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.

Prioritizing putative cancer drivers with expression-associated functional impact

Hanmin Guo

hmguo@bicmr.pku.edu.cn

Peking University

Alexander Eckehart Urban

Stanford University

Wing Hung Wong

Stanford University

Method Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11697846/v1

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

Additional Declarations: No competing interests reported.

Prioritizing putative cancer drivers with expression-associated functional impact

Hanmin Guo[1,2,#], Alexander Eckehart Urban[3,4,5], Wing Hung Wong[6,7,#]

[1]Beijing International Center for Mathematical Research, Peking University, Beijing 100871, China

[2]Department of Biostatistics, Peking University, Beijing 100871, China

[3]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine, Stanford, California 94305, USA

[4]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

[5]Maternal and Child Health Research Institute, Stanford University School of Medicine, Stanford, California 94305, USA

[6]Department of Statistics, Stanford University, Stanford, California 94305, USA

[7]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, California 94305, USA

[#] To whom correspondence should be addressed:

Hanmin Guo (hmguo@bicmr.pku.edu.cn)

Wing Hung Wong (whwong@stanford.edu)

Abstract

Identifying cancer drivers remains a central challenge in cancer genomics. Most existing approaches rely exclusively on somatic mutation profiles to infer positive selection, which limits their ability to detect rare but biologically important driver events. Here, we introduce IMED (Integrative Mutation and Expression analysis for Driver detection), a Bayesian framework that jointly models somatic mutation profiles and gene expression data to estimate mutation-associated expression effects and latent cellular mixture proportions. IMED prioritizes candidate driver mutations associated with substantial expression changes in their linked genes, providing complementary evidence beyond mutation patterns alone. Through extensive simulations, we showed that IMED achieved higher sensitivity than existing methods, particularly for infrequently mutated driver genes, while maintaining appropriate control of false positives. Applying IMED to three hematological malignancies, we identified 159 candidate driver genes, supported by strong enrichment in established cancer gene sets, validation using AI-based scoring models, and replication in independent datasets. A broader analysis across 29 TCGA cancer types further prioritized putative drivers with low mutation recurrence that were not detected by existing methods. Together, these results highlight the potential of integrative modeling of mutation-associated expression changes for uncovering functionally important cancer drivers, particularly for rare driver events.

Introduction

Cancer is a complex genetic disease driven by the accumulation of somatic mutations that disrupt critical cellular processes, leading to uncontrolled cell proliferation and tumorigenesis[1]. Recent advances in high-throughput sequencing have enabled large-scale efforts, such as The Cancer Genome Atlas (TCGA)[2] and the International Cancer Genome Consortium (ICGC)[3], to systematically catalog millions of somatic mutations across a wide range of cancer types. Despite these advances, a central challenge in cancer genomics persists: distinguishing the relatively small number of driver mutations that confer a selective growth advantage to tumor cells, from a vast background of passenger mutations with little or no functional impact. Accurate identification of these drivers is essential not only for elucidating the molecular mechanisms of oncogenesis but also for guiding the development of targeted therapies and advancing precision oncology.

Historically, the identification of cancer drivers has relied primarily on detecting signals of positive selection within the somatic mutation landscape[4]. This approach rests on the premise that genes driving tumorigenesis will exhibit specific deviations from the expected patterns of neutral mutagenesis. For example, MuSiC[5], MutSigCV[6], and driverMAPS[7] prioritize drivers by identifying genes mutated at significantly higher frequencies than the background rate. Complementary methods search for alternative signatures of selection and can be considered as extended frequency-based models, including biases in mutation composition (dNdScv[8], 20/20+[9]), positional clustering (OncodriveCLUST[10], TUSON[11]), and high functional impact (OncodriveFML[12]). While these methods have successfully cataloged major global drivers like *TP53* and *KRAS*, they face critical limitations. First, valid statistical inference requires an accurate background mutation model[6]. Although genomic covariates such as DNA replication timing and chromatin state are often incorporated to refine background rate estimates, these global adjustments frequently fail to capture local fluctuations in mutability. Consequently, variations in how different algorithms model this background noise often lead to inconsistent predictions and false positives[9]. More importantly, these frequency-based methods lack the statistical power to detect rare drivers (i.e., genes mutated in only a small percentage of patients but are critical to the tumor's survival). For instance, if a gene is infrequently mutated in 3 out of 500 patients, it is unlikely to reach statistical significance against the background noise, regardless of the precision of covariate adjustment. This leaves a vast landscape of low-frequency, yet functionally critical, driver events largely undetected.

To complement mutation-centric views, transcriptomic data provides valuable information about the functional state of cancer cells. Indeed, several methods have combined somatic mutations with transcriptomic changes and molecular networks to prioritize candidate cancer drivers. For example, DriverNet links somatic mutations to expression outliers through molecular interaction network and prioritizes altered genes across a cohort[13]. DawnRank prioritizes patient-specific drivers based on their effects on the expression of downstream genes in regulatory networks[14]. Moonlight combines differential expression, inferred regulatory networks, and biological process information to predict oncogenic,

tumor-suppressive, or context-dependent dual gene roles[15]. Although these approaches leverage transcriptomic information to prioritize cancer drivers, they do not explicitly model the cellular composition of bulk tumor samples. Bulk RNA-seq profiles usually reflect mixtures of normal and malignant cells, and variations in tumor purity and cellular composition can obscure the true tumor-specific expression signal, potentially complicating the interpretation of mutation-associated expression changes. These limitations underscore the need for sophisticated statistical frameworks that explicitly account for cellular heterogeneity when jointly modeling somatic mutations and gene expression data to prioritize candidate cancer drivers.

In this work, we proposed Integrative Mutation and Expression analysis for Driver detection (IMED), a Bayesian framework that jointly models somatic mutation profiles and gene expression data while accounting for cellular heterogeneity. IMED integrates mutant-allele read counts to perform cell type deconvolution (distinguishing normal cells, early-cancer cells, and late-cancer cells), while simultaneously estimating subject-specific cellular mixture proportions and the effect sizes of somatic mutations in cancer cells. Unlike approaches that primarily characterize mutation-associated perturbations in downstream transcriptional networks, IMED focuses on expression changes in mutation-linked genes. By using tumor expression data as a quantitative readout of the functional impact of somatic mutations, we can prioritize functionally consequential mutations rather than relying solely on frequency. Through extensive simulations, we demonstrated that IMED achieves higher sensitivity than existing methods, especially for infrequently mutated driver genes, while maintaining relatively low false discovery rate (FDR). Application to three blood cancers and a broader analysis of 29 TCGA cancer types further demonstrated its ability to prioritize additional candidate drivers with low mutation recurrence. Computational validation in independent datasets and AI-based scoring suggest that many candidate genes identified by IMED are likely to be true drivers.

Results

Method Overview

Our goal is to identify driver mutations and their regulated driver genes by jointly modeling somatic mutation profiles and gene expression data. Here we provide an overview of the proposed framework (**Figure 1**); technical details are described in the “Methods” section. We assume that each tumor biopsy is composed of three major cell types: normal cells, early-cancer cells, and late-cancer cells. Normal cells do not harbor somatic mutations. Early-cancer cells contain a small number of driver mutations together with early-arising passenger mutations. These mutations are inherited by late-cancer cells, which additionally accumulate a large number of late-arising passenger mutations due to genomic instability during tumor progression. Because driver mutations arise early and are present in both early- and late-cancer cells, they are expected to be observed in a larger fraction of tumor cells and therefore have higher variant allele fraction (VAF), defined as the ratio of alternative allele reads count to total read depth. More importantly, we assume that driver mutations exert functional effects by significantly disrupting the expression of

their target genes in the cells in which they occur (e.g., upregulating oncogenes such as *MYC* or downregulating tumor suppressor genes such as *TP53*). These expression changes confer a selective advantage and promote clonal expansion. We use a Bayesian framework to infer the posterior probability that a somatic mutation is a driver, jointly with subject-specific cell type proportions. Mutations and genes with posterior probability greater than 0.5 are classified as drivers.

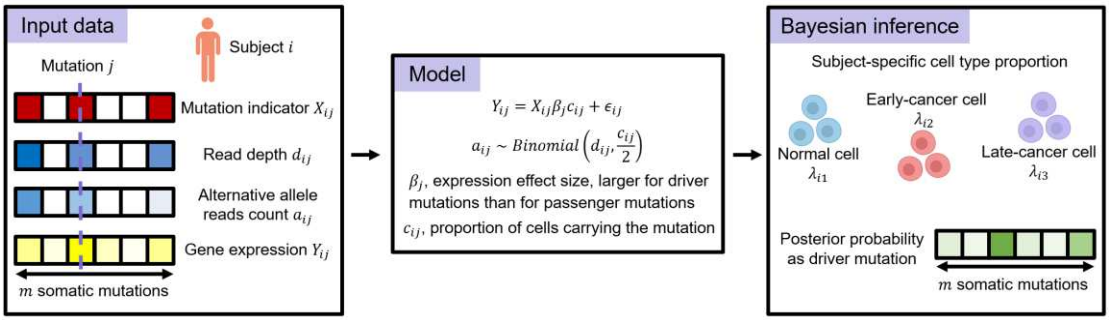


Figure 1. IMED integrates somatic mutation profiles, mutant-allele read counts, and gene expression data to prioritize putative driver mutations and linked genes.

