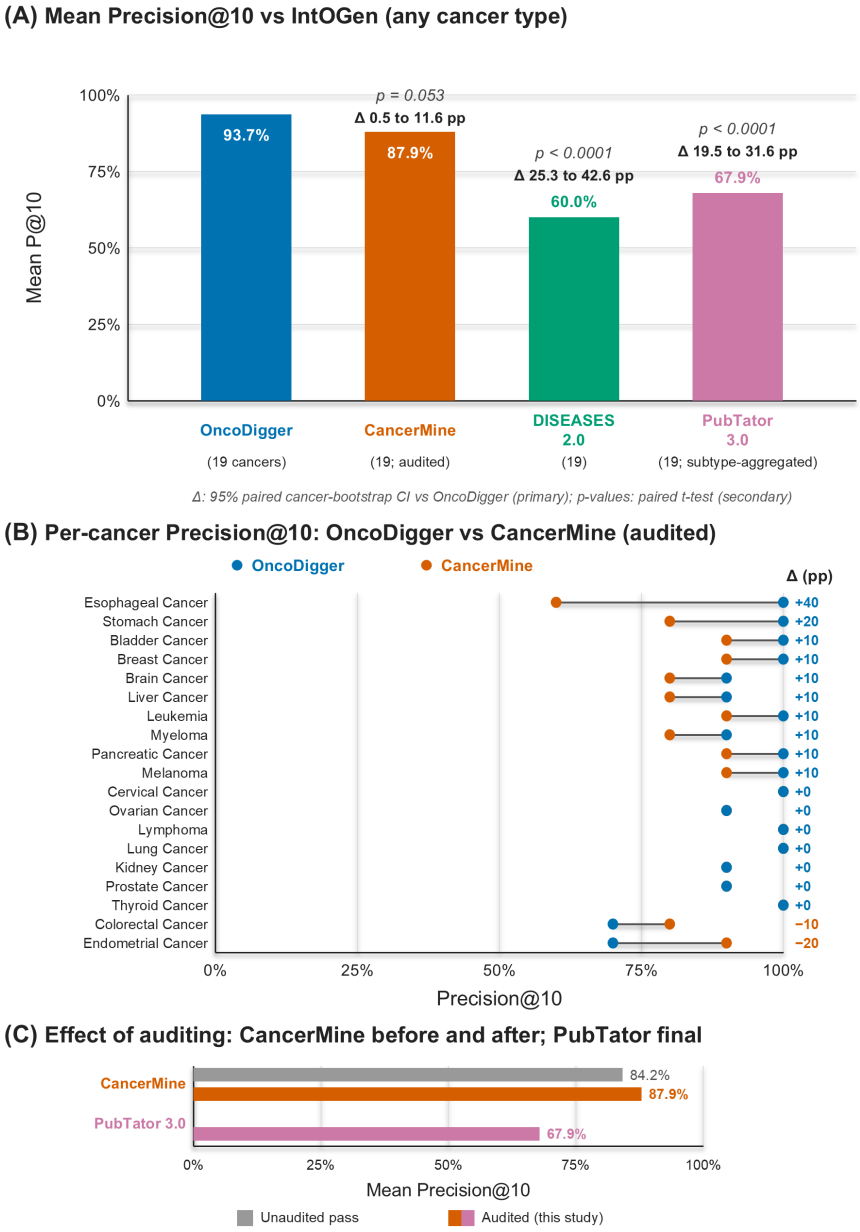
## 5.3 Exact cancer-type driver concordance

Exact cancer-type driver concordance was 78.9% (150/190, 95% CI 74.2-83.2%). Of the 40 non-matching top-10 hits, 28 (70.0%) are genes IntOGen lists as drivers in a different study cancer, often through secondary-tissue patterns (for example BRCA1, a driver in Breast and Ovarian Cancer, and CHEK2, a driver in Breast Cancer, also surfacing in Prostate Cancer through germline predisposition, Pritchard et al., 2016). The remaining 12 (30.0%) are nine genes IntOGen does not list as a significant driver in any cancer. Each has a plausible biological or clinical explanation for its literature prominence that does not require a somatic point-mutation signal in the evaluated cohorts (Supplementary Results S2, Table S4).

Endometrial and Prostate Cancer scored lowest, each at 60%. A `driver mutation` query variant gave 95.8% any-cancer precision and 80.5% exact concordance (+2.1 and +1.6 points, paired t-test p=0.042 and 0.19), so the result is not sharply wording-dependent (Supplementary Results S3). Recall of IntOGen's full catalogue is limited (Results 5.5). Exact matches concentrated toward the top of the ranking (78.9% of ranks 1 to 10 versus 35.8% at ranks 21 to 50, Supplementary Results S1).

## 5.4 Comparison with CancerMine, DISEASES 2.0 and PubTator 3.0

After auditing, CancerMine reached 87.9% Precision@10 against IntOGen (167/190, Supplementary Results S4), against 93.7% for OncoDigger, a 5.8-point difference whose bootstrap interval only just excludes zero (95% CI 0.5 to 11.6 pp; paired t-test p=0.053, Wilcoxon p=0.12). Exact concordance differed more, 78.9% versus 60.0%, a significant 18.9-point difference (95% CI 13.2-23.7 pp, $p=1.9\text{x}10^{-6}$), measured against a catalogue that also incorporates CGC and CancerMine knowledge in its post-processing (Results 5.9, Supplementary Table S11). DISEASES 2.0 and PubTator 3.0 reached 60.0% and 67.9% Precision@10, both significantly lower ($p<10^{-5}$ each, Supplementary Results S4), though not like-for-like because both capture broader gene-disease associations. Their lower agreement here reflects a scope mismatch with a mutation-focused query and not a judgment of either resource. Genes not listed as IntOGen drivers (for example PDCD1 and CTLA4 in Melanoma) are often still biologically meaningful. PubTator's aggregation across subtype entities lets broadly co-mentioned genes (TP53 in 18/19 cancers) accumulate large unnormalised totals, a disclosed limitation that was not corrected because that would score PubTator by a method it does not use. Figure 4 summarises the benchmark.




• *PTGDR artifact (1,329 mis-tagged "-AS1" sentences) + DOID mapping errors incl. Brain Cancer subtypes*

• *Single-parent-MeSH mapping missed subtype-tagged relations; fixing it also resolved Liver/ALS contamination*

**Figure 4. Comparison with published literature-mining resources.** (A) Mean IntOGen-driver Precision@10 for OncoDigger, audited CancerMine, DISEASES 2.0 and subtype-aggregated PubTator 3.0 (all 19 cancer categories included). (B) Per-cancer OncoDigger-CancerMine comparison. (C) Effect of auditing on CancerMine (before and after) and the final audited PubTator value. PubTator's pre-correction value is omitted as unverifiable (Supplementary Table S3).

## 5.5 Literature weight of unrecovered annotations

To test whether the top-50 recovery of about one quarter (26.3%) of IntOGen's catalogue reflects a ranking failure on well-supported genes, all 1,740 pairs were scored for rank and supporting-paper count (Methods 2.5). Support tracked rank sharply. Median papers fell from 84.5 in the top 10 to 13.0 at ranks 11 to 50 and 3.0 beyond rank 50 (Spearman rho=-0.937, $p<10^{-300}$), in all 19 cancers, and only 1.7% of non-top-10 pairs matched or exceeded their cancer's weakest top-10 gene (Repository Table RT2). Of the 978 pairs that never surfaced, 790 have a lexicon-detected mention elsewhere in the corpus and 188 have none (Supplementary Results S10). Restricting recall to pairs that are corpus-present and supported by at least 2 of IntOGen's 7 methods raised recall to 21.1% at rank 10, 31.4% at rank 20, 50.9% at rank 50 and 70.3% at any rank (Supplementary Results S11). Sampling corpus text for the 191 pairs still unrecovered assigned most to an identifiable category and left three unclassified (Supplementary Results S12, Repository Table RT4, Section 7). Figure 5 summarises this analysis.

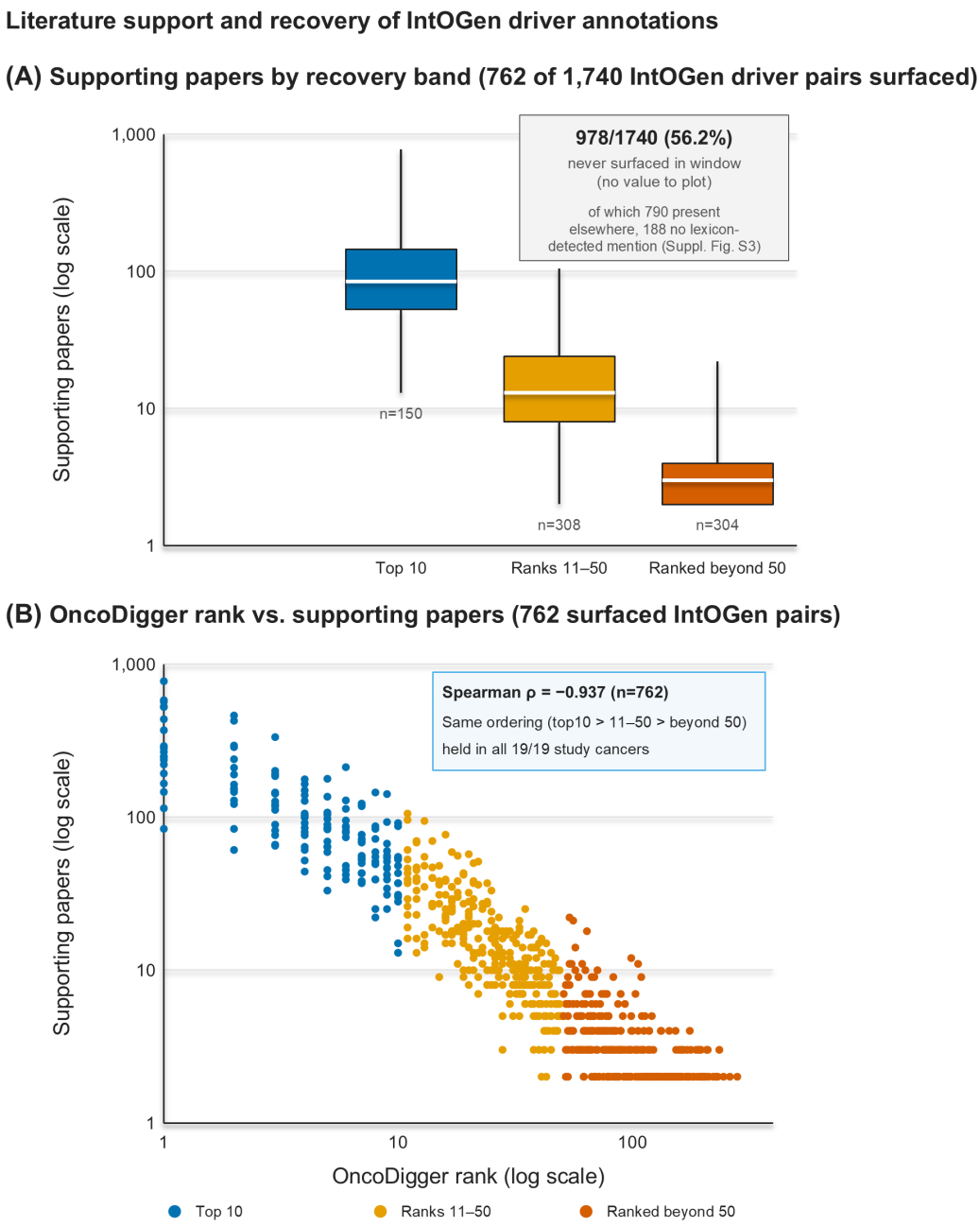



**Figure 5. Literature support and recovery of IntOGen driver annotations.** (A) Unweighted supporting-paper count by recovery band (OncoDigger top 10, n=150; ranks 11–50, n=308; ranked beyond 50, n=304) for the 762 IntOGen driver gene-cancer pairs that surfaced within the standard analysis window. The other 978 of the 1,740 pairs never surfaced, have no value to plot and are summarized in the panel annotation; of these, the large majority have a lexicon-detected mention elsewhere in the full corpus and a minority have none (Supplementary Figure S3, Repository Table RT12). (B) OncoDigger rank versus supporting-paper count (log-log) for the 762 pairs that did surface, coloured by band, with Spearman ρ=–0.937 and the same recovery-band ordering (top 10 > ranks 11–50 > beyond 50) holding in all 19 study cancers individually. Panel B is a mechanistic ranking diagnostic, not an independent validation, since supporting-paper count is one multiplicative component of the OncoDigger score it is plotted against (Methods 2.5).

## 5.6 Practical ChatGPT baseline

On exact concordance a single ChatGPT free-tier response (one response, with no repeated sampling and so no estimate of run-to-run variability) reached 76.3% (145/190), not significantly different from OncoDigger's 78.9% (OncoDigger minus ChatGPT +2.6 pp, 95% CI -5.3 to 11.6 pp; these statistics are descriptive because the comparator is a single response). It stated that it was grounded in TCGA/cBioPortal mutation-frequency data (Cerami et al., 2012; Gao et al., 2013), the same kind of evidence IntOGen is computed from; whether it actually retrieved and used those data was not verified. Its gene-level Precision@10 was lower at 78.9% (OncoDigger minus ChatGPT +14.7 pp, 95% CI 7.4 to 21.6 pp; higher in 12 cancers, tied in 6, lower in 1; Supplementary Results S4), skewed toward large passenger genes (TTN, MUC16, Supplementary Results S4).

## 5.7 Ranking depth

IntOGen-driver enrichment persisted below rank 10, with mean cumulative Precision@k of 97.9% at k=5, 93.7% at 10, 88.7% at 20 and 81.4% at 40. Non-overlapping intervals showed the same gradient (93.7% at ranks 1 to 10 down to 66.6% at ranks 31 to 50, Supplementary Results S5). Three of the 950 top-50 positions are occupied by directly identified residual ambiguous matches (TTR, FH and APC, Methods 3.3), which may affect the depth values to a small, unquantified degree (Figure 6).

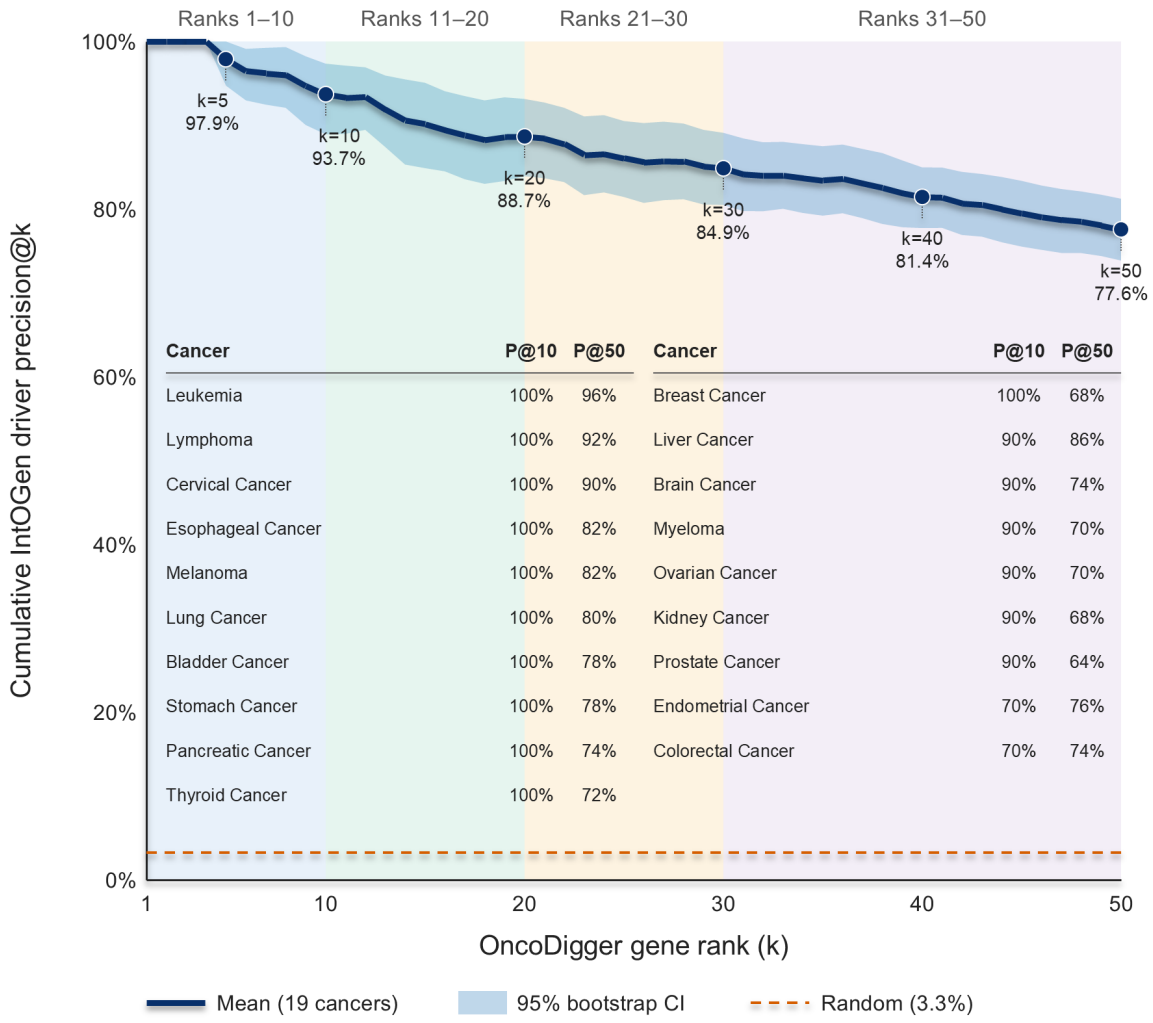



**Figure 6. IntOGen-driver enrichment across ranking depth.** Cumulative IntOGen-driver Precision@k through rank 50 across 19 cancers. The bold line is the mean and the shaded region the 95% bootstrap confidence interval. The dashed 3.3% line is an illustrative genome-wide random-gene reference (633 IntOGen drivers/19,250 protein-coding genes), not a matched random-ranking control. Background rank ranges are descriptive only. The inset table gives each cancer's own Precision@10 and Precision@50.

## 5.8 Post hoc decomposition of retrieval, vocabulary and scoring-formula contributions

A 2x2 factorial (Methods 3.4) crossed document retrieval (BM25 pool versus complete corpus) with vocabulary state (cleaned versus true-raw) across all 19 cancers using flat mention-count ranking, which isolates the interaction from the scoring formula (Table 2).

## Table 2. Retrieval x vocabulary factorial against IntOGen, flat mention-count ranking, all 19 cancers (Methods 3.4, Results 5.8)

| Retrieval | Vocabulary | Any-cancer P@10 | Exact concordance |
|---|---|---|---|
| Full corpus | True-raw | 20/190 = 10.5% | 2/190 = 1.1% |
| Full corpus | Cleaned | 108/190 = 56.8% | 62/190 = 32.6% |
| BM25 pool | True-raw | 28/190 = 14.7% | 10/190 = 5.3% |
| BM25 pool | Cleaned | 175/190 = 92.1% | 145/190 = 76.3% |

Interaction = (cell 4 - cell 3) - (cell 2 - cell 1): +31.1 percentage points any-cancer and +39.5 pp exact concordance. Confidence intervals and per-cancer counts are in Results 5.8 (`scripts/analysis/interaction_all19_bootstrap.py`).

The matched comparison is the key qualification. With BM25 retrieval and the cleaned vocabulary fixed, counting (92.1% any-cancer, 76.3% exact) and the full formula (93.7%, 78.9%) differ by +1.6 pp (95% cancer-bootstrap CI 0.0 to 3.2, formula better in 3 cancers, equal in 16 and worse in none, paired t-test p=0.083, Wilcoxon p=0.083) and +2.6 pp (CI 0.0 to 5.3, better in 6, equal in 12 and worse in 1, p=0.056 and 0.059), which is 3 and 5 of 190 top-10 slots. The formula's numerical advantage over counting is therefore small and not statistically significant at the 0.05 level here, which does not show that the two are equivalent. Retrieval and vocabulary cleaning interacted positively, +31.1 pp on any-cancer Precision@10 beyond the additive prediction (95% cancer-bootstrap CI 21.1 to 41.6 pp) and +39.5 pp on exact concordance (CI 31.1 to 47.9 pp). The any-cancer term was positive in 16 of 19 cancers (zero in Lung and Colorectal Cancer and negative in Ovarian Cancer) and the exact term in 18 of 19 (Supplementary Figure S4). Document selection alone improved any-cancer Precision@10 by 4.2 pp (14.7% versus 10.5%) and vocabulary cleaning alone by 46.3 pp (56.8% versus 10.5%), and together they reached 92.1% against an additive prediction of 61.1%. Cleaning raised the BM25-pool condition from 28 to 175 of 190 and the full-corpus condition from 20 to 108, and the full-corpus condition counts a gene's mentions in every paper of the collection, so it is the condition most exposed to false matches that retrieval does not filter out. This is an inference from this experiment and not an established property of BM25. The interaction is driven largely by severe alias collisions in the unfiltered vocabulary and is a property of this pipeline and benchmark scale and not a general synergy measure. Ten ambiguous single-token aliases, including ordinary English function words, account for most of it, and removing only these recovers 82.6% of the full any-cancer gap (Supplementary Results S7, exploratory).

A second experiment held retrieval at the BM25 pool and compared the full formula (Methods 3.2) with flat counting under both vocabulary states. The formula's marginal contribution was +1.6 pp under the cleaned vocabulary and +21.6 pp under the true-raw vocabulary. Under true-raw vocabulary, rank-weighting alone contributed essentially nothing over counting, while enrichment alone (90.5%) far exceeded the combined formula (36.3%). Positional and frequency weighting were not separately isolated (Supplementary Results S7). A three-way factorial was not run (Methods 3.4), so these are two separate experiments specific to this query and discovery mode.

