PLOS Computational Biology

METHODS

PICDGI: A framework for predicting cancer driver genes through dynamic gene-gene interaction modeling of single-cell data

Komlan Atitey[1], **Benedict Anchang**[1,2*]

1 Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina, United States of America, **2** Center for Cancer Research, National Cancer Institute, Bethesda, Maryland, United States of America

* benedict.anchang@nih.gov

OPEN ACCESS

Citation: Atitey K, Anchang B (2026) PICDGI: A framework for predicting cancer driver genes through dynamic gene-gene interaction modeling of single-cell data. PLoS Comput Biol 22(4): e1014143. https://doi.org/10.1371/journal.pcbi.1014143

Editor: Pingzhao Hu, Western University, CANADA

Received: September 2, 2025

Accepted: March 19, 2026

Published: April 27, 2026

Data availability statement: The raw LUAD dataset is publicly available in the NCBI Gene Expression Omnibus (GEO) under accession number GSE131907. For external validation, we analyzed an independent pediatric AML scRNA-seq dataset spanning Diagnosis,

Abstract

Identifying cancer driver genes (CDGs) remains a central challenge in cancer genomics, as frequency-based mutation approaches often miss rare but functionally important regulators. We present PICDGI, a computational framework that predicts driver-like regulatory genes by integrating dynamic gene-gene interaction modeling with single-cell RNA sequencing (scRNA-seq) data. Rather than relying on DNA mutation calls, PICDGI infers functional driver activity from time-resolved expression patterns and latent regulatory influence among genes during tumor progression. Methodologically, PICDGI employs a time-varying state-space model with variational Bayesian inference and Markov Chain Monte Carlo (MCMC) sampling to estimate evolving gene interaction effects. The posterior distributions capture both the magnitude and uncertainty of each gene's inferred regulatory influence. From these, PICDGI derives a driver coefficient that quantifies the strength and reliability of each gene's contribution to progression-associated expression dynamics, enabling the prioritization of impactful regulators over neutral passengers. Applied to lung adenocarcinoma (LUAD) scRNA-seq data, PICDGI recovered known oncogenes and tumor suppressors and nominated novel candidate drivers, including *JPH1* and *CHEK1*, which are implicated in calcium signaling, mitochondrial regulation, and DNA repair. These genes exhibit trajectory-aligned activity consistent with tumor evolution and immune-modulatory processes. Comparative analysis using Moran's I statistics in Monocle 3 showed that PICDGI-prioritized genes display stronger progression-associated dynamics than genes selected by spatial autocorrelation alone. We further validated PICDGI on an independent pediatric acute myeloid leukemia (AML) scRNA-seq cohort, where it consistently recovered known drivers and relapse-associated regulatory programs under fixed model parameters. By integrating interaction-informed dynamic modeling with single-cell resolution data, PICDGI provides a generalizable and biologically grounded framework for identifying rare and context-specific regulatory drivers of cancer progression, with broad applicability across tumor types.

End-of-Induction, and Relapse stages. The AML dataset is publicly available in GEO under accession number GSE235923. All data analyzed in this study are publicly accessible through the referenced repositories. Additionally, processed data and supporting R Source files for downstream analysis using PICDGI are available at https://github.com/NIEHS/PICDGI.

Funding: This research was supported by the Intramural Research Program of the National Institutes of Health, specifically the National Institute of Environmental Health Sciences and the National Cancer Institute (grant number 1ZIAES103350-05 to AB). KA and AB are employees of the National Institutes of Health and receive salary support from the Intramural Research Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Identifying which genes truly drive cancer progression is a central challenge in cancer biology. Most existing approaches focus on how often mutations occur across patients, which can overlook rare but functionally important drivers. We developed PICDGI, a computational method that integrates single-cell RNA sequencing with dynamic, interaction-aware modeling to identify cancer driver genes. Unlike traditional mutation-based tools, PICDGI evaluates how genes influence one another over time, while accounting for uncertainty in these regulatory effects, allowing it to capture the evolving gene networks that shape tumor heterogeneity and immune evasion at single-cell resolution. Using lung adenocarcinoma as a test case, we show that PICDGI recovers known driver genes and highlights new candidates involved in processes such as DNA repair and mitochondrial regulation. Beyond this application, PICDGI provides a broadly applicable framework for studying disease evolution from dynamic single-cell data and for uncovering regulatory targets that may inform personalized treatment strategies across cancer types.

1 Introduction

Cancer arises due to multiple genetic alterations, including mutations in oncogenes (OGs) and tumor suppressor genes (TSGs) [1]. OGs promote uncontrolled cell growth through gain-of-function mutations while TSGs drive oncogenesis when they lose their protective function. Together, these cooperate to promote cancer development [2] (Fig 1A). Traditionally, somatic mutations are classified as either drivers, which are causally implicated in cancer progression, or passengers, which are considered biologically neutral. Distinguishing between these two categories remains a significant challenge due to the heterogeneity of somatic mutations and the contamination from non-tumor cells in the clinical samples [3]. In this study, we focus on a subset of cancer driver genes that we refer to as immunoregulatory cancer driver genes. These genes contribute to tumor initiation and progression through intrinsic oncogenic or tumor-suppressive functions, while also influencing the tumor microenvironment and immune-cell regulatory programs. Such genes may be associated with modulation of cytotoxic immune-cell activity, cytokine signaling, antigen presentation, or other pathways that shape anti-tumor immune responses. By modeling dynamic gene-gene interactions across both tumor and immune compartments, PICDGI is designed to identify genes that exhibit coordinated regulatory influence across malignant and immune contexts during cancer evolution.

Many computational methods rely on mutation recurrence to predict cancer driver genes (CDGs), assuming that frequently mutated genes are more likely to be drivers [4]. Tools such as MutSigCV [5], OncodriveFM [6], OncodriveFML [7], and OncodriveCLUST [8] have successfully identified recurrent drivers from bulk sequencing data. However, these methods often struggle with rare drivers [9], which are easily

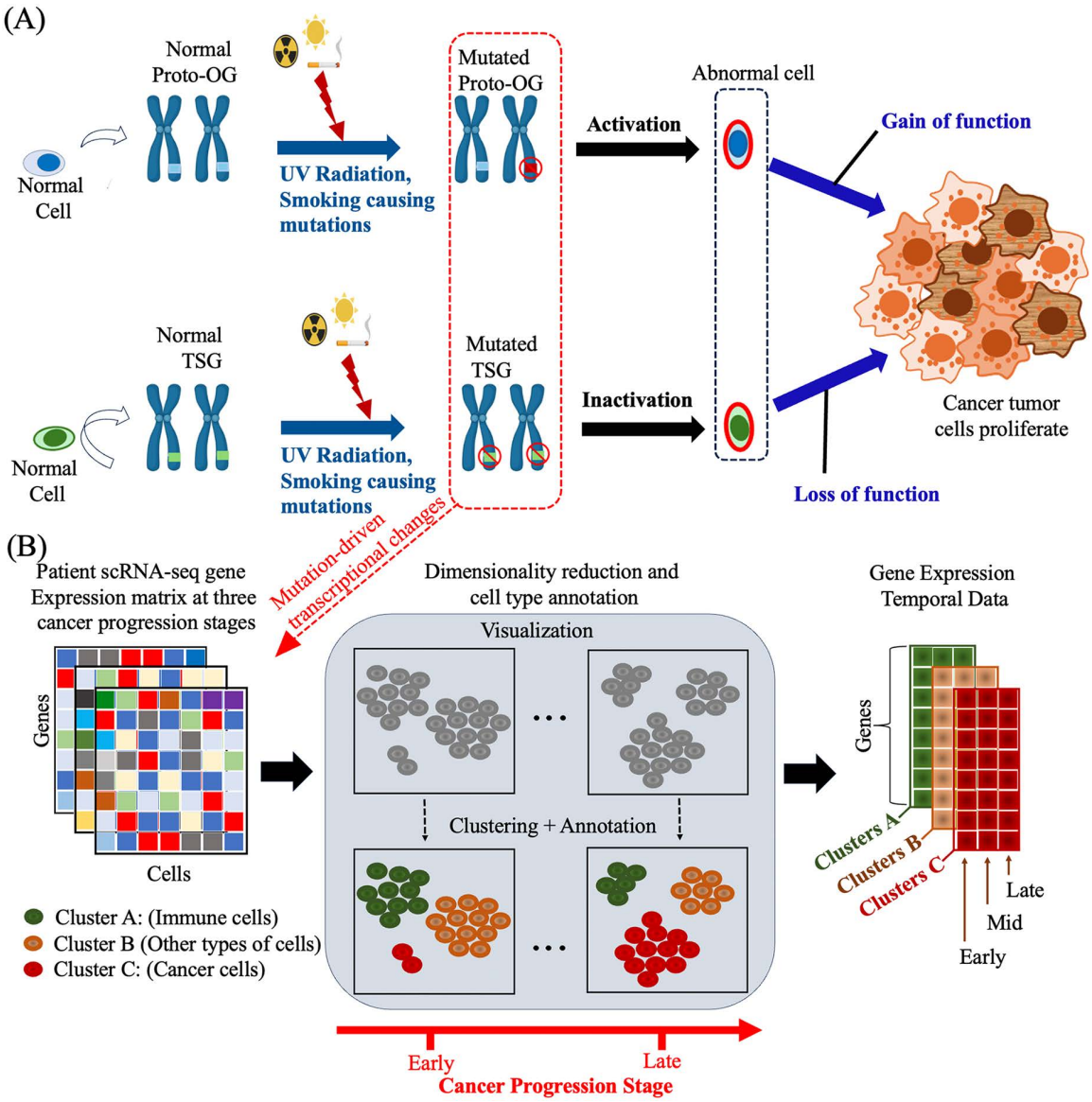

Fig 1. From environmental mutations to the emergence of cellular heterogeneity in cancer progression. Schematic representation of how environmental factors contribute to mutations that drive cancer development. Mutations in Proto-oncogenes (Proto-OG), and or tumor suppressor genes (TSG) impair their normal protective roles, leading to emergence of cancerous cells. Mutations induced by factors such as UV radiation and smoking can activate OGs (upper pathway) or the inactivation of TSGs (lower pathway). These mutations disrupt normal cellular regulation, leading to uncontrolled cell proliferation and tumor formation, which in turn cause widespread changes in gene expression. (B). Overview of single-cell gene expression heterogeneity. ScRNA-seq data are collected from cancer patients at different stages of progression for example Early, Mid, and Late. The processed expression matrices were visualized using a nonlinear dimensionality reduction method to denoise data, reduce complexity, and improve cluster interpretability for cell type identification. Clustering and annotation are used to reveal distinct cell populations, including immune cells, cancer cells, and other cell types. For each identified cluster (Cluster A, Cluster B, Cluster C), time-series gene expression vectors are derived from the three stages, representing dynamic changes in expression during cancer progression.