Simulation Results

We first carried out simulations to evaluate whether IMED controls false positive findings. We simulated somatic mutation profiles and bulk-level gene expression data for 300 cancer patients (**Methods**). For patients carrying a given mutation, we perturbed the expression level of the corresponding gene, assuming that driver mutations have much larger functional impact on gene expression than passenger mutations. FDR of IMED was well-calibrated (<0.05) in all simulation settings with varying number of driver genes, mean number of driver mutations per subject, and proportion of coding mutations (**Figures 2a-c**). In addition, IMED accurately estimated subject-specific cell type proportions across these simulation settings (**Supplementary Figures 1-3**). We further benchmarked IMED against two existing driver detection methods. The first method, dNdScv, relies solely on somatic mutation profiles and identifies driver genes based on an excess of nonsynonymous mutations relative to synonymous mutations[8]. Using a q-value threshold of 0.05, dNdScv generally controlled the FDR at its nominal level (**Figures 2a-c**). Another method, DriverNet, integrates somatic mutation profiles and gene expression data through molecular interaction network, to identify driver genes that are connected to genes with outlier expression[13]. However, DriverNet showed substantial FDR inflation under the original simulation setting (**Supplementary Figure 4**). To provide a more favorable setting for DriverNet, we restricted true driver genes to be highly connected genes in the molecular interaction network and used a more stringent q-value threshold of 0.01, yielding the number of detected driver genes comparable with that of dNdScv (**Methods**). Although these modifications reduced DriverNet's FDR, substantial inflation persisted (**Figures 2a-c**).

Next, we compared sensitivity of the three methods in detecting driver genes. IMED achieved considerably higher sensitivity than dNdScv (**Figures 2d-f**). Notably, under

simulation settings with many driver genes, or low driver mutation burden per subject, or the vast majority of driver mutations fall within non-coding region, driver mutations were often sparsely distributed across patients, limiting the ability of dNdScv to detect signals of positive selection. In contrast, by using gene expression data as a readout of functional impact of mutations, IMED could identify driver genes even when their mutation recurrence was low. Although DriverNet could also detect some infrequently mutated driver genes by integrating mutation and gene expression data, its specificity and sensitivity remained consistently lower than those of IMED.

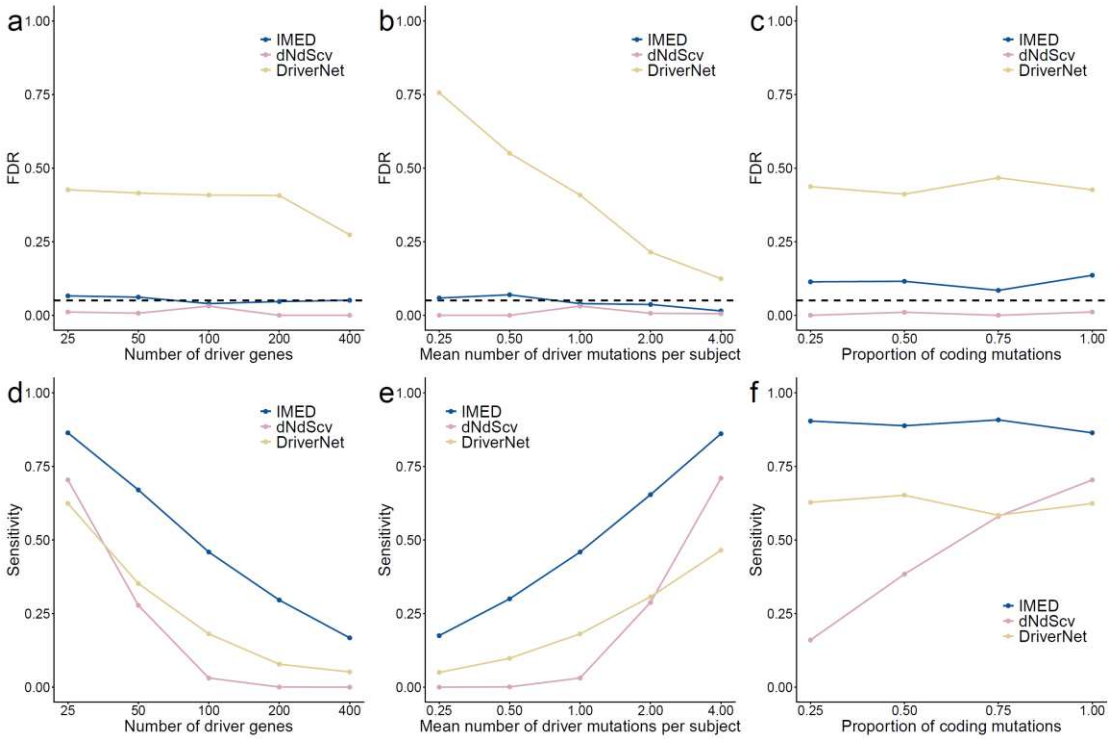


Figure 2. Comparison of FDR and sensitivity between IMED, dNdScv, and DriverNet across varying simulation parameters. Three parameters were considered: the number of driver genes (M), the mean number of driver mutations per subject (λ), and the proportion of coding mutations (p). (a, d) Varying M with fixed $\lambda = 1$ and $p = 1$. (b, e) Varying λ with fixed $M = 100$ and $p = 1$. (c, f) Varying p with fixed $M = 25$ and $\lambda = 1$. Panels (a-c) show FDR, with the nominal level of 0.05 indicated by a dashed line, and panels (d-f) show sensitivity.

Application of IMED in two blood cancers from TCGA

We applied IMED to perform driver detection in two hematological malignancies (DLBC: lymphoid neoplasm diffuse large B-cell lymphoma, n = 36; LAML: acute myeloid leukemia, n = 93) using TCGA data[2]. IMED identified 35 driver mutations across 28 driver genes in DLBC, including 26 genes not detected by either dNdScv or DriverNet (**Figure 3a**; **Supplementary Tables 1-2**). For LAML, IMED identified 14 driver mutations across 12 genes, of which 10 were novel relative to these two methods (**Figure 3a**; **Supplementary Tables 3-4**). In contrast, dNdScv identified only 8 driver genes in DLBC and 7 in LAML, while DriverNet identified 12 and 6, respectively, much fewer than IMED in both cancer types. This is consistent with the simulation results that IMED has much higher statistical

power in identifying rare cancer drivers. Notably, most of the driver mutations were associated with significant up-regulation of the affected genes in mutation carriers compared with non-carriers, suggesting that these mutations may promote tumorigenesis through transcriptional activation[16,17] (**Supplementary Tables 2** and **4**).

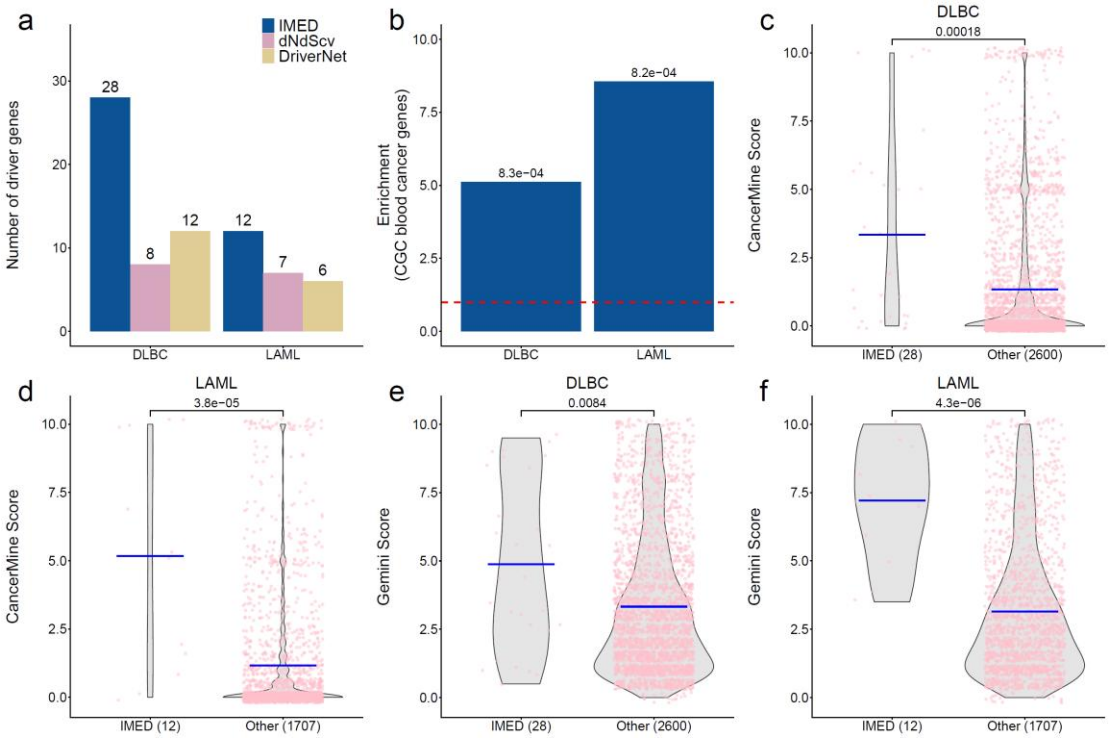


Figure 3. IMED prioritized novel driver genes in lymphoid neoplasm diffuse large B-cell lymphoma (DLBC) and acute myeloid leukemia (LAML). (a) Number of driver genes identified by IMED, dNdScv, and DriverNet. (b) Enrichment of IMED-identified driver genes in a curated list of 297 CGC blood cancer driver genes. (c-d) IMED-identified driver genes show significantly higher CancerMine Scores than background genes in DLBC and LAML. (e-f) IMED-identified driver genes show significantly higher Gemini Scores than background genes in DLBC and LAML.