## 5.9 Sensitivity to IntOGen's literature-informed admission rules (post hoc)

Restricting the reference to the 1,780 IntOGen rows meeting the non-CGC statistical rule (Methods 2.6) lowered OncoDigger's any-cancer Precision@10 from 93.7% to 85.3% (162/190) and its exact concordance from 78.9% to 71.6% (136/190), and every comparator fell as well (Table 3). The comparative findings held. OncoDigger's exact-concordance advantage was +19.5 points over CancerMine (95% CI 13.2 to 25.8) and +35.3 points over the pooled control (95% CI 27.9 to 42.6). The ChatGPT response (+3.2 points, 95% CI -5.8 to 13.2) remained indistinguishable from OncoDigger on exact concordance, whereas the BM25-matched flat-counting control trailed OncoDigger by 4.2 points (95% CI 1.1 to 7.9, against 2.6 points under the full catalogue), and the pooled control's any-cancer precision still exceeded OncoDigger's (90.0% versus 85.3%). Removing IntOGen's CancerMine filter, which removes 142 rows none of which touches a top-10 pair, left 178/190 and 150/190 unchanged. The lower absolute concordance is not a measure of literature-derived bias, because restricting the reference changes which genes count as drivers (Supplementary Results S13, Supplementary Tables S9 to S11 and Repository Table RT5).

## Table 3. Concordance with IntOGen under the full catalogue and under the statistical-rule restriction (Methods 2.6, Results 5.9)

| System | Any-cancer P@10, full catalogue | Any-cancer P@10, statistical rule | Exact concordance, full catalogue | Exact concordance, statistical rule |
|---|---|---|---|---|
| OncoDigger | 178/190 = 93.7% | 162/190 = 85.3% | 150/190 = 78.9% | 136/190 = 71.6% |
| Pooled cancer-blind control | 190/190 = 100.0% | 171/190 = 90.0% | 101/190 = 53.2% | 69/190 = 36.3% |
| CancerMine | 167/190 = 87.9% | 152/190 = 80.0% | 114/190 = 60.0% | 99/190 = 52.1% |
| ChatGPT (one response) | 150/190 = 78.9% | 147/190 = 77.4% | 145/190 = 76.3% | 130/190 = 68.4% |
| DISEASES 2.0 | 114/190 = 60.0% | 96/190 = 50.5% | 82/190 = 43.2% | 63/190 = 33.2% |
| PubTator 3.0 | 129/190 = 67.9% | 114/190 = 60.0% | 83/190 = 43.7% | 64/190 = 33.7% |
| BM25-matched flat counting (clean vocabulary) | 175/190 = 92.1% | 158/190 = 83.2% | 145/190 = 76.3% | 128/190 = 67.4% |
| Frequency baseline (aliases) | 118/190 = 62.1% | 113/190 = 59.5% | 78/190 = 41.1% | 61/190 = 32.1% |
| Frequency baseline (primary symbols) | 121/190 = 63.7% | 115/190 = 60.5% | 73/190 = 38.4% | 61/190 = 32.1% |

Statistical rule: catalogue rows with combined q<=0.05 and >=2 individually significant methods (1,780 of 4,478 rows). A sensitivity analysis, not an independent reference.

### 5.10 Literature-cutoff-matched comparison with CancerMine (post hoc)

Collections truncated at 2022 retain 1,317,084 of 1,640,505 abstracts (80.3%) and those truncated at 2023 retain 1,400,746 (85.4%). The top 10 changed in 12 of 190 slots at 2022 and 10 at 2023, and performance changed little. OncoDigger reached 93.7% any-cancer Precision@10 (178/190) and 76.8% exact concordance (146/190) at 2022, and 93.7% (178/190) and 77.4% (147/190) at 2023, against 93.7% and 78.9% on the full collections (Supplementary Table S12). Its exact-concordance advantage over CancerMine was +16.8 points (95% CI 10.5 to 21.6) at 2022 and +17.4 points (95% CI 10.5 to 22.1) at 2023, against +18.9 points on the full collections, and the any-cancer difference was +5.8 points (95% CI 0.5 to 11.6) at both cutoffs, as on the full collections. The advantage over CancerMine is therefore not explained by access to post-2023 literature, although the comparison still differs in method (Section 7).

---

## 6 Discussion

OncoDigger showed strong agreement with a cohort-computed driver-gene reference while producing rankings that differed substantially among cancers. It identifies genes associated with mutation-focused literature and does not establish that a gene is mutated in the retrieved studies. It measures **cancer-specific literature prominence** and not biological importance directly, and IntOGen concordance measures agreement with a driver catalogue and not independent proof of biological truth (Results 5.9).

The controls clarify where performance comes from. Cancer-specific separation contributes **cancer differentiation** (a pooled index reached 100% Precision@10 but returned an identical list for every cancer), although this comparison characterizes the complete cancer-aware versus cancer-blind strategy and not index separation alone (Results 5.1). Within the cancer-specific collections, document retrieval and vocabulary quality control were the principal determinants of **IntOGen-driver enrichment**, interacting rather than adding, which is a property of this pipeline and benchmark scale and not a claim of general synergy (Results 5.8). The scoring formula's contribution depends on vocabulary quality. With a cleaned vocabulary it adds only a small gain over counting that is not statistically significant, and with an unfiltered vocabulary the performance that remains traces to its enrichment term (Results 5.8). The background count in that term matters in practice, because replacing the engine's approximate count with the exact number of collection papers changed 22 of 190 top-10 slots and the exact concordance by 8 slots (Supplementary Methods S1).

Exact concordance shows greater agreement with IntOGen's cancer-type calls than a cancer-blind strategy achieves (53.2% for the pooled control despite 100% any-cancer precision). Most of the 40 non-matching pairs are genes IntOGen lists in a different cancer or relationships outside somatic point-mutation detection, and are not necessarily ranking errors (Results 5.3). CancerMine was the most informative comparator, with significantly lower exact concordance and a smaller Precision@10 difference whose interval only just excludes zero, though its methods differ. DISEASES and PubTator address broader gene-disease relationships, so their larger gaps need more caution. The audit corrected real problems on both sides, and OncoDigger's own output needed the same treatment (Methods 3.3).

IntOGen also draws on the CGC and CancerMine (Methods 2.6), so the concordance figures do not establish wholly independent validation. Every comparative conclusion was preserved under the statistical-rule restriction (Results 5.9), but that reference is not independent either, and a literature-independent check such as raw mutation recurrence in the cohorts was not performed.

OncoDigger is best viewed as a **literature-ranking and evidence-prioritization system**, not a clinical variant-interpretation tool or a replacement for expert-curated driver-gene references. It reaches this performance from freely accessible PubMed abstracts alone. The results support separating cancer literature and cleaning the gene vocabulary. Query relevance and background enrichment improve recovery over unmatched counting baselines but add only a small numerical gain over counting matched on BM25 retrieval and cleaned vocabulary, which is not statistically significant and does not show that the two are equivalent (Results 5.8). Recall is the main limit of the approach: the top 50 contain 26.3% of IntOGen's 1,740 driver pairs, or 50.9% of the pairs that are detectable in the corpus and backed by at least two IntOGen methods (Results 5.5), so it prioritizes a short list and does not enumerate a cancer's drivers.

---

## 7 Scope and limitations

**What the ranking measures.** OncoDigger ranks genes that co-occur with mutation-focused retrieved literature and does not extract gene-specific mutation relations, so a ranked gene is not necessarily reported as mutated in the retrieved abstracts (for example PARP1 in Prostate Cancer, MGMT in Brain Cancer and IGH in Myeloma, Supplementary Table S4). The proportion of supporting abstracts that report an alteration in the ranked gene has not been measured.

**Recall gap (Results 5.5).** Never-surfaced IntOGen pairs concentrate in IntOGen's single-method, lower-significance tier (76.6% versus 40.6% of surfaced pairs, Supplementary Results S9), which does not make them false driver calls, and "never surfaced" is not "absent from the literature". Only 10.8% of pairs have no lexicon-detected mention in the full corpus, and three genes (H3-3A, H3C2, H4C9) are missing from the vocabulary under current symbols (Supplementary Results S10). Three pairs with substantial literature remain unresolved after the strictest filter (Supplementary Results S12).

**Query, corpus and reference scope.** The study benchmarks one mutation-focused query against one primary reference standard, and the `driver mutation` check shows wording robustness within it. Each corpus pools molecular subtypes into one ranking, so top-10 lists are subtype-blended. OncoDigger does not exclude abstracts found in more than one cancer, so a paper can appear in several collections and rankings, and the 19 corpora are not fully independent (for example 1,388 papers, 5.2% of the Myeloma corpus, are also indexed under Leukemia), though this did not inflate the two cancers' ranking similarity (Jaccard 0.111, below the 0.165 mean). OncoDigger's corpora are PubMed abstracts only, unlike the full-text-indexing comparators (Table 1), and the live OncoDigger website differs from this frozen release. The pooled control is a cancer-blind negative control and not a cancer-conditioned pooled search, so it does not test whether query-time conditioning of a single pooled index would match separate collections.

**Metastatic-site literature.** The collections include records indexed under the /secondary MeSH subheading, among them studies of metastases to the named site (16.0% of the Liver collection, 10.5% of Brain, 9.8% of Lung, 8.5% of Kidney, 8.4% of Melanoma and 2.4% or less of the others). In a post hoc analysis across all 19 cancers these records were removed and the rankings recomputed with the unchanged pipeline. Six of 190 top-10 positions changed, in five cancers, and any-cancer Precision@10 moved from 93.7% (178/190) to 93.2% (177/190) while exact concordance stayed at 78.9% (150/190). Overall concordance is therefore robust to this exclusion, although individual rankings are sensitive to metastatic-site literature. The /secondary qualifier is an imperfect proxy that also removes relevant primary-cancer studies, so the collections should not be read as containing only primary tumours of the named cancer type.

**CancerMine and IntOGen.** CancerMine's fixed 2023-03-01 snapshot (Zenodo 7689627; Lever, 2023) cannot incorporate later literature, but a rerun on truncated collections left OncoDigger's advantage at a similar magnitude (Results 5.10). IntOGen's post-processing removes candidate genes absent from the CGC and an 80-gene whitelist that have no CancerMine literature reference, so the comparison with CancerMine is not fully independent of IntOGen, although removing this gate changed no top-10 result (Results 5.9, intogen-plus, 2024). IntOGen detects somatic point-mutation signal only, so fusion- or germline-driven genes can be undetectable (Results 5.3), and its method weights are calibrated using CGC membership, so it is independent in its evidence but not fully in its method (Results 5.9).

**Short aliases.** To avoid collisions with common words, the vocabulary does not match aliases shorter than four characters unless they are the gene's own symbol. This also excludes protein-product names such as p53 and p21 (about 5,200 papers in the Breast collection contain the token p53, against about 2,050 for TP53), which may omit some gene-associated evidence. The effect of the exclusion on the rankings has not been quantified.

**Gene-recognition ambiguity.** The ambiguity audit targeted all genes appearing in the reported rankings and controls, with replacement entries checked after correction. It is a bounded audit and not a guarantee that no ambiguous term remains elsewhere in the vocabulary, and the term readings were made by a language-model agent that had seen the manuscript (samples of 6 to 30 abstracts per term) without blinded human review. A separate, systematic post hoc review of the top-10 supporting abstracts is described below and in Supplementary Results S16. Eight real gene names with mixed, collection-dependent meanings (KIT, MET, APC, POLE, DDIT3 through CHOP, MIA, BAD and FAS) and the family-name aliases HDAC, VEGFR, HSP90, PARP and PI3K were retained, because the audit adopted no cancer-specific recognition rule and did not resolve family-name mapping. A targeted check of the 30 gene entries that newly entered a top-50 list after the rebuild found two strongly contaminated entries, TTR in Colorectal Cancer (rank 50) and FH in Breast Cancer (rank 49), and one mixed case, APC in Melanoma (rank 49). They were retained and disclosed and not corrected, to avoid a further post hoc revision of the vocabulary. None entered a top 10, and counting all three as non-drivers would lower mean Precision@50 by 0.1 point (77.6% to 77.5%) and leave Precision@10 unchanged, but the count of three is not a verified bound on every ambiguous match at greater ranking depths (Supplementary Methods S1).

**Post hoc audit of the top-10 genes.** A separate, post hoc, language-model-assisted review of the supporting abstracts of all 190 top-10 gene-cancer pairs (Supplementary Results S16) judged 89 of 7,808 randomly or exhaustively sampled gene-abstract credits (1.14%) to credit the wrong gene, and 44 as uncertain. Errors were concentrated: seven pairs exceeded 10%, through the aliases PI3K (PIK3CA) and HIF-1 (SETD2). Removing the identified aliases and rescoring with exact background counts left top-10 membership and the 178/190 and 150/190 concordance unchanged. This shows robustness to the tested alias mappings only, not that no other recognition error exists. The review used one unblinded reader and has no independent human validation, and it did not assess missed mentions or whether an abstract reports a mutation of the ranked gene.

**Capitalization filter (Methods 3.3).** The capitalization-based false-positive filter can admit a common word that is capitalized only by title-casing convention. A check of MET and REST across all 19 corpora found a small effect (3 of 4,615 retained MET postings and 11 of 202 retained REST postings, `scripts/analysis/title_case_blind_spot_check.py`), which is not corrected here.

**Post hoc analyses.** The decomposition (Results 5.8) is exploratory. The two full-corpus/real-formula cells were not evaluated, the 10-alias mechanism was identified by inspecting results, and all of it used a single query and discovery mode. A post hoc, exploratory proximity weighting that compares weights against IntOGen is reported only in Supplementary Results S15.

## Data and code availability

OncoDigger rankings were generated with the engine source pinned at commit 4e39fc0 of the private application repository (2026-05-31), vendored in `reference_engine_code/`, together with the analysis scripts, the exact background count (Methods 3.2) and vocabulary policy v5 (Methods 3.3) applied to it. The companion repository, linked below, contains all of them. Twelve further supplementary tables are supplied as CSV files with the release (Repository Tables RT1 to RT12, `data/supplementary_tables/`, index in `INDEX.md`), together with the vocabulary-policy history (Repository note VH1, `docs/VOCABULARY_POLICY_HISTORY.md`).

github.com/jfataphd/OncoDigger-gene-prioritization, release `v1.3`, archived at Zenodo (https://doi.org/10.5281/zenodo.23774746; the concept DOI 10.5281/zenodo.23269249 resolves to the latest version)

The repository contains the corrected gene vocabulary, quality-control rules, ranking code, settings and derived result tables used here, including the IntOGen statistical-support sensitivity analysis (`scripts/analysis/intogen_statistical_support_rescoring.py`) and the as-run row-level audit scripts and outputs behind Supplementary Results S13 (`scripts/analysis/independence_audit/`).

The 19 PubMed collections (about 27 GB) are not redistributed. The release provides, for each collection, the PubMed identifiers with checksums (`data/provenance/pmid_manifests/`) and a script that rebuilds a collection from them by downloading the records from NCBI (`scripts/reproduce/rebuild_corpus_from_pmids.py`); because PubMed records are revised after the fact, a rebuild can differ slightly from the databases used here. Analyses that do not read the databases run from the deposited files (`docs/REPRODUCTION.md`). The release also contains the gene-recognition audit (`data/derived/gene_recognition_audit/`), which holds PubMed identifiers and titles only.

The continuously developed web interface is available at the following address.

`oncodigger.com`

The live site may differ from the frozen manuscript pipeline and should not be used to reproduce the publication results unless it has been updated to the same release.

The IntOGen driver gene compendium used throughout this paper, for both Precision@10 and exact cancer-type driver concordance, is release 2024.09.20 (`Compendium_Cancer_Genes.tsv`, IntOGen, 2024; Martínez-Jiménez et al., 2020), licensed CC0 (public domain) and included in the companion repository at `data/external/intogen/` with no access restrictions. Every derived result in this paper (rankings, precision scores, cancer-type mappings) is fully available in the repository.

---

## CRediT author contribution statement

**Jimmie E. Fata.** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing – Original Draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

## Declaration of competing interests

The author is the creator and developer of OncoDigger, the software evaluated in this study, and declares no financial competing interests. OncoDigger is a freely accessible website with no commercial licensing, subscription or sales, and none is planned.

## Funding

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

## Use of generative AI

OncoDigger and the research objectives of this study were conceived by Jimmie E. Fata. Large language models, principally Claude (Anthropic) and ChatGPT (OpenAI), were used extensively to assist with Python programming, statistical and computational analysis, development and refinement of analytical approaches, validation scripts, figure generation, and manuscript drafting and editing. AI-assisted suggestions included substantive contributions to the design and implementation of the benchmarking and sensitivity analyses. The author directed the research, evaluated the proposed approaches, interpreted the findings, reviewed the AI-assisted outputs against the underlying data and scientific sources, and made all final scientific and editorial decisions. Analysis and validation scripts are provided with the companion materials to facilitate independent examination. The author assumes full responsibility for the accuracy, integrity and reproducibility of the study.

---

## References

Bryant HE, Schultz N, Thomas HD, et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature.* 2005;434(7035):913–917. PMID: 15829966.

Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature.* 2013;499(7456):43–49. PMID: 23792563.

Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2012;2(5):401–404. PMID: 22588877.

Chen L, Liu H, Friedman C. Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics.* 2005;21(2):248–256. PMID: 15333458.

Farmer H, McCabe N, Lord CJ, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature.* 2005;434(7035):917–921. PMID: 15829967.

Fontaine JF, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Génie: literature-based gene prioritization at multi genomic scale. *Nucleic Acids Res.* 2011;39(Web Server issue):W455–W461. PMID: 21609954.

Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* 2013;6(269):pl1. PMID: 23550210.

Gorlov IP, Pikielny CW, Frost HR, et al. Gene characteristics predicting missense, nonsense and frameshift mutations in tumor samples. *BMC Bioinformatics.* 2018;19(1):430. PMID: 30453881.

Grissa D, Junge A, Oprea TI, Jensen LJ. Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration. *Database (Oxford).* 2022;2022:baac019. PMID: 35348648.

Hegi ME, Diserens AC, Gorlia T, et al. MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med.* 2005;352(10):997–1003. PMID: 15758010.

Herman JG, Umar A, Polyak K, et al. Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc Natl Acad Sci U S A.* 1998;95(12):6870–6875. PMID: 9618505.

Idos G, Hampel H, Valle L. Lynch Syndrome. 2004 Feb 5 [Updated 2026 Sep 17]. In: Adam MP, Feldman J, Mirzaa GM, et al., editors. *GeneReviews.* Seattle (WA): University of Washington; 1993–2026. PMID: 20301390.

IntOGen. *IntOGen cancer driver genes compendium, release 2024.09.20 (IntOGen-Drivers-20240920).* Barcelona Biomedical Genomics Laboratory; 2024. https://www.intogen.org/download. PMID: not applicable (dataset).

intogen-plus. *intogen-plus pipeline, tag v2024* [software]. Barcelona Biomedical Genomics Laboratory; released 2024-09-19. https://github.com/bbglab/intogen-plus/releases/tag/v2024. PMID: not applicable (software).

Jourquin J, Duncan D, Shi Z, Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. *BMC Genomics.* 2012;13(Suppl 8):S20. PMID: 23282288.

Kamihara J, Schultz KA, Rana HQ. FH Tumor Predisposition Syndrome. 2006 Jul 31 [Updated 2025 May 8]. In: Adam MP, Feldman J, Mirzaa GM, et al., editors. *GeneReviews.* Seattle (WA): University of Washington; 1993–2026. PMID: 20301430.

Kowdley KV. Iron, hemochromatosis, and hepatocellular carcinoma. *Gastroenterology.* 2004;127(5 Suppl 1):S79–S86. PMID: 15508107.

Kuehl WM, Bergsagel PL. Early genetic events provide the basis for a clinical classification of multiple myeloma. *Hematology Am Soc Hematol Educ Program.* 2005:346–352. PMID: 16304402.

Lai PT, Wei CH, Luo L, Chen Q, Lu Z. BioREx: improving biomedical relation extraction by leveraging heterogeneous datasets. *J Biomed Inform.* 2023;146:104487. PMID: 37673376.

Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013;499(7457):214–218. PMID: 23770567.

Lever J. *CancerMine data, release of 2023-03-01* [dataset]. Zenodo; 2023. DOI: 10.5281/zenodo.7689627. PMID: not applicable (dataset).

Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods.* 2019a;16(6):505–507. PMID: 31110280.

Lever J, Jones MR, Danos AM, et al. Text-mining clinically relevant cancer biomarkers for curation into the CIViC database. *Genome Med.* 2019b;11(1):78. PMID: 31796060.

Martínez-Jiménez F, Muiños F, Sentís I, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer.* 2020;20(10):555–572. PMID: 32778778.

Meindl A, Hellebrand H, Wiek C, et al. Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. *Nat Genet.* 2010;42(5):410–414. PMID: 20400964.

Metzler D, Croft WB. A Markov random field model for term dependencies. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM; 2005:472–479. PMID: not indexed in PubMed. DOI: 10.1145/1076034.1076115.

Morgan AA, Lu Z, Wang X, et al. Overview of BioCreative II gene normalization. *Genome Biol.* 2008;9(Suppl 2):S3. PMID: 18834494.

National Library of Medicine. PubMed User Guide: MeSH Terms [mh]. https://pubmed.ncbi.nlm.nih.gov/help/ (accessed 2026-10-29). PMID: not applicable (online documentation).

O'Connell FP, Pinkus JL, Pinkus GS. CD138 (syndecan-1), a plasma cell marker: immunohistochemical profile in hematopoietic and nonhematopoietic neoplasms. *Am J Clin Pathol.* 2004;121(2):254–263. PMID: 14983940.

Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* 2017;45(D1):D833–D839. PMID: 27924018.

Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: text mining and data integration of disease-gene associations. *Methods.* 2015;74:83–89. PMID: 25484339.

Pritchard CC, Mateo J, Walsh MF, et al. Inherited DNA-repair gene mutations in men with metastatic prostate cancer. *N Engl J Med.* 2016;375(5):443–453. PMID: 27433846.

Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. *Found Trends Inf Retr.* 2009;3(4):333–389. PMID: not indexed in PubMed. DOI: 10.1561/1500000019.

Sayers EW, Beck J, Bolton EE, et al. Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* 2025;53(D1):D20–D29. PMID: 39526373.

Seal RL, Braschi B, Gray K, McClay J, Tweedie S, Bruford EA. Genenames.org: the HGNC and PGNC resources in 2026. *Nucleic Acids Res.* 2026;54(D1):D1098–D1107. PMID: 41287213.

Tao T, Zhai C. An exploration of proximity measures in information retrieval. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM; 2007:295–302. PMID: not indexed in PubMed. DOI: 10.1145/1277741.1277794.

Topalian SL, Hodi FS, Brahmer JR, et al. Safety, activity, and immune correlates of anti–PD-1 antibody in cancer. *N Engl J Med.* 2012;366(26):2443–2454. PMID: 22658127.

Tuason O, Chen L, Liu H, Blake JA, Friedman C. Biological nomenclatures: a source of lexical knowledge and ambiguity. *Pac Symp Biocomput.* 2004:238–249. PMID: 14992507.