https://doi.org/10.1371/journal.pcbi.1014143.g001

misclassified as passengers due to sampling bias, sequencing noise, or tumor purity effects [10]. Additional reliance on somatic mutation data introduces biases, limits discovery to well-studied genes, and makes it difficult to assess functional consequences without experimental validation [11].

To overcome these limitations, researchers have turned to single-cell transcriptomics. scRNA-seq provides high-resolution profiling of individual cells, uncovering cellular heterogeneity and enabling refined models of tumor evolution [12]. Building on this, time-series and trajectory inference tools such as RNA velocity [13], scVelo [14], and Waddington-OT [15] have enabled prediction of cell-state transitions and global population dynamics. Meanwhile, gene regulatory network (GRN) reconstruction methods like GRNBoost2 [16], SCODE [17], and Dyngen [18] model interactions underlying state changes, though they often rely on linear assumptions or dense temporal sampling not feasible in tumors. Similarly, methods such as PseudotimeDE [19] identify temporally varying genes but do not directly connect dynamic regulation to driver gene prioritization.

Network-based and impact-based methods such as ActiveDriver [20], DawnRank [21], DriverNet, PNC [22], and SCS [23] integrate prior pathway knowledge to identify impactful genes. Comparative evaluations [24] show that some, like ActiveDriver, perform well across multiple cancers; including LUAD, but remain limited to specific mutation types or pre-defined gene sets [25]. More recently, multi-omics frameworks such as IMI-driver [26] and CSDGI [27] integrate diverse data modalities to improve driver discovery, yet most do not explicitly incorporate temporal gene-gene interaction dynamics. Together, these efforts highlight progress but also underscore persistent challenges: bias toward recurrent mutations, neglect of dynamic tumor evolution; dependence on known gene sets; and difficulty validating novel candidates [20,28]. Importantly, most current methods adopt a static view of tumors, overlooking how non-stationary gene-gene interactions shape heterogeneity, therapeutic resistance, and immunosuppression [29].

To address these limitations, we introduce PICDGI (Predicting Immunoregulatory Cancer Driver Genes via Gene-Gene Interactions), a Bayesian framework that integrates scRNA-seq data with dynamic gene interaction modeling to prioritize functionally relevant CDGs. Methodologically, PICDGI builds on variational Bayesian inference [30,31] combined with MCMC sampling to infer non-stationarity regulatory effects over tumor progression. The model derives a driver coefficient from the posterior distribution, quantifying each gene's evolving influence on tumor growth and immunoregulatory processes. This work compliments our earlier study [32], where we modeled interactions of canonical drivers (e.g., EGFR, KRAS, TP53) using an algorithm called DEGBOE. PICDGI generalizes this approach into a four-step pipeline: 1.) Cancer progenitor identification by integrating scRNA-seq data across tumor stages to construct average temporal profiles (Fig 1B). 2.) Modeling dynamic, nonstationary gene-gene interactions along tumor progression. 3.) Bayesian inference of regultory influence on tumor evolution, and 4.) Computing driver coefficients to prioritize candidate CDGs based on dynamic regulatory impact.

We applied PICDGI to nine scRNA-seq datasets from three LUAD patients [33]. Among the top 30 predicted CDGs, 62% overlapped with known OGs and TSGs [2], validating recovery of established drivers. The remaining 38% represent novel candidates for further validation. Functional evaluation against Moran's I statistics [34] in Monocle 3 showed that PICDGI-prioritized genes exhibited stronger expression dynamics and higher tumor-associated expression levels [35,36] reinforcing their role as high-confidence drivers. We further validated PICDGI on an independent pediatric acute myeloid leukemia (AML) scRNA-seq cohort, where it consistently identified known drivers and relapse-associated regulatory programs using the same model settings without any re-tuning.

In this study, we use the term "driver gene" in a functional rather than strictly genomic sense. PICDGI does not analyze DNA mutation calls nor does it attempt to infer sequence-level mutation events. Instead, the framework identifies genes that exhibit driver-like regulatory influence based on their dynamic expression behavior across tumor progression. Thus, PICDGI captures functional regulatory drivers, which are genes whose time-dependent transcriptional influence promotes cancer progression and immune suppression, even in the absence of detectable somatic mutations. The following sections detail the methodological framework and demonstrate its application to LUAD single-cell datasets, followed by external validation in an independent AML cohort.

2 Materials and methods

In this study, we assume that single-cell gene expression dynamics reflect the functional regulatory consequences of oncogenic processes that drive cancer progression, rather than directly measuring DNA-level mutation events.

Additionally, we consider the heterogeneity in gene expression across individual cells captures true biological diversity such as sub-clonal structures, lineage differentiation, and dynamic cellular states, rather than being merely the result of technical noise.

2.1 Overview and rationale

This study aims to identify immunoregulatory CDGs by leveraging time-resolved scRNA-seq data within a dynamic modeling framework called PICDGI. Unlike conventional differential expression or pseudotime trajectory methods, PICDGI explicitly models gene-gene interactions as stochastic, time varying processes and infers latent regulatory trajectories using Bayesian inference. This enables the identification of genes that conditionally regulate other genes over time, including those with immunoregulatory effects in the tumor microenvironment [37,38].

The framework consists of four major components; each directly tied to the observed scRNA-seq data:

1. **Time-dependent gene expression:**

The scRNA-seq data are preprocessed and normalized, dimensionally reduced, and annotated into cell types across distinct cancer stages. We then construct temporal gene expression matrices for each annotated cell type, forming the basis for modeling dynamic gene activity.

2. **Cancer originating cell identification:**

Cancer originating cells are identified as the most likely cells of origin based on trends in cell population expansion and cancer cell fraction (CCF), expression of cancer-associated programs, and stage-wise persistence. These cells provide the primary context for modeling regulatory evolution and serve as the reference lineage for downstream driver inference.

3. **Stochastic modeling of gene-gene interactions:**

The temporal expression of each gene is modeled as a nonstationary stochastic process using a time-varying fractional Autoregressive Moving Average model (ARMA) model. Gene-gene interactions are captured through latent variables representing the regulatory influence of one gene on another during progression.

4. **Bayesian inference and driver scoring:**

Variational Bayesian inference and MCMC sampling are used to approximate the joint posterior distribution of mutation states and gene-gene interaction effects. From this posterior, a driver coefficient (DrCoef) is computed as a squared signal-to-noise ratio, quantifying both the strength and stability of each gene's inferred regulatory impact on tumor progression. Genes with high DrCoef values are prioritized as candidate functional drivers because they exhibit strong and reliable influence on expression trajectories over time.

By modeling latent regulatory interactions across time and evaluating their differential impact on immune versus tumor compartments, PICDGI provides a principled approach to uncover functionally significant cancer drivers that may evade detection through static or marginal analysis.

Throughout this study, all modeling in PICDGI is based exclusively on scRNA-seq expression data. No genomic mutation calls are used as input. References in the text to "mutation events," "mutation states," or "driver genes" correspond to latent regulatory influence variables inferred from transcriptional dynamics, not observed DNA-level alterations. Thus, the Driver Coefficient quantifies functional regulatory impact, not genomic mutation status.

2.2 The PICDGI algorithm

In PICDGI, the term gene mutation state denotes a latent regulatory activity process inferred from expression dynamics. It does not correspond to observed DNA-level mutations, but rather represents a probabilistic variable used to model

time-varying gene influence within regulatory networks. The PICDGI algorithm consists of four main steps to model and infer cancer driver-like genes (CDG) from single-cell data using a time-aware, probabilistic framework: (1) identifying cancer originating cells and summarizing gene expression temporal data from reduced scRNA-seq data, (2) modeling gene expression trajectories as nonstationary, time-varying stochastic processes, allowing regulatory influences between genes to change across progression (3) using Variational Bayesian inference combined with MCMC sampling to estimate posterior distributions over latent regulatory influences. and (4) Computing driver coefficients from posterior mean and variance to quantify each gene's regulatory impact and stability over time, and genes are ranked accordingly as candidate drivers of cancer progression.