Next, we compared the identified driver genes with a curated list of 297 known blood cancer driver genes from the COSMIC Cancer Gene Census[18] (CGC version 102; **Methods**). IMED-identified driver genes showed strong enrichment for known blood cancer drivers in both DLBC (enrichment = 5.12, p = 8.3e-4) and LAML (enrichment = 8.55, p = 8.2e-4) (**Figure 3b**). These results demonstrate that IMED not only robustly recovers known blood cancer drivers, but also identifies biologically plausible novel driver genes that are missed by existing mutation frequency-based approaches.

To further assess biological relevance, we evaluated literature support for the identified driver genes using CancerMine[19], a knowledgebase that applies text-mining algorithms to systematically extract associations between genes and cancer types from the biomedical literature (**Methods**). IMED-identified driver genes showed significantly higher blood cancer relevance scores than background genes in both DLBC (Wilcoxon Rank Sum Test;

all drivers: p = 1.8e-4; **Figure 3c**) and LAML (p = 3.8e-5; **Figure 3d**). We observed consistent results using an independent scoring system derived from Gemini[20], a large language model developed by Google, which provides an alternative literature-based assessment of gene-cancer relevance (p = 8.4e-3 and 4.3e-6; **Figure 3e-f**).

IMED identified multiple novel drivers that provide new insights into the molecular mechanisms underlying hematological malignancies. Notably, several IMED-prioritized candidates have been reported as cancer drivers in solid tumors but not in blood cancers according to the CGC database. These include *GRIN2A* (Posterior Probability as driver, PP = 1), *PTPN13* (PP = 1), *CACNA1D* (PP = 0.98) and *NR2F2* (PP = 0.68). The identification of these genes in DLBC suggests that oncogenic pathways originally characterized in other cancer contexts, such as chromatin regulation, calcium signaling, and phosphatase-mediated signal transduction, may also play previously underappreciated roles in DLBC tumorigenesis[21-24]. Additionally, IMED identified novel driver genes that are not in CGC. For example, gene *DTX3L* (PP = 0.99) is frequently overexpressed in DLBC and has been implicated in lymphoma-associated inflammatory signaling and DNA-repair mechanisms that may contribute to tumor-cell survival and treatment resistance[25,26]. Another gene, *IRF8* (PP = 0.88) is a hematopoietic transcription factor essential for B-cell and myeloid development. *IRF8* regulates lineage-associated genes such as *CD37*, and its expression correlates with variable immunophenotypic and clinical features in DLBC[27]. It is also a susceptibility gene and potential therapeutic target in acute myeloid leukemia[28].

In LAML, IMED prioritized one candidate gene *PIK3CB* (PP = 0.68) that has been reported as squamous cell carcinoma and non-small cell lung cancer driver in CGC but not related to blood cancers. Interestingly, knockdown of *PIK3CB* was reported to impair IGF-1 stimulated AKT activation, cell growth, and survival in AML cells[29]. Additionally, gene *EGR1* (PP = 1) is a tumor suppressor located at 5q31.2, which is commonly lost through chromosome 5q deletion/haploinsufficiency in myeloid malignancies[30]. It can suppress leukemogenic phenotypes by promoting myeloid differentiation, inhibiting cell proliferation, and inducing apoptosis[31]. Another gene *DLC1* (PP = 0.98) is a tumor suppressor that is frequently downregulated or epigenetically silenced in AML. Restoration of *DLC1* suppresses RhoA GTPase signaling, thereby inhibiting proliferation and invasion of myeloid cells[32].

Application of IMED in acute lymphoblastic leukemia from MP2PRT-ALL

To evaluate the generalizability of our approach in hematological malignancies, we applied IMED to a large cohort of acute lymphoblastic leukemia (ALL) patients from MP2PRT-ALL (n = 1,356)[33]. IMED identified 161 driver mutations across a total of 119 driver genes (**Figure 4a** and **Supplementary Table 5**), including 103 novel genes that were missed by either dNdScv or DriverNet (**Supplementary Table 6**). In contrast, dNdScv identified only 73 driver genes, while DriverNet identified 19, much fewer than IMED.

We next sought to validate the biological relevance of these candidates by comparing them against established cancer gene sets. The 119 IMED-identified drivers showed strong enrichment for the 72 known ALL driver genes from CGC (enrichment = 11.28, p = 1.5e-5; **Figure 4b**), suggesting that these novel candidates are biologically plausible. Second, given that chromosomal translocations are a hallmark mechanism in ALL, we investigated the overlap between our candidates and 48 known ALL translocation-associated genes (**Methods**). IMED candidates were significantly enriched for these genes (enrichment = 11.28, p = 4.3e-4; **Figure 4c**), indicating that IMED can identify driver mutations that mimic the expression-disrupting effects associated with gene fusions. Finally, IMED-identified drivers showed higher ALL relevance score in CancerMine database (p = 5.5e-4; **Figure 4d**) and Gemini model (p = 1.6e-4; **Figure 4e**) compared with background genes, providing multi-source evidence for their role in leukemogenesis.

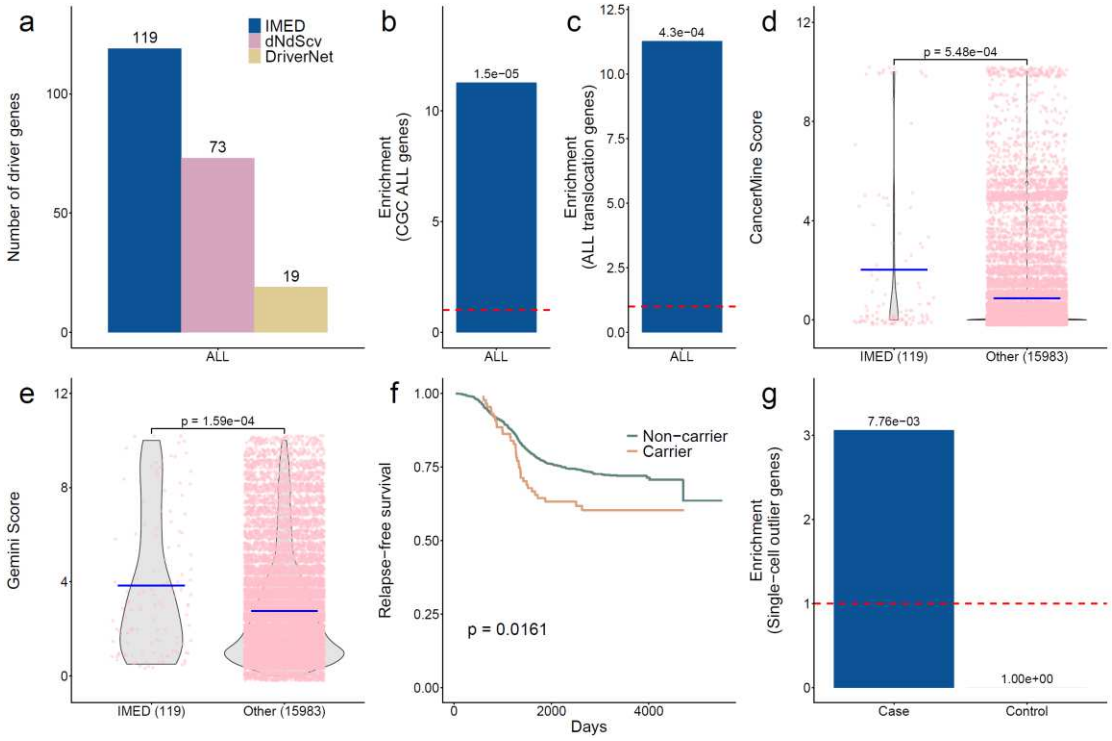


Figure 4. IMED prioritized novel driver genes in acute lymphoblastic leukemia (ALL). (a) Number of driver genes identified by IMED, dNdScv, and DriverNet. (b-c) Enrichment of IMED-identified driver genes among CGC-curated ALL genes and ALL translocation-associated genes. (d-e) IMED-identified ALL driver genes show significantly higher CancerMine Scores and Gemini Scores than background genes. (f) Driver mutation carriers (n = 87) show significantly higher recurrence rates and poorer survival compared with non-carriers (n = 1,269). (g) Enrichment of IMED-identified driver genes among genes with outlier expression in independent single-cell datasets.