Wei CH, Allot A, Lai PT, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. *Nucleic Acids Res.* 2024;52(W1):W540–W546. PMID: 38572754.

---

# Supplementary Methods S1. Corpus reconstruction, vocabulary quality control and reproducibility

**Supplementary Table S1. Cancer-specific literature collections.** Coverage is the percentage of each collection represented in the scoring pool (up to 1,000 papers, 861 for Cervical Cancer and 817 for Myeloma). Vocabulary size is the number of distinct word tokens indexed for that collection, and abstract length is the mean number of words per abstract.

| Cancer | Papers | Coverage of pool | Vocabulary size | Mean abstract length (words) |
|---|---|---|---|---|
| Breast Cancer | 250,061 | 0.40% | 427,569 | 226 |
| Lung Cancer | 178,908 | 0.56% | 339,571 | 225 |
| Colorectal Cancer | 168,174 | 0.59% | 327,311 | 229 |
| Liver Cancer | 127,252 | 0.79% | 273,114 | 223 |
| Prostate Cancer | 110,937 | 0.90% | 239,258 | 235 |
| Leukemia | 106,922 | 0.94% | 232,482 | 201 |
| Brain Cancer | 103,180 | 0.97% | 220,319 | 216 |
| Lymphoma | 79,022 | 1.27% | 171,981 | 202 |
| Melanoma | 66,447 | 1.50% | 184,880 | 209 |
| Pancreatic Cancer | 65,223 | 1.53% | 168,545 | 218 |
| Stomach Cancer | 64,372 | 1.55% | 163,064 | 225 |
| Ovarian Cancer | 61,538 | 1.63% | 164,033 | 221 |

| Cancer | Papers | Coverage of pool | Vocabulary size | Mean abstract length (words) |
|---|---|---|---|---|
| Cervical Cancer | 50,875 | 1.69% | 137,776 | 234 |
| Kidney Cancer | 46,739 | 2.14% | 125,170 | 219 |
| Thyroid Cancer | 38,640 | 2.59% | 105,443 | 228 |
| Esophageal Cancer | 37,925 | 2.64% | 108,665 | 228 |
| Bladder Cancer | 35,620 | 2.81% | 110,528 | 226 |
| Myeloma | 26,591 | 3.07% | 88,809 | 209 |
| Endometrial Cancer | 22,079 | 4.53% | 75,563 | 231 |

**Corpus build and reconstruction (Methods 2.1).** The 2026-09-27 build is the 2026-05-08 baseline (1,608,566 abstracts) plus an incremental PubMed fetch for publication dates 2026/05/01 to 2026/09/27, net of records removed during the refresh. The vocabulary audits and manual samples of Methods 3.3 (the 40-posting MET sample and the short-symbol checks) were read on the 2026-05-08 text, and the corrections derived from them are vocabulary-level. The raw abstracts needed by the audit scripts for the 27 September build were reconstructed from the 2026-05-08 raw files and the incremental fetch and verified against the databases. All 1,640,505 records match on title and abstract after HTML-entity decoding, except three whose text differs between the database and the reconstructed raw file, namely PubMed 42004856 (Pancreatic Cancer collection), 42101311 (Esophageal Cancer) and 42018436 (Thyroid Cancer). In each, the two versions carry different wording of the same record (a changed title word and rephrased abstract sentences in the first, a trailing keyword block in the second, minor phrasing in the third), consistent with a PubMed revision between fetches. Rankings are computed from the database text, and the raw files are read only by the audit and ablation scripts, where each record can change at most one posting. Neither version of any of the three records mentions a top-50 gene symbol except BRAF (Thyroid Cancer), which appears in both. Analyses with their own dates are documented separately, namely the date-truncated sensitivity analyses (Methods 2.7) and the frozen external snapshots and live-service counts (CancerMine, PubTator, DISEASES, the ChatGPT response, Supplementary Methods S2 to S4).

**Supplementary Table S2. Exact corpus construction queries (Methods 2.1).** Each collection was built via the NCBI PubMed E-utilities API using the query `(<MeSH term>[mh]) AND english[la] AND hasabstract`, restricted to publication dates 2000/01/01-2026/09/27, with the cancer-specific MeSH term below. Unlike the PubTator subtype-mapping limitation corrected in Methods 2.4, this single-term query does not under-count subtype-specific literature that is MeSH-indexed under the exploded term. PubMed's `[mh]` tag explodes by default to every narrower term in the MeSH tree (National Library of Medicine, 2026, e.g. `liver neoplasms[mh]` already includes `carcinoma, hepatocellular` and other child terms, confirmed via NCBI E-utilities, 217,667 results with explosion versus 199,945 with `[mh:noexp]`), whereas PubTator's bulk relation file carries only the specific disease identifier each paper was tagged with and has no equivalent automatic hierarchy at query time. All 19 collections were retrieved and incrementally refreshed from the same live corpus infrastructure (`data/external/PROVENANCE.md` documents the equivalent identity anchors for external comparator datasets, and the repository deposits, for each of the 19 collections, the PMID list with the SHA-256 of the list and of the database (`data/provenance/pmid_manifests/`) and the build record with paper counts, refresh interval and database hashes (`data/provenance/corpus_build_2026-09-27.json`). The databases themselves and the exact retrieval timestamps are held by the corresponding author. A script rebuilds a collection from its PMID list by downloading those records from NCBI (`scripts/reproduce/rebuild_corpus_from_pmids.py`); because PubMed records are revised after the fact, a rebuild can differ slightly from the databases used here. Every analysis stage that does not read the databases

runs from the deposited files (`docs/REPRODUCTION.md`).

| Cancer | MeSH term |
|---|---|
| Bladder Cancer | urinary bladder neoplasms |
| Brain Cancer | brain neoplasms |
| Breast Cancer | breast neoplasms |
| Cervical Cancer | uterine cervical neoplasms |
| Colorectal Cancer | colorectal neoplasms |
| Endometrial Cancer | endometrial neoplasms |
| Esophageal Cancer | esophageal neoplasms |
| Kidney Cancer | kidney neoplasms |
| Leukemia | leukemia |
| Liver Cancer | liver neoplasms |
| Lung Cancer | lung neoplasms |
| Lymphoma | lymphoma |
| Melanoma | melanoma |
| Myeloma | multiple myeloma |
| Ovarian Cancer | ovarian neoplasms |
| Pancreatic Cancer | pancreatic neoplasms |
| Prostate Cancer | prostatic neoplasms |
| Stomach Cancer | stomach neoplasms |
| Thyroid Cancer | thyroid neoplasms |

**Pool-size sensitivity (Methods 2.1).** Relative to the 1,000-paper analysis pool used throughout this paper, mean top-10 Jaccard similarity was 0.763 at a 250-paper pool, 0.952 at 750 papers and 0.887 at 2,000 papers. Adjacent-pool comparisons (250/500, 500/750, 750/1,000, 1,000/1,500, 1,500/2,000) gave 0.808, 0.869, 0.952, 0.914 and 0.943 respectively (mean 0.897) (`scripts/analysis/pool_size_jaccard.py`). Cervical Cancer and Myeloma have only 861 and 817 abstracts containing a query word, so pools of 1,000 papers or more give them identical top-10 lists. Excluding the two, the similarity to the 1,000-paper pool was 0.756 at 250 papers, 0.947 at 750 papers and 0.873 at 2,000 papers, and the adjacent comparisons gave 0.816, 0.884, 0.947, 0.904 and 0.936 (mean 0.897) (`scripts/analysis/pool_size_jaccard_unsaturated.py`). Both views support 1,000 papers as a reasonable operating point, with some change in top-10 membership at the smaller pool sizes, rather than an isolated or cherry-picked choice.

**MET positive gene-context terms (Methods 3.3).** The pre-identified positive gene-context terms used to classify MET's 40-posting robustness sample, and to build the stricter 1,268-posting condition, are frozen in `scripts/analysis/met_disambiguation_check.py`'s `POSITIVE_TERMS` list, namely "c-met," "proto-oncogene," "protooncogene," "receptor tyrosine kinase," "hepatocyte growth factor," "hgf," "amplification," "exon 14," "crizotinib," "capmatinib," "tepotinib," "savolitinib," "kinase inhibitor," "juxtamembrane," and the compounds "met receptor/gene/oncogene/mutation/inhibitor/amplif/kinase/signaling/signalling/pathway/alterations," plus "copy number." A posting was classified "genuine" if it matched one of these terms and none of the negative-phrase patterns (Methods 3.3). This is the same rule used to build the 1,268-posting positive-context-only condition.

**Vocabulary exclusions and the original term-level suppressions (Methods 3.3).** Seven symbols (MICE, DCR, PC, C2, FBN1, SGCG, WDHD1) are excluded from all rankings (`GENE_EXCLUDE`). MICE, DCR, PC and C2 produce common-word or abbreviation collisions with no safer alias. FBN1, SGCG and WDHD1 were inherited from the live application's earlier quality-control policy and were tested post hoc (`scripts/analysis/gene_exclusion_readmission_test.py`) by rerunning the primary ranking for all 19 cancers with the exclusion lifted. Admitting every term of the three genes changed no top-10 slot and 208 of 950 top-50 slots, all through WDHD1, which entered the top 50 of nine cancers (at ranks 18 to 47) and no top 10, because its alias AND-1 normalises to the phrase "and 1" (about 6,800 of 250,061 Breast abstracts contain it, against 2 for the symbol WDHD1). Admitting only the approved symbol and approved full name of the three genes left all 950 top-50 slots unchanged, so the exclusions act as an alias-collision control and do not alter the reported rankings. A first set of 13 symbols (GC, HR, HCCS, BPIFA4P, MB, ELN, CP, UBC, CAMP, EFS, IVD, SCT and SON) is handled by term-level suppression, which removes only the bare ambiguous symbol and leaves the gene's full approved name and every other alias matchable under the same identity. Four of them (GC, HR, HCCS and BPIFA4P) were identified first and are also indexed under safer aliases (VDBP, KDM3D, CCHL and LATH). The other nine came from a post hoc sweep of every symbol of four characters or fewer in the top-50 rankings (66 candidates), each verified against live PubMed text, for example MB against MDA-MB cell lines, megabase and medulloblastoma, and CP against chronic phase and chronic pancreatitis. INS and TPO were not confirmed and were left unchanged. The evidence for each symbol is in Repository note VH1 (`docs/VOCABULARY_POLICY_HISTORY.md`). The later v4 and v5 corrections extend this set (below), so the final policy excludes 7 genes and suppresses 31 bare symbols and 51 single-word aliases at term level. Four of the 31 symbols (GC, HR, HCCS and BPIFA4P) have no entry in this paper's base vocabulary, so 78 of the 82 declared terms change the matcher. The live application's `gene_quality.py` policy differs from this one (it is broader and does not yet include the v4 and v5 additions; Repository note VH1). The analysis scripts that read the application's unfiltered lookup table (Methods 3.4, Supplementary Results S7) import `gene_quality.py` policy version 2026-09-27-v1 (commit `d592370` of the live-application repository), vendored in `reference_engine_code/gene_quality_test_engine/`.

**History of the first nine corrections (Methods 3.3, Results 5.1-5.3).** The nine symbols above were first corrected by whole-gene exclusion, as the original seven were, and then refined to term-level suppression, because exclusion treated them more harshly than the four symbols identified first; SCT alone among them recovers a genuine match once only its bare symbol is suppressed, since "secretin" independently earns Pancreatic Cancer rank 42 on the 2026-09-27 build. This correction was made before the paper adopted IntOGen as its primary reference. All numbers reported in this paper come from the final pipeline of Methods 3.2 and 3.3, and the only earlier figures quoted are the v1.1 release values in the background-count paragraph below (details in Repository note VH1).

**Background count (Methods 3.2).** The relevance score uses B, the number of collection papers in which a gene is found. The engine's built-in count (`entity_background_papers`) took the smallest postings count among a gene's matched single-word terms, contributed nothing from matched multi-word terms (a gene matched only as a phrase, for example NKX2-1, had a count of zero, which the caller replaced by the supporting count) and used a different recogniser from the pool matcher, so it understated B for most genes. This was found in a post hoc review of the formula (`exploratory/formula_study/`) and corrected before the analyses reported here were run. The count used for every ranking in this paper is exact: the engine's own per-paper matching code (single tokens, then phrases by anchor token and substring test, with longest-match masking and alias suppression, excluding the query words and the gene's own label) was run over every paper of the collection once per query, counting each paper once per gene (`scripts/analysis/exact_background_counts.py` and `exploratory/background_rebuild/`; specification `docs/BACKGROUND_COUNT_SPEC.md`, written before the

recomputation). Two tests were required before any rerun was used, and both passed in the Myeloma and Lung Cancer pilot. The matcher run over the retrieved pool reproduced the engine's supporting count for every candidate gene (224 of 224 and 135 of 135), and supporting_papers ≤ background_papers ≤ collection size held for every gene. The same two conditions, and 0 < enrichment ≤ collection size / pool size, were checked at every candidate gene of every rerun, which stops on any violation, and no run stopped. In the 19 main rankings, 3,955 gene-cancer candidates were evaluated. The engine's count was lower than the exact count for 3,202 of them (81.0%), equal for 753 and higher for none, and it was below the supporting count itself for 227 (`data/provenance/hook_stats_main_ranking.json`). The formula, BM25 settings, pool sizes, vocabulary and its corrections were not changed. Every analysis that uses B was rerun with the exact count (the pooled control, the cutoff-matched runs, the vocabulary-state runs and the sensitivity analyses), and analyses that do not use B, namely flat counts and the external rankings, were unchanged. Replacing the engine's count with the exact count changed 22 of 190 top-10 slots in 14 cancers and 84 of 950 top-50 slots (measured before the later v4 and v5 vocabulary corrections, which changed no top-10 slot and 30 further top-50 positions; the v1.1 and final lists differ in 22 top-10 and 101 top-50 slots), raised exact concordance from 142 to 150 of 190 (paired difference +4.2 points, 95% cancer-bootstrap CI +1.6 to +6.8) and any-cancer Precision@10 from 177 to 178 of 190 (+0.5 points, CI -1.1 to +2.1), lowered the mean pairwise top-10 Jaccard similarity between cancers from 0.187 to 0.165, and lowered the pooled control's exact concordance from 111 to 101 of 190. The earlier analysis (release v1.1 of the companion repository, Zenodo DOI 10.5281/zenodo.23269250) used the engine's count and is superseded by the numbers in this paper.

**Collision screen on the exact-background rankings (Methods 3.3).** The predeclared screen of the vocabulary audit (`scripts/analysis/vocab_collision_risk_screen.py`; a gene is RED if fewer than 50% of its matched papers contain the standalone capitalised symbol, AMBER if 50% to 85% do or fewer than 30% of its matches lie near a generic qualifier word, and GREEN otherwise, with the same thresholds and the same generic qualifier list for every gene) was rerun on the 347 genes of the exact-background top-50 lists, before the v4 and v5 corrections, and gave 268 GREEN, 61 AMBER and 18 RED. Of the 36 genes new to any top-50 list (listed in Repository note VH1), 19 are IntOGen drivers. Two were RED (RHO and SHE) and eight were AMBER (AXIN2, CCS, FAM107B, ID3, MT-ATP6, PTPRD, SMO and TRAF2). SHE, RHO and CCS were read and corrected in policy v4 (below). Conversely, all sampled abstracts for ID3, SMO, PTPRD, AXIN2 and TRAF2 (10, 10, 7, 5 and 8) were genuine mentions of the gene (for example mutations in Burkitt lymphoma, the hedgehog pathway component Smoothened in medulloblastoma, and PTPRD inactivation in melanoma), so they were retained. The abstracts were drawn at random with a fixed seed from the collection in which the gene ranks and were read by a language-model agent that had seen the manuscript, so this is not independent validation. FAM107B and MT-ATP6 each have a single matched paper and were not read.

**Vocabulary policy v4 and v5 (Methods 3.3).** After the exact-background rerun the collisions were corrected in two bounded steps under one frozen, versioned policy (`scripts/analysis/vocab_policy.py`, version 2026-10-16-v5). Every cleaned-vocabulary analysis reads its lists from it, and the true-raw arms do not apply it. Step 1 (v4) suppressed the bare symbols SHE, RHO, CCS and NODAL and the aliases RANK of TNFRSF11A, nudE of NDE1 and Lobe of AKT1S1. The sampled abstracts were the pronoun (8 of 8), Rho GTPases and Spearman's rho (0 genuine of 8 and of 30), other abbreviations for CCS (0 genuine of 8, 3 of 30 genuine), the adjective "nodal" (0 of 60), log-rank and rank-sum (0 of 14), nude mice (0 of 30) and anatomical lobes (0 of 30) (`data/derived/vocab_rescoring/vocab_correction_v4_evidence.csv`). Step 2 (v5) came from a vocabulary-wide screen that used only the collections' token counts (no matcher, ranking or reference standard). An alias was a candidate if, in at least one collection, it occurred in at least 200 papers and in at least five times as many papers as its gene's own symbol (133 candidates). A symbol was a candidate if it is an ordinary word among the 20,000 most common English words (the public list `20k.txt` of

github.com/first20hours/google-10000-english, which `vocab_inventory_screen.py` fetches and checks against a recorded checksum) and occurred in at least 300 papers of some collection (32 candidates). The audit covered the 602 genes appearing in any reported ranking or control (119 of the 165 candidates; `data/derived/vocab_rescoring/vocab_inventory_screen_candidates.csv`). The decision rule was fixed before the rebuild: suppress the term if it refers to the gene in at most 10% of sampled papers in every collection examined, retain it if it is a genuine biological name, and retain it as collection-dependent if it is a real gene name that is also a common word with a mixed reading. Of the 122 decisions (`data/derived/vocab_rescoring/vocab_policy_v5_table.csv`, with the collection read, the genuine fraction and the other meaning for each), 69 are suppressions (18 symbols: SHE, RHO, CCS, NODAL, ADM, ARC, CAD, COIL, ESD, GEM, IDE, IDS, MAX, MOS, NHS, POR, PTS and REST; and 51 aliases, for example blast of CD48, race of AMACR, flap of ALOX5AP, killer of TNFRSF10B, fast of FASTK, delta of YY1, wave of WASF1, warts of LATS1, spin of GTF2I, clip of POMC and CD74, sera of PHGDH, nail of CD244, crest of SS18L1 and genesis of FOXD3), 45 are retentions of genuine names (for example HER2, aromatase, PSMA, albumin, insulin, survivin, CA125 of MUC16, CD25 of IL2RA, NRF2 of NFE2L2, HDAC, HSP90, VEGFR, PARP and PI3K, and the symbols AFP, MGMT, RET, ATM, TYR and SRC) and 8 are collection-dependent retentions (KIT, MET, APC, POLE, MIA, BAD, FAS and CHOP of DDIT3). Collection dependence was documented and not turned into a recognition rule. For example CHOP is a gene symbol in solid-tumour collections and a chemotherapy regimen in Lymphoma, where DDIT3 ranks 50th. The family-name aliases HDAC, VEGFR, HSP90, PARP and PI3K were retained and are disclosed: whether a family name should map to one gene is a different ambiguity from an ordinary-word collision, and this audit does not resolve it. The term readings were made by a language-model agent that had seen the manuscript, from seeded samples (`random.seed(42)`) drawn from the postings of the collection in which the term is most prominent: 6 abstracts for every term, 100 screened with gene-context patterns for aliases and 14 to 30 read for symbols and mixed terms. They are not independent validation. A preflight test of the matcher (`scripts/analysis/test_vocab_policy_v5.py`, 195 checks, no failures) confirmed that each suppressed term stops matching its gene, that each suppressed gene stays reachable through another term or its full name, that genes sharing a token are unaffected and that retained terms still match.

**Effect of the v4 and v5 corrections, and the stopping rule (Methods 3.3, Results 5.8).** Rebuilding every cleaned-vocabulary result under v5 left the top 10 identical in all 19 cancers (190 of 190 slots), so the headline Precision@10 (178/190) and exact concordance (150/190), the pooled-control figures (190/190 and 101/190) and the CancerMine and ChatGPT comparisons did not change. Thirty of the 950 top-50 positions changed in the primary ranking, and the precision at deeper ranks and the recall figures moved slightly (Results 5.7, Supplementary Results S5 and S6). The controls that count genes over the whole collection were affected most. The full-corpus cleaned condition of the factorial (Table 2) rose from 82 to 108 of 190 any-cancer hits and the any-cancer interaction fell from +44.2 to +31.1 pp, consistent with false aliases having inflated the full-corpus flat counts, which BM25 retrieval does not filter out (an inference from these results), and the frequency-only control rose from 44.7% to 62.1%. The matched-counting advantage of the full formula fell from +2.1 to +1.6 pp (p=0.083). The 30 gene entries that entered a top-50 list after the correction were then checked in the same way. TTR (Colorectal Cancer, rank 50) occurs in 111 papers of that collection, and all 10 sampled abstracts used it for time to recurrence. FH (Breast Cancer, rank 49) was family history in 18 of 20 sampled abstracts. APC (Melanoma, rank 49) was the gene in 7 of 20 sampled abstracts and otherwise antigen-presenting cells, the anaphase-promoting complex, activated protein C or annual percentage change. MKI67, CTLA4, CD34, BAX and BCAR3 were genuine gene mentions in their samples (2 to 10 abstracts each). No further vocabulary revision was made for Paper 1: these three were disclosed (Section 7) and the frozen policy was not changed after the results were seen. The live application's `gene_quality.py` policy does not yet include the v4 and v5 additions.