2.2.1 ***PICDGI identifies cancer progenitor cells for discovering cancer driver.*** Let the gene expression profile of cell j be denoted by $g^j \in \mathbb{R}^N$, where N is the number of genes. For a given gene i, let $\left\{g_i^j\right\}_{j=1}^{M}$ denote its expression levels across M cells in a cluster C at a given biological stage. The cluster-level mean expression of gene i is computed as:

$$\bar{g}_i = \frac{1}{M}\sum_{j=1}^{M} g_i^j \tag{1}$$

To model temporal dynamics, we define an ordered set of biological sampling stages. In this study, cancer progression stages are denoted by $S = \{S_1, S_2, \ldots, S_k\}$, such that $S_1 < S_2 < \ldots < S_k$ where each stage corresponds to a clinically or experimentally defined time point (e.g., diagnosis, treatment response, relapse, or early/advanced disease). These biological stages are mapped to ordered model time indices $T = \{T_1, T_2, \ldots, T_k\}$ used for dynamic inference.

Given N genes, C clusters, and time points T, we compute the mean $\bar{g}_i$ for each cluster and each time point T_k. This results in a temporal gene expression dataset of dimension $N \times k \times C$ (Fig 1B) enabling joint modeling of gene dynamics across cell populations and disease progression.

Cancer progenitor cells are identified using two criteria: (1) The abundance of cancer originating cells, denoted as ξ_c, exhibits a consistent stage-dependent trend across progression stages [39,40], satisfying:

$$\xi_c\left(S_k\right) \text{ exhibits a consistent stage dependent trend across } k \tag{2}$$

(2) For each cell type C_i, we compute its cancer cell fraction $CCF(C_i)$ using marker genes. For each cluster C_i, the CCF is defined as:

$$CCF\left(C_i\right) = \frac{1}{M_i}\sum_{j=1}^{M_i} \mathbb{I}\left(g^j \in \text{marker} - \text{positive}\right) \tag{3}$$

where M_i is the number of cells in cluster C_i and $\mathbf{I}(\cdot)$ is an indicator function equal to 1 if the gene expression profile g^j of cell j exceeds a defined threshold for a known cancer marker gene. In this analysis, a cell is considered marker-positive (and thus potentially cancerous) if it shows any non-zero expression of the selected gene (e.g., *EPCAM*). A progenitor cell type C_p must have higher CCF than all other cell types:

$$CCF\left(C_p\right) > CCF\left(C_j\right), \qquad \forall C_j \neq C_p \tag{4}$$

By jointly applying Eqs (2) and (3), we identify cancer originating cells. Specifically, for each cluster C_i across patients and stages $S_1, S_2, \ldots, S_K$, we calculate $\xi_{C_i}(S_k)$, the fraction of that cluster at stage S_k. A valid progenitor population must exhibit consistent stage-dependent trend across stages (Eq. 2) and the highest CCF among clusters (Eq. 3). This

classification allows estimation of *CCF* by counting the proportion of marker-positive cells within each cluster exhibiting tumor-associated expression signatures. The final progenitor cell population C_p is defined as the cluster that (1) shows increasing/decreasing abundance or enrichment across cancer stages and (2) has the highest average *CCF* among all clusters. This dual criterion ensures that the selected cluster both demonstrates stage-associated expansion during progression and displays strong tumor-like expression, consistent with a likely cancer-originating population. The above described stage-aware aggregation framework is agnostic to cancer type and is applicable to both solid tumors (e.g., LUAD) and hematologic malignancies (e.g., AML), provided that ordered sampling stages are available.

2.2.2 ***Modeling nonstationary, discrete time-varying genetic events.*** We model gene mutations as nonstationary, discrete-time, integer-valued stochastic processes. This modeling follows five key steps, outlined below

Step 1: *Modeling gene expression dynamics as stochastic processes using the ARMA model*

We treat gene mutations as nonstationary, discrete-time, integer-valued stochastic processes, where event counts fluctuate over time or space [41,42]. In our framework, a gene system consists of N different distributions or populations, each represented by a latent variable $x \in \mathbb{R}^+ = [0, \infty]$. Their evolution is driven by gene interaction effects (Fig 2A-2C).

Following Grenier (1983) [43], we model the nonstationary signal of a single gene distribution at time k denoted as x_k using a finite-order, time-varying Autoregressive Moving Average (ARMA) process. We formulate the ARMA model as:

$$x_k = \sum_{i=1}^{m} \phi_{i,(k-i)} x_{k-i} + \sum_{j=1}^{n} \lambda_{j,(k-j)} u_{k-j} + u_k \tag{5}$$

where $x_k \in \mathbb{R}^N$ is the gene mutation state at time k, and $u_k \in \mathbb{R}^N$represents the innovation (input) or error process, capturing random mutation events that drive changes in gene populations over time. In Eq. (5), the indices i and j denote the autoregressive and moving-average "lag orders", respectively. Consequently, $k-i$ and $k-j$ refer to earlier time points of the same gene signal, rather than to different genes. Gene-gene regulatory influences are incorporated later through the global interaction matrix defined in Section 2.2.3. Unlike gene expression, which reflects transcript abundance, u_k models additional stochastic mutation signals beyond past history, allowing the ARMA process to account for the nonstationary nature of mutation dynamics. The coefficients, ϕ, and λ are the autoregressive and moving average coefficients, respectively, which capture the autocorrelation of the output process x_k. The input u_k is assumed to be a zero-mean Gaussian error process [44], correlated over time to allow for a wide range of memory decay properties in the time series.

Step 2: *State-space representation*

Equation (5) express the system in stacked state-space form, where the concatenated state vector $\boldsymbol{x}_k = \{x_{1:k}\}$ and the innovation process $\boldsymbol{u}_k = \{u_{1:k}\}$ capture the gene expression or mutation dynamics across all stages up to time k. We model the system as a linear transformation of innovations:

$$\boldsymbol{x}_k = \boldsymbol{\Theta}_k \boldsymbol{u}_k \tag{6}$$

Both $\boldsymbol{x}_k \in \mathbb{R}^{N \times k}$ and $\boldsymbol{u}_k \in \mathbb{R}^{N \times k}$ are concatenated multivariate vectors, reflecting the simultaneous modeling of N interacting gene distributions across k time points. The transfer matrix $\boldsymbol{\Theta}_k \in \mathbb{R}^{(N \times k) \times (N \times k)}$ defined as $\boldsymbol{\Theta}_k = \boldsymbol{\phi}_k^{-1} \boldsymbol{\Lambda}_k$, governs the relationship between the innovations and the system states, encoding both the observability and the controllability of the evolving gene network [32]. The matrices $\boldsymbol{\phi} = \{\phi_1, \phi_2, \cdots, \phi_m\}$, and $\boldsymbol{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ represent the ARMA coefficients (See S1 Text).

Step 3: *Modeling non-stationarity with fractional gaussian noise*

To account for the model's non-stationarity, the innovation process $\boldsymbol{u}_k$ is modeled as fractional Gaussian noise with a mean $E[\boldsymbol{u}_k] = 0$ representing stationary increments [45]. The increment process $\Delta u(k) = u(k) - u(k-1)$ is characterized