Given the high degree of heterogeneity in ALL, we investigated whether IMED-identified novel drivers could refine patient stratification. We found that patients carrying these driver mutations (n = 87) showed a significantly higher relapse rate and poorer survival compared with non-carriers (n = 1,269) (**Figure 4f**). This association suggests that the expression-

associated driver mutations identified by IMED may mark a more aggressive disease subtype. Potential mechanisms for this poor prognosis include transcriptional amplification of growth-related pathways leading to increased proliferation[17], as well as the induction of immunosuppressive microenvironments that facilitate immune evasion[34]. These findings imply that IMED-derived drivers have prognostic utility and could guide the development of more personalized treatment strategies.

We also performed validation using independent single-cell RNA-seq (scRNA-seq) datasets from two pediatric ALL cohorts (total n = 58)[35,36]. Although matched genomic sequencing was unavailable for these samples, we hypothesized that if the IMED-identified genes are true drivers, their expression dysregulation (the “transcriptional footprint”) should be detectable in independent leukemic blasts even without direct mutation calling. We identified 307 genes showing outlier expression events in leukemic blasts within the single-cell data (**Methods**). These outliers showed significant overlap with our IMED-identified drivers (enrichment = 3.06, p = 7.8e-3; **Figure 4g**). As a negative control, no significant overlap was observed with genes showing outlier expression in healthy single-cell samples[35-37] (total n = 55) (**Figure 4g**). This confirms that the dysregulation of IMED candidates is specific to the malignant state and reproducible across platforms.

Many IMED-prioritized novel drivers are biologically meaningful and provide new insights into the molecular mechanisms underlying ALL. Notably, IMED identified 17 genes that have been reported as solid tumor drivers but not ALL drivers in the CGC list (**Supplementary Table 5**), and 9 of them were missed by dNdScv or DriverNet. For example, *CHD4* (PP = 0.71) encodes a core component of the nucleosome remodeling and deacetylase (NuRD) complex, which regulates gene transcription and cell proliferation. *CHD4* has been shown to be essential for AML cell proliferation and disease maintenance, while recurrent *CHD4* mutations have also been identified in T-cell ALL, suggesting its potential involvement in lymphoid leukemogenesis[38,39]. Another gene *FAT3* (PP = 0.56) encodes an atypical cadherin involved in regulating cell polarity, intercellular interactions, and cellular differentiation[40]. Interestingly, recurrent somatic mutations in *FAT3* have been reported in early T-cell precursor ALL[41]. IMED also identified novel ALL driver genes that are not in CGC. For example, *CENPE* (PP = 0.94) encodes a kinesin motor protein essential for chromosome alignment, segregation, and mitotic checkpoint regulation[42]. Pharmacological inhibition of *CENPE* has demonstrated potent antiproliferative activity in ALL cell lines, suggesting its potential relevance to leukemia cell proliferation and survival[43]. *SRSF6* (PP = 0.69) encodes a serine/arginine-rich splicing factor that regulates alternative mRNA splicing, particularly exon skipping. Notably, *SRSF6* is upregulated at the protein level in T-cell ALL, and its depletion inhibits leukemic cell growth[44]. Another gene *MEIS1* (PP = 0.55) encodes a homeobox transcription factor that regulates hematopoietic stem cell self-renewal and differentiation. Its aberrant overexpression promotes leukemogenesis, particularly in *KMT2A*-rearranged acute leukemia, by supporting oncogenic transcriptional programs[45,46]. Together, these findings provide independent evidence supporting the utility of IMED for identifying functionally important driver genes in ALL tumorigeneses.

Application of IMED in 29 TCGA cancer types

To further demonstrate the broad utility of IMED, we applied it to 29 cancer types from TCGA. Across these cancer types, IMED identified a total of 1,149 driver genes, much more than the 342 and 374 identified by dNdScv and DriverNet, respectively (**Figure 5a** and **Supplementary Table 7**). IMED detected the greatest number of driver genes in 22 of the 29 cancer types. Notably, IMED recovered a large proportion of the driver genes identified by dNdScv (**Supplementary Figure 5**). Moreover, these IMED-identified driver genes showed strong enrichment for the known cancer drivers in the CGC (median enrichment = 3.53, **Figure 5b**), supporting the biological relevance of our findings. Interestingly, IMED-identified driver genes showed significantly lower mutation recurrence than those identified by dNdScv (p = 5.6e-8, **Figure 5c**) or DriverNet (p = 1.2e-3, **Figure 5d**). These results are consistent with our simulation results and analyses of hematological malignancies, further demonstrating that IMED can not only recover driver genes identified by existing methods but also improve the detection of rarely mutated, potentially important cancer drivers.

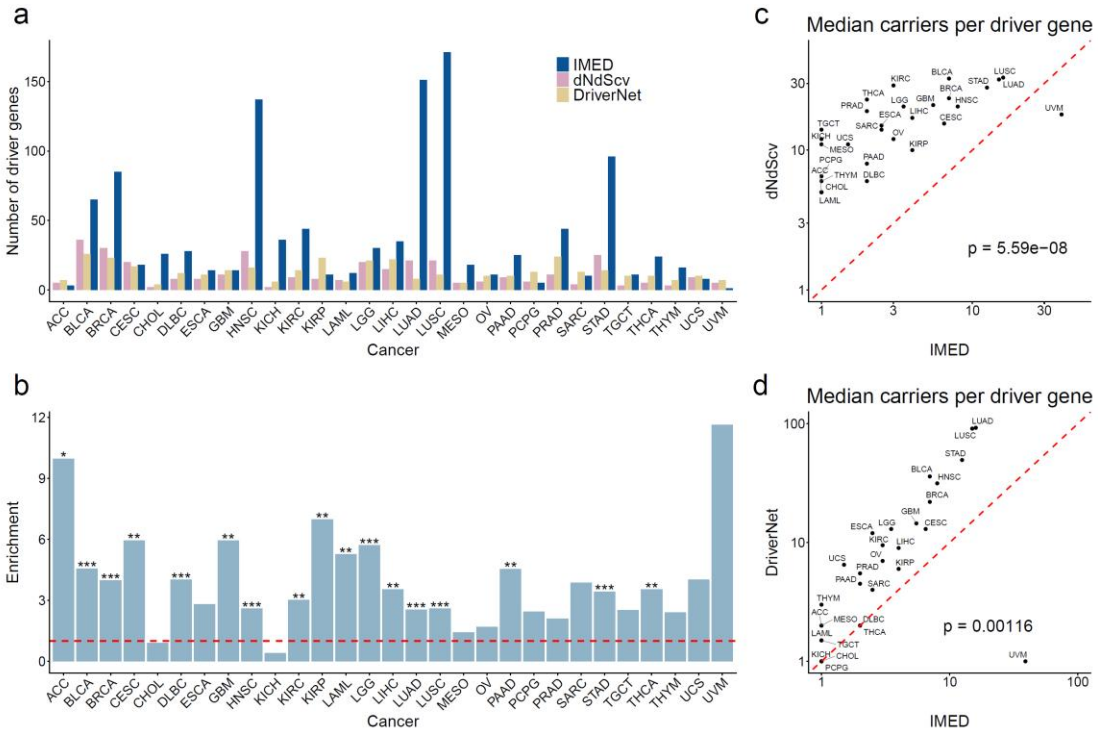


Figure 5. IMED prioritized novel driver genes across 29 TCGA cancer types. (a) Number of driver genes identified by IMED, dNdScv, and DriverNet. (b) Enrichment of IMED-identified driver genes among CGC-curated cancer driver genes. Statistical significance is indicated as follows: *p<0.05, **p<0.01, and ***p<0.001. (c-d) Median number of patients carrying driver mutations per driver gene across 29 cancer types, comparing IMED with (c) dNdScv and (d) DriverNet. Statistical significance was assessed using an exact binomial test.

Discussion

We have developed IMED, an integrative statistical framework that jointly models paired somatic mutation profiles and gene expression data to prioritize candidate cancer drivers while accounting for cellular heterogeneity. Through extensive simulations, we demonstrated that IMED substantially improves statistical power for driver detection compared with dNdScv and DriverNet, while maintaining well-calibrated control of false positives. The improved performance of IMED can be attributed to its joint modeling of mutation-associated gene expression changes and cellular mixture proportions, which enables quantitative assessment of the functional impact of somatic mutations that cannot be captured by DNA sequence level information alone. Applications to three hematological malignancies and a broader analysis of 29 TCGA cancer types identified multiple biologically plausible candidate driver genes, including many with low mutation recurrence that were not detected by the comparison methods. These findings highlight the potential of integrating somatic mutations with transcriptomic data to prioritize candidate cancer drivers for further biological investigation.