**Symbol-presence eligibility filter (Methods 3.3).** In the ranking step (`rank_genes` in `scripts/analysis/run_query_planes.py`) the engine's gene table is read in order, a gene is skipped if it is one of the 7 excluded genes or has already been listed, and it is also skipped if its own lower-case approved symbol has no posting in the cancer's collection (`SELECT COUNT(DISTINCT pmid) FROM postings WHERE token = <symbol>`), until 50 genes are listed. The rule comes from the first version of the script and was not tuned against IntOGen. It is an eligibility rule that requires symbol evidence somewhere in the collection and not a validated recognition correction: a gene with a single literal symbol occurrence and many alias matches elsewhere passes, and a gene that the literature mentions only through a legitimate full name or alias is excluded. In the final analysis (`scripts/analysis/symbol_gate_diagnostic.py`, `data/derived/symbol_gate_diagnostic.csv`: the primary query rerun on the frozen build with policy v5 and the exact background, which reproduces all 950 listed positions) the rule removed 18 candidate entries in 10 cancers, at engine positions 16 to 54 and none in a top 10 (Lung 6, Thyroid 3, Esophageal 2, and Breast, Colorectal, Prostate, Melanoma, Liver, Endometrial and Brain 1 each): AFAP1, APOB, BANK1 (admitted by the alias "bank"), BCAR3, BRAP, H3C6, IGLL2P, IGLL3P, KALRN, NLRP2, PDZK1, PRIMA1, SART3, SPEF1 and TMEM200B. Sixteen of the 18 fell inside the 50 places, so the lists were filled from further down the engine's table and the replacements entered at ranks 47 to 50 in 10 cancers (16 genes, 9 of them IntOGen drivers in some cancer, against none of the 16 removed). Without the rule the mean cumulative Precision@k would be lower by 0.26, 0.53, 0.92 and 0.95 points at k=20, 30, 40 and 50 and unchanged at k=10, so the depth figures of Results 5.7 are slightly more favourable than they would be without it. The rule is applied wherever rankings come from this step (the main, driver-mutation, cutoff-matched, pooled-control and proximity rankings and the relevance rankings of the matched controls). It is not applied to the flat-count rankings of the retrieval-by-vocabulary factorial (Table 2) or to the frequency-only controls, which rank by paper counts. The live application does not apply it.

**Vendored engine source (Methods 3.1).** The "Primary Evidence" content-type multiplier is a term-and-entity-weighted function, not a single fixed value, defined in the OncoDigger application source (`src/oncodigger/search/discovery_modes.py`, commit `4e39fc068102`, 2026-05-31, of the author's private live-application repository, not independently resolvable by readers). Because that repository continues to be actively developed past this commit and is not public, an exact copy of this file and the other engine source files this pipeline depends on (BM25 retrieval, entity aggregation, HGNC vocabulary loading) is vendored in this repository at `reference_engine_code/`, the actual citable and reproducible source for this paper's engine code.

**Original wording of passages condensed in the main text (Methods 2.2, 3.2, 3.3, Section 4, Results 5.2 and 5.8, Discussion and Section 7).** The main text was shortened for length. The passages below are reproduced as they were written before the shortening, so that every detail, number and script path remains in the supplement.

*Methods 2.2, first paragraph.* The query `mutation mutations` was applied to all 19 cancers and scored against IntOGen (release 2024.09.20, 633 driver genes; IntOGen, 2024; Martínez-Jiménez et al., 2020), used as an **external, cohort-computed calibration reference** and not ground truth. Its construction also incorporates prior cancer-gene knowledge (the Cancer Gene Census and CancerMine), so a post hoc sensitivity analysis restricts it to its statistical admission rule (Methods 2.6, Results 5.9). IntOGen's gene membership, cancer-type assignments and significance calls were never used to tune OncoDigger's ranking algorithm, scoring parameters or vocabulary rules, with the exception of the post hoc proximity analysis in Supplementary Results S15, in which weights were compared against it. One audit (Methods 3.3) used IntOGen membership only to select genes for vocabulary-collision scrutiny, and each correction was determined by reading PubMed text. The later v4 and v5 corrections were selected from the genes appearing in the reported rankings and controls and decided by a rule fixed before the rebuild, and whether a correction raised or lowered concordance was not a criterion. A secondary `driver mutation driver mutations` query served as a wording sensitivity check.

*Methods 2.2, statistics paragraph.* Precision@10 is the percentage of a cancer's top 10 genes listed as drivers anywhere in IntOGen's 633-gene compendium. Confidence intervals use a case bootstrap (19 cancers resampled with replacement, 2,000 times), because the cancer type and not the gene is the experimental unit (a Wilson interval is in Supplementary Results S1). Where paired t-tests and Wilcoxon tests are reported they use the 19 per-cancer values and are secondary to the bootstrap intervals. No multiplicity correction is applied across the comparators, apart from the Bonferroni correction of the per-cancer Spearman correlations, so the weaker comparisons are descriptive. The two exact-concordance differences, against CancerMine ($p=1.9x10^{-6}$) and against the pooled control ($p=4.4x10^{-7}$), would survive a Bonferroni correction for 20 tests. The 19 cancers are not independent, because they share genes, biology and some abstracts, so the intervals describe variation across these 19 categories and not across independent biological replicates. Random selection would hit an IntOGen driver about 3.3% of the time (633/19,250 protein-coding genes; the denominator is from HGNC, Seal et al., 2026). Rank AUC and a Spearman correlation summarised quality across each top 50 (`scripts/analysis/secondary_stats.py`).

*Methods 3.2, computational requirements.* **Computational requirements.** The analyses were run on one Windows workstation with 16 logical processors and 63 GB of RAM. The 19 processed collections occupy about 27 GB of SQLite databases. Computing the exact background counts and the 19-cancer ranking for the first time took 86 minutes with 6 worker processes (5,160 s), and the pooled-corpus control over 1,514,148 papers took 75 minutes (4,480 s). Count tables are cached by content, so a repeat of an unchanged ranking is much faster. Peak memory was not recorded, and the main process was observed at about 4 GB. Query latency of the web application is not measured here.

*Methods 3.3, MET and REST audit.* A further audit applied a two-step verification (a standalone capitalized match, then a pre-identified alternate meaning) to every confirmed IntOGen driver gene in the top-50 rankings (195 genes on the 2026-09-27 build). Two required correction, MET (Lung Cancer, rank 9) and REST (Thyroid Cancer, rank 41). Of MET's 4,105 raw postings, 1,733 (42.2%) passed both checks under the final recognition rule and 2,372 (57.8%) were excluded. All 136 REST postings were excluded. The false-positive mentions were redacted from the text and the engine was rerun. MET stayed at rank 9, leaving both headline metrics unchanged, and REST was removed from Thyroid Cancer's ranking. The bare symbol REST was later also suppressed globally in policy v5, so in the final pipeline REST enters no ranking with or without the redaction and the redaction changes no ranking list (MET's supporting count falls from 67 to 58 at an unchanged rank 9, and FH's in Kidney Cancer from 90 to 88 at an unchanged rank 5).

*Methods 3.3, MET sample and strict rerun.* A 40-posting sample of MET postings that passed the rule without a positive gene-context term found 33/40 (82.5%) genuine and four further false-meaning patterns, which were added to the rule. This estimate is diagnostic and not independent validation. As a stricter check, the pipeline was rerun keeping only the 1,268 postings (30.9% of 4,105) with both a capitalized match and a positive context term (`scripts/analysis/vocab_corrected_rerun_met_strict.py`). MET then fell to rank 11 and left the top 10, replaced by NRAS, an IntOGen driver in Lung Cancer, so both headline metrics were unchanged (178/190 and 150/190). The headline is therefore not sensitive to MET's retention rule (Supplementary Methods S1).

*Methods 3.3, collision screen and policy v5.* The audits above were done on rankings computed with the engine's approximate background count (Methods 3.2). The exact count changed which genes reach the top 50, so 36 genes appeared in a top-50 list that had been in no list before, and the predeclared collision screen of Supplementary Methods S1, rerun on the corrected lists (347 genes), flagged SHE, RHO and CCS, all collisions with ordinary words or abbreviations. A whole-vocabulary screen of single-word aliases and ordinary-word symbols (`scripts/analysis/vocab_inventory_screen.py`) then found 165 candidate terms, 119 of which belong to the 602 genes appearing in any reported ranking or control. Each candidate was classified once under a

rule written before the rebuild and independent of IntOGen. A term was suppressed if it referred to the gene in at most 10% of the sampled abstracts in every collection examined, and retained if it is a genuine biological name, including names that are also dictionary words (insulin, albumin, survivin). The abstracts were drawn at random with a fixed seed from the collection in which the term is most prominent (6 for every term, 100 screened with gene-context patterns for aliases, 14 to 30 read for symbols and mixed terms) and were read by a language-model agent that had seen the manuscript, so the readings are not independent validation. The result is the frozen policy v5 (Supplementary Methods S1, `data/derived/vocab_rescoring/vocab_policy_v5_table.csv`): 69 terms suppressed (51 aliases and 18 symbols, including SHE, RHO, CCS and NODAL), 45 terms retained as genuine names, and 8 real gene names with mixed, collection-dependent meanings (KIT, MET, APC, POLE, DDIT3 through its alias CHOP, MIA, BAD and FAS) retained globally and disclosed, because no cancer-specific recognition rule was adopted. The family-name aliases HDAC, VEGFR, HSP90, PARP and PI3K were also retained, since mapping a family name to one gene is a different ambiguity from an ordinary-word collision and the audit did not resolve it. Every cleaned-vocabulary analysis was rebuilt under v5, including the exact-background rankings. The top 10 was identical in all 19 cancers and 30 of 950 top-50 positions changed. Whether a correction raised or lowered concordance was not a criterion for any decision, the frozen table was not revised after the results were seen, and the true-raw control arms (Methods 3.4) remain unfiltered. After the rebuild the 30 gene entries that newly entered a top-50 list were checked, and three ambiguous matches were found (TTR in Colorectal Cancer, FH in Breast Cancer and APC in Melanoma); they were disclosed and left in place (Section 7, Supplementary Methods S1).

*Results 5.2, symbol-only frequency control.* Restricting the frequency-only control to primary gene symbols raised it to 63.7%, narrowing the difference from OncoDigger to 30.0 points.

*Results 5.8, odds-ratio equivalents.* (odds-ratio equivalents about 6.0 for any-cancer membership and 1.3 for exact concordance)

*Results 5.8, minimally corrected vocabulary with the full formula.* Scored with the full formula, that minimally corrected vocabulary reached 83.7% any-cancer and 65.3% exact concordance, against 93.7% and 78.9% with the final vocabulary.

*Results 5.8, interpretation of rank-weighting versus enrichment (clause).* , perhaps because rank-weighting rewards mentions in highly ranked documents whether or not they are genuine gene references, whereas enrichment penalises words that are common throughout the collection

*Section 4, frozen snapshot tag.* (git tag `manuscript1.1-v1.0-vocab-audit-2026-10-04`, retained in the author's private development history and not resolvable by readers)

*Discussion, first paragraph.* , and IntOGen concordance measures agreement with a driver catalogue that combines somatic-mutation statistics with CGC-informed admission rules, not independent proof of biological truth (Results 5.9).

*Discussion, last paragraph.* The approach gives a reproducible way to prioritize mutation-focused literature without treating literature prominence as biological truth.

*Section 7, metastatic-site literature.* **Metastatic-site literature.** The collections are built with MeSH searches that include records indexed under the /secondary subheading, among them studies of metastases to the named site. These records are 16.0% of the Liver collection, 10.5% of Brain, 9.8% of Lung, 8.5% of Kidney, 8.4% of Melanoma (where they concern metastatic melanoma itself) and 2.4% or less of the others. In a post hoc sensitivity analysis across all 19 cancers, these records were removed from the frozen collections, the corpus statistics rebuilt and the rankings recomputed with the unchanged pipeline. Six of 190 top-10 positions changed,

in five cancers (two genes in Liver and one each in Brain, Kidney, Breast and Ovarian Cancer), and the other 14 cancers kept identical top-10 sets. Any-cancer Precision@10 changed from 93.7% (178/190) to 93.2% (177/190) and exact concordance stayed at 78.9% (150/190). Overall concordance is therefore robust to this exclusion, although individual rankings are sensitive to metastatic-site literature. The /secondary qualifier is an imperfect proxy. It also removes relevant primary-cancer studies and misses metastasis-related records indexed another way, and the change in concordance does not show that the removed papers were misattributed. The collections should not be read as containing only primary tumours of the named cancer type.

*Section 7, CancerMine and IntOGen.* **CancerMine and IntOGen.** CancerMine's fixed 2023-03-01 snapshot (Zenodo 7689627; Lever, 2023) cannot incorporate later literature, but a rerun on truncated collections left OncoDigger's advantage at a similar magnitude (Results 5.10). IntOGen's post-processing removes candidate genes absent from the CGC and an 80-gene whitelist that have no CancerMine literature reference, so the CancerMine comparison is not fully independent of IntOGen, although removing this gate changed no top-10 result (Results 5.9, intogen-plus, 2024). IntOGen detects somatic point-mutation signal only, so fusion- or germline-driven genes can be undetectable (Results 5.3). Its methods are also calibrated using CGC membership (method weights, a CGC-only FDR rescue and filter exemptions, intogen-plus, 2024), so it is independent in its evidence but not fully in its method (Results 5.9).

*Section 7, short aliases.* **Short aliases.** To avoid collisions with common words, the vocabulary does not match aliases shorter than four characters unless they are the gene's own symbol. This also excludes some widely used names, such as p53, p21, p16, p65 and p63, which are often used for the protein products of TP53, CDKN1A, CDKN2A, RELA and TP63 rather than for the genes themselves (about 5,200 papers in the Breast collection contain the token p53, against about 2,050 for TP53). Excluding them may omit some gene-associated evidence, whereas including them without contextual disambiguation would add protein-level mentions, such as immunostaining, that do not indicate a mutation. The effect of the exclusion on the rankings has not been quantified.

*Section 7, gene-recognition ambiguity.* **Gene-recognition ambiguity.** The ambiguity audit targeted all genes appearing in the reported rankings and controls, with replacement entries checked after correction. It is a bounded audit and not a guarantee that no ambiguous term remains elsewhere in the vocabulary, and the term readings were made by a language-model agent that had seen the manuscript (samples of 6 to 30 abstracts per term) without blinded human review. A separate, systematic post hoc review of the top-10 supporting abstracts is described in the next paragraph and Supplementary Results S16. Eight real gene names with mixed, collection-dependent meanings (KIT, MET, APC, POLE, DDIT3 through CHOP, MIA, BAD and FAS) were retained globally, and the family-name aliases HDAC, VEGFR, HSP90, PARP and PI3K were retained as well; mapping a family name to one gene is a separate ambiguity that the audit did not resolve. After the rebuild, a targeted check of the 30 gene entries that newly entered a top-50 list found two strongly contaminated entries, TTR in Colorectal Cancer (rank 50; 10 of 10 sampled abstracts used the abbreviation for time to recurrence) and FH in Breast Cancer (rank 49; 18 of 20 sampled abstracts concerned family history), and one mixed case, APC in Melanoma (rank 49; a mixture of the gene and unrelated uses such as antigen-presenting cells and annual percentage change). FH is a genuine gene in Kidney Cancer, where a collection-specific correction applies (Methods 3.3), so it was not suppressed globally. These cases were retained and disclosed and not corrected, to avoid a further post hoc revision of the vocabulary. None entered a top 10 and the primary concordance results do not depend on them. The ranking step also excludes genes whose own symbol never occurs in the collection, which removed 18 candidates below rank 15 and slightly raises the depth figures (Supplementary Methods S1). The count of three describes the entries identified by this check and is not a verified bound on every ambiguous match at greater ranking depths. Counting all three as non-drivers would lower mean Precision@50 by 0.1 point (77.6% to 77.5%) and leave Precision@10 unchanged.

*Methods 2.5.* To test whether IntOGen pairs absent from OncoDigger's top 10 or 50 reflect weak literature coverage and not a ranking failure, all 1,740 gene-cancer pairs were scored for rank at any depth and for supporting-paper count by rerunning the search engine against the corpus databases at an extended entity limit (`scripts/analysis/intogen_literature_weight_audit.py`). Support count is one of the three multiplied inputs to the relevance score, so it is not independent of rank. Genes that never appeared in the ranking were also checked against the vocabulary, because a gene absent from it cannot be recognized (Results 5.5, Supplementary Results S10). The extended ranking run reproduced the main run's top 50 exactly for all 458 IntOGen-driver pairs present in both lists, with no rank discrepancy.

*Methods 2.6.* IntOGen's construction incorporates prior cancer-gene knowledge through the Cancer Gene Census (CGC) and CancerMine (intogen-plus, 2024). To assess sensitivity to these procedures, all systems' top-10 lists were rescored, unchanged, against the IntOGen entries that satisfy its stricter non-CGC statistical admission rule (combined q<=0.05 and at least 2 individually significant methods, which retains 1,780 of the 4,478 catalogue rows), using the same bootstrap (`scripts/analysis/intogen_statistical_support_rescoring.py`). This post hoc analysis is not an independent reference, because CGC-calibrated method weights shape every q-value and cannot be removed without rerunning IntOGen on the raw cohorts, which was not done (Supplementary Results S13).

*Methods 2.7.* CancerMine's snapshot ends on 2023-03-01, whereas OncoDigger's collections run to September 2026. To bound the effect of newer literature, the unchanged pipeline was rerun on collections truncated at publication year 2022 and 2023, with collection size, inverse document frequency, average document length and enrichment background counts recomputed on the truncated set (`scripts/analysis/build_date_truncated_corpora.py`, `scripts/analysis/cutoff_matched_rescoring.py`). Rankings were scored against the same IntOGen reference and compared with CancerMine using the same bootstrap. The MET, REST and FH redaction of Methods 3.3 was applied to the truncated collections as well and changed no top-10 membership (Supplementary Results S14). Truncating by publication year is a literature-age sensitivity analysis and not a reconstruction of PubMed on 2023-03-01, because later revisions of earlier records are included.

*Methods 3.4.* After the vocabulary audits, a post hoc analysis examined how much of the pipeline's IntOGen agreement depends on retrieval, vocabulary cleaning and the relevance formula, and whether they interact, under a frozen written protocol (`FROZEN_ANALYSIS_SPEC_retrieval_vocab_decomposition.md`, companion repository). Two experiments were run across all 19 cancers using production code. The first crossed document retrieval (BM25 pool versus complete corpus) with vocabulary state (cleaned versus true-raw) using flat mention-count ranking, which requires retrieval to be unaffected by vocabulary (confirmed byte-identical for all 19 cancers, `scripts/analysis/retrieval_vocab_independence_check.py`). The second held retrieval at the BM25 pool and compared the full relevance formula with flat counting under both vocabulary states. A three-way factorial was not run, because the enrichment term equals 1 for every gene when the pool is the full corpus. Both experiments used the case bootstrap of Methods 2.2.

*Methods 2.1, second paragraph.* All primary results were computed on one frozen build, the 2026-09-27 release (19 SQLite databases), whose file hashes are recorded in `data/provenance/corpus_build_2026-09-27.json` in the companion repository. The manual vocabulary audits of the first correction round (Methods 3.3) were read on the earlier 2026-05-08 text of the same records, and the corrections derived from them were applied unchanged. The term readings of the later v4 and v5 corrections were drawn from the 2026-09-27 collections. Supplementary Methods S1 describes how the build was assembled and verified. A post hoc sensitivity analysis that excludes records indexed under the cancer term's /secondary MeSH subheading is described in Section 7 (scripts and outputs in `data/derived/metastatic_exclusion/`).

*Methods 2.4.* OncoDigger was compared with CancerMine (Lever et al., 2019a), DISEASES 2.0 (Pletscher-Frankild et al., 2015; Grissa et al., 2022) and PubTator 3.0 (Wei et al., 2024). DisGeNET and CIViCmine were excluded as scope-mismatched (Piñero et al., 2017; Lever et al., 2019b). These resources differ from OncoDigger in gene recognition, source data and scoring, so a score difference reflects the whole system and not corpus separation. They were downloaded between 2026-09-26 and 2026-09-30 (identities in `data/external/PROVENANCE.md`) and audited and corrected before any number was final (Supplementary Table S3, Repository Table RT10, Supplementary Methods S2 to S4). The audit found a real entity-recognition error in CancerMine, where all 1,329 sentences credited to PTGDR were unrelated lncRNA mentions, so PTGDR was removed from the benchmark and the author notified. PubTator genes were ranked by the sum of PubTator's per-disease publication counts across each cancer's matched subtypes. The API returns counts and not PMID lists, so a paper tagged under two matched subtypes of one cancer can be counted more than once. A single ChatGPT free-tier response (27 September 2026) listing each cancer's top 10 most frequently mutated genes was evaluated as an informal practical baseline and not as a competing system. It is one response with no repeated sampling, so its run-to-run variability is not estimated. It stated that it used live web searches grounded in TCGA/cBioPortal data (Cerami et al., 2012; Gao et al., 2013).

*Methods 3.3, first paragraph.* The HGNC vocabulary was cleaned before benchmarking. 143 aliases colliding with another gene's own approved symbol were removed, aliases doubling as ordinary English words were filtered out, a longest-match rule was applied to 8,203 nested-name cases (for example EGF is not counted inside `epidermal growth factor receptor`), 7 genes were excluded entirely, and 31 bare symbols and 51 single-word aliases were suppressed at term level, so that only the ambiguous term is removed and the gene's approved name and its other terms stay matchable (frozen policy 2026-10-16-v5; term lists and the evidence for each decision in Supplementary Methods S1). The result is an audited, ambiguity-filtered vocabulary and not an ambiguity-free one (Section 7). The ranking step also applies a symbol-presence eligibility filter: a gene whose own approved symbol has no posting anywhere in a cancer's collection is not reported, even when the engine recognised it through a full name or an alias (Supplementary Methods S1).

*Section 4, first paragraph (after the first pass).* The environment was fixed for reproducibility (`PYTHONHASHSEED=0`, per-connection caches cleared between cancers), and repeated runs gave byte-identical top-50 lists and rank order. Most data and figures came from a frozen internal codebase snapshot, kept separate from the public web application. The MET, REST and FH rules (Methods 3.3) and the decomposition (Methods 3.4) postdate this snapshot and used the same code. The companion repository (Data and code availability) carries the code, the derived outputs and a build-provenance record that ties every output to the corpus build and engine version that produced it. *Section 7, what the ranking measures.* **What the ranking measures.** OncoDigger ranks genes that co-occur with mutation-focused retrieved literature and does not extract gene-specific mutation relations, so a ranked gene is not necessarily reported as mutated in the retrieved abstracts. PARP1 in Prostate Cancer, MGMT in Brain Cancer and IGH in Myeloma illustrate this (Supplementary Table S4), where the literature concerns drug targeting, promoter methylation or translocation breakpoints and not mutation of the ranked gene itself. The proportion of supporting abstracts that report an alteration in the ranked gene has not been measured.

*Section 7, query, corpus and reference scope.* **Query, corpus and reference scope.** The study benchmarks one mutation-focused query against one primary reference standard, and the `driver mutation` check shows wording robustness within it. Each corpus pools molecular subtypes into one ranking, so top-10 lists are subtype-blended. The 19 corpora are not fully independent by construction (for example 1,388 papers, 5.2% of the Myeloma corpus, are also indexed under Leukemia, 365 of them through the Plasma Cell Leukemia term), though this did not inflate the two cancers' ranking similarity (Jaccard 0.111, below the 0.165 mean). OncoDigger's corpora are PubMed abstracts only, unlike the full-text-indexing comparators (Table 1), and the live OncoDigger website

differs from this frozen release.