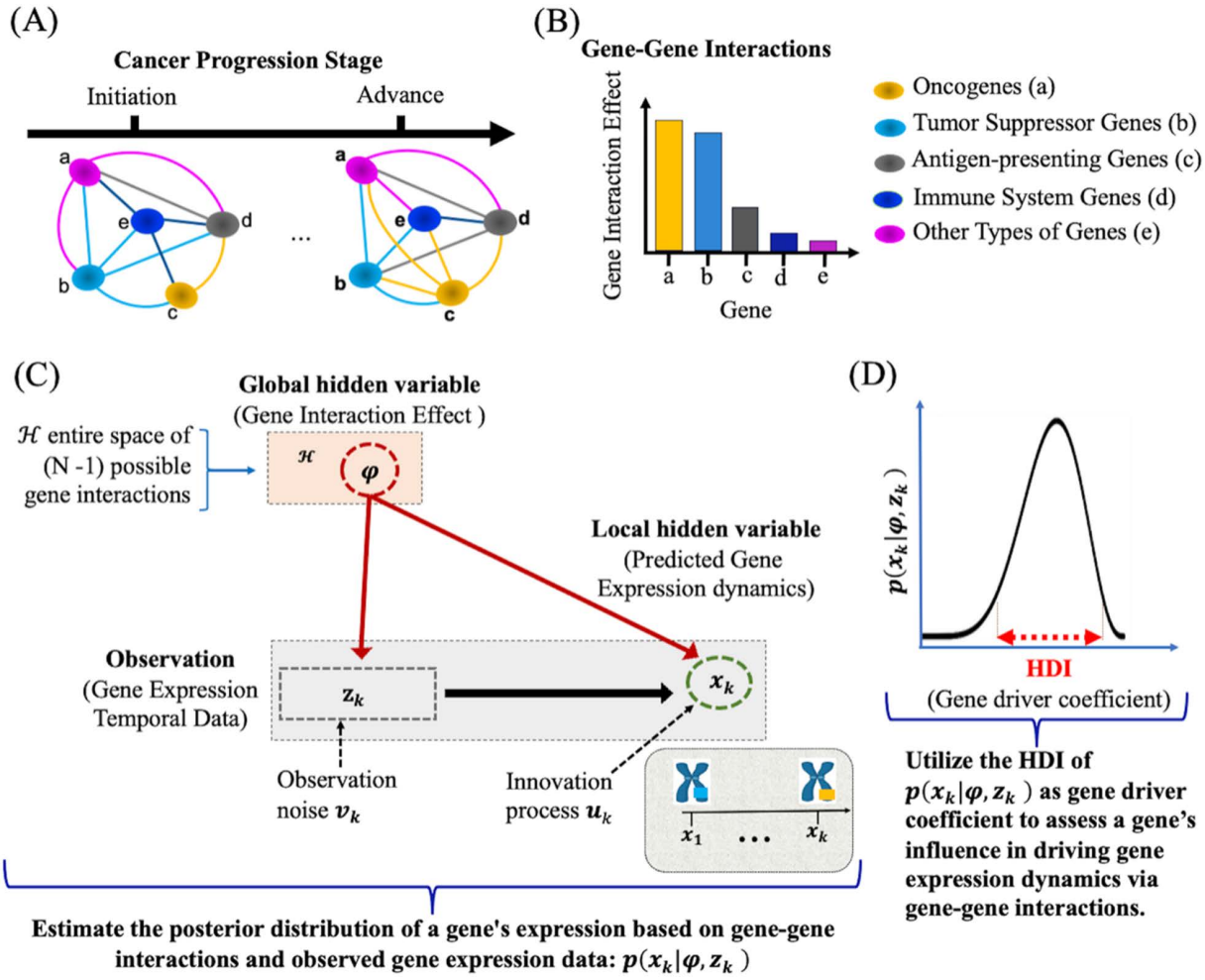

Fig 2. PICDGI framework. (A) Representation of gene-gene interaction effects (GIE) in cancer progression. Nodes denote genes and edges denote regulatory interactions, with statistical variability in interactions contributing to genetic heterogeneity. Five categories of genes are considered, with their interaction effects differing by type. **(B)** Illustration of GIE strength: oncogenes (OGs) and tumor suppressor genes (TSGs) are expected to exert stronger effects on network dynamics compared with other gene classes. **(C)** Computational formulation of PICDGI. The model links observed temporal gene expression data to hidden variables at two levels: (i) local hidden variables (e.g., gene-specific mutations and expression fluctuations) and (ii) global hidden variables capturing the overall GIE structure across the network. **(D)** Inference procedure. The effect of a gene on driving mutations in other genes is quantified through the highest density interval (HDI) of the posterior distribution over gene expression dynamics, integrating both temporal patterns and estimated gene–gene interactions.

https://doi.org/10.1371/journal.pcbi.1014143.g002

by the Hurst exponent H, governing long-range dependence in time series [46]. Following Chiang *et al.*[47], we define the autocovariance function $\gamma_{u_k}(\tau)$ of the increments as:

$$\gamma_{\boldsymbol{u}}(\tau) = E\left[u_{k+\tau}u_k\right] = \frac{\sigma_u^2}{2}\left[|\tau+1|^{2H} - 2|\tau|^{2H} + |\tau-1|^{2H}\right] \tag{7}$$

We constrain the variance σ_u and the Hurst exponent H to $0.5 < H < 1$ to preserve non-stationary properties [48]. For sequence lengths $k = \{1, 2, \cdots K\}$ (where K=3, as shown in Fig 1B), the autocorrelation function simplifies to: $\rho_u(\tau) = \frac{1}{2}\left[|\tau+1|^{2H} - 2|\tau|^{2H} + |\tau-1|^{2H}\right]$ yielding $\gamma_{\boldsymbol{u}}(\tau) = \sigma_u^2.\rho_u(\tau)$, which parameterizes the noise process using only the Hurst exponent H.

Step 4: *Covariance matrix for the innovation process*
The covariance matrix $\boldsymbol{C}_{\boldsymbol{u}_k}$ for the zero-mean Gaussian vector $\boldsymbol{u}_k$ is:

$$\boldsymbol{C}_{\boldsymbol{u}_k} = \sigma_u^2 \boldsymbol{R}_{\boldsymbol{u}_k} \tag{8}$$

where $\boldsymbol{R}_{\boldsymbol{u}_k}$ is the $K \times K$ correlation matrix, defined as a Toeplitz matrix:

$$\boldsymbol{R}_{\boldsymbol{u}_k} = \begin{pmatrix} \rho_u(0) & \rho_u(1) & \cdots & \rho_u(K-1) \\ \rho_u(1) & \rho_u(0) & \cdots & \rho_u(K-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho_u(K-2) & \rho_u(K-1) & \cdots & \rho_u(1) \\ \rho_u(K-1) & \rho_u(K-2) & \cdots & \rho_u(0) \end{pmatrix} \tag{9}$$

The correlation structure reflects long-range dependencies in the innovation process, governed by the Hurst exponent H. Heatmaps of $\boldsymbol{C}_{\boldsymbol{u}_k}$ for different H values illustrate how persistence increases with H (Fig 3). Based on these analyses, the optimal Hurst exponent is chosen as $H = 0.6$ (Fig 3, S2 Text).

Step 5: *Covariance matrix for the state process*
In this model, the transfer matrix $\boldsymbol{\Theta}_k$ is typically dense, meaning that all entries can potentially be nonzero. This density reflects the complex dependencies between different gene populations, where each gene's mutation dynamics can be influenced by innovations from multiple other genes. Thus, $\boldsymbol{\Theta}_k$ encodes both direct and indirect interactions between genes. Using Equation 6, we define the covariance matrix for the zero-mean Gaussian-distributed variable $\boldsymbol{x}_k$ as:

$$\boldsymbol{C}_{\boldsymbol{x}_k} = \boldsymbol{\Theta}_k \boldsymbol{C}_{\boldsymbol{u}_k} \boldsymbol{\Theta}_k^T \tag{10}$$

The evolution of the i-th gene population, where $i \in \{1,\ 2,\ \ldots,\ N\}$, follows a Gaussian process defined as:

$$\boldsymbol{x}_k \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}\right), \tag{11}$$

where $\boldsymbol{\mu}_{\boldsymbol{x}_k} = 0$ represents the zero mean and $\boldsymbol{C}_{\boldsymbol{x}_k}$ is the covariance matrix governing the system's dynamics.

2.2.3 ***Modeling gene-gene interaction.*** We define a generative model where observed gene expression arises from latent mutation states and global interaction parameters. Tumor cellular complexity emerges from gene interplay, rather than the actual individual genes alone. To model this complexity, we introduce a local hidden variable x_k representing gene mutation states at time k and a global hidden variable φ representing the gene interaction effects at time k. The observed time-series gene expression vector, z_k is modeled with additive Gaussian noise v_k, resulting in:

$$z_k = \mathfrak{h}\left(x_k, \varphi\right) + v_k, \qquad \mathfrak{h}\left(x_k, \varphi\right) \sim \mathcal{GP}\left(0, \kappa\left(\left(x_k, \varphi\right), \left(x_k', \varphi'\right)\right)\right) \tag{12}$$

with $v_k \sim \mathcal{N}(0, \sigma_v^2 I)$. Here, $\mathfrak{h}(.)$ is a nonlinear generative function governed by a probabilistic graphical model $p\left(z_k | x_k, \varphi\right)$, where z_k depends on both the latent mutation state (x_k) and interaction coefficients (φ). Specifically, $z_k \in \mathbb{R}^N$ is the observed gene expression vector, $x_k \in \mathbb{R}^N$ is the latent gene mutation signal, and $\varphi \in \mathbb{R}^{N \times N}$ is the global interaction matrix with entries modeled as gamma-distributed random variables. The kernel function $\kappa\left(\left(x_k, \varphi\right), \left(x_k', \varphi'\right)\right)$ encodes similarity between gene states and interaction patterns at different times or across cells, enabling a flexible and nonparametric mapping. x_k' and φ' correspond to an alternative latent mutation state and interaction matrix. The noise term $v_k \sim \mathcal{N}(0, \sigma_v^2 I)$ accounts for biological and technical variation. This $\mathcal{GP}$ model supports highly nonlinear, context-dependent relationships while providing uncertainty quantification. To ensure tractability, we use a variational Bayesian approach to approximate the joint posterior $p\left(\boldsymbol{x}_k, \varphi | \boldsymbol{z}_k\right)$ as described in the following section.