We previously developed the carrier statistic, which leverages gene expression measured in disease-relevant tissues to quantify the functional impact of rare germline variants and thereby improves statistical power in rare variant association studies[47]. Somatic mutations, by contrast, occur independently across subjects and are therefore inherently rare. Here, we generalize the carrier statistic framework to the somatic setting and demonstrate that tumor gene expression data can similarly be used to quantify the functional impact of somatic mutations and enhance cancer driver detection. Importantly, incorporating expression data enables assessment of the functional impact of not only coding variants but also noncoding variants, which constitute the vast majority of human genomic variations. Although noncoding variants are known to play a critical role in gene regulation, they are largely excluded from existing driver detection methods because frequency-based approaches, such as dNdScv, primarily rely on protein-level deleteriousness annotations (e.g., nonsynonymous vs. synonymous mutations in coding regions)[8]. As a result, our framework will be particularly well suited for analyzing somatic mutations from whole-genome sequencing (WGS) data. In addition, IMED prioritizes many novel driver mutations that are associated with significant up-regulation of gene expression, consistent with transcriptional activation. We should note that even when such mutations occur in coding regions, they are often overlooked by protein-centric approaches, which are primarily designed to prioritize disruption of protein structure rather than regulatory mechanisms.

IMED prioritizes mutations associated with substantial changes in gene expression as candidate cancer drivers. However, some driver mutations act through mechanisms that do not necessarily alter the mRNA abundance of their linked genes, such as changes in protein activity, stability, splicing, or molecular interactions. Consequently, IMED may overlook driver events that lack detectable expression changes. To address this limitation, the framework can incorporate a dNdScv-informed prior that captures gene-level evidence of positive selection (**Methods**). Incorporating this prior allowed us to identify additional 809 candidate driver genes across 29 TCGA cancer types (**Supplementary Table 8**). More broadly, this framework can accommodate prior information from other complementary

driver detection methods, potentially improving its ability to prioritize candidates acting through diverse molecular mechanisms.

Although our method incorporates regulatory mechanisms to prioritize driver mutations with large functional impact on gene expression, it could further benefit from the integration of additional data types. For example, incorporating HiC, HiChIP, or other 3D chromosomal interaction data could help prioritize mutations in regulatory regions that are more likely to influence gene expression, thereby further improving the statistical power of driver detection. Second, single-cell and spatial transcriptomics technologies are now becoming increasingly prevalent in cancer genomics. These data offer the opportunity to localize the regulatory effects of somatic mutations within specific cellular contexts, such as distinct cell types and anatomical regions. Integrating such data into the IMED framework could substantially enhance its ability to resolve context-specific driver mechanisms. Finally, our method focuses on the direct effects of mutations on their target genes. One important future direction is to extend the framework to incorporate gene regulatory network, allowing quantification of the total functional impact of mutations, including their propagated effects on the downstream genes. As multi-omics datasets continue to accumulate alongside genome sequencing data, we anticipate that IMED will provide an increasingly powerful approach for prioritizing cancer drivers and elucidating the molecular mechanisms underlying tumorigenesis.

Methods

Statistical model of IMED

Suppose there are n subjects with m somatic mutation variants across the genome. Let I_j and $\tilde{I}_j$ denote the indicator of whether mutation j is a driver mutation and early mutation (i.e., a mutation presented in early-cancer cells), respectively. Note that driver mutations must be early mutations. Let $\boldsymbol{\lambda}_i = (\lambda_{i1}, \lambda_{i2}, \lambda_{i3})^T$ denote the subject-specific cell type proportions, representing the proportion of normal cells, early-cancer cells, and late-cancer cells in subject i respectively, where $\lambda_{i1}, \lambda_{i2}, \lambda_{i3}$ are non-negative and $\lambda_{i1} + \lambda_{i2} + \lambda_{i3} = 1$. Let Y_{ij} denote the expression level (e.g., transcripts per million, TPM) of the gene linked to mutation j in subject i. Here $\boldsymbol{Y}_j = \left(Y_{1j}, Y_{2j}, \dots, Y_{nj}\right)^T$ is standardized using the linear transformation $\frac{x-\mu}{\sigma+0.001}$, where μ and σ denote the mean and standard deviation of expression among subjects without any mutation in the corresponding gene, and a small constant of 0.001 was added to the denominator to avoid division by zero. We also assume mutations linked to the same gene within the same subject have been de-duplicated, retaining a single representative mutation. We consider the following linear model:

$$Y_{ij} = X_{ij}\beta_j c_{ij} + \epsilon_{ij}, \quad (1)$$

where X_{ij} is the mutation indicator taking value of 1 if subject i has mutation j and otherwise 0, β_j is the expression effect size of the mutation j, c_{ij} is the proportion of cells carrying the mutation j in subject i, and ϵ_{ij} is non-genetic effect. Here $c_{ij} = \lambda_{i2}\tilde{I}_j + \lambda_{i3}$ because early mutations are present in both early- and late-cancer cells, while late mutations will be present in only late-cancer cells. We assume expression effect size β_j is random, where

$$\beta_j | I_j = 1 \sim N(0, \sigma_{DM}^2), \quad (2)$$

$$\beta_j | I_j = 0 \sim N(0, \sigma_{PM}^2), \quad (3)$$

and σ_{DM}^2 is much larger than σ_{PM}^2 because driver mutations have much larger functional impact on cancer gene expression than passenger mutations. Hyperparameters σ_{DM}^2 and σ_{PM}^2 are set as 25 and 1 throughout the analysis in the paper. We assign a uniform Dirichlet prior on the subject-specific cell type proportions and a Bernoulli prior to the driver mutation and early mutation indicators $I_j, \tilde{I}_j$:

$$\boldsymbol{\lambda}_i \sim Dirichlet(1,1,1), \quad (4)$$

$$\tilde{I}_j \sim Bernoulli(0.05), \quad (5)$$

$$I_j | \tilde{I}_j = 1 \sim Bernoulli(0.05). \quad (6)$$

Further, let a_{ij} and d_{ij} denote the alternative allele reads count and the read depth for mutation j in subject i, respectively. Then $a_{ij} \sim Binomial\left(d_{ij}, \frac{c_{ij}}{2}\right)$. The factor 2 in the denominator is because we assume all somatic mutations are heterozygous.

Bayesian inference

Let $\boldsymbol{Y} = \left(Y_{ij}\right)$, $\boldsymbol{X} = \left(X_{ij}\right)$, $\boldsymbol{a} = \left(a_{ij}\right)$, and $\boldsymbol{d} = \left(d_{ij}\right)$ be the observed data matrices where $i = 1, \dots, n$ and $j = 1, \dots, m$. Let $\boldsymbol{I} = (I_1, \dots, I_m)^T$, $\tilde{\boldsymbol{I}} = (\tilde{I}_1, \dots, \tilde{I}_m)^T$, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_n)^T$ be the unknown parameters to be inferred. Due to the high dimensionality of the parameter

space and the lack of a closed-form posterior distribution, we use Gibbs sampling, an efficient Markov Chain Monte Carlo (MCMC) method, to generate samples from the posterior distribution $P(\boldsymbol{I},\tilde{\boldsymbol{I}},\boldsymbol{\lambda}|\boldsymbol{Y},\boldsymbol{X},\boldsymbol{a},\boldsymbol{d})$. Let $\boldsymbol{I}_{-j}$ and $\tilde{\boldsymbol{I}}_{-j}$ denote the vector of $\boldsymbol{I}$ and $\tilde{\boldsymbol{I}}$ excluding the j-th element, and let $\boldsymbol{\lambda}_{-i}$ denote the matrix of $\boldsymbol{\lambda}$ excluding the i-th row. Furthermore, let $\boldsymbol{I}^{(t)},\tilde{\boldsymbol{I}}^{(t)},\boldsymbol{\lambda}^{(t)}$ denote the random sample in iteration t. Then, in each iteration, we update the driver mutation indicator of mutation j by

$$P\left(I_j^{(t+1)}=1\middle|\boldsymbol{I}_{-j}^{(t)},\tilde{\boldsymbol{I}}^{(t)},\boldsymbol{\lambda}^{(t)},\boldsymbol{Y},\boldsymbol{X},\boldsymbol{a},\boldsymbol{d}\right)=\frac{1}{1+\frac{P\left(I_j=0\middle|\tilde{I}_j^{(t)}\right)}{P\left(I_j=1\middle|\tilde{I}_j^{(t)}\right)}*\prod_{i:X_{ij}=1}\frac{P\left(Y_{ij}|I_j=0,\tilde{I}_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}{P\left(Y_{ij}|I_j=1,\tilde{I}_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}} \quad (8)$$