*Section 7, overlap between cancer collections.* **Overlap between cancer collections.** OncoDigger does not exclude abstracts that can be found in more than one cancer, so the same paper can appear in the collections and rankings of more than one cancer.

*Discussion, IntOGen independence paragraph.* IntOGen is derived primarily from tumour-sequencing cohorts but also draws on the CGC and CancerMine (Methods 2.6), so the concordance figures do not establish wholly independent validation. Under the statistical-rule restriction every comparative conclusion was preserved (Results 5.9), but that reference is not independent either, and a literature-independent check such as raw mutation recurrence in the cohorts was not performed.

*Results 5.3, recall sentence.* Recall of IntOGen's full 1,740-pair catalogue rose from 8.6% at rank 10 to 14.6% at rank 20 and 26.3% at rank 50, limited by the fixed analysis window, uneven literature coverage and vocabulary limits (Results 5.5, Supplementary Results S9 to S12).

*Methods 2.2, second paragraph.* The ranking is a retrieval-and-enrichment procedure that ranks genes occurring in the abstracts a mutation-focused query retrieves. It performs no gene-specific relation extraction and does not test whether an abstract reports a mutation in the ranked gene, so a gene can rank because it is discussed alongside mutations in other genes (for example as a drug target, biomarker or methylation marker, Supplementary Table S4). Precision@10 against IntOGen therefore measures agreement of the top-ranked genes with somatically mutated driver genes and not the proportion of supporting abstracts that report a mutation in the gene.

## Supplementary Methods S2. Benchmark-audit details

**Supplementary Table S3. How each comparison was audited.**

| Resource | Scoring used here | Audit | Finding | Correction / interpretation | Final result |
|---|---|---|---|---|---|
| CancerMine | Citation counts summed across matched disease terms and three roles | Disease terms checked; supporting sentences inspected for suspicious PTGDR result; cross-resource crosswalk audit (Results 5.4) checked for missed subtype disease terms | 3 disease-term mapping mistakes; all 1,329 PTGDR-supporting sentences were unrelated AS1 lncRNA mentions; Brain Cancer's mapping was separately found to be missing ependymoma/medulloblastoma terms present in CancerMine's own data | 3 terms reassigned; PTGDR excluded; Brain Cancer mapping extended to include the missing subtype terms | 84.2% before correction; 87.9% after correction (against IntOGen) |

| Resource | Scoring used here | Audit | Finding | Correction / interpretation | Final result |
|---|---|---|---|---|---|
| DISEASES 2.0 | Maximum matched association z-score per cancer | Overlapping disease terms and ontology definitions checked | Gallbladder/bladder substring collision; 32 `[organ] lymphoma`/`[organ] melanoma` routing issues in the benchmark mapping; unresolved MYOM2/LOC102723407 pattern | Mapping rules corrected; unresolved pattern retained as an open flag | 60.0% (114/190) |
| PubTator 3.0 | Per-disease distinct-PMID Gene-Disease relation counts, summed across parent + subtype MeSH entities (a paper tagged under two subtypes can be counted twice) | Single-parent-ID mapping found to miss subtype-tagged relations; Liver Cancer bucket (single ambiguous parent code) dominated by ALS-related genes and papers | Single-parent mapping omitted subtype-tagged evidence and produced an ALS-contaminated Liver Cancer ranking | Parent + subtype mapping restored subtype evidence and eliminated the Liver/ALS artifact; Liver Cancer included without exclusion once queried via its own specific subtypes | 67.9% (129/190) after correction (see the note below the table) |
| ChatGPT free tier | Single generated answer | Prompt, response and access conditions saved | Not a bulk/versioned dataset; the exact underlying free-tier model could not be determined | No correction; practical baseline only | 78.9% gene-level Precision@10 (145/190 = 76.3% on exact cancer-type concordance, Results 5.6) |

**PubTator pre-correction value.** The exact pre-correction percentage for PubTator was not separately preserved and is not restated here to avoid citing an unverified figure. A live reconstruction of the original single-parent-MeSH methodology against PubTator's current API (`pubtator_v1_single_parent_reconstruction.py`) returned 67.9% (129/190), indistinguishable from the corrected value rather than confirming the originally reported figure, which indicates that PubTator's underlying relations data has changed since the original benchmark window. The pre-correction value is therefore omitted from Figure 4C.

**CancerMine deduplication check (Methods 2.4).** CancerMine's citation-sum scoring was checked against a fully deduplicated alternative (distinct supporting PMIDs per gene per cancer, from the sentence-level evidence file, Zenodo 7689627; Lever, 2023). Deduplication changed almost nothing. Top-10 membership changed in only 4/19 cancers, pooled Precision@10 was nearly identical (167/190=87.9% citation-sum vs. 166/190=87.4% unique-PMID, `scripts/analysis/intogen_cancermine_double_counting_check.py`), and exact cancer-type concordance likewise (114/190=60.0% vs. 113/190=59.5%, `scripts/analysis/intogen_cancermine_dedup_exact_concordance_check.py`). Dual-diagnosis terms (e.g. adult T-cell leukemia/lymphoma) were allowed to count toward both relevant cancers (Repository Table RT1). OncoDigger and CancerMine were compared per cancer using a paired t-test, Wilcoxon signed-rank test, and paired 2,000-resample case bootstrap, with effect sizes and CIs treated as the primary result.

## Supplementary Methods S3. CancerMine disease-name mapping

The complete unabridged 251 cancer-term pairs are stored with the manuscript data, and the per-cancer mapping summary is Repository Table RT1.

## Supplementary Methods S4. PubTator disease mapping

PubTator 3.0's entity recognition and identifier mapping (Wei et al., 2024), upstream of its BioREx relation-extraction model (Lai et al., 2023), tags subtype mentions (for example "glioblastoma") under their own MeSH identifiers, separate from the broad parent term, so mapping a cancer to its parent descriptor alone systematically under-counts its relations. Checked against the PubTator3 API on 2026-09-30, EGFR carries 1,604 distinct-PMID relations under Glioblastoma's own identifier (D005909) and FLT3 carries 3,199 under Acute Myeloid Leukemia's (D015470), none reachable through the parent codes Brain Neoplasms (D001932) or Leukemia (D007938). Each cancer was therefore mapped to its parent descriptor plus every matched subtype entity in PubTator's disease vocabulary, found through its entity-autocomplete API with the clinically recognised subtype names of that cancer (`SUBTYPE_QUERIES` in `pubtator_benchmark_v2_subtype_aggregated.py`). Two limitations remain. The API returns counts and not PMID lists, so a paper tagged under two matched subtypes of one cancer can be counted more than once (Methods 2.4), and subtype coverage is only as complete as the curated terms.

The disease entities queried for each cancer are listed in Repository Table RT10 (`data/supplementary_tables/RT10_PubTator_disease_entity_mapping.csv`).

Full per-cancer entity mapping and audit trail are in `data/derived/pubtator_v2_entity_mapping.csv` (`scripts/analysis/pubtator_benchmark_v2_subtype_aggregated.py`).

**Supplementary Table S4. Individual audit of the 12 gene-cancer pairs not detected as a driver anywhere by IntOGen (Results 5.3).** Established-role sources are Lynch syndrome (Idos et al., 2004), FH/HLRCC (Kamihara et al., 2006), HFE/hemochromatosis (Kowdley, 2004), RAD51C (Meindl et al., 2010), IGH translocations in myeloma (Kuehl and Bergsagel, 2005), MGMT/temozolomide (Hegi et al., 2005) and PARP1 synthetic lethality (Bryant et al., 2005; Farmer et al., 2005). The roles in the table are plausible biological explanations from these sources, not demonstrated causes of the nonconcordance in the retrieved abstracts.

| Cancer | Gene | Mechanism category | Established role |
|---|---|---|---|
| Kidney Cancer | FH | Germline predisposition | Causes hereditary leiomyomatosis and renal cell cancer (HLRCC) |
| Colorectal Cancer | MLH1 | Germline predisposition | Lynch syndrome mismatch-repair gene |
| Endometrial Cancer | MLH1 | Germline predisposition | Lynch syndrome mismatch-repair gene |
| Colorectal Cancer | MSH2 | Germline predisposition | Lynch syndrome mismatch-repair gene |
| Endometrial Cancer | MSH2 | Germline predisposition | Lynch syndrome mismatch-repair gene |
| Colorectal Cancer | MSH6 | Germline predisposition | Lynch syndrome mismatch-repair gene |

| Cancer | Gene | Mechanism category | Established role |
|---|---|---|---|
| Endometrial Cancer | MSH6 | Germline predisposition | Lynch syndrome mismatch-repair gene |
| Liver Cancer | HFE | Germline predisposition | Variants cause hereditary hemochromatosis (iron overload), which predisposes to hepatocellular carcinoma, particularly in patients with cirrhosis |
| Ovarian Cancer | RAD51C | Germline predisposition | Homologous-recombination repair gene; germline mutations predispose to breast and ovarian cancer |
| Myeloma | IGH | Translocation/fusion partner | Recurrent partner locus in myeloma translocations (e.g. t(4;14), t(11;14)), rather than a protein-coding driver identified through recurrent somatic point mutations |
| Brain Cancer | MGMT | Epigenetic biomarker | Promoter methylation status predicts temozolomide response |
| Prostate Cancer | PARP1 | Therapeutic biomarker | Synthetic-lethality drug target discussed alongside BRCA mutation status |

All 9 unique genes (12 gene-cancer pairs) were individually checked against their established clinical or biological role rather than assumed to fit a general pattern, and each has an established biological or clinical role that can explain its prominence in mutation-focused cancer literature, consistent with IntOGen's disclosed scope limit (Methods 2.3). Absence from the IntOGen compendium does not establish that these genes lack somatic driver activity, particularly the mismatch-repair genes, which can also be inactivated somatically.

# Supplementary Methods S5. IntOGen cancer-type mapping

IntOGen's release 2024.09.20 (`Compendium_Cancer_Genes.tsv`, CC0 licence) assigns each driver gene to the specific cohort(s) in which it reached significance, each cohort labelled with one of 86 cancer-type codes, combined across its seven driver-detection methods into one significance call per gene per cohort (Martínez-Jiménez et al., 2020). These codes were mapped to the study's 19 cancer categories conservatively. A code was included when it denoted the study cancer or a related histology or anatomical site assigned to only one study cancer (for example UTUC, upper-tract urothelial carcinoma, to Bladder Cancer; CHOL, cholangiocarcinoma, to Liver Cancer; MDS, myelodysplastic syndromes, to Leukemia; WT, Wilms tumour, to Kidney Cancer), and codes that could plausibly span more than one study cancer, or that duplicated a more specific code already included, were excluded rather than guessed at (full code list and exclusion reasons in Supplementary Table S5, scoring script `scripts/analysis/intogen_full_rescoring.py`). A gene-cancer pair counted as an exact match only when IntOGen listed the gene as a significant driver specifically within a cohort mapped to that exact cancer. This mapping and matching rule is the single source used for every exact-concordance and recall figure reported for every system in this paper (OncoDigger, the pooled control, CancerMine and the ChatGPT baseline). Because some included codes are broader than the corresponding PubMed collection, exact concordance was recomputed with codes removed (`scripts/analysis/mapping_code_sensitivity.py`). Removing UTUC, MDS and CHOL together lowers OncoDigger's exact concordance from 150 to 149 of 190 (the only pair lost is BRAF in Liver Cancer, which rests on CHOL). Removing every included code that is not the main histology of its cancer (those three plus UM, WT, PANET, UCS, CEAD, PAST, MBL, EPM, ATRT, LIHB and SCLC; the script also lists CML, which has no rows in this release) lowers it to 147 of 190 (BRAF in Liver Cancer, MEN1 in Pancreatic Cancer and KRAS in Brain Cancer are lost). Any-cancer concordance is unaffected by the mapping.

**Supplementary Table S5. IntOGen cancer-type codes included in each study cancer, with excluded codes classified as ambiguous, duplicative, or outside the study's scope.**

| Study cancer | IntOGen codes included |
|---|---|
| Breast Cancer | BRCA |
| Lung Cancer | LUAD, LUSC, NSCLC, SCLC |
| Colorectal Cancer | COAD, READ, COADREAD |
| Prostate Cancer | PRAD |
| Melanoma | MEL, SKCM, UM |
| Bladder Cancer | BLCA, UTUC |
| Kidney Cancer | CCRCC, CHRCC, PRCC, RCC, WT |
| Pancreatic Cancer | PAAD, PANET |
| Liver Cancer | HCC, CHOL, LIHB |
| Stomach Cancer | STAD |
| Esophageal Cancer | ESCA, ESCC |
| Ovarian Cancer | OVT |
| Endometrial Cancer | UCEC, UCS |
| Cervical Cancer | CESC, CEAD |
| Thyroid Cancer | WDTC |
| Brain Cancer | GB, GBM, HGGNOS, LGGNOS, PAST, MBL, EPM, ATRT |

| Study cancer | IntOGen codes included |
|---|---|
| Leukemia | ALL, AML, CLLSLL, MDS |
| Lymphoma | BL, DLBCLNOS, NHL, MLYM |
| Myeloma | PCM |

Of the 86 codes present in the data, 50 are included above. The remaining 36 were excluded, in two distinct categories. Eight are ambiguous or duplicative of a more specific code already included, namely EGC (spans Stomach and Esophageal Cancer), LNM (spans Leukemia, Lymphoma and Myeloma), and the generic LUNG, PANCREAS, STOMACH, BLADDER, PROSTATE and SKIN buckets. The remaining 28 denote cancer types with no corresponding study cancer at all, rather than an ambiguity within this study's scope, namely ACC, ACYC, ANGS, ANSC, BCC, CSCC, ES, GBC, GIST, HNSC, LIPO, LMS, MGCT, MT, NBL, NETNOS, NPC, OS, PGNG, PLMESO, RBL, RMS, SACA, SARCNOS, SIC, SOFT_TISSUE, THYM and VULVA (sarcomas, gastrointestinal stromal tumours, head-and-neck and non-melanoma skin cancers, and other rare or paediatric tumour types not among this study's 19 cancers). One further code, CML (chronic myelogenous leukemia), is present in the IntOGen cohort table and is mapped to Leukemia in the scoring script, but its single cohort has no driver genes in this release, so it is absent from the 86 codes above and affects no result.

---

# Supplementary Results S1. Additional cancer-specificity statistics

**Headline-precision detail (Results 5.2).** A Wilson CI treating all 190 observations as independent gives 89.3-96.4%, narrower than the primary 95% cancer-bootstrap CI (88.9-97.4%) used throughout this paper, since it ignores that genes cluster within 19 cancers rather than spreading across 190 truly independent trials (Methods 2.2). Mean rank AUC's per-cancer range (0.530-0.886) reflects IntOGen's broader, cohort-computed driver catalogue making the top-50 sorting task somewhat harder than a smaller, more selective reference would. The per-cancer Spearman rank-vs-driver rho ranged from -0.046 to -0.497 in absolute value (`scripts/analysis/secondary_stats.py`, `data/derived/suppl_rank_spearman_intogen.csv`). The frequency-only controls apply the same term-level suppressions as the main pipeline (Methods 3.3).

**Exact-concordance and recall detail (Results 5.3).** Among the 28 driver-elsewhere non-matches, BRCA1 (Pancreatic and Prostate Cancer), PALB2 (Breast, Pancreatic and Prostate Cancer) and CHEK2 (Prostate Cancer) illustrate the secondary-tissue pattern (germline-predisposition relevance in a tissue where IntOGen's somatic signal is weak or absent), and CD274 in Lung Cancer is an IntOGen driver only in Lymphoma. By rank interval, exact matches fall from 78.9% at ranks 1-10 to 54.7% at ranks 11-20 and 35.8% at ranks 21-50. Cervical Cancer has the highest Recall@50 of any cancer (39.6%, Repository Table RT11).

**Literature-weight detail (Results 5.5).** Supporting-paper IQRs by band were 53-145 for top 10 (n=150), 8-24 for ranks 11-50 (n=308), and 2-4 beyond rank 50 (n=304) (Kruskal-Wallis H=582.5, $p=3.3\text{x}10^{-127}$, Repository Table RT2). Only 27 of 1,590 non-top-10 pairs (1.7%) matched or exceeded the weakest top-10 gene's supporting-paper count for their cancer, 25 of them at ranks 11-20 (the others at ranks 21 and 25), by up to 2.3 times the threshold (e.g. MYC in Lymphoma, 96 papers at rank 11 vs. a 42-paper threshold, full list in Repository Table RT2). Endometrial and Prostate Cancer, the lowest-concordance cancers (6/10 exact hits each), recover 6 of their 79 and 95 mapped driver genes respectively within the top 10 (Repository Table RT11), at the rates their 60% exact-concordance figures already report.

NPM1 and FLT3 reached the top 20 only in Leukemia, and CDK12 only in Prostate Cancer, illustrating the pipeline's capture of lineage-specific as well as pan-cancer drivers. The most similar cancer pair by top-10 Jaccard was Colorectal and Liver Cancer (0.54), followed by ten pairs at 0.43 (for example Liver/Stomach, Cervical/Lung and Endometrial/Ovarian).

The pooled top 10 was EGFR, KRAS, TP53, BRAF, PIK3CA, NRAS, BRCA1, POLE, BRCA2 and FLT3, all genuine IntOGen drivers (100% any-cancer Precision@10) but identical for every cancer (zero differentiation, treating Leukemia the same as Prostate Cancer). Because a pooled index has one ranking, its pairwise cancer Jaccard similarity is 1.0 by construction. OncoDigger's corresponding mean pairwise Jaccard was 0.165.

Separation-recovered genes included VHL, PBRM1 and BAP1 for Kidney Cancer, NPM1, DNMT3A and TET2 for Leukemia, and MYD88, CD79B and EZH2 for Lymphoma, not obscure findings. Across the 100 IntOGen-driver gene-cancer pairs recovered only through corpus separation, median supporting PubMed count was 317 papers (IQR 169-719, minimum 27; live NCBI counts queried 2026-10-16, `scripts/analysis/cancer_specific_drivers_pubmed_validation.py`). 83 of the 112 separation-recovered genes (74.1%) were also confirmed IntOGen drivers for that specific cancer.

The pooled control's inflated any-cancer score traces to famous pan-cancer genes. IntOGen recognises TP53 and KRAS as significant drivers in 19/19 study cancers each, so the pooled list's cancer-type-specificity score is elevated by the breadth of these particular genes rather than by any cancer-type-aware mechanism. A perfect general any-cancer score is achievable simply by returning famous pan-cancer genes everywhere, while genuine cancer-type-specific accuracy still requires separating the literature by cancer.

CancerMine's corrected mean pairwise top-10 Jaccard was 0.207, down from 0.233 before audit correction. DISEASES was 0.160, and PubTator (subtype-aggregated, all 19 cancers) was 0.184. These values show why cancer specificity cannot be inferred from architecture alone, since DISEASES was as differentiated as OncoDigger (0.165), and PubTator and CancerMine only slightly less so, by raw cross-cancer overlap even though they address broader tasks.

**Collection size.** Precision@10 was not associated with collection size across the 11-fold range of the 19 collections (Spearman rho=0.05, permutation p=0.84; Kendall tau-b=0.03, p=0.87). The lowest value, 70%, occurred in Colorectal and Endometrial Cancer. Each cancer's own Precision@10 is given in the inset table of Figure 6, and because Precision@10 takes only multiples of 10% many cancers share a value.

## Supplementary Results S2. Detailed exact-concordance decomposition

Across 190 OncoDigger gene-cancer observations, scored against IntOGen (Methods 2.3), the breakdown is as follows.

- 150/190 (78.9%) were confirmed IntOGen drivers in the exact cancer OncoDigger ranked them in.
- 28/190 (14.7%) were confirmed IntOGen drivers, but in a different one of the 19 study cancers.
- 12/190 (6.3%) were not detected as significant drivers by IntOGen in any of its 86 cancer-type codes.
- Among the 178 observations confirmed as IntOGen drivers somewhere, 150/178 (84.3%) retained exact cancer-type concordance.

The IntOGen non-matches concentrate in genes with a plausible mechanistic reason to evade point-mutation-based detection, namely TERT (5), EGFR (3), the germline-predisposition genes PALB2 (3) and

BRCA1 and BRCA2 (2 each), and the Lynch-syndrome mismatch-repair genes MLH1, MSH2 and MSH6 (2 each). The 12 observations not detected as a driver anywhere by IntOGen (9 unique genes, namely FH, HFE, IGH, MGMT, MLH1, MSH2, MSH6, PARP1, RAD51C) were each individually checked against their established clinical/biological role, rather than assumed to fit a general pattern. For all 9, literature prominence in these pairings may reflect biological relationships not captured by somatic point-mutation driver detection. FH (Kidney Cancer), MLH1, MSH2 and MSH6 (Colorectal and Endometrial Cancer), HFE (Liver Cancer) and RAD51C (Ovarian Cancer) are germline-predisposition genes. MLH1, MSH2 and MSH6 are Lynch syndrome mismatch-repair genes (Idos et al., 2004), FH causes hereditary leiomyomatosis and renal cell cancer (Kamihara et al., 2006), HFE variants cause hereditary hemochromatosis, which predisposes to hepatocellular carcinoma, particularly in patients with cirrhosis (Kowdley, 2004), and germline RAD51C mutations predispose to breast and ovarian cancer (Meindl et al., 2010). Mismatch-repair genes can also be inactivated somatically in sporadic tumours, for example by MLH1 promoter methylation (Herman et al., 1998), so their absence from the IntOGen compendium reflects how that reference classifies them and is not evidence that they lack somatic activity. Because FH is not an IntOGen-confirmed driver in any cohort, it was not among the 23 genes flagged for vocabulary-collision scrutiny by the IntOGen-driver-guided audit (Methods 3.3), and FH is also a common clinical abbreviation ("family history") unrelated to the fumarate hydratase gene, so its supporting evidence was checked directly rather than assumed correct by analogy with the audited genes. A random sample of 40 of FH's 297 raw postings in the 2026-05-08 Kidney Cancer corpus was read in full. 36 (90.0%) were genuine fumarate hydratase gene references, and 4 (10.0%) were false, all sharing a single pattern, "favorable histology (FH)," a Wilms tumor clinical-staging abbreviation rather than "family history" as initially anticipated. This pattern matched 31 of FH's 305 raw postings in the 2026-09-27 corpus. Redacting them directly from the title and abstract text, not only from the retrieval index, and re-running the full retrieval and scoring engine reduced FH's supporting-paper count from 90 to 88 and left Kidney Cancer's top-10 gene list, including FH's own rank-5 position, unchanged (`scripts/analysis/vocab_collision_pipeline.py`, `scripts/analysis/vocab_corrected_rerun.py`, `vocab_corrected_rescoring.py`). IGH (Myeloma) is a recurrent translocation partner locus in myeloma (e.g. t(4;14), t(11;14)), rather than a protein-coding driver identified through recurrent somatic point mutations (Kuehl and Bergsagel, 2005). MGMT (Brain Cancer) and PARP1 (Prostate Cancer) are biomarker- or epigenetically-driven rather than mutation-driven. MGMT promoter methylation predicts temozolomide response (Hegi et al., 2005), and PARP1 is a synthetic-lethality drug target discussed alongside BRCA mutation status (Bryant et al., 2005; Farmer et al., 2005, full per-gene classification in Supplementary Table S4). These results do **not** show that IntOGen should have detected all 40 pairings. They show that IntOGen's statistical driver-detection framework is a narrower criterion, restricted to somatic point-mutation signal, than the existence of mutation-focused literature for a gene-cancer pairing, related but not equivalent evidence, with its own disclosed scope limits (Methods 2.3).