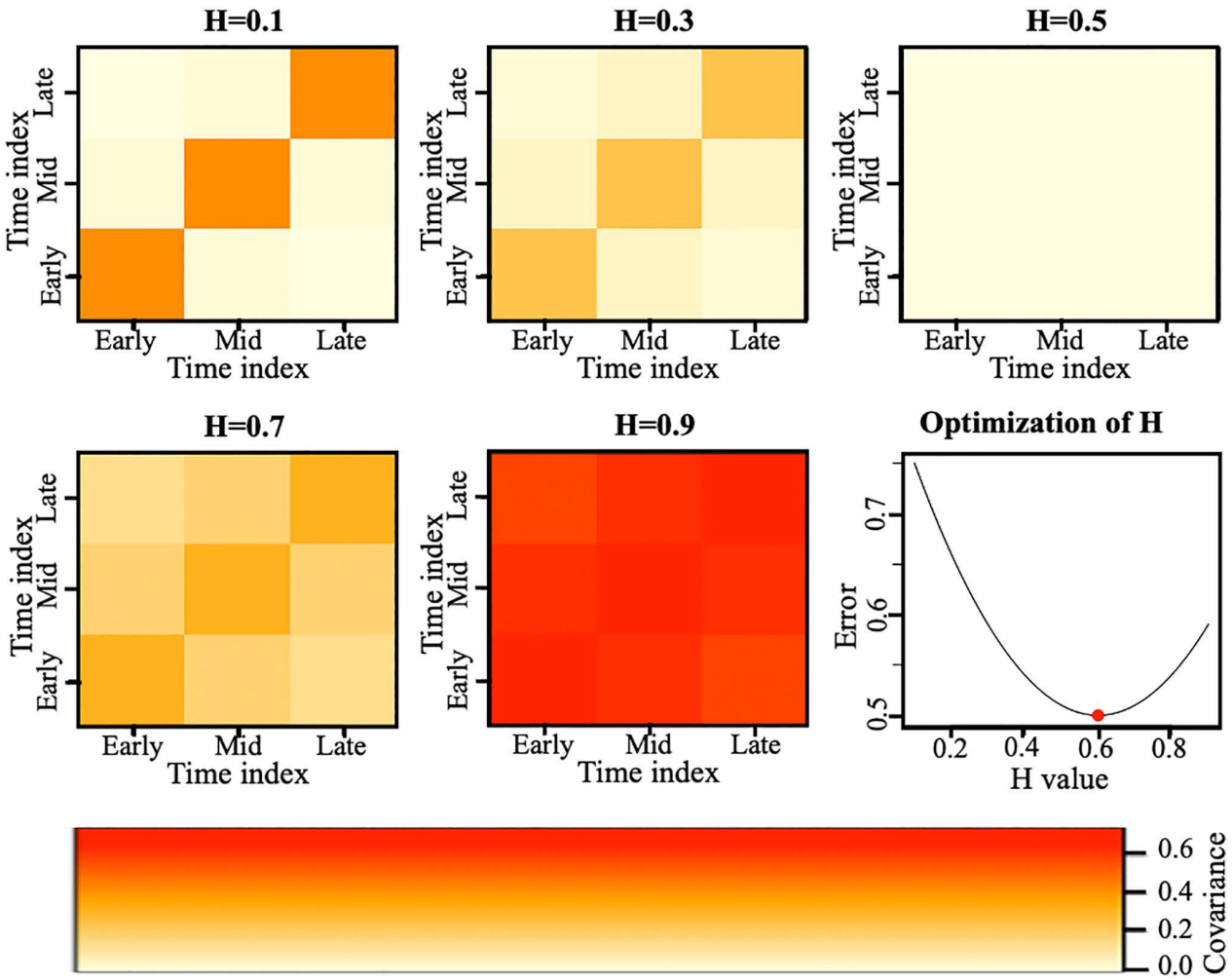

Fig 3. Heatmap visualization of covariance structures across hurst exponents. Heatmaps of covariance matrices for the innovation (error generating) process illustrating the Influence of the Hurst Exponent on long-range dependence over time. For H = 0.1 andH = 0.3, covariance is highly localized along the diagonal, with weak long-range dependence. At H = 0.5, the covariance matrix is more uniform, balancing local and global dependence. As H increases to 0.7 and 0.9, covariance spreads further, indicating stronger long-range dependence. The optimal H is the value that minimizes the error between the estimated and observed covariance matrices, ensuring the best alignment with the observed covariance structure.

https://doi.org/10.1371/journal.pcbi.1014143.g003

Briefly the model includes: (1) observations z_k, time-series gene expression data; (2) global hidden variables φ, capturing gene interaction effects during cancer progression; and (3) local hidden variables x_k, representing gene mutation states (Fig 2C). Specifically, we define $\boldsymbol{z}_k = z_{k=1:3}$ as the set of observed expression vectors, $\boldsymbol{\varphi} = \{\varphi_{1:N}\}$, as the global hidden variables and $\boldsymbol{x}_k = x_{k=1:3}$ as the local hidden variables. Applying Bayes' rule [32], we compute the joint posterior distribution [49] as:

$$p\left(\boldsymbol{x}_k, \varphi | \boldsymbol{z}_k\right) \propto p\left(\boldsymbol{z}_k | \boldsymbol{x}_k, \varphi\right) p\left(\boldsymbol{x}_k, \varphi\right) \tag{13}$$

where $p\left(\boldsymbol{z}_k | \boldsymbol{x}_k, \varphi\right)$ is the likelihood, and $p\left(\boldsymbol{x}_k, \varphi\right)$ is the prior distribution. Due to the computational intractability of the joint posterior, an approximation inference approach is used. This formulation enables us to capture how mutations (x_k)

and gene-gene interactions (φ) jointly shape observed gene expression (z_k), allowing robust inference of context-specific driver gene effects during tumor progression [50].

2.2.4 ***Bayesian inference in time-varying gene mutation for cancer progression.*** The probabilistic formulation in Equations (5) through (13) models latent gene mutation trajectories over the constructed time derived from single-cell data. The observed input z_k corresponds to estimated average gene expression levels for a given gene at time points $k = 1$ (Early), 2 (Mid), and 3 (Late). These observed trajectories are used to infer hidden cancer drivers via variational Bayesian inference.

We use variational Bayesian inference to approximate the joint posterior distribution of both local and global hidden variables, thereby making the joint posterior density function computationally tractable. This process consists of two key steps: variational Bayesian inference and the mean field approximation of the variational free energy [51] (S3 Text), which is closely related to maximizing the Evidence Lower Bound (ELBO). We derive an approximate probability density function that captures gene mutation dynamics during cancer progression, while explicitly incorporating the influence of gene-gene interactions as:

$$q\left(\boldsymbol{x}_k|\varphi\right) = \frac{1}{\mathcal{M}}\mathcal{N}\left(\boldsymbol{x}_k; \mu_{\boldsymbol{x}_k}, \boldsymbol{C}_{\boldsymbol{x}_k}\right) \exp\left[-\sum\nolimits_k \left(z_k - x_k\right)^T \langle\varphi\rangle \left(z_k - x_k\right)\right] \tag{14}$$

with $\langle\varphi\rangle$ denotes the expected value of the global hidden variable φ, and $\mathcal{M}$ is a normalization constant. The covariance $\boldsymbol{C}_{\boldsymbol{x}_k}$ in the approximated distribution $q\left(\boldsymbol{x}_k|\varphi\right)$ captures the gene dynamics and is used to compute the Hurst exponent (Equation 7), which measures long-term memory in the mutation process. Full derivations are provided in S3 Text.

To assess the contribution of gene-gene interactions to PICDGI's performance, we conducted a comparative evaluation between two models as described in S4 Text. The first model served as a baseline and assumed that gene expression trajectories are mutually independent, thereby excluding any interaction structure. The second, interaction-aware model incorporated a structured gene-gene interaction matrix directly into the posterior formulation. Across simulated datasets designed to mirror the sparse temporal resolution of our LUAD application, the interaction-aware model consistently outperformed the baseline. The independence model, constrained by its inability to represent regulatory coupling among genes, exhibited inflated prediction errors and poorly calibrated posterior estimates. In contrast, the interaction-aware formulation accurately recovered underlying expression dynamics, yielding posterior predictions that aligned closely with the simulated ground truth. This improvement was reflected in both a substantial reduction in mean squared error (MSE) and markedly lower negative log-posterior values. These findings demonstrate that the interaction parameters are identifiable under conditions similar to those of our empirical data and that modeling gene dependencies is essential for generating biologically coherent predictions. Because the driver coefficient in PICDGI is derived directly from the inferred interaction effects, this robustness is particularly important: it ensures that the genes prioritized by PICDGI reflect stable and meaningful regulatory influences rather than artifacts of model misspecification.

2.2.5 ***Gene driver coefficient calculation for PICDGI.*** While variational Bayesian inference (VBI) offers computational efficiency for high-dimensional latent variable models, it is known to potentially underestimate posterior uncertainty due to the mean-field independence assumption. To address this limitation and enhance the accuracy of downstream inference, we adopted a hybrid inference strategy.

In the initial phase, VBI was employed to approximate the joint posterior distribution $p\left(\boldsymbol{x}_k, \varphi|\boldsymbol{z}_k\right)$, allowing efficient estimation of gene mutation dynamics and interaction effects across time-series scRNA-seq data. To improve the precision of the driver gene identification, we then applied Markov Chain Monte Carlo (MCMC) sampling, drawing 2000 samples from the posterior distribution $q\left(\boldsymbol{x}_k|\varphi\right)$. These samples were used to estimate the driver coefficient (DrCoef) [52]. The 95% highest density interval (HDI) of the posterior $q\left(\boldsymbol{x}_k|\varphi\right)$ was then computed, providing the range of the most probable true gene effects [53]. This step ensures that credible intervals and posterior variances are accurately quantified, thereby improving the robustness of driver gene identification. By combining VBI for initial scalability with MCMC for final inference precision, our hybrid approach effectively balances computational efficiency and statistical reliability [54]. This makes it

particularly well-suited for modeling complex gene-gene interactions in large-scale single-cell RNA sequencing (scRNA-seq) datasets.

To compute the 95% HDI, we conditioned on the interval δ, which contains the most credible values of DrCoef. The driver coefficient is formally defined as:

$$DrCoef = \left(\frac{E[\boldsymbol{x}_k, \varphi | \boldsymbol{z}_k]}{sd(\boldsymbol{x}_k, \varphi | \boldsymbol{z}_k)} \right)^2 \tag{15}$$

where $\hat{\delta}$ represents the estimated posterior mean of the effect size, and $sd(\delta)$ denotes the standard deviation of the posterior samples. This formulation captures both the magnitude and stability of gene effects, facilitating a more reliable identification of key driver genes in complex biological systems. In S5 Text, we illustrate the construction of the driver coefficient using a toy example with four hypothetical genes. For each gene, PICDGI yields a posterior distribution over its regulatory effect size, approximated here by a normal distribution with mean μ_g and standard deviation σ_g. The supplement S5 Text, S1 Fig shows the posterior densities, with dashed lines indicating the posterior mean and dotted lines indicating zero effect. The S5 Text, S2 Fig in the supplement displays barplots of the corresponding DrCoef values, defined as $(\mu_g/\sigma_g)^2$. Genes with strong, well-constrained effects (large μ_g, small σ_g) obtain high DrCoef values, whereas genes with either small effects or high uncertainty receive lower DrCoef values. This schematic demonstrates how posterior mean and variance jointly determine the ranking of genes in PICDGI.