Where $\frac{P\left(I_j=0\middle|\tilde{I}_j^{(t)}\right)}{P\left(I_j=1\middle|\tilde{I}_j^{(t)}\right)}=\begin{cases}\frac{1-p_j^{(2)}}{p_j^{(2)}},\ if\ \tilde{I}_j^{(t)}=1\\ \infty,\ if\ \tilde{I}_j^{(t)}=0\end{cases}$, $\frac{P\left(Y_{ij}|I_j=0,\tilde{I}_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}{P\left(Y_{ij}|I_j=1,\tilde{I}_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}=\exp\left\{\frac{1}{2}\log\left(\sigma_{i1|j}^2\right)-\frac{1}{2}\log\left(\sigma_{i0|j}^2\right)+Y_{ij}^2\left(\frac{1}{2\sigma_{i1|j}^2}-\frac{1}{2\sigma_{i0|j}^2}\right)\right\}$, $\sigma_{i1|j}^2=c_{ij}^{(t)2}\sigma_{DM}^2+1$, $\sigma_{i0|j}^2=c_{ij}^{(t)2}\sigma_{PM}^2+1$, and $c_{ij}^{(t)}=\lambda_{i2}^{(t)}\tilde{I}_j^{(t)}+\lambda_{i3}^{(t)}$. The update for the early mutation indicator of mutation j is given by

$$P\left(\tilde{I}_j^{(t+1)}=1\middle|\tilde{\boldsymbol{I}}_{-j}^{(t)},\boldsymbol{I}^{(t)},\boldsymbol{\lambda}^{(t)},\boldsymbol{Y},\boldsymbol{X},\boldsymbol{a},\boldsymbol{d}\right)=\frac{1}{1+\frac{P\left(\tilde{I}_j=0\right)*P\left(I_j^{(t)}|\tilde{I}_j=0\right)}{P\left(\tilde{I}_j=1\right)*P\left(I_j^{(t)}|\tilde{I}_j=1\right)}*\prod_{i:X_{ij}=1}\frac{P\left(a_{ij}|\tilde{I}_j=0,d_{ij},\boldsymbol{\lambda}_i^{(t)}\right)}{P\left(a_{ij}|\tilde{I}_j=1,d_{ij},\boldsymbol{\lambda}_i^{(t)}\right)}*\frac{P\left(Y_{ij}|\tilde{I}_j=0,I_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}{P\left(Y_{ij}|\tilde{I}_j=1,I_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}} \quad (9)$$

Where $\frac{P\left(\tilde{I}_j=0\right)*P\left(I_j^{(t)}|\tilde{I}_j=0\right)}{P\left(\tilde{I}_j=1\right)*P\left(I_j^{(t)}|\tilde{I}_j=1\right)}=\begin{cases}0,\ if\ I_j^{(t)}=1\\ \frac{1-p_j^{(1)}}{p_j^{(1)}}*\frac{1}{1-p_j^{(2)}},\ if\ I_j^{(t)}=0\end{cases}$, $\frac{P\left(a_{ij}|\tilde{I}_j=0,d_{ij},\boldsymbol{\lambda}_i^{(t)}\right)}{P\left(a_{ij}|\tilde{I}_j=1,d_{ij},\boldsymbol{\lambda}_i^{(t)}\right)}=\frac{\lambda_{i3}^{(t)a_{ij}}*\left(2-\lambda_{i3}^{(t)}\right)^{d_{ij}-a_{ij}}}{\left(\lambda_{i2}^{(t)}+\lambda_{i3}^{(t)}\right)^{a_{ij}}*\left(2-\lambda_{i2}^{(t)}-\lambda_{i3}^{(t)}\right)^{d_{ij}-a_{ij}}}$, $\frac{P\left(Y_{ij}|\tilde{I}_j=0,I_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}{P\left(Y_{ij}|\tilde{I}_j=1,I_j^{(t)},\boldsymbol{\lambda}_i^{(t)}\right)}=\exp\left\{\frac{1}{2}\log\left(\sigma_{i1|j}^2\right)-\frac{1}{2}\log\left(\sigma_{i0|j}^2\right)+Y_{ij}^2\left(\frac{1}{2\sigma_{i1|j}^2}-\frac{1}{2\sigma_{i0|j}^2}\right)\right\}$, $\sigma_{i1|j}^2=\left(\lambda_{i2}^{(t)}+\lambda_{i3}^{(t)}\right)^2Var\left(\beta_j|I_j\right)+1$, $\sigma_{i0|j}^2=\lambda_{i3}^{(t)2}Var\left(\beta_j|I_j\right)+1$. The update for the subject-specific cell type proportions $\boldsymbol{\lambda}_i$ is given by

$$P\left(\boldsymbol{\lambda}_i^{(t+1)}|\boldsymbol{I}^{(t)},\tilde{\boldsymbol{I}}^{(t)},\boldsymbol{\lambda}_{-i}^{(t)},\boldsymbol{Y},\boldsymbol{X},\boldsymbol{a},\boldsymbol{d}\right)\propto P(\boldsymbol{\lambda}_i)*\prod_{j:X_{ij}=1}P\left(a_{ij}|\tilde{I}_j^{(t)},d_{ij},\lambda_{ij}\right)*P\left(Y_{ij}|I_j^{(t)},\tilde{I}_j^{(t)},\lambda_{ij}\right)=P(\boldsymbol{\lambda}_i)*w_i(\boldsymbol{\lambda}_i) \quad (10)$$

Where $P\left(a_{ij}|\tilde{I}_j^{(t)},d_{ij},\lambda_{ij}\right)=\binom{d_{ij}}{a_{ij}}\left(\frac{c_{ij}^{(t)}}{2}\right)^{a_{ij}}\left(1-\frac{c_{ij}^{(t)}}{2}\right)^{d_{ij}-a_{ij}}$, $c_{ij}^{(t)}=\lambda_{i2}\tilde{I}_j^{(t)}+\lambda_{i3}$, $P\left(Y_{ij}|I_j^{(t)},\tilde{I}_j^{(t)},\lambda_{ij}\right)=\frac{1}{\sqrt{2\pi\sigma_{ij}^2}}\exp\left\{-\frac{Y_{ij}^2}{2\sigma_{ij}^2}\right\}$, $\sigma_{ij}^2=\left(\lambda_{i2}\tilde{I}_{ij}^{(t)}+\lambda_{i3}\right)^2Var\left(\beta_j|I_j\right)+1$. We use importance sampling method to update $\boldsymbol{\lambda}_i^{(t+1)}$ in Equation (10) based on 100 random

samples from $P(\boldsymbol{\lambda}_i) \sim Dirichlet(1,1,1)$ with weights given by $w_i(\boldsymbol{\lambda}_i)$.

We perform MCMC sampling using two independent chains, each consisting of 500 iterations, with the first 100 iterations discarded as burn-in. Posterior estimates are obtained by averaging the posterior distributions from the two chains. After posterior inference, we estimated the effect sizes of driver mutations as follows. Let $\hat{\lambda}_{i2}$ and $\hat{\lambda}_{i3}$ denote the posterior means of λ_{i2} and λ_{i3}, respectively, and let $\hat{c}_{ij} = \hat{\lambda}_{i2} + \hat{\lambda}_{i3}$. Substituting $\hat{c}_{ij}$ into Equation (1) and treating it as known, we obtained the ordinary least squares (OLS) estimate of β_j as $\hat{\beta}_j = \frac{\sum_i X_{ij}\hat{c}_{ij}Y_{ij}}{\sum_i X_{ij}^2 \hat{c}_{ij}^2}$. The posterior probability that a gene is a driver can be calculated as $1 - \prod_{j=1}(1 - PP_j)$, where PP_j denotes the posterior probability that the j-th mutation in that gene is a driver mutation.

dNdScv-informed prior

Different driver detection approaches leverage complementary types of information. For example, dNdScv uses genetic sequence information classified by their translational consequence without incorporating gene expression data. Consequently, dNdScv may identify driver genes missed by IMED when their expression-level signals are weak. To improve sensitivity for driver detection, we incorporate dNdScv results as prior information in IMED. Specifically, the dNdScv results are used to assign prior probabilities to the driver mutation and early mutation indicators as follows:

$$\tilde{I}_j \sim Bernoulli\left(p_j^{(1)}\right), \tag{5}$$

$$I_j | \tilde{I}_j = 1 \sim Bernoulli\left(p_j^{(2)}\right). \tag{6}$$

$$p_j^{(1)} = p_j^{(2)} = \begin{cases} 0.5, & if\ \ q_j^{dNdScv} \le 0.05 \\ 0.2, & if\ \ q_j^{dNdScv} > 0.05\ \ and\ \ p_j^{dNdScv} \le 0.05 \\ 0.05, & if\ \ p_j^{dNdScv} > 0.05 \end{cases}. \tag{7}$$

Here $p_j^{dNdScv}, q_j^{dNdScv}$ are the p-value and q-value of the gene linked to mutation j from dNdScv. Note that this strategy is not limited to dNdScv and can be extended to incorporate prior information from other complementary driver-detection methods.

Simulations

We conducted simulations to generate input data for IMED and evaluate its driver detection performance. Specifically, we randomly selected M driver genes from background protein-coding genes that were connected to at least 100 genes, with M varying from 25 to 400. Gene-gene interactions were quantified using the latest version of the Reactome protein functional interaction network, following the approach described in the DriverNet manuscript[48]. This selection criterion was deliberately designed to favor DriverNet by restricting candidate driver genes to highly connected genes, thereby reducing its FDR, while having no impact on the performance of IMED or dNdScv (**Supplementary Figure 4**). Further, for each of 300 simulated cancer patients, we first generated subject-specific cell type proportion $(\lambda_1, \lambda_2, \lambda_3)$ independently from a Dirichlet distribution $Dirichlet(1,1,1)$.