## Supplementary Results S3. Sensitivity to `driver mutation`

Replacing the primary query with `driver mutation driver mutations` produced mean top-10 Jaccard similarity of 0.848 to the primary rankings (range 0.667–1.0). IntOGen any-cancer precision changed from 93.7% to 95.8% (+2.1 percentage points, 95% cancer-bootstrap CI 0.5 to 4.2, paired t-test p=0.042, Wilcoxon p=0.046, four more top-10 slots of 190), and exact cancer-type driver concordance changed from 78.9% to 80.5% (+1.6 percentage points, CI -0.5 to 3.7, p=0.19, Wilcoxon p=0.18), a difference not distinguishable from no difference (`scripts/analysis/driver_mutation_sensitivity.py`). The rankings for this query were computed with the exact background count for the query `driver mutation driver mutations` and without the MET, REST and FH redaction. This supports robustness to wording within the mutation-focused domain only.

# Supplementary Results S4. Detailed comparator statistics

## CancerMine

- Corrected Precision@10 (IntOGen). 87.9% (167/190)
- Unique-PMID deduplicated alternative Precision@10 (post-correction). 87.4% (166/190)
- OncoDigger-CancerMine difference. 5.8 pp
- 95% bootstrap CI. 0.5–11.6 pp
- Paired t-test. p=0.053
- Wilcoxon signed-rank. p=0.12
- OncoDigger higher / tie / lower. 10 / 7 / 2 cancers
- CancerMine exact cancer-type driver concordance (IntOGen). 60.0% (114/190)
- Exact-concordance difference. 18.9 pp
- 95% bootstrap CI. 13.2–23.7 pp
- Paired t-test. $p=1.9\times10^{-6}$
- Wilcoxon. $p=2.1\times10^{-4}$
- Exact-concordance higher / tie / lower. 18 / 0 / 1 cancers
- Retention among CancerMine's own IntOGen any-cancer hits. 114/167 = 68.3%

CancerMine achieved 10/10 confirmed IntOGen drivers in Cervical Cancer, Lung Cancer, Lymphoma and Thyroid Cancer. Its lowest cancer score was 60% in Esophageal Cancer.

Before correction, the CancerMine benchmark was 84.2% against IntOGen (the value from the original unaudited run, which the repository scripts do not regenerate, because only the corrected mapping and the PTGDR exclusion of `scripts/analysis/cancermine_benchmark.py` are kept). Most of the change to 87.9% resulted from removing PTGDR after the confirmed 1,329-sentence normalization artifact (Results 5.4), and three disease-term assignments were also corrected. A cross-resource crosswalk audit then found Brain Cancer's disease-term mapping was separately missing ependymoma and medulloblastoma terms that CancerMine's own data already carried. Extending the mapping to include them brought two additional genes (GLI1, SHH) into that cancer's top 10 and moved its score from 90% to 80%. CancerMine had 10/10 confirmed IntOGen drivers in 4 cancers, compared with 11 for OncoDigger.

## DISEASES 2.0

- Precision@10 (IntOGen). 60.0% (114/190)
- OncoDigger higher in 17 cancers, tied in 2
- Paired t-test $p=1.2\times10^{-6}$
- Wilcoxon $p=2.8\times10^{-4}$
- 95% bootstrap CI. 25.3–42.6 pp
- Mean pairwise top-10 Jaccard. 0.160

Illustrative Melanoma genes not listed as IntOGen drivers were CTLA4, PDCD1, CD8A, CD4, IFNG, PMEL and IL2, while the confirmed IntOGen drivers were BRAF, CD274 and GNA11. Illustrative Myeloma genes not listed as IntOGen drivers were CD38, SDC1, CD19, IL6 and CRBN, with MYOM2 and LOC102723407 treated separately as an unresolved pattern. PDCD1/PD-1 is an established immunotherapy target (Topalian et al., 2012) and SDC1/CD138 an established plasma-cell marker (O'Connell et al., 2004), illustrating why a gene outside the reference catalogue does not mean biologically irrelevant.

## PubTator 3.0

- Precision@10 across all 19 cancers (subtype-aggregated, IntOGen). 67.9% (129/190)
- OncoDigger higher in 17, tied in 2
- Paired t-test $p=2.0\times10^{-7}$
- Wilcoxon $p=2.5\times10^{-4}$
- 95% bootstrap CI. 19.5–31.6 pp
- Mean pairwise top-10 Jaccard. 0.184 across all 19 cancers

Thyroid Cancer examples not listed as IntOGen drivers included TG, SLC5A5 and LGALS3 (TSHR is itself a confirmed IntOGen driver in Thyroid Cancer). Lymphoma example KRT20 was supported by papers genuinely concerning lymphoma. Subtype-entity aggregation's own disclosed limitation (Results 5.4) is that TP53 occupied a top-10 slot in 18 of 19 cancers and AKT1 in 17 of 19, since summed relation counts with no background normalisation reward genes broadly co-mentioned across many disease terms, not only cancer-type-specific ones.

Under the earlier, since-corrected single-parent-MeSH-ID mapping, the ambiguous Liver Cancer bucket (`Liver Neoplasms`, D008113) was dominated by SOD1 (397 supporting PMIDs), TARDBP (250), C9ORF72 (230) and FUS (120), an unrelated ALS/neurodegeneration pattern also present in VAPB, TBK1, UBQLN2 and NEFL. Eight top-gene evidence sets were checked at the title/PMID level and confirmed unrelated to liver cancer. This did not recur once Liver Cancer was queried via its own specific subtype entities (Hepatocellular Carcinoma, Cholangiocarcinoma) instead of the single ambiguous parent code.

## ChatGPT practical baseline

- Gene-level Precision@10 (IntOGen). 78.9%
- OncoDigger higher / tie / lower. 12 / 6 / 1 cancers
- Paired t-test p=0.00071
- Wilcoxon p=0.0059
- 95% bootstrap CI. 7.4–21.6 pp
- Exact cancer-type driver concordance (IntOGen). 76.3% (145/190)
- Difference from OncoDigger. +2.6 pp (OncoDigger marginally higher, not significant)
- 95% bootstrap CI. –5.3 to 11.6 pp
- Paired t-test p=0.56
- Wilcoxon p=0.63
- Exact-concordance higher / tie / lower for OncoDigger. 9 / 2 / 8 cancers

- Retention among each system's own IntOGen any-cancer hits. OncoDigger 84.3%, CancerMine 68.3%, ChatGPT 96.7%

TTN appeared in 15/19 ChatGPT lists, MUC16 in 12/19, and LRP1B, RYR2, SYNE1 and FAT4 in 5–6 each. Leukemia and Lymphoma were the two 10/10 ChatGPT cancer lists.

## Supplementary Results S5. Ranking-depth detail

Cumulative mean IntOGen-driver Precision@k is as follows.

- k=5. 97.9% (95% CI 94.7%–100%)
- k=10. 93.7% (95% CI 88.9%–97.4%)
- k=15. 90.2%
- k=20. 88.7%
- k=30. 84.9%
- k=40. 81.4%
- k=50. 77.6%

Non-overlapping rank intervals are as follows.

- ranks 1–10. 178/190 = 93.7%
- ranks 11–20. 159/190 = 83.7%
- ranks 21–30. 147/190 = 77.4%
- ranks 31–50. 253/380 = 66.6%

Rank-membership correlations ranged from -0.046 to -0.497 across cancers. Bladder, Cervical, Endometrial, Esophageal, Leukemia, Liver, Lymphoma and Melanoma remained 100% IntOGen-driver precision through ranks 11–20. Ovarian Cancer and Prostate Cancer fell to 50% in that interval.

These are descriptive rank intervals, not claims of biological novelty or discovery timing.

## Supplementary Results S6. IntOGen driver recall

Precision (Results 5.3) asks whether OncoDigger's top-10 genes are IntOGen drivers specifically confirmed for that cancer. The reverse question, recall, asks how much of IntOGen's complete cancer-type-resolved driver catalogue OncoDigger's ranking actually contains. This uses the same top-50 rankings as the main analysis (`data/canonical/precision_at_k_detail.csv`) and the full IntOGen compendium, mapped to the 19 study cancers using the same conservative mapping as the primary analysis (Methods 2.3, Supplementary Methods S5).

Across the 19 cancers, IntOGen's conservative mapping yields 1,740 gene-cancer driver pairs. Pooled recall increased from 8.6% at rank 10 (150/1,740) to 14.6% at rank 20 (254/1,740) and 26.3% at rank 50 (458/1,740). The unweighted per-cancer macro-average follows a similar pattern, 9.8% at rank 10, 16.0% at rank 20, 27.5% at rank 50 (median 8.5%/15.2%/26.8%). These absolute recall values are modest mainly because the denominator spans the full top-50 depth across every cancer, not because OncoDigger's ranking recovers few genes in absolute terms.

Recall is far from uniform across cancers (Repository Table RT11, `data/supplementary_tables/RT11_IntOGen_recall_by_cancer.csv`). Cervical Cancer has the *highest* Recall@50 of any cancer (39.6%), followed by Myeloma (35.8%) and Leukemia (32.8%), and Lung Cancer has the lowest (17.6%).

Exact matches are nonetheless strongly concentrated toward the top of the ranking rather than spread evenly. Of the 190 available positions, 78.9% are exact matches at ranks 1–10, 54.7% at ranks 11–20, and 35.8% at ranks 21–50. Match density therefore declines with rank depth, consistent with a prioritization ranking rather than a flat or random one.

The complete list of IntOGen driver genes not recovered at any rank up to 50, per cancer, is available in `data/derived/intogen_rescoring/intogen_recall_missed_genes.csv`.

## Supplementary Results S7. Genuinely unfiltered vocabulary sensitivity check on the primary pipeline

An earlier version of this check used a "raw, uncleaned vocabulary" baseline that, on closer inspection (Results 5.8, Methods 3.4), still carried the vocabulary-auditing module's own common-word filter and was therefore not genuinely unfiltered. It is replaced here with the actual unfiltered comparison, built from the same vocabulary-auditing module's raw lookup table with every filtering rule removed (69,228 single-token aliases and 11,150 phrase anchors, versus the much smaller filtered vocabulary this paper otherwise uses), run through the identical BM25 retrieval, matching, and scoring pipeline for all 19 cancers (`scripts/analysis/true_raw_vs_cleaned_all19.py`, `data/derived/intogen_rescoring/true_raw_vs_cleaned_all19.csv`).

Pooled Precision@10 against IntOGen fell from 93.7% (178/190) with this paper's vocabulary QC applied to 36.3% (69/190) with the genuinely unfiltered vocabulary, a 57.4-percentage-point drop, present in all 19 cancers individually (Supplementary Table S6). The drop is driven overwhelmingly by a small number of specific alias collisions, not a general property of filtering intensity. Four genes (BRIP1, CD44, SPI1, WWOX) occupy a top-10 slot in every one of the 19 cancers under the unfiltered vocabulary, and a fifth (WAS) in 18 of 19, traced to 10 ambiguous single-token aliases, ordinary English function words and two biomedical abbreviations ("of," "in," "for," "was," "all," "not," "as," "an," "cml," "phl"). Removing only these 10 entries, scored with the real relevance formula rather than the flat mention-count ranking used in the main factorial, recovers 82.6% of the full any-cancer gap and 75.0% of the exact-concordance gap between true-raw (36.3%/24.2%) and cleaned (93.7%/78.9%) vocabulary (83.7%/65.3% with only this minimal correction, BM25 pool, all 19 cancers, Results 5.8).

**Formula-decomposition detail (Results 5.8).** The formula's exact-concordance advantage over flat counting is +2.6 pp under cleaned vocabulary (93.7%/78.9% vs. 92.1%/76.3%) and +18.9 pp under true-raw vocabulary (36.3%/24.2% vs. 14.7%/5.3%). Decomposing the formula's two remaining multiplicative terms under true-raw vocabulary (rank-weighted mention position and enrichment, supporting-paper-count term held fixed), rank-weighting alone reached 14.2%/4.7% (vs. flat counting's 14.7%/5.3%, essentially no improvement), while enrichment alone reached 90.5%/73.2%. Whether the combined formula's underperformance relative to enrichment alone is caused by rank-weighting's positional component specifically, or by the additional frequency weighting it implicitly reintroduces, was not separately isolated and is left as disclosed future work.

**Supplementary Table S6. Genuinely unfiltered vocabulary sensitivity against IntOGen, primary pipeline, Precision@10 by cancer.**

| Cancer | Cleaned (this paper's pipeline) | True-raw (unfiltered) |
|---|---|---|
| Breast Cancer | 100% | 50% |
| Lung Cancer | 100% | 30% |
| Melanoma | 100% | 30% |
| Bladder Cancer | 100% | 40% |
| Pancreatic Cancer | 100% | 40% |
| Stomach Cancer | 100% | 30% |
| Esophageal Cancer | 100% | 20% |
| Cervical Cancer | 100% | 40% |
| Thyroid Cancer | 100% | 50% |
| Leukemia | 100% | 40% |
| Lymphoma | 100% | 30% |
| Prostate Cancer | 90% | 40% |
| Kidney Cancer | 90% | 20% |
| Liver Cancer | 90% | 40% |
| Ovarian Cancer | 90% | 40% |
| Brain Cancer | 90% | 30% |
| Myeloma | 90% | 30% |
| Colorectal Cancer | 70% | 50% |
| Endometrial Cancer | 70% | 40% |

## Supplementary Results S8. Parameter-tuning sensitivity check

Git history for the scoring parameters (BM25 k1 and b, title/MeSH match multipliers, review down-weight, review multiplier, analysis-pool size) in the author's private live-application repository shows all were set once in a single commit (`eb4da47`, 2026-05-04) and never subsequently modified. This is evidence against iterative tuning against validation performance. To test this more directly, the primary "mutation, mutations" query was re-run for all 19 cancers at a range of alternate values for each parameter, varying one parameter at a time while holding all others at their reported value, and pooled Precision@10 against IntOGen's any-cancer driver set was rescored at each setting (`scripts/analysis/intogen_parameter_sensitivity.py`, `data/derived/intogen_rescoring/intogen_parameter_sensitivity.csv`).

Across all 24 unique configurations tested (23 parameter perturbations plus the shared baseline, which is repeated within each parameter block in the table below, giving 29 displayed rows, k1 = 1.0-2.0, b = 0.5-1.0, title-match multiplier = 1.0-1.5, MeSH-match multiplier = 1.0-1.35, review multiplier = 0.5-1.0, analysis-pool size = 500-2,000 papers), pooled Precision@10 ranged only from 93.2% (177/190) to 94.7% (180/190), a spread of 1.6 percentage points around the reported baseline of 93.7% (178/190), with no configuration producing a sharp peak or a value outside this narrow band (Repository Table RT3). The result showed little sensitivity to the parameter variations examined. A stable local sweep does not exclude every form of tuning.

# Supplementary Results S9. IntOGen's own confidence signals for never-surfaced pairs

Results 5.5 shows that most IntOGen driver pairs never entered OncoDigger's ranking at any evaluated depth. To check whether this reflects a uniform sample of IntOGen's driver catalogue or a concentration in its weaker-evidence calls, three signals already present in IntOGen's own compendium were compared between the pairs OncoDigger's ranking surfaced and the pairs it never surfaced. None of these three signals is used anywhere else in this paper's scoring. They are how many of IntOGen's seven driver-detection methods (dNdScv, OncodriveFML, OncodriveCLUSTL, HotMAPS, CBaSE, MutPanning and smRegions) agreed on that gene's significance in that cohort, the combined q-value, and the percentage of the cohort carrying a mutation in that gene (`scripts/analysis/intogen_confidence_breakdown.py`).

Never-surfaced pairs are concentrated in IntOGen's single-method, lower-confidence tier. 76.6% rest on agreement from only one of the seven methods, versus 40.6% of surfaced pairs. Median q-value was $8.2\times10^{-4}$ for never-surfaced pairs versus $6.7\times10^{-9}$ for surfaced pairs, roughly five orders of magnitude weaker, and median percentage of the cohort mutated was 3.3% versus 6.5%. All three differences are significant in exploratory pair-level Mann-Whitney U tests ($p<10^{-35}$ for each, Supplementary Figure S2, Supplementary Table S7), which treat gene-cancer pairs rather than the 19 cancer types as the observation unit and so do not account for within-cancer clustering, unlike this paper's primary cancer-bootstrap analyses (Methods 2.2). This does not establish that any individual never-surfaced gene is a false positive in IntOGen's own catalogue, since IntOGen's own methods already apply multiple-testing correction within each call. It does show that the recall gap is not spread evenly across IntOGen's driver catalogue. It concentrates in the calls IntOGen itself supports most thinly, which is also where large-gene background-mutation-rate artefacts of the kind already discussed for TTN and MUC16 (Results 5.6, Lawrence et al., 2013, and Gorlov et al., 2018) are most likely to occur.

**Supplementary Table S7. IntOGen confidence signals, surfaced vs never-surfaced driver pairs (Results 5.5, Supplementary Results S9). Mann-Whitney p-values are exploratory pair-level comparisons without adjustment for dependence within cancers.**

| Metric | Surfaced (n=762) | Never surfaced (n=978) | Mann-Whitney p |
|---|---|---|---|
| Median number of IntOGen methods agreeing | 2 | 1 | $2.2\times10^{-70}$ |
| Median q-value | $6.7\times10^{-9}$ | $8.2\times10^{-4}$ | $1.9\times10^{-72}$ |
| Median % of cohort mutated | 6.5% | 3.3% | $2.1\times10^{-36}$ |
| Share backed by only 1 method | 40.6% | 76.6% | — |

# Supplementary Results S10. Full-corpus presence of never-surfaced pairs

Results 5.5 defines "never surfaced" as not entering OncoDigger's ranking within the standard BM25-selected analysis pool (up to 1,000 papers per cancer) used throughout this paper (Methods 2.5). This is narrower than the full corpus. Each cancer's full collection ranges from 22,079 to 250,061 abstracts (Supplementary Table S1), so the analysis pool this check uses covers well under 5% of most corpora. To separate genuine literature absence from an artefact of this analysis window, each never-surfaced pair's raw mention count was queried directly against the full corpus index, bypassing BM25 retrieval and entity-table extraction entirely (`scripts/analysis/intogen_full_corpus_presence_check.py`).

Only 188 pairs (10.8% of all 1,740) have zero lexicon-detected mentions anywhere in the full corpus, meaning this paper's gene vocabulary found no occurrence of that gene's recognized symbols or aliases in any abstract of that cancer's corpus. This is not the same as confirmed biological or literature absence. Detection depends on vocabulary completeness and specificity, and this paper's own vocabulary audit independently confirms both false negatives (H3-3A, H3C2 and H4C9 missing from the vocabulary entirely under their current HGNC symbols, below) and false positives (WAS and MAX's raw full-corpus counts dominated by common-word collisions, below) in this same detection step. These 188 pairs are more precisely described as having no lexicon-detected mentions, not as confirmed absent from the literature. The remaining 790 (45.4% of all 1,740) have at least one lexicon-detected mention somewhere in the corpus but did not reach the analysis pool (789 in the magnitude analysis of Supplementary Figure S3B, which excludes the WAS collision described below). The same vocabulary caveat applies in the other direction here, since some of these 790 detected mentions may themselves be false-positive alias collisions rather than genuine gene mentions. This presence is typically sparse, with a median of 5 mentions across the whole corpus, so most of these pairs would likely remain unranked even under a substantially larger analysis pool. The practical conclusion is not that recall would approach 100% with more compute, but that the raw "never surfaced" figure should not be read as "no literature support." A smaller group has more substantial full-corpus presence (49 pairs with ≥50 mentions, 20 with ≥100), listed in `data/derived/intogen_rescoring/intogen_full_corpus_presence_check.csv`.

Two false-alias collisions were found by this check itself. WAS (Brain Cancer) returned 69,892 mentions, because "was" is a common English auxiliary verb, not a genuine mention of the WAS (Wiskott-Aldrich syndrome) gene. MAX (Brain Cancer) returned 439 mentions, confirmed by direct inspection of a sample of the underlying titles (Supplementary Results S12) to be dominated by "maximum [tolerated dose/uptake/...]" and the unrelated soybean species *Glycine max*, not the MAX proto-oncogene. Both are the same class of vocabulary bug already documented for MB, ELN and seven other symbols (Methods 3.3), not previously caught because this specific check had not been run before. Both are excluded from the magnitude statistics above (but retained in the pair counts, since they do have a nonzero raw count).

The full-corpus presence breakdown (762 surfaced within the standard analysis window, 790 never surfaced but with lexicon-detected mentions elsewhere in the full corpus, 188 with none; of 1,740 pairs) is Repository Table RT12 (`data/supplementary_tables/RT12_full_corpus_presence_breakdown.csv`).

# Supplementary Results S11. Recall against IntOGen's higher-confidence driver calls

Supplementary Results S9 and S10 establish two independent reasons a pair may never surface without reflecting a ranking failure. It may be backed by only one of IntOGen's seven methods (overwhelmingly a CGC-rescue case, Supplementary Table S7) rather than agreement from at least two methods, or it may have little to no presence in the analysed corpus at all (Repository Table RT12). Combining both as explicit filters and recomputing recall shows how much of the raw recall figures reported in Results 5.5 and Supplementary Results S6 is attributable to these two factors (`scripts/analysis/intogen_cleaned_recall.py`).