Having illustrated how PICDGI quantifies and ranks gene-level regulatory influence, we also contrast this framework with existing cancer-driver discovery methods to clarify its distinct modeling assumptions and data requirements. Traditional cancer-driver discovery tools such as *MutSigCV, OncodriveFM, OncodriveCLUST, DriverNet, DawnRank*, and related approaches operate exclusively on bulk sequencing data and rely on mutation recurrence or static network information rather than dynamic, time-resolved single-cell gene expression. Because PICDGI infers regulatory influence from temporal interaction trajectories in scRNA-seq data, these tools are not directly comparable and do not provide a meaningful benchmarking reference. A detailed summary of these methods, their required input data types, and their modeling assumptions is provided in S6 Text.

2.3 Trajectory analysis for identifying key genes in cancer development

Genes with trajectory-dependent expression act as CDGs by influencing progression at different stages. For example, they may promote early proliferation and later facilitate metastasis by altering tumor microenvironment interactions. Such genes regulate pathways like DNA repair and immune evasion in a stage-specific manner. Techniques such as scRNA-seq and time-dependent data analysis reveal their role in cancer transitions [35], supporting both targeted therapy and improved prediction of disease progression [55].

Building on this, we sought to identify genes with trajectory-dependent expression patterns by integrating scRNA-seq datasets across different times from cancer patient. To achieve this, we applied a statistical test commonly used in spatial data analysis [56] to detect genes exhibiting expression variations along developmental time-dependent trajectories throughout cancer progression. Specifically, within Monocle 3, we used the *principalGraphTest()* function, which utilizes Moran's I test [57] to detect differentially expressed genes along trajectories. Moran's I measures spatial autocorrelation by capturing relationships between data points through a nearest neighbor graph [58], making it ideal for large scRNA-seq datasets. The Moran's I statistic is defined as:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{i,j} \left(x_i - \bar{x}\right) \left(x_j - \bar{x}\right)}{\sum_i \left(x_i - \bar{x}\right)^2} \tag{16}$$

where N is the number of cells, x represents gene expression, x_i and x_j are gene expression values for cells i and j, and $\bar{x}$ is the mean across all cells. The weight matrix $w_{i,j}$ is based on a nearest neighbor graph, with diagonal elements set to

zero and off-diagonal elements defined as $w_{i,j} = \frac{1}{\vartheta_i}$, where ϑ_i is the number of nearest neighbors. W is the sum of all $w_{i,j}$ values, ensuring proper autocorrelation normalization.

3 Results

In this study, we analyzed a subset of nine scRNA-seq datasets generated by Kim et al. [33], who profiled 208,506 cells across 44 patients to investigate LUAD progression from normal lung tissue to metastasis. Their study revealed a cancer cell population that deviates from the normal differentiation trajectory and dominates during metastatic stages. To focus on modeling increasing cancer progression across normal, tumor, and metastatic tissues within the same individuals, we selected three patients; P0019, P0006, and P0008 (hereafter referred to as patient 1, patient 2, and patient 3, respectively). Our selection criteria prioritized patients who had matched samples available from all three tissue stages (normal lung, primary tumor, and metastatic brain), enabling reconstruction of complete progression trajectories. The datasets for these patients included: for normal lung tissue, 42,996 cells (patient 1), 3,871 cells (patient 2), and 3,381cells (patient 3). For primary lung tumor samples, there were 45,150 cells (patient 1), 4,362 cells (patient 2), and 3,766 cells (patient 3). For metastatic brain tissue, the datasets included 29,061 cells (patient 1), 3,301 cells (patient 2), and 5,731cells (patient 3). Across all nine samples, approximately 29,634 genes were profiled at high sequencing depth. Details about sample preparation and sequencing protocols are available in S7 Text. By focusing on patients with complete progression trajectories, our analysis captures stage-specific gene expression dynamics critical for understanding LUAD evolution from early tumorigenesis to brain metastasis, thereby extending the findings of Kim et al [33].

The number of cells captured in the LUAD single-cell datasets varied substantially across the three patients, with Patient 1 contributing markedly higher cell counts at all tissue stages compared with Patients 2 and 3. This imbalance reflects technical variability inherent to single-cell sequencing workflows, including differences in tissue dissociation efficiency, viability of recovered cells, microfluidic capture rates, and sequencing depth, rather than any biological disparity among patients. Importantly, PICDGI analyzes each patient independently and operates on cluster-level mean expression profiles rather than raw cell frequencies. As a result, differences in total cell numbers do not influence the inferred temporal gene-expression trajectories or the resulting driver-gene estimates. We further verified that the inferred driver coefficients are stable under down-sampling, confirming that the variation in cell counts does not bias the model's performance or the interpretation of progression dynamics.

3.1 Identification of epithelial cells as cancer progenitors for PICDGI LUAD analysis

Cell clustering, annotation, and identification of epithelial cells were performed using Seurat R package. PICDGI subsequently uses these epithelial temporal expression profiles as the progenitor population for dynamic modeling of cancer progression.

We analyzed scRNA-seq data from three LUAD patients, each sampled at three distinct stages of cancer progression: Early (normal lung tissue), Mid (primary lung tumor), and Late (metastatic brain tissue), resulting in nine single-cell datasets in total. Following quality control and filtering procedures, we retained 42,996 cells for the Early stage, 45,150 cells from the Mid stage, and 29,061 cells from the Late stage for downstream analysis. Using unsupervised clustering and marker-based cell type annotation with Seurat R package [59], we identified key immune and non-immune cell types, including dendritic cells (DC), mast cells, T cells, B cells, NK cells, fibroblasts, endothelial cells, ependymal cells, oligodendrocytes, and epithelial cells (Fig 4A, S7 Text, S1–S9 Figs).

To determine candidate cancer progenitor cell types, we tracked changes in cell type abundance and calculated the cancer cell fraction (CCF) across the three stages. Epithelial cells showed consistent expansion in relative proportion from Early, Mid, and Late stages, satisfying the originating criteria defined in Equations (2)-(4). CCF was assessed using *EpCAM (CD326),* a widely used epithelial tumor-associated marker marker [60,61]. Epithelial cells exhibited the highest CCF values, which increased with disease progression (Fig 4B), supporting their role as LUAD progenitor cells likely harboring driver mutations.

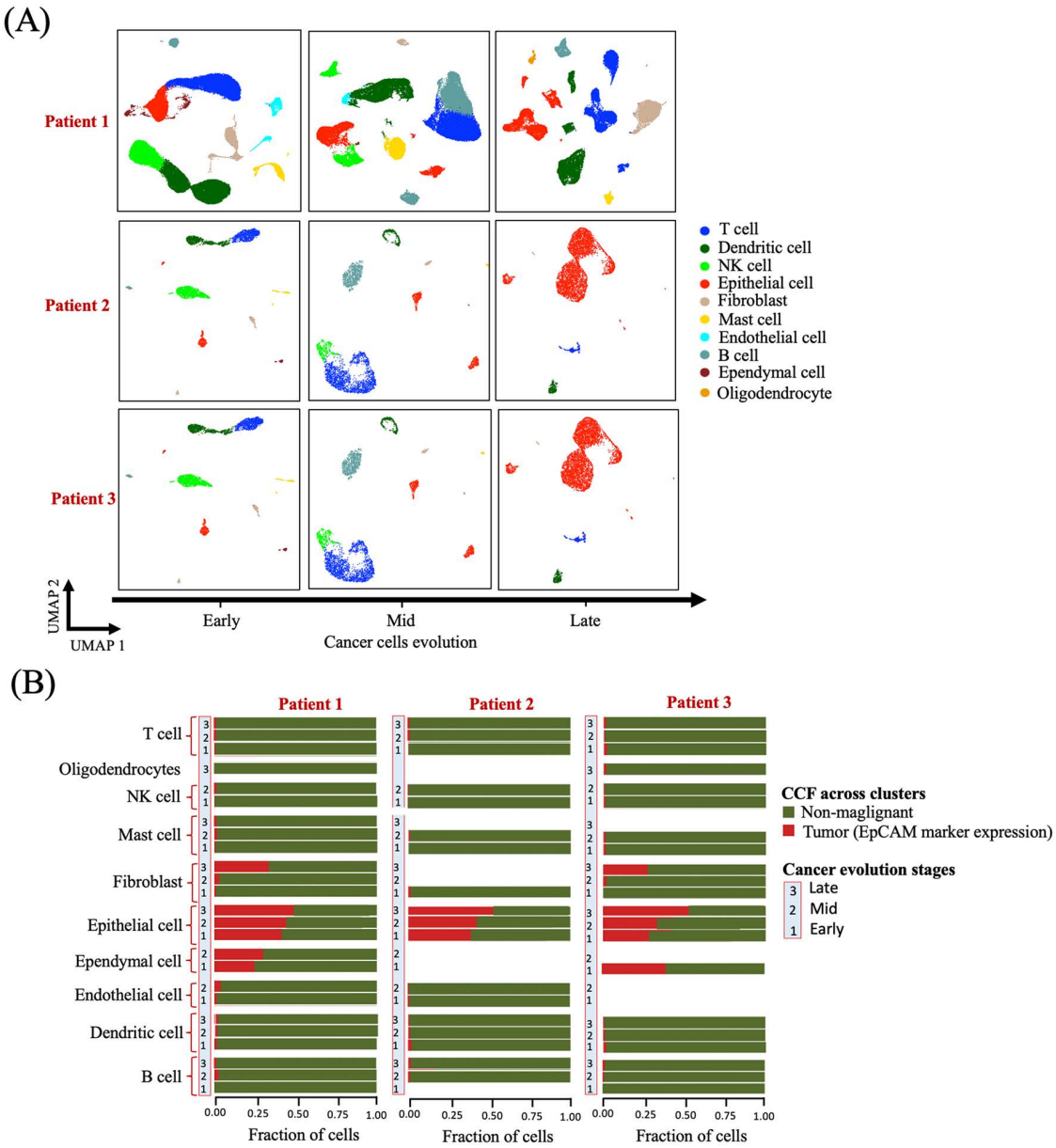

Fig 4. Overview of single cells from the lung tissues of three patients. (A) t-SNE plots showing profiles of single cells from each tissue origin for three patients. In the first row (patient 1), 42,996, 45,150, and 29,061 cells are shown, respectively. In the second row (patient 2), 3,871, 4,362, and 3,301 cells are shown, respectively. In the third row (patient 3), 3,381, 3,766, and 5,731 cells are shown, respectively. Plots are color-coded by major cell lineages and gene expression counts. **(B)** Fractions of cells originating from tumor versus non-malignant lung tissues across cell types. Tumor-origin cell fractions vary by cell type and LUAD stage across patients, with epithelial cells consistently exhibiting the highest tumor fractions, increasing with LUAD progression.