We then simulated somatic mutation profiles for these patients. Each patient was assumed to harbor 100 somatic mutations, and the number of driver mutations per patient was independently drawn from a Poisson distribution $Poisson(\lambda)$, with λ varing from 0.25 to 4. Driver mutations were randomly assigned to the M driver genes, while the remaining passenger mutations were assigned uniformly at random across all 19,215 background protein-coding genes. All driver mutations were set as early mutations, while each passenger mutation had a probability of 0.1 of being an early mutation. For all somatic mutations, the read depth was fixed at 150x sequence coverage. The alternative allele reads count was independently generated from a Binomial distribution $Binomial\left(150, \frac{c}{2}\right)$, where $c = \lambda_2 + \lambda_3$ for early mutations and $c = \lambda_3$ for late mutations. Finally, we simulated bulk-level gene expression data. Background expression levels were independently drawn from a standard normal distribution $N(0,1)$ for each gene and each cancer patient. For each mutation, the expression of the corresponding gene in the affected patient was perturbed by adding βc to the background expression, where $\beta \sim N(0,25)$ for driver mutations and $\beta \sim N(0,1)$ for passenger mutations. To mimic the propagation of transcriptomic perturbations through gene-gene interactions, we further perturbed the expression of the genes connected to mutation-harboring genes by adding $\beta c f$ to their background expression, where $f \sim N(0,0.25)$,. Each simulation setting was repeated for 100 times.

Because dNdScv identifies driver genes based on an excess of nonsynonymous mutations relative to synonymous mutations, it requires specification of both the trinucleotide context (e.g., ACC→ATC) and the functional annotation of each mutation. To apply dNdScv in our simulations, we adopted the same underlying genetic architecture described above, with additional assumptions tailored to its input requirements. Specifically, we assumed that all mutations are single-nucleotide variants, and that only a proportion p of somatic mutations occurred within coding regions, with p varying from 0.25 to 1. For passenger mutations, we randomly selected a trinucleotide context from all possible contexts given the sequence of the corresponding gene. For driver mutations, we randomly selected a trinucleotide context from the subset of contexts annotated as nonsynonymous (e.g., essential splice, missense, nonsense, or stop-loss). Under this construction, all driver mutations were classified as nonsynonymous, creating a simulation setting that is deliberately biased in favor of dNdScv.

We evaluated the performance of different methods using two metrics: FDR and sensitivity. FDR was defined as the proportion of falsely identified driver genes among all identified ones. Sensitivity was defined as the proportion of truly identified driver genes among all underlying drivers.

Implementation of different methods

The existing method dNdScv was implemented using the R package dndscv (v0.0.1.0). All parameters were set to their default values, with the hg38 reference genome, the precomputed trinucleotide context-dependent substitution model 192r_3w, and

epigenomic covariates enabled. Genes with qglobal_cv < 0.05 were identified as driver genes.

DriverNet was implemented using the R package DriverNet (v1.34.0). For each gene, patients with expression levels more than two standard deviations above or below the mean expression across all patients were classified as outliers. The latest version of the Reactome protein functional interaction network was used as the influence graph[48]. The number of randomization tests for calculating empirical p-value was set to 500. All other parameters were set to their default values. In simulations, DriverNet showed substantial FDR inflation when using the conventional q-value threshold of 0.05 without restricting the connectivity of underlying true driver genes (**Supplementary Figure 4**). To create a simulation setting more favorable to DriverNet, we restricted candidate driver genes to those highly connected genes with at least 100 interaction partners. Although this restriction reduced the FDR, substantial inflation persisted. We therefore used a more conservative q-value threshold of 0.01, which yielded a number of detected driver genes comparable to that identified by dNdScv. This threshold was used for DriverNet in simulation and real data analyses presented in the main text.

Hematological malignancies data preparation

We downloaded somatic simple nucleotide variations (Masked Somatic Mutation files) detected from whole-exome sequencing (WES) studies for 29 TCGA cancer types from the Genomic Data Commons (GDC) Data Portal[2]. COAD, READ, SKCM, and UCEC with exceptionally high tumor mutation burden (TMB) or high TMB variability were excluded from the analyses. Somatic mutations identified from WES and called by MuTect2 for ALL in the MP2PRT-ALL[33] cohort were also obtained from the GDC Data Portal. In addition, RNA-seq gene expression data from tumor tissue were downloaded from the GDC Data Portal. Only subjects with both somatic mutation profiles and gene expression data were included in the analysis.

We applied stringent filtering procedures to remove likely false-positive mutations. Somatic mutations with tumor read depth < 30 or normal read depth < 10 were excluded. To minimize germline contamination, mutations with a variant allele fraction (VAF) > 0.1 in normal tissue were also removed. Mutations linked to the same gene within the same subject were de-duplicated, retaining a single representative mutation. For hematological malignancies, we retained all coding variants and only noncoding variants located within ENCODE bone marrow candidate cis-regulatory elements (cCREs), as these mutations are more likely to affect gene expression. For other cancer types, all mutations passing the initial quality-control filters were retained.

Gene expression levels were quantified using transcripts per million (TPM), and only autosomal protein coding genes with a maximal TPM > 1 across subjects were included. To reduce skewness, TPM values were log-transformed as $\log(TPM + 0.001)$ when necessary. Expression values were then standardized using the linear transformation $\frac{x-\mu}{\sigma+0.001}$, where μ and σ denote the mean and standard deviation of expression among

subjects without any mutation in the corresponding gene. A small constant of 0.001 was added to the denominator to avoid division by zero. Finally, IMED was applied to integrate somatic mutation profiles and gene expression data.

Validation against established cancer gene sets

We defined a set of known blood cancer driver genes as the union of genes annotated in the CGC (version 102)[18] as being associated with leukemia, myeloproliferative neoplasm, lymphoma, or myeloma. This resulted in 297 "known" blood cancer driver genes. In addition, 72 genes annotated in CGC with cancer types abbreviated as "ALL", "B-ALL", "ETP-ALL", "pre-B ALL", or "T-ALL" were defined as known ALL driver genes. Among these, 48 genes involved in gene fusions with annotated translocation partners were further defined as known ALL translocation-associated driver genes. Enrichment of IMED-identified drivers within these established cancer gene sets was assessed using Fisher's Exact Test.

We further validated IMED-identified drivers using CancerMine[19], a literature-based database that systematically catalogs gene-cancer associations via text mining. For each gene, we computed a cancer relevance score as the sum of citation counts linking the gene to all cancer types, and a blood cancer relevance score as the sum of citation counts restricted to blood cancers. Each score was capped at 10, and the final relevance score was defined as the average of the cancer relevance score and the blood cancer relevance score. A Wilcoxon Rank Sum Test was then used to compare the relevance score distribution between IMED-identified driver genes and remaining background genes.

In addition, we validated IMED-identified driver genes using Gemini[20], a large language model developed by Google. For each gene, we queried Gemini to obtain a cancer relevance score using the prompt: "For each gene, can you give a cancer relevance score? 10 is fully relevant, 0 is fully irrelevant. Only output the score." Blood cancer relevance scores were obtained using an analogous prompt with "blood cancer relevance score" substituted for "cancer relevance score." Due to context length constraints, genes were queried in batches of 100. Gemini 2.5 Flash was used without personal memory or conversation recall. The final Gemini-based relevance score was computed as the average of the cancer relevance score and the blood cancer relevance score. Differences in score distributions between IMED-identified driver genes and background genes were assessed using the Wilcoxon rank-sum test.

Validation with independent scRNA-seq data

We obtained three independent scRNA-seq datasets from pediatric ALL patients and healthy controls: Chen et al. ($n_{case} = 18,\ n_{ctrl} = 5$)[35] and Dunsmore et al. ($n_{case} = 40,\ n_{ctrl} = 5$)[36] and Zeng et al. ($n_{ctrl} = 40$)[37], yielding a total of 58 cases and 55 controls. Genes with outlier expression event were identified separately within the case and control groups. For each subject, we aggregated gene expression across leukemic blast cells to obtain subject-level pseudo-bulk expression profiles quantified in TPM, and further transformed as $\log(TPM + 0.001)$. Only autosomal protein coding genes with a maximal TPM > 1 across subjects were retained for downstream analysis. For each gene, we

computed a leave-one-out z-score for subject i as $z_i = \frac{EXPR_i - mean(EXPR_{-i})}{sd(EXPR_{-i}) + 0.001}$, where $EXPR_i$ denotes the gene expression level for subject i, and $mean(EXPR_{-i})$ and $sd(EXPR_{-i})$ represent the mean and standard deviation of gene expression levels across all other subjects. A small constant of 0.001 was added to the denominator to avoid division by zero. Genes with a maximal absolute z-score over 15 across subjects were defined as showing outlier expression events. Using this procedure, we identified 307 outlier genes in the case group and 6 outlier genes in the control group. We then assessed the overlap between outlier genes identified from the scRNA-seq datasets and IMED-identified driver genes. Enrichment of IMED-identified driver genes among outlier genes was evaluated using Fisher's Exact Test.