**Supplementary Table S8. Recall under progressively stricter corpus-detectability and IntOGen-support criteria.**

| Filter | Pairs | Any rank | Rank 10 | Rank 20 | Rank 50 |
|---|---|---|---|---|---|
| None (raw, as reported in Results 5.5) | 1,740 | 43.8% | 8.6% | 14.6% | 26.3% |
| Remove pairs with zero full-corpus mentions | 1,552 | 49.1% | 9.7% | 16.4% | 29.5% |
| Remove zero-mention pairs **and** require ≥2 of IntOGen's seven methods | 644 | 70.3% | 21.1% | 31.4% | 50.9% |

Restricting to the 644 pairs that are both detectable in the corpus and supported by at least two of IntOGen's own driver-detection methods, a higher-evidence subset defined for this sensitivity analysis and not IntOGen's own formal tier classification (most pairs this excludes enter IntOGen's compendium through its single-method CGC-rescue route, Supplementary Results S9), raises recall from 26.3% to 50.9% at rank 50, nearly twice the original value, and from 43.8% to 70.3% at any evaluated rank, approximately 60% higher. This does not mean the raw recall figures are incorrect. It means most of the apparent gap concentrates in pairs that are either undetectable in this analysis pool or only weakly, CGC-dependently supported by IntOGen itself, rather than in pairs representing a genuine failure to rank independently well-evidenced drivers. The remaining gap under the strictest filter, 29.7% of these 644 pairs (191) still never surfaced, is addressed directly in Supplementary Results S12.

# Supplementary Results S12. Root-cause classification of the remaining recall gap

The 191 pairs that are both detectable in the full corpus and backed by ≥2 of IntOGen's seven methods, yet still never entered OncoDigger's ranking at any evaluated depth (Supplementary Results S11), were classified, rather than left as an unexplained residual, by two different procedures (`scripts/analysis/intogen_miss_root_cause.py`): 157 by full-corpus mention count alone, and 34 by a sample of paper titles drawn directly from the corpus text (bypassing BM25 retrieval and ranking).

For the 34 of these 191 pairs with substantial full-corpus presence (≥20 mentions), up to 20 paper titles mentioning the gene in that cancer's corpus were sampled directly and checked for mutation-focused framing (a

title containing "mutat*"). The remaining 157 pairs have fewer than 20 full-corpus mentions and were classified as low-volume without a title sample, consistent with the sparse-presence pattern already established for the broader never-surfaced set (Supplementary Results S10).

The 191 pairs were classified as low corpus volume (157, 82.2%), low `mutat*` occurrence in sampled titles (30, 15.7%), a vocabulary artefact (MAX in Brain Cancer, 1, 0.5%) and unresolved (3, 1.6%: SF3B1 in Breast Cancer, ARID2 in Lung Cancer and CREBBP in Lung Cancer, each with at least 20 full-corpus mentions and at least 20% of sampled titles containing `mutat*`, which still did not rank); the full table is Repository Table RT4. 188 of the 191 pairs remaining under the strictest filter (98.4%) could be assigned to an identifiable diagnostic category associated with limited recoverability. Most have literature too sparse to plausibly rank, some have substantial literature whose sampled titles rarely contain `mutat*`, and one case reflects a newly-identified vocabulary collision rather than genuine gene literature. These classifications identify plausible limitations on recoverability, but they do not experimentally prove that each factor caused the corresponding ranking failure. Only three pairs (1.6%) remain unclassified after this audit.

## Supplementary Results S13. Sensitivity to IntOGen's literature-informed admission rules

IntOGen is computed from cohort mutation data, but its catalogue admission uses the Cancer Gene Census (CGC), which is literature-curated, and a CancerMine snapshot (Methods 2.6). An audit of the intogen-plus code (tag `v2024`; intogen-plus, 2024) against the released 2024.09.20 tables found that 2,969 of the 4,478 catalogue rows (66.3%) could only have been admitted for a CGC gene, through at least one of three routes that overlap (2,695 rows with fewer than two individually significant methods, 1,558 Tier 2 rescues and 431 CGC filter exemptions), and that at least 473 of the 633 genes are certainly CGC genes (a lower bound, since CGC membership is not released). The CancerMine gate removes 142 rows, all with zero CancerMine papers and all passing on statistics alone, so it only removes candidates and never admits a gene or alters a q-value. None of the 142 corresponds to a top-10 pair. The audit reproduced the published 178/190 and 150/190 exactly, and was performed read-only, with no IntOGen re-run.

**Admission rules and the statistical-rule reference.** In the intogen-plus code (tag `v2024`; intogen-plus, 2024), the CGC sets the weights of the seven driver-detection methods, waives the requirement for any individually significant method for CGC genes, admits CGC-only genes at a relaxed FDR (Tier 2, q<0.25) and exempts CGC genes from two filters, while a CancerMine snapshot (2021-12-07) removes candidates that are neither CGC genes nor on an 80-gene whitelist and have no CancerMine sentence. CGC membership is licence-restricted and not released, so it cannot be applied directly. The statistical-rule reference therefore keeps the catalogue rows that meet the admission rule IntOGen applies to a non-CGC gene, combined q<=0.05 and >=2 individually significant methods (the `combination` entry is the weighted p-value combination, not an individual method). This removes 2,698 of the 4,478 catalogue rows (60.2%) and leaves 1,780 rows, 385 genes and 682 gene-cancer pairs. A row-level audit of the 190 OncoDigger pairs additionally classed each matched pair as A (q<=0.05, >=2 methods, no CGC-only filter exemption), B (q<=0.05 but admitted through the CGC waiver or a CGC-only filter exemption) or C (q>0.05, present only through Tier 2 rescue), and tested the CancerMine gate by counterfactual, adding back the 142 rows it removes (`scripts/analysis/independence_audit/`).

**Supplementary Table S9. Reference definitions.**

| Reference | Catalogue rows | Genes | Gene-cancer pairs | OncoDigger any-cancer | OncoDigger exact |
|---|---|---|---|---|---|
| Published (all catalogue rows) | 4,478 | 633 | 1,740 | 178/190 = 93.7% | 150/190 = 78.9% |
| Statistical rule (q<=0.05 and >=2 individually significant methods) | 1,780 | 385 | 682 | 162/190 = 85.3% | 136/190 = 71.6% |
| Class A (statistical rule, also no CGC-only filter exemption) | audit only | audit only | audit only | 161/190 = 84.7% | 133/190 = 70.0% |
| Published plus the 142 CancerMine-gate rows (counterfactual) | 4,620 | 764 | 1,840 | 178/190 = 93.7% | 150/190 = 78.9% |

**Supplementary Table S10. Evidence class of OncoDigger's 190 top-10 pairs.**

| Level | Hits | A | B | C | Not a hit at this level |
|---|---|---|---|---|---|
| Exact cancer-matched | 150 | 133 | 15 | 2 | 40 (28 driver only in another cancer, 12 not in the catalogue) |
| Any cancer | 178 | 161 | 14 | 3 | 12 |

The A, B and C counts classify the hits at each level, so the class B counts differ between the two rows. Class A rows meet the non-CGC statistical rule with no CGC-only exemption, class B have q<=0.05 but depend on the CGC waiver or a CGC filter exemption, class C have q>0.05 and exist only through Tier 2 rescue. Row classes, pair classes and per-pair evidence are in `data/derived/intogen_independence_audit/paper1_190_pair_evidence.csv`.

**Reconciliation with the statistical-rule reference.** The 136 restricted-reference exact hits are the 133 class A pairs plus 3 class B pairs that are CGC-flagged for the cancer but still pass the statistical rule (Cervical PTEN and ERBB2, Melanoma KIT). Likewise the 162 restricted-reference any-cancer hits are the 161 class A pairs plus 1 class B pair (Myeloma MYC). Class A (the stricter evidence class) and the statistical-rule reference are not expected to coincide. Exact concordance fell by 3 hits in Ovarian Cancer, by 2 in Breast, Stomach and Thyroid Cancer and by 1 in Cervical, Colorectal, Kidney and Pancreatic Cancer and Leukemia, and was unchanged in ten cancers (Bladder, Brain, Endometrial, Esophageal, Liver, Lung, Lymphoma, Melanoma, Myeloma, Prostate; Repository Table RT5).

**Supplementary Table S11. Paired difference, OncoDigger minus comparator, in percentage points (95% cancer-bootstrap CI); cancers where OncoDigger is better / tied / worse.**

| Comparator | Exact, published | Exact, statistical | Any-cancer, published | Any-cancer, statistical |
|---|---|---|---|---|
| Pooled cancer-blind control | +25.8 (+20.0 to +32.1); 19/0/0 | +35.3 (+27.9 to +42.6); 18/1/0 | -6.3 (-11.1 to -2.6); 0/11/8 | -4.7 (-10.0 to -0.5); 3/8/8 |
| CancerMine | +18.9 (+13.2 to +23.7); 18/0/1 | +19.5 (+13.2 to +25.8); 16/2/1 | +5.8 (+0.5 to +11.6); 10/7/2 | +5.3 (-1.6 to +12.1); 9/6/4 |
| ChatGPT (one response) | +2.6 (-5.3 to +11.6); 9/2/8 | +3.2 (-5.8 to +13.2); 9/1/9 | +14.7 (+7.4 to +21.6); 12/6/1 | +7.9 (+1.1 to +14.2); 10/6/3 |
| DISEASES 2.0 | +35.8 (+26.3 to +44.7); 17/2/0 | +38.4 (+30.0 to +46.8); 18/1/0 | +33.7 (+25.3 to +42.6); 17/2/0 | +34.7 (+25.3 to +44.2); 16/3/0 |
| PubTator 3.0 | +35.3 (+25.3 to +43.7); 16/2/1 | +37.9 (+29.5 to +45.8); 19/0/0 | +25.8 (+19.5 to +31.6); 17/2/0 | +25.3 (+17.9 to +31.6); 17/1/1 |
| BM25-matched flat counting (clean vocabulary) | +2.6 (+0.0 to +5.3); 6/12/1 | +4.2 (+1.1 to +7.9); 7/11/1 | +1.6 (+0.0 to +3.2); 3/16/0 | +2.1 (-0.5 to +4.7); 6/11/2 |
| Frequency baseline (aliases) | +37.9 (+30.0 to +46.3); 19/0/0 | +39.5 (+31.6 to +48.4); 19/0/0 | +31.6 (+24.2 to +38.9); 19/0/0 | +25.8 (+17.4 to +34.2); 16/3/0 |
| Frequency baseline (primary symbols) | +40.5 (+33.2 to +47.9); 19/0/0 | +39.5 (+32.6 to +46.3); 19/0/0 | +30.0 (+23.2 to +36.3); 18/1/0 | +24.7 (+16.8 to +32.6); 16/2/1 |

Bootstrap intervals in this section come from `intogen_statistical_support_rescoring.py` (19 cancers resampled, 2,000 replicates, seed 42) and can differ by up to about 0.5 percentage points from the same intervals in Supplementary Results S1, which use a separate implementation.

## Supplementary Results S14. Literature-cutoff-matched comparison

OncoDigger was rerun on collections truncated at publication year 2022 and 2023 (Methods 2.7) and scored against the same IntOGen references as the full-collection analysis.

**Supplementary Table S12. Concordance with IntOGen for OncoDigger on year-truncated collections, CancerMine and the pooled control.**

| System | Any-cancer P@10, full catalogue | Any-cancer P@10, statistical rule | Exact concordance, full catalogue | Exact concordance, statistical rule |
|---|---|---|---|---|
| OncoDigger, collections to 2022 | 178/190 = 93.7% | 161/190 = 84.7% | 146/190 = 76.8% | 129/190 = 67.9% |
| OncoDigger, collections to 2023 | 178/190 = 93.7% | 162/190 = 85.3% | 147/190 = 77.4% | 130/190 = 68.4% |
| OncoDigger, full collections (this paper) | 178/190 = 93.7% | 162/190 = 85.3% | 150/190 = 78.9% | 136/190 = 71.6% |
| CancerMine (snapshot 2023-03-01) | 167/190 = 87.9% | 152/190 = 80.0% | 114/190 = 60.0% | 99/190 = 52.1% |
| Pooled cancer-blind control | 190/190 = 100.0% | 171/190 = 90.0% | 101/190 = 53.2% | 69/190 = 36.3% |

Paired difference, OncoDigger minus CancerMine, in percentage points (95% cancer-bootstrap CI): exact concordance, full catalogue +16.8 (+10.5 to +21.6) at the 2022 cutoff and +17.4 (+10.5 to +22.1) at 2023; exact

concordance, statistical rule +15.8 (+8.4 to +22.6) and +16.3 (+9.5 to +22.6); any-cancer, full catalogue +5.8 (+0.5 to +11.6) and +5.8 (+0.5 to +11.6). The truncated runs share 178/190 (2022) and 180/190 (2023) top-10 slots with the full-collection lists (`data/derived/intogen_rescoring/cutoff_matched_overlap.csv`). The reference is not truncated, so both systems are scored against a reference that contains post-cutoff evidence.

The MET, REST and FH redaction of the main analysis (Methods 3.3) was applied to the truncated Lung, Thyroid and Kidney collections as well, so publication year is the only difference from the main analysis. The false-positive records inside each cutoff were 1,646 and 1,836 of MET's 2,372, 124 and 130 of REST's 136, and 29 and 29 of FH's 31, at the 2022 and 2023 cutoffs (`scripts/analysis/cutoff_matched_vocab_correction.py`). The correction changed no top-10 membership at either cutoff. It swapped FH and PBRM1 in the 2022 Kidney Cancer ranking and MET and NRAS in the 2023 Lung Cancer ranking, and every count in Results 5.10 and Supplementary Table S12 is identical with and without it.

---

## Supplementary Results S15. Proximity-adjusted ranking (post hoc, exploratory)

This analysis was designed after the IntOGen results had been seen, compares weights against the same reference it is judged on, and is not part of the main results. It is reported as a negative result: a proximity multiplier did not improve agreement with IntOGen on the rankings computed with the exact background count. The scripts are `scripts/analysis/proximity_boost_run.py` and `scripts/analysis/proximity_summary.py`, the outputs are in `data/derived/proximity/`, and the complete tables are Repository Tables RT6 to RT9.

**Method.** For each gene the pipeline credits, a proximity score P is the mean over its supporting papers of 1/(1 + d), where d is the token distance between the closest mention of the gene (approved symbol, alias or full name in the vocabulary of Methods 3.3) and the closest query token (`mutation` or `mutations`) in the same segment. Title and abstract are separate segments, adjacent tokens give d = 1, and a paper in which no segment contains both contributes 0. The adjusted relevance is the pipeline's relevance multiplied by (1 + w x P) at full precision. Retrieval, evidence selection, the background, the vocabulary and the order of ties are unchanged, and w = 0 reproduces the main ranking exactly (0 of 190 top-10 and 0 of 950 top-50 slots differ). Four weights (0.25, 1, 3 and 10) were examined and all are reported, and none was chosen before the results were seen. The stored text has no punctuation, so sentence boundaries are not used. Distance-sensitive term-dependency features are established in document ranking (Metzler and Croft, 2005; Tao and Zhai, 2007); here the distance is between an entity and a query word. Because the weights were compared with IntOGen, this analysis is an exception to the statement in Methods 2.2 that IntOGen was never used to tune the ranking, so two selection-aware summaries are reported, leave-one-cancer-out and an out-of-bag bootstrap (5,000 resamples). The runs use the September 27 collections with the MET, REST and FH corrected text and the frozen vocabulary. The calculation is the standalone module `reference_engine_code/proximity.py` (sha256 prefix d8103d24c3cf), which the OncoDigger application also uses. As a vocabulary sensitivity, a second set of runs removes single-word aliases credited to two or more genes (unless the term is a gene's own approved symbol) and the family-name aliases PI3K, AKT and PARP.

**Results.** With w = 3, the number of top-10 genes that are IntOGen drivers in the same cancer was 149/190 (78.4%) against 150/190 (78.9%) for the unmodified ranking, a paired difference of −0.5 percentage points (95% cancer-bootstrap CI −2.1 to +1.1), with one cancer better and two worse. Drivers in any cancer were 178/190 (93.7%) for both rankings, and under the statistical-rule reference of Results 5.9 the exact count fell from 136 to

134. Of the top-10 genes 180 were retained, 10 left and 10 entered. With w = 0.25, 1 and 10 the exact matches were 150, 151 and 147, so no weight raised the unmodified ranking by more than one top-10 slot, and w = 10 lowered it by 1.6 points (CI -3.7 to +0.5; one cancer better and four worse). Against the matched frequency-counting control of Results 5.8 (145/190 exact, 175/190 any) the w = 3 ranking gained +2.1 points on exact concordance (95% cancer-bootstrap CI -0.5 to +4.7, with 6 cancers better, 11 equal and 2 worse) and +1.6 points on any-cancer concordance (CI +0.0 to +3.2). That is no more than the unmodified ranking's own difference from counting (+2.6 and +1.6 points, Results 5.8), so the multiplier adds nothing to it. Because w = 1 was the best weight examined, any gain would be optimistic. Choosing the weight on 18 cancers and applying it to the 19th gave a leave-one-cancer-out gain of +0.0 points (w = 1 was chosen in 18 folds and w = 0.25 in one), and the out-of-bag bootstrap gave −0.2 points (95% interval −2.5 to +0.0), positive in 0% of resamples, with an optimism of 0.8 points. These use the same collections and the same reference, so they are internal checks and not independent validation. Median P differed descriptively by IntOGen status (0.31 for the 150 pairs that are drivers in this cancer, 0.29 for 28 drivers in another cancer and 0.24 for 12 genes IntOGen does not list), but the distributions overlap, the smallest group is small, and P is computed from the same papers as the relevance score, so this is not an independent test. With the alias rule, 20 top-10 slots and 356 of 950 top-50 slots changed in position (only 2 of 190 top-10 genes and 19 of 950 top-50 genes were replaced; mean Jaccard similarity 0.98 and 0.96) and the exact count of the unmodified ranking rose to 151/190. With that vocabulary w = 3 gave 150/190 (-0.5 points against 151, CI -2.1 to +1.1), so the absence of a gain does not depend on the alias rule.

**Limitations.** The analysis is exploratory. It was designed after the IntOGen results were seen, compares weights against the reference it is judged on, uses the 19 collections that defined the benchmark, and was evaluated for mutation queries on genes only. Distance is measured in tokens to the two query words, so nearby mentions may belong to different statements. Averaging over supporting papers is a design choice and was not compared with other aggregation rules. Proximity understates alterations described with other words, such as amplification or fusion, and cannot distinguish a gene that is altered from one that is merely mentioned near the word. It measures agreement with IntOGen only and does not show that the abstracts behind a boosted gene contain more mutation evidence.

## Supplementary Results S16. Post hoc audit of gene recognition in the top-10 evidence

This audit was designed after the rankings were frozen and after the IntOGen comparison had been made. It used one language-model reader, was not blinded and has no independent human validation. It is reported to show where gene recognition failed in the top-10 evidence and whether the failures affect the conclusions. The vocabulary and the rankings were not changed.

**Scope and sampling.** For each of the 19 cancers, every paper in the 1,000-paper retrieval pool that credited one of the top-10 genes was extracted with the frozen pipeline (190 gene-cancer pairs; 22,071 gene-abstract credits). The reader saw the title and the passages around the matched expression (about two per abstract) and classified the identity of the match as correct, incorrect or uncertain. For Breast and Lung Cancer, the sample per gene was the top 15 papers by pool rank, 25 random papers and all alias-only papers. For the other 17 cancers it was all matched papers if 65 or fewer, otherwise 40 random papers (seed 42) plus targeted papers (alias-only matches, capped at 40, and papers in known ambiguity contexts). In all, 8,489 gene-abstract credits were read (5,954 distinct abstracts within a cancer, because one abstract can credit several top-10 genes): 7,808 from random or all-matched sampling, 476 from targeted-only sampling and 205 from the earlier Breast and Lung protocol (164 top-ranked, 41 alias-only). Percentages below use the 7,808 random or all-matched credits unless stated. The targeted-only and earlier-protocol credits are counted in the sampling above and are not included in those

percentages.

**Retrospective decision framework.** The criteria for acting on the audit were written down after all 19 cancers had been read and before any counterfactual was run. They were an overall identity accuracy of at least 95% with no major outliers, a targeted review and a counterfactual rescoring for any pair above 10% incorrect, and correction before submission of any error that changes the primary benchmark. They were not preregistered.

**Identity results.** Among the 7,808 random or all-matched credits, 89 (1.14%) were judged incorrect and 44 (0.56%) uncertain. This describes the reviewed credits and is not a verified error rate for all 22,071 credits. A lower bound for the mechanisms found does not depend on sampling: a defined alias rule, applied to every pool abstract and checked on the abstracts read, identified 213 of 22,071 credits (0.97%) as incorrect. These rule-identified credits were not each read individually. Thirty-two of the 190 pairs had at least one such credit, 21 pairs had more than 2%, 14 more than 5% and 7 more than 10%. The seven pairs above 10% are rates for the identified alias mechanism and not exhaustive estimates of every recognition error for that gene, because an unknown mechanism in unread abstracts would not be seen: PIK3CA through "PI3K" in Thyroid (15/54, 27.8%), Stomach (14/91, 15.4%), Endometrial (23/189, 12.2%), Esophageal (10/84, 11.9%), Liver (6/51, 11.8%) and Cervical Cancer (16/146, 11.0%), and SETD2 through "HIF-1" in Kidney Cancer (13/77, 16.9%). Bladder PIK3CA was exactly 10.0% (10/100). Smaller mechanisms were the syndrome name "HNPCC" credited to MLH1 or MSH2 (7.5% to 9.7% of credits in Colorectal and Endometrial Cancer), the family term "PARP" credited to PARP1 in Prostate Cancer (56 of 76 credits judged uncertain, not incorrect), "kit" as an assay term (KIT, Melanoma), B-cell receptor read as BCR (Leukemia), "tert-butyl" read as TERT (Bladder), and gene names nested inside longer gene names (BRCA2 within PALB2 and similar; six abstracts in all).

**Identity is not mutation evidence.** A credit was counted as an error only if the matched term did not refer to the credited gene. Credits that refer to the right gene without being mutation evidence (for example PD-L1 expression credited to CD274, HER2 status labels credited to ERBB2, or anti-EGFR therapy context credited to EGFR) were not counted as errors. A rule-assisted classification of evidence type was drafted, but an independent re-judgement of 72 abstracts agreed with it exactly in only 61% (44 of 72), so no evidence-type percentages are reported.