https://doi.org/10.1371/journal.pcbi.1014143.g004

To ensure that epithelial lineage markers were accurately represented in our dataset, we performed an additional ambient RNA-correction step using SoupX prior during progenitor-cell identification [62]. Ambient RNA contamination is common in droplet-based scRNA-seq of solid tumors and can result in misleading detection of epithelial transcripts in non-epithelial lineages. After correction, EpCAM expression was restricted to epithelial clusters, confirming that its presence in

immune populations originated from background contamination rather than true biological signal. We further verified that epithelial cells uniquely exhibited a consistent increase in abundance across cancer progression and retained the highest *CCF* values. Although *CCF* was computed for all annotated cell types, only epithelial-derived *CCF* values were used in subsequent analyses. Full details of this validation is provided in S8 Text.

3.2 PICDGI predicts gene expression levels of cells considering gene-gene interactions

Genes encode proteins that regulate essential functions like cell growth [63], and mutations can disrupt protein function and drive cancerous transformations [64]. Specific gene mutations can alter proteins in ways that promote tumorigenesis, making it crucial to understand not only gene expression levels but also the interactions between genes that influence these levels. Thus, modeling gene expression dynamics from scRNA-seq data is essential for identifying cancer driver genes (CDGs) whose effects are mediated through gene-gene interactions.

To achieve this, we applied the PICDGI framework to model gene expression dynamics in individual epithelial cells, driving LUAD progression across the three patients. To evaluate performance, we calculated Pearson's correlation coefficient (ρ) and the coefficient of determination (R^2) between observed and predicted gene expression (TGE and PGE), followed by statistical significance testing. Results showed positive Pearson correlation coefficients (ρ) with p-values < 0.05, indicating a strong linear relationship between TGE and PGE (Fig 5A-5C). Correlations for each patient at different LUAD stages are shown in Table 1. Fig 5D displays the distribution of ρ values across genes, further validating prediction consistency. The focus on epithelial cells is critical for identifying cancer drivers, as PICDGI relies on predicted gene expression patterns in these cells, where gene interactions influence disease progression.

3.3 PICDGI identifies cancer driver genes based on the influence of gene-gene interactions

To enhance PICDGI's accuracy in identifying cancer driver genes (CDGs), we derived predicted gene expression from the latent hidden variables (gene mutation states and gene-gene interaction matrix). This approach filters noise and captures the underlying gene expression patterns, even in noisy or low-quality scRNA-seq measurements. Using Bayesian analysis, we quantified uncertainty through the 95% HDI of the posterior distribution, reflecting the conditional effect of each gene while accounting for gene interactions driving mutations. These predicted expression values enable the reliable identification of CDGs and help distinguish observed driver mutations from passengers, which have minimal impact on cancer progression.

We ranked the top 30 genes with the highest coefficients for the three patients (Fig 6A, S7 Text,S10A–S11A Figs), finding that 63.33%, 63.33%, and 60% were previously identified as CDGs (Markers, OGs, or TSGs) (S9 Text). The remaining uncharacterized genes may represent potential CDGs for further validation. In particular, genes *TP53INP1, CA12,* and *LCNL1* were predicted as key drivers for cancer progression in patients 1, 2, and 3, respectively. *TP53INP1*, a tumor suppressor gene, is downregulated in cancers and collaborates with *p53* to regulate cell death and migration [65]. *CA12*, involved in pH regulation, is overexpressed in cancers and may be a novel prognostic marker [50,66]. *LCNL1* affects lung cancer susceptibility, particularly in never-smokers, highlighting its potential role in cancer risk [67].

To examine more closely the biological relevance of the driver genes prioritized by PICDGI, we performed cross-patient pathway enrichment analysis on the union of predicted drivers and their inferred modulatory partners. This analysis revealed that PICDGI-identified genes converge on pathways central to LUAD progression, including cell-cycle regulation, metabolic rewiring, autophagy, and microenvironmental remodeling, with distinct yet coherent patterns observed across all three patients. These findings provide an additional layer of validation by demonstrating that the inferred drivers map to biological programs known to intensify during tumor evolution and metastatic expansion. A detailed presentation of these pathway enrichment results, including cancer-focused and Gene Ontology analyses for each patient, is provided in S10 Text.

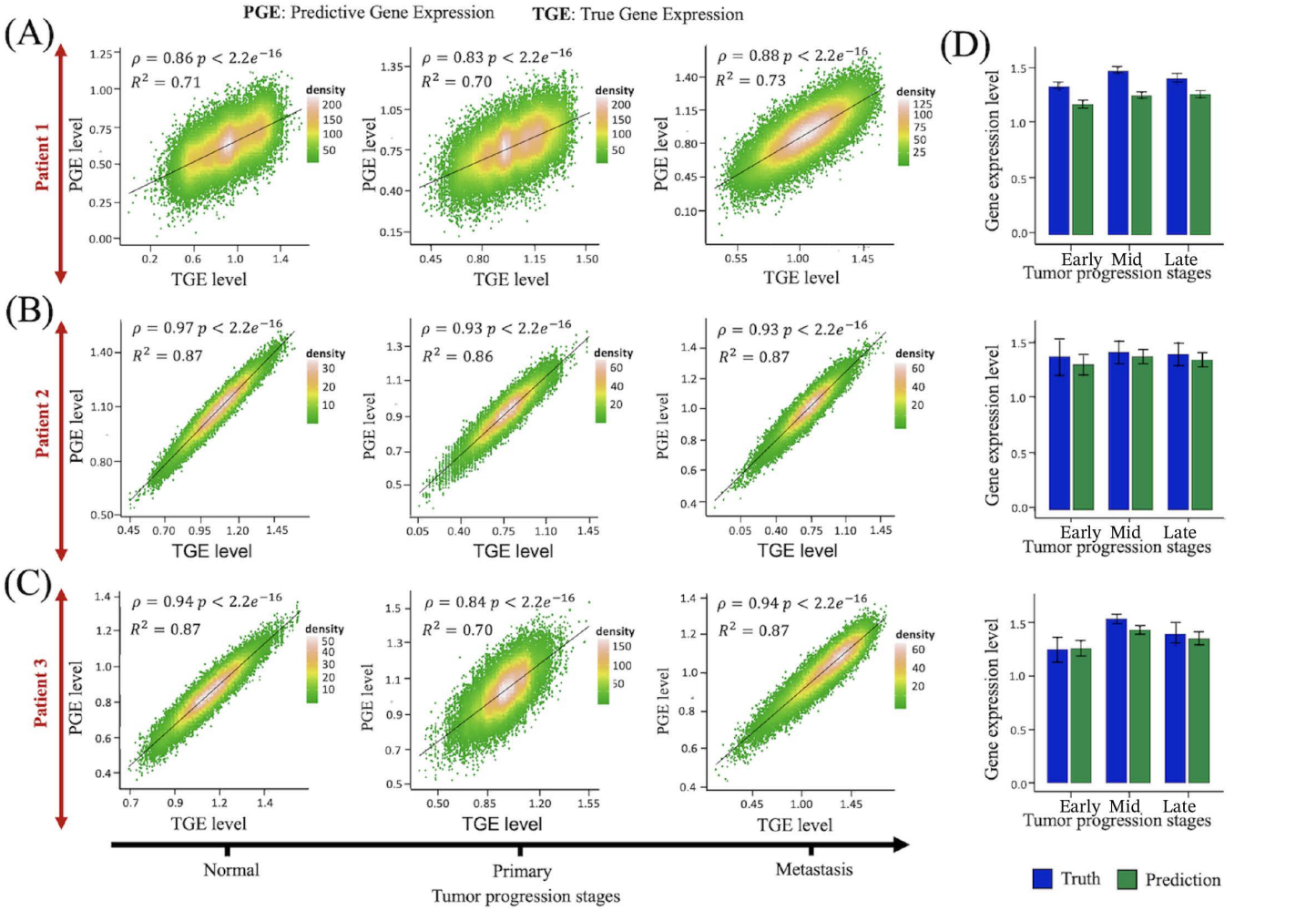

Fig 5. Predicted vs. observed gene expression levels in epithelial cells. (A-C) Scatterplots illustrating the performance of the PICDGI framework in predicting epithelial cell gene expression across the Early, Mid, and Late stages of LUAD progression for Patients 1, 2, and 3, respectively. Each plot shows the relationship between true gene expression (TGE) and predicted gene expression (PGE), with Pearson's correlation coefficient (ρ), coefficient of determination (R^2), and corresponding p-value computed using a two-sided t-test. **(D)** Summary of predictive accuracy across stages. Barplots display the mean Pearson correlation coefficients (ρ) ± SEM (Standard Error of the Mean) for the comparison between TGE and PGE at each of the three time points; Early, Mid, Late for each patient. These summary statistics complement the scatterplots by providing an aggregated view of model performance across genes. From top to bottom, the panels correspond to Patient 1, Patient 2, and Patient 3.

https://doi.org/10.1371/journal.pcbi.1014143.g005

Table 1. Pearson's correlation coefficient (ρ) and coefficient of determination (R^2).