Data availability

Molecular and clinical data for 29 TCGA cancer types and for ALL in MP2PRT-ALL were downloaded from the GDC Data Portal, https://portal.gdc.cancer.gov; Single-cell RNA-seq data for pediatric ALL patients and healthy controls were downloaded from Human Tumor Atlas Network https://humantumoratlas.org/, Childhood Cancer Data Initiative Hub, https://ccdi.cancer.gov/, and the GitHub website of BoneMarrowMap atlas, https://github.com/andygxzeng/BoneMarrowMap; CGC annotation, https://cancer.sanger.ac.uk/cosmic/; Functional interaction network was downloaded from Reactome database, https://reactome.org/download-data; Google Gemini model, https://gemini.google.com/; CancerMine database, https://cancermine.ai4biomed.org/; ENCODE candidate cis-regulatory elements, https://screen.encodeproject.org/.

Code availability

IMED is available at https://github.com/ghm17/IMED.

Acknowledgements

This work was partially supported by the Stanford Maternal and Child Health Research Institute. The results shown here are part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Author contributions

H.G. and W.H.W. designed the idea. H.G. performed data analysis. H.G., A.E.U., and W.H.W. interpreted the results. H.G. and W.H.W. wrote the manuscript. A.E.U. and W.H.W. supervised the study.

Fundings

This research was partially supported by grants from the National Institutes of Health (R01 HG010359 to W.H.W.; R01 MH116529 to A.E.U.; and P50 HG007735 to W.H.W. and A.E.U.) and the National Science Foundation (DMS 2310788 to W.H.W.).

Declaration of interests

The authors declare no competing interests.

References

1. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: the next generation. *cell* **144**, 646–674 (2011).
2. Weinstein, J.N. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature genetics* **45**, 1113–1120 (2013).
3. Consortium, I.C.G. International network of cancer genome projects. *Nature* **464**, 993 (2010).
4. Martínez-Jiménez, F. *et al.* A compendium of mutational cancer driver genes. *Nature Reviews Cancer* **20**, 555–572 (2020).
5. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome research* **22**, 1589–1598 (2012).
6. Lawrence, M.S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
7. Zhao, S. *et al.* Detailed modeling of positive selection improves detection of cancer driver genes. *Nature communications* **10**, 3399 (2019).
8. Martincorena, I. *et al.* Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041. e21 (2017).
9. Tokheim, C.J., Papadopoulos, N., Kinzler, K.W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proceedings of the National Academy of Sciences* **113**, 14330–14335 (2016).
10. Tamborero, D., Gonzalez-Perez, A. & Lopez-Bigas, N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics* **29**, 2238–2244 (2013).
11. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
12. Mularoni, L., Sabarinathan, R., Deu-Pons, J., Gonzalez-Perez, A. & López-Bigas, N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome biology* **17**, 128 (2016).
13. Bashashati, A. *et al.* DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome biology* **13**, R124 (2012).
14. Hou, J.P. & Ma, J. DawnRank: discovering personalized driver genes in cancer. *Genome medicine* **6**, 56 (2014).
15. Colaprico, A. *et al.* Interpreting pathways to discover cancer driver genes with Moonlight. *Nature communications* **11**, 69 (2020).
16. Di Malta, C. *et al.* Transcriptional activation of RagD GTPase controls mTORC1 and promotes cancer growth. *Science* **356**, 1188–1192 (2017).
17. Lin, C.Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
18. Tate, J.G. *et al.* COSMIC: the catalogue of somatic mutations in cancer. *Nucleic acids research* **47**, D941–D947 (2019).
19. Lever, J., Zhao, E.Y., Grewal, J., Jones, M.R. & Jones, S.J. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nature methods* **16**, 505–507 (2019).

20. Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
21. Wei, X. *et al.* Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nature genetics* **43**, 442–446 (2011).
22. Mcheik, S., Aptecar, L., Coopman, P., D'hondt, V. & Freiss, G. Dual role of the PTPN13 tyrosine phosphatase in cancer. *Biomolecules* **10**, 1659 (2020).
23. Phan, N.N. *et al.* Voltage-gated calcium channels: Novel targets for cancer therapy. *Oncology letters* **14**, 2059–2074 (2017).
24. Qin, J., Tsai, S.Y. & Tsai, M.-J. The critical roles of COUP-TFII in tumor progression and metastasis. *Cell & bioscience* **4**, 58 (2014).
25. Juszczynski, P. *et al.* BAL1 and BBAP are regulated by a gamma interferon-responsive bidirectional promoter and are overexpressed in diffuse large B-cell lymphomas with a prominent inflammatory infiltrate. *Molecular and cellular biology* **26**, 5348–5359 (2006).
26. Wang, L., Sun, X., He, J. & Liu, Z. Functions and molecular mechanisms of deltex family ubiquitin E3 ligases in development and disease. *Frontiers in cell and developmental biology* **9**, 706997 (2021).
27. Elfrink, S. *et al.* IRF8 is a transcriptional activator of CD37 expression in diffuse large B-cell lymphoma. *Blood Advances* **6**, 2254–2266 (2022).
28. Sharma, A. *et al.* Constitutive IRF8 expression inhibits AML by activation of repressed immune response signaling. *Leukemia* **29**, 157–168 (2015).
29. Park, S. *et al.* Role of the PI3K/AKT and mTOR signaling pathways in acute myeloid leukemia. *haematologica* **95**, 819 (2009).
30. Joslin, J.M. *et al.* Haploinsufficiency of EGR1, a candidate gene in the del (5q), leads to the development of myeloid disorders. *Blood, The Journal of the American Society of Hematology* **110**, 719–726 (2007).
31. Krishnaraju, K., Hoffman, B. & Liebermann, D.A. The zinc finger transcription factor Egr-1 activates macrophage differentiation in M1 myeloblastic leukemia cells. *Blood, The Journal of the American Society of Hematology* **92**, 1957–1966 (1998).
32. Ullmannova-Benson, V. *et al.* DLC1 tumor suppressor gene inhibits migration and invasion of multiple myeloma cells through RhoA GTPase pathway. *Leukemia* **23**, 383–390 (2009).
33. Chang, T.-C. *et al.* Genomic determinants of outcome in acute lymphoblastic leukemia. *Journal of Clinical Oncology* **42**, 3491–3503 (2024).
34. George, S. *et al.* Loss of PTEN is associated with resistance to anti-PD-1 checkpoint blockade therapy in metastatic uterine leiomyosarcoma. *Immunity* **46**, 197–204 (2017).
35. Chen, C. *et al.* Single-cell multiomics reveals increased plasticity, resistant populations, and stem-cell–like blasts in KMT2A-rearranged leukemia. *Blood, The Journal of the American Society of Hematology* **139**, 2198–2211 (2022).
36. Dunsmore, K.P. *et al.* Children's Oncology Group AALL0434: a phase III randomized clinical trial testing nelarabine in newly diagnosed T-cell acute lymphoblastic leukemia. *Journal of Clinical Oncology* **38**, 3282–3293 (2020).
37. Zeng, A.G. *et al.* Single-cell transcriptional atlas of human hematopoiesis reveals genetic and hierarchy-based determinants of aberrant AML differentiation. *Blood Cancer Discovery* **6**, 307–324 (2025).
38. Heshmati, Y. *et al.* The chromatin-remodeling factor CHD4 is required for maintenance of

childhood acute myeloid leukemia. *Haematologica* **103**, 1169 (2018).
39. O'Connor, D. *et al.* The clinicogenomic landscape of induction failure in childhood and young adult T-cell acute lymphoblastic leukemia. *Journal of Clinical Oncology* **41**, 3545–3556 (2023).
40. Le Pabic, P., Ng, C. & Schilling, T.F. Fat-Dachsous signaling coordinates cartilage differentiation and polarity during craniofacial development. *PLoS genetics* **10**, e1004726 (2014).
41. Neumann, M. *et al.* Whole-exome sequencing in adult ETP-ALL reveals a high rate of DNMT3A mutations. *Blood, The Journal of the American Society of Hematology* **121**, 4749–4752 (2013).
42. Wood, K.W., Sakowicz, R., Goldstein, L.S. & Cleveland, D.W. CENP-E is a plus end–directed kinetochore motor required for metaphase chromosome alignment. *Cell* **91**, 357–366 (1997).
43. Lock, R.B. *et al.* Initial testing of the CENP-E inhibitor GSK923295A by the pediatric preclinical testing program. *Pediatric blood & cancer* **58**, 916–923 (2012).
44. Zhou, Y. *et al.* Posttranslational regulation of the exon skipping machinery controls aberrant splicing in leukemia. *Cancer discovery* **10**, 1388–1409 (2020).
45. Wong, P., Iwasaki, M., Somervaille, T.C., So, C.W.E. & Cleary, M.L. Meis1 is an essential and rate-limiting regulator of MLL leukemia stem cell potential. *Genes & development* **21**, 2762–2774 (2007).
46. Smith, A.L. *et al.* Enhancer heterogeneity in acute lymphoblastic leukemia drives differential gene expression in patients. *Blood* **146**, 2073–2087 (2025).
47. Guo, H., Urban, A.E. & Wong, W.H. Prioritizing disease-related rare variants by integrating gene expression data. *PLOS Genetics* **20**, e1011412 (2024).
48. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome biology* **11**, R53 (2010).

Supplementary Files

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

- SupplementaryFigures.docx
- SupplementaryTables.xlsx