**Counterfactual rescoring (read-only).** In all, 17 gene-cancer interventions were evaluated in 12 cancers. Fifteen addressed aliases shown to credit the wrong gene (identity interventions: PIK3CA in 10 cancers, SETD2 in Kidney and MLH1 and MSH2 in Colorectal and Endometrial Cancer) and two were broader policy-sensitivity tests (PARP1 in Prostate Cancer and ERBB2 in Breast Cancer). They were carried out in 15 rescoring runs, because the HNPCC alias was removed from MLH1 and MSH2 together in each of the two cancers. Each run removed only the stated alias-to-gene mapping from the frozen v5 lexicon, kept symbols and full names, and recomputed the supporting-paper count and the exact background count B (table columns Supporting and Exact B), enrichment, relevance and ranks with the same corpora, pool, formula and other rules. Each unmodified baseline first reproduced the frozen v5 top-50 order. The policy tests are not corrections, because they also remove correct mentions (for example "PARP-1" is tokenised to "parp").

**Supplementary Table S13. Read-only gene-recognition counterfactuals for 17 gene-cancer interventions in 12 cancers.** Supporting-paper counts, exact background counts (B), relevance scores and ranks are shown before and after each change. Identity tests remove aliases shown to credit the wrong gene; policy tests remove broader terms that also include legitimate mentions.

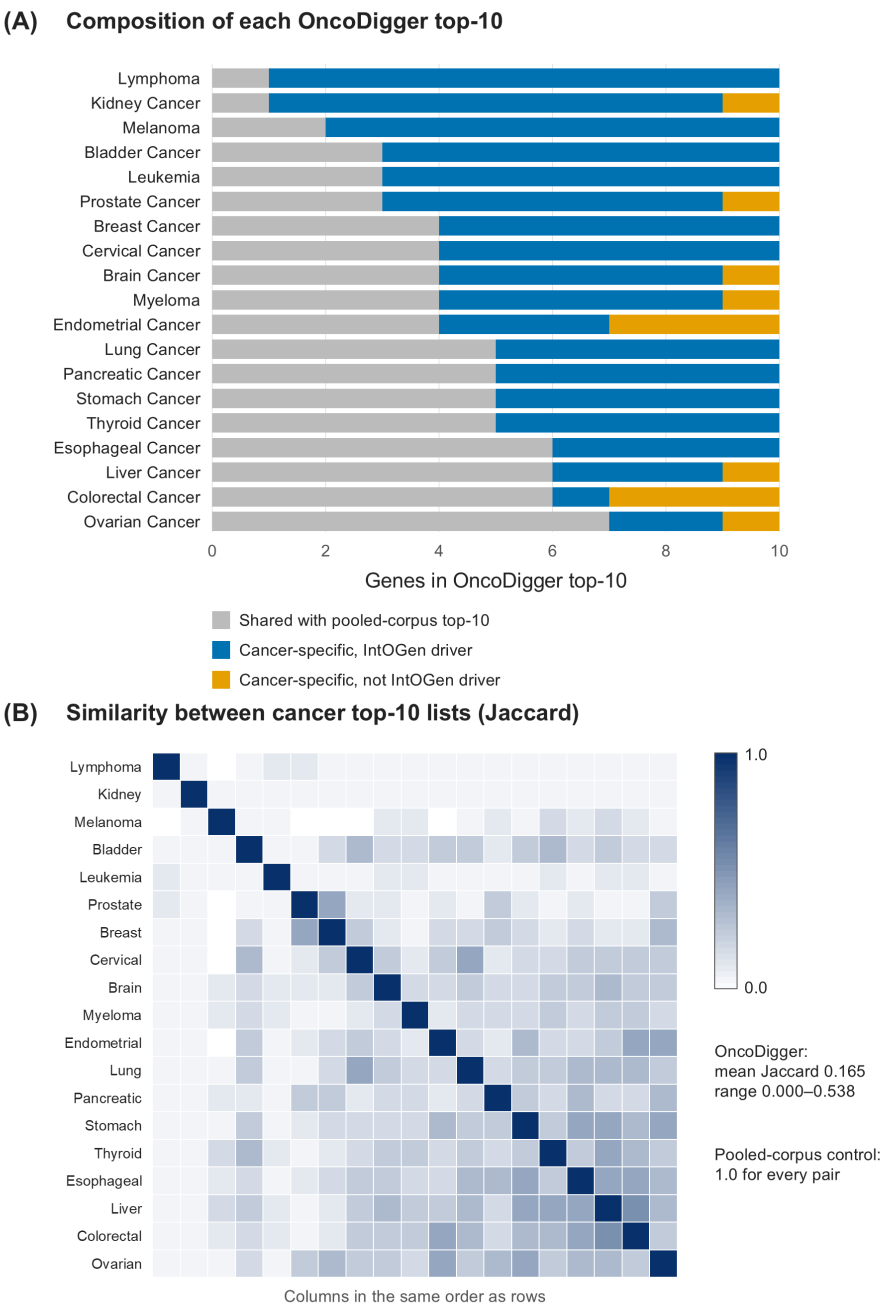
| Cancer | Gene | Alias removed | Test | Rank | Supporting | Exact B | Relevance | Top-10 change |
|---|---|---|---|---|---|---|---|---|
| Breast | PIK3CA | PI3K | Identity | 3 to 3 | 201 to 190 | 4,921 to 1,546 | 179.0 to 241.3 | none |
| Lung | PIK3CA | PI3K | Identity | 6 to 5 | 79 to 72 | 3,073 to 682 | 36.2 to 56.5 | none |
| Colorectal | PIK3CA | PI3K | Identity | 6 to 4 | 128 to 121 | 2,803 to 973 | 97.9 to 133.8 | none |
| Colorectal | MLH1 | HNPCC | Identity | 5 to 6 | 150 to 125 | 3,144 to 2,395 | 107.0 to 92.3 | none |
| Colorectal | MSH2 | HNPCC | Identity | 7 to 9 | 136 to 113 | 2,578 to 1,824 | 81.5 to 69.5 | none |
| Prostate | PARP1 | PARP | Policy | 9 to outside top 50 | 76 to - | 1,014 to - | 37.5 to - | PARP1 to RB1 |
| Bladder | PIK3CA | PI3K | Identity | 4 to 4 | 100 to 90 | 508 to 141 | 60.3 to 83.4 | none |
| Kidney | SETD2 | HIF-1 | Identity | 7 to 6 | 77 to 64 | 511 to 201 | 42.8 to 48.3 | none |
| Liver | PIK3CA | PI3K | Identity | 8 to 5 | 51 to 45 | 1,991 to 156 | 25.3 to 56.8 | none |
| Stomach | PIK3CA | PI3K | Identity | 4 to 4 | 91 to 77 | 1,298 to 201 | 49.1 to 70.8 | none |
| Esophageal | PIK3CA | PI3K | Identity | 2 to 2 | 84 to 74 | 469 to 118 | 52.8 to 73.5 | none |
| Endometrial | PIK3CA | PI3K | Identity | 2 to 2 | 189 to 166 | 599 to 281 | 140.6 to 161.6 | none |
| Endometrial | MLH1 | HNPCC | Identity | 6 to 6 | 161 to 148 | 589 to 526 | 100.0 to 92.9 | none |
| Endometrial | MSH2 | HNPCC | Identity | 8 to 9 | 124 to 111 | 423 to 360 | 72.5 to 65.6 | none |
| Cervical | PIK3CA | PI3K | Identity | 1 to 1 | 146 to 130 | 563 to 162 | 199.1 to 236.2 | none |
| Thyroid | PIK3CA | PI3K | Identity | 9 to 9 | 54 to 39 | 576 to 133 | 19.6 to 21.8 | none |
| Breast | ERBB2 | HER2 aliases | Policy | 5 to 15 | 178 to 36 | 28,885 to 2,564 | 63.6 to 11.2 | ERBB2 to ATM |

**Results.** In none of the 15 identity interventions (Supplementary Table S13) did a gene enter or leave the top 10, and IntOGen concordance was unchanged (178/190 any-cancer, 150/190 exact). Removing "PI3K" lowered PIK3CA's supporting counts (for example 54 to 39 in Thyroid Cancer) and, much more, its background counts (576 to 133), so its relevance rose in all ten cancers. Its rank was unchanged in seven cancers and higher in three (Lung 6 to 5, Colorectal 6 to 4, Liver 8 to 5). Removing "HIF-1" moved SETD2 from 7 to 6 in Kidney Cancer (FLCN 6 to 7). Removing "HNPCC" lowered Colorectal MLH1 from 5 to 6 and MSH2 from 7 to 9, and Endometrial MSH2 from 8 to 9 (Endometrial MLH1 unchanged at 6). Scores are therefore not stable to these changes even though top-10 membership is. The two policy tests did change top-10 membership. Strict ERBB2 recognition in Breast Cancer moved ERBB2 from rank 5 to 15 and brought ATM into the top 10 with no change in the benchmark. Removing "PARP" in Prostate Cancer removed PARP1 (rank 9) from the top 50 and brought RB1 into the top 10, raising concordance by one pair on each metric (179/190 any-cancer, 151/190 exact). These two tests are reported for completeness, not as corrections or as a basis for changing the vocabulary, and the vocabulary was not changed although the PARP test raised agreement with IntOGen. What the interventions

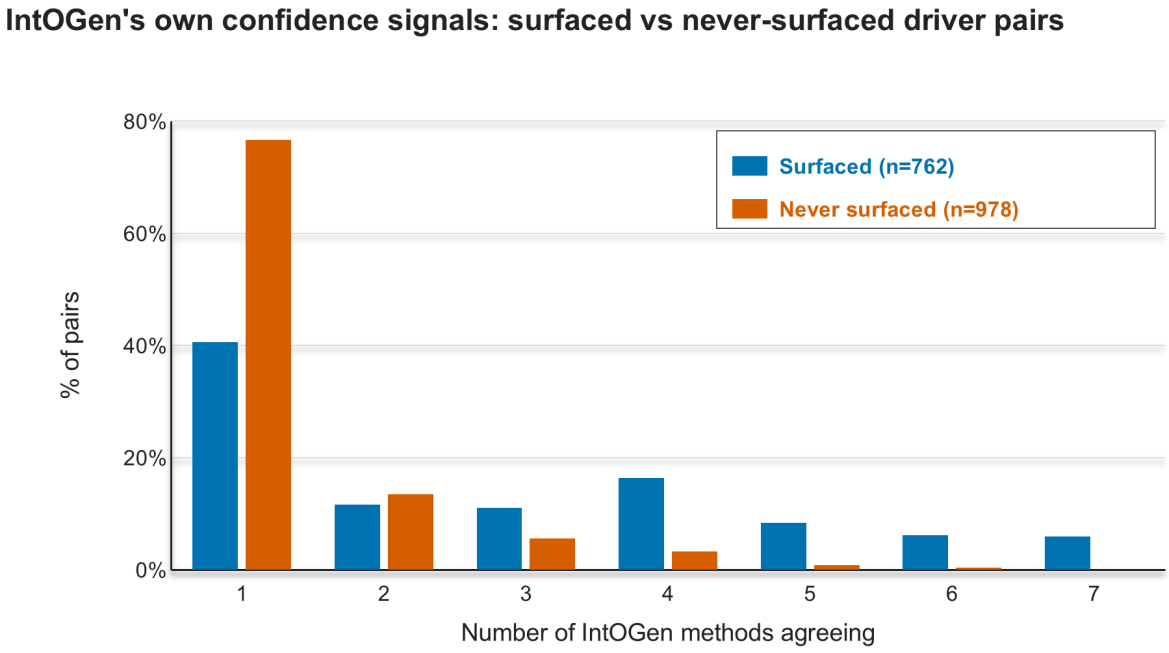
show is narrower than robustness to every recognition error: removing the identified problematic alias mappings under the tested policies did not change top-10 membership. Other undiscovered recognition errors, or a different evidence-filtering policy, could still change the rankings.

**Limitations.** One non-blinded language-model reader judged mainly from the matched passages, and there is no independent human validation. Only false positives among credits of the top-10 genes were assessed, not missed mentions or genes ranked 11 to 50. Sampling differed between Breast and Lung and the other 17 cancers, targeted strata reflect risks identified by the reader, and the pooled 1.14% describes the reviewed random or all-matched credits and is not an unbiased estimate for all 22,071 credits.

# Supplementary Figures

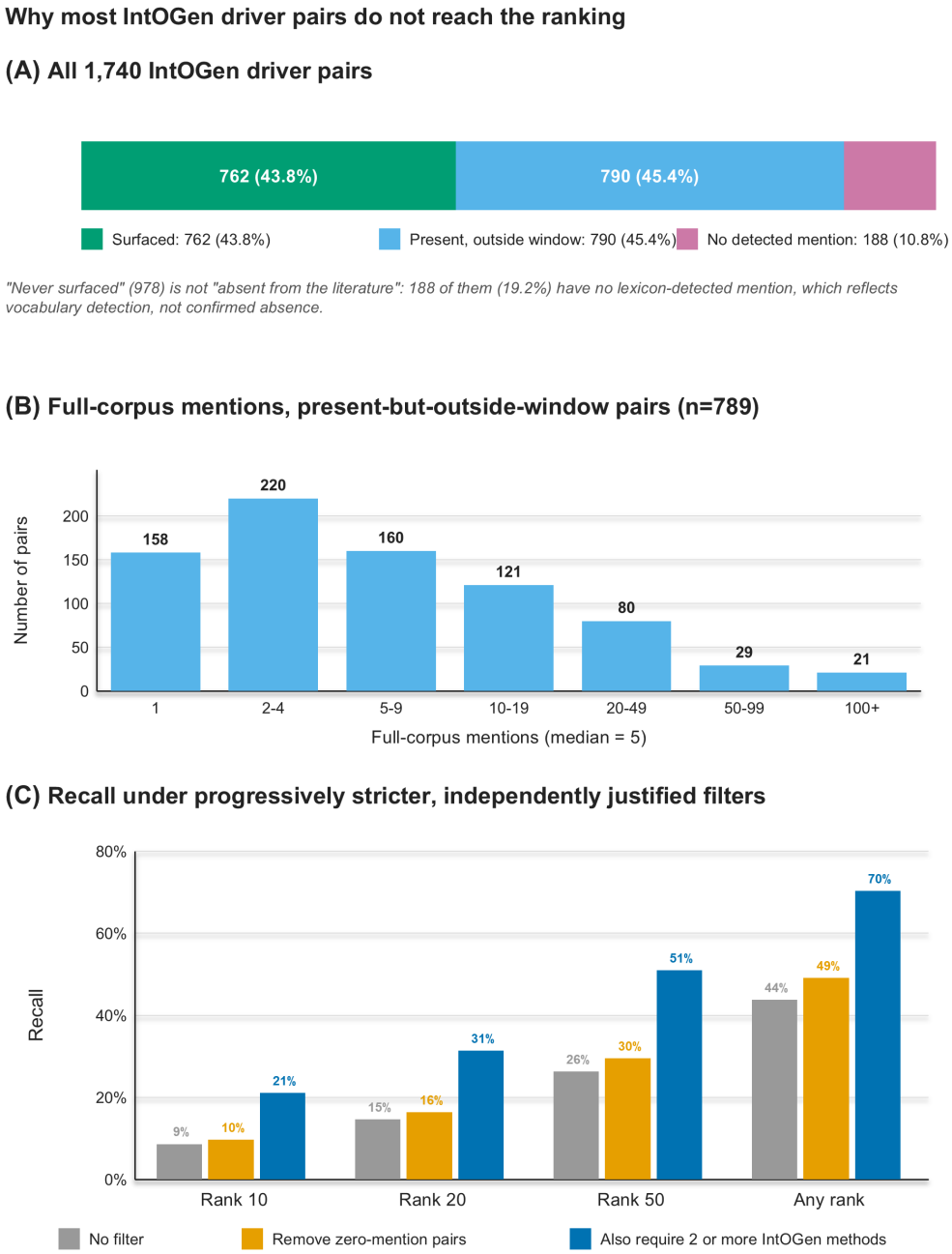



**Supplementary Figure S1. Cancer differentiation relative to the pooled control.** (A) Composition of each OncoDigger top 10 as genes shared with or absent from the pooled global top 10. (B) Pairwise Jaccard similarity among the 19 cancer top-10 lists. Mean OncoDigger Jaccard=0.165, pooled control=1.0.

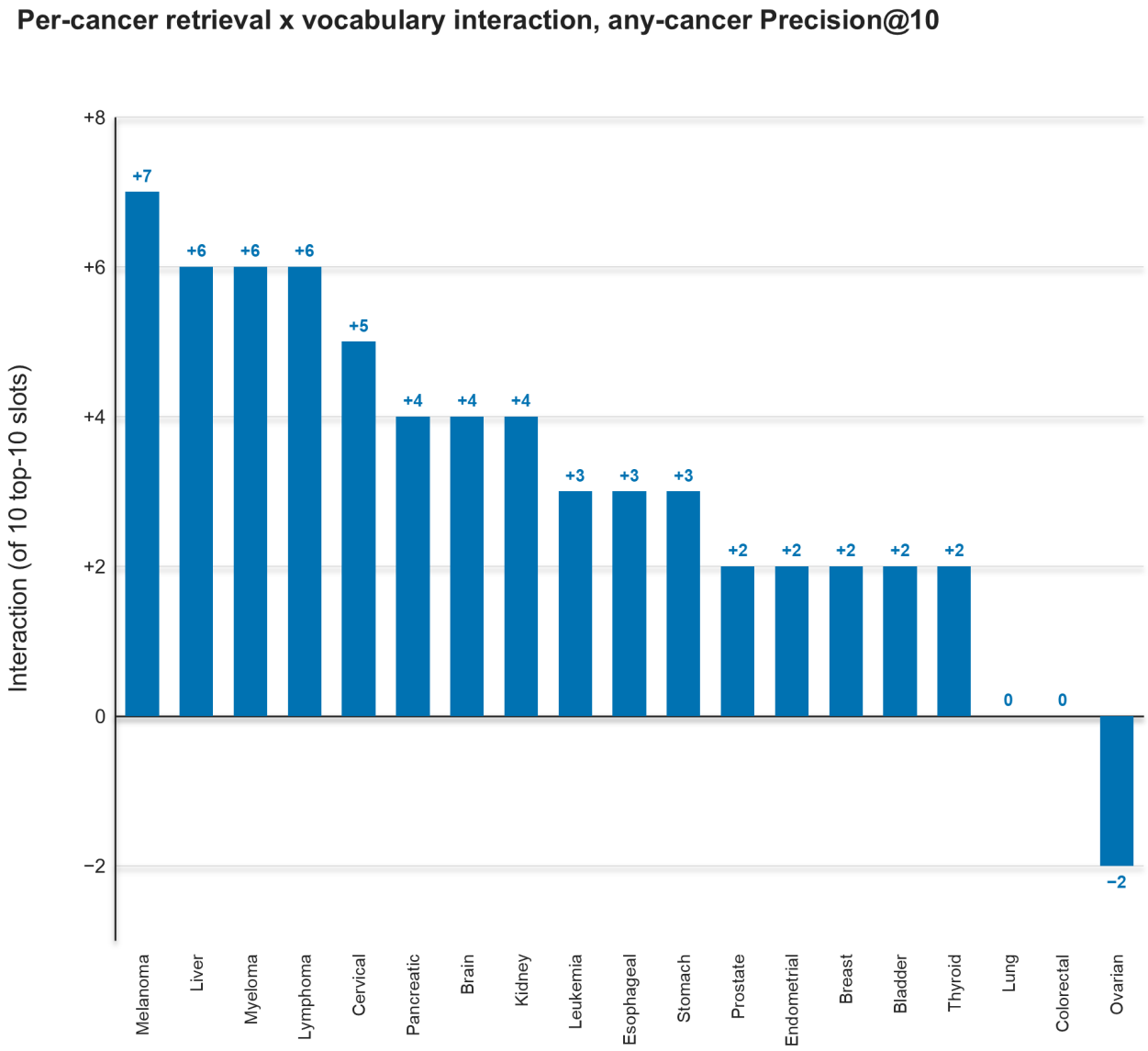



*Never-surfaced pairs skew toward IntOGen's single-method, lower-significance, lower-mutation-frequency tier (Mann-Whitney $p<10^{-35}$ for all three comparisons); 76.6% rest on only one of IntOGen's seven driver-detection methods, versus 40.6% of surfaced pairs (Supplementary Table S12).*

**Supplementary Figure S2. IntOGen's own confidence signals for never-surfaced driver pairs.** Distribution of IntOGen method-count agreement (1–7 of its 7 driver-detection methods), surfaced versus never-surfaced pairs. Never-surfaced pairs concentrate in IntOGen's single-method, lower-significance, lower-mutation-frequency tier (median q-value and median percentage of cohort mutated, Supplementary Results S9, Supplementary Table S7).




**Supplementary Figure S3. Full-corpus presence of never-surfaced driver pairs and recall under stricter filters.** (A) All 1,740 driver pairs split into surfaced within the standard window (762, 43.8%), present in the full corpus but outside the window (790, 45.4%) and with no lexicon-detected mention anywhere in the full corpus (188, 10.8%). Neither "never surfaced" nor "no lexicon-detected mention" means "absent from the literature", since detection depends on vocabulary completeness and specificity (Supplementary Results S10, Repository Table RT12). (B) Full-corpus mention count of the 789 present-but-outside-window pairs (excludes WAS in Brain Cancer, a confirmed false-alias collision), log-binned; most have only a handful (median 5), so a larger analysis pool alone would not recover them. (C) Recall at ranks 10, 20, 50 and any rank after removing pairs with zero full-corpus mentions and then pairs backed by only one of IntOGen's seven methods. Under the strictest filter recall roughly doubles at rank 50 (26.3% to 50.9%) and rises about 60% at any rank (43.8% to 70.3%), so measured recall depends on corpus detectability and on the IntOGen support required; this does not establish why individual pairs were not recovered (Supplementary Results S11, Supplementary Table S8).




**Supplementary Figure S4. Per-cancer retrieval x vocabulary interaction.** Interaction term (BM25-pool + cleaned vocabulary) minus (BM25-pool + true-raw vocabulary) minus (full-corpus + cleaned vocabulary) plus (full-corpus + true-raw vocabulary), per cancer, flat mention-count ranking against IntOGen any-cancer Precision@10. Positive in 16 of 19 cancers (zero in Lung and Colorectal Cancer, negative in Ovarian Cancer; range -2 to +7 of 10 top-10 slots), pooled interaction +31.1 percentage points (95% cancer-bootstrap CI +21.1 to +41.6 pp, Results 5.8, Table 2).