Stage	Patient 1		Patient 2		Patient 3	
	ρ	R^2	ρ	R^2	ρ	R^2
Early	0.80	0.71	0.97	0.87	0.94	0.87
Mid	0.76	0.69	0.93	0.86	0.74	0.62
Late	0.85	0.73	0.93	0.87	0.94	0.87

https://doi.org/10.1371/journal.pcbi.1014143.t001

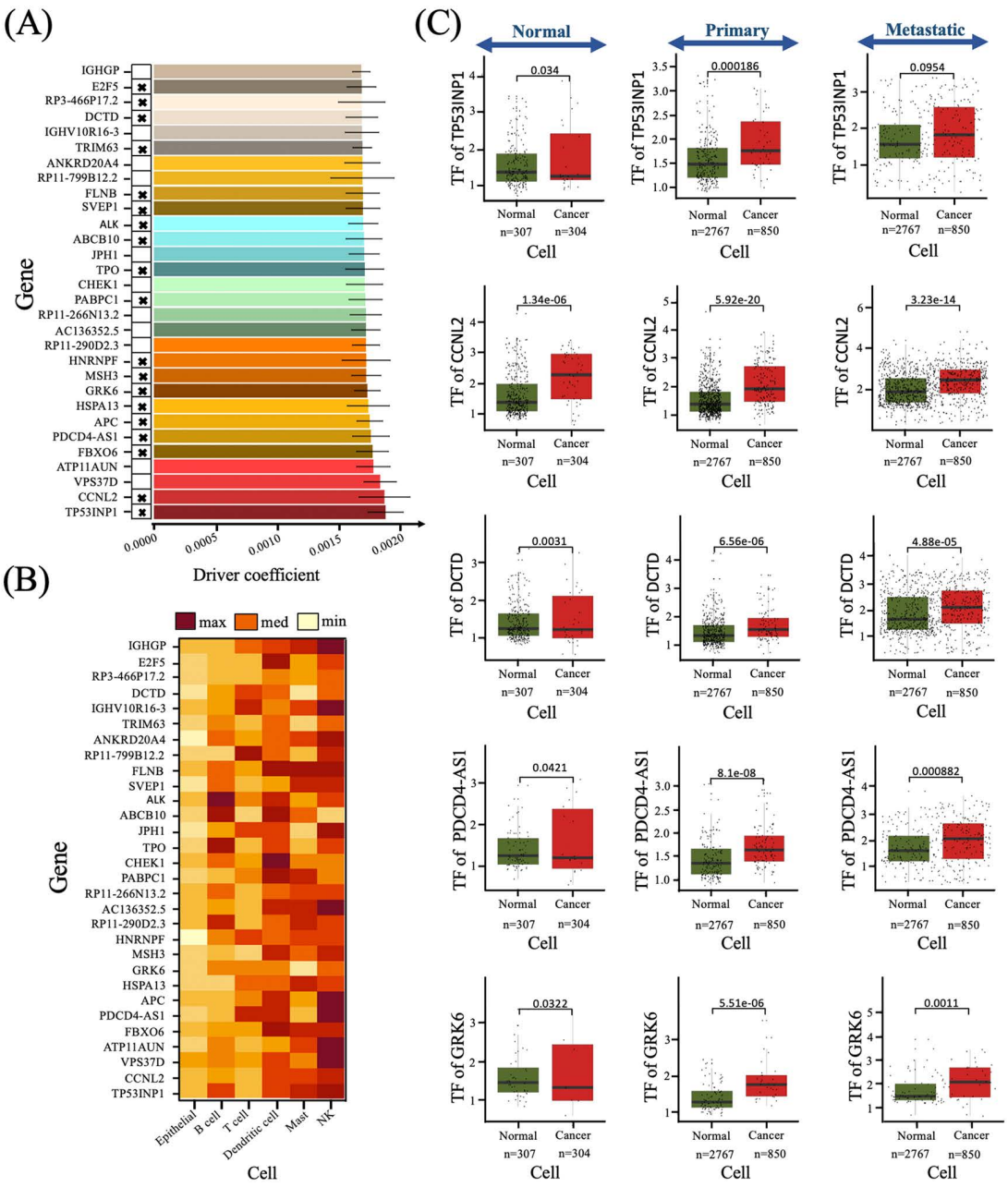

Fig 6. Cancer driver genes with the highest driver coefficient. (A) Barplot showing the driver coefficients of epithelial cell genes derived from patient 1 gene expression data using the PICDGI framework. Data are presented as mean +/- SEM (Standard Error of the Mean). Black cross marks indicate genes previously identified as oncogenes (OGs) or tumor suppressor genes (TSGs). **(B)** Heatmap showing PICDGI-derived DrCoef values for the top 30 epithelial driver genes (selected based on panel A) recalculated independently within each annotated immune cell type from patient 1 single-cell data. DrCoef values in this panel are computed using cell-type specific models, enabling assessment of the regulatory influence of epithelial-identified driver genes across immune compartment. **(C)** Boxplots comparing transcription factor (TF) expression and TF activity between normal epithelial and cancer cells for two representative TFs showing discordance between differential activity and differential expression. P-values for differential TF activity and expression were calculated using a t-test and Wilcoxon rank-sum test, respectively. Boxplot elements indicate the median (horizontal line), interquartile range (box), and whiskers extending to 1.5 × interquartile range.

https://doi.org/10.1371/journal.pcbi.1014143.g006

3.4 PICDGI reveals that top epithelial cancer drivers exhibit strong immunoregulatory influence

The immune system is crucial for detecting and eliminating abnormal cells, including cancer cells. However, tumor progression is frequently associated with the establishment of an immunosuppressive microenvironment that enables immune evasion. In lung cancer, multiple regulatory genes have been implicated in modulating immune signaling and suppressing anti-tumor responses [68,69]. To explore the potential immunomodulatory roles of PICDGI-identified drivers, we examined the 30 highest-ranked epithelial driver genes in each of the three patients (Fig 6B, S7 Text, S10B–S11B Figs). Although these genes were selected based on their high DrCoef values in epithelial cells, we observed that many of them exhibited equal or even higher DrCoef values in immune cell populations, particularly in NK cells.

Importantly, elevated DrCoef in NK cells does not indicate epithelial identity within immune cells. Rather, it reflects strong dynamic regulatory influence within the NK-cell transcriptional network across disease stages. The consistently high DrCoef values observed in NK cells suggest substantial regulatory rewiring in this compartment during tumor progression. Given the known role of NK cells in anti-tumor immunity, this pattern is consistent with tumor-driven modulation of NK-cell function and may reflect mechanisms of immune evasion active in advanced-stage LUAD (Fig 4A).

To ensure methodological consistency, DrCoef values for immune cells were computed using the identical PICDGI pipeline applied to epithelial cells. After identifying the top 30 epithelial candidate driver genes, we evaluated their dynamic activity across major immune populations, including T cells, B cells, dendritic, mast, and NK cells. Although immune cells are not tumor-initiating populations, they are critical components of the tumor microenvironment and engage in continuous crosstalk with cancer cells. Several of the top-ranked genes are known to participate in immune signaling or stress-response pathways, supporting a dual role in tumor progression and immune modulation.

Among the top-ranked genes in each patient, we further assessed transcription factor (TF) activity for five representative genes in both normal and cancer epithelial cells. These included *TP53INP1, CCNL2, DCTD, PDCD4-AS1*, and *GRK6* (patient 1); *ALG1, C9orf16, GPX1, CA12*, and *LINC01620* (patient 2); and *NRBF2, C9orf16, SFR1, CA12*, and *PITPNC1* (patient 3). Differential TF activity between normal and tumor epithelial cells indicates regulatory reprogramming during tumor progression (Fig 6C, S7 Text, S10C–S11C Figs), potentially affecting transcriptional control, downstream target engagement, and cellular proliferation dynamics.

The variability in top-ranked transcriptional regulators across patients likely reflects both biological heterogeneity and technical variability. Biologically, lung cancer exhibits substantial inter-patient heterogeneity in mutational landscape, tumor subtype, and microenvironmental context [70,71]. Technical factors inherent to scRNA-seq, including sampling bias, dropout effects, and batch variation, may also contribute [72]. However, given the reproducible recovery of coherent regulatory programs within each patient and in external validation datasets, biological heterogeneity is likely the dominant contributor to the observed differences [73].

3.5 Comparison of PICDGI and Monocle 3 in identifying cancer progression genes

We compared the genes prioritized by PICDGI with those identified by Moran's I test implemented in Monocle 3, a tool shown by Cao & Spielmann et al. [58], to identify variable genes in scRNA-seq data. While Moran's I assesses spatial autocorrelation of gene expression across temporal trajectories [74], PICDGI incorporates dynamic modeling of nonstationary gene interactions to prioritize putative driver genes. Comparing the two approaches allows benchmarking PICDGI against a well-established method for identifying biologically variable genes in scRNA-seq data.

For this comparison, we integrated scRNA-seq data across the three time points per patient, creating a progression LUAD single-cell landscape. Using PCA for Slingshot and UMAP for Monocle3 [18], we inferred cell-type trajectories and performed temporal analysis for clustering and visualization [75] (Fig 7A).

The analysis revealed that the genes identified by the PICDGI framework exhibited higher average expression levels than those identified by Monocle 3 across the three patients thus positioning them as strong candidates for true cancer drivers [35,36] (Fig 7, S7 Text, S12 - S13 Figs). This difference stems from Monocle 3's non-spatial trajectory inference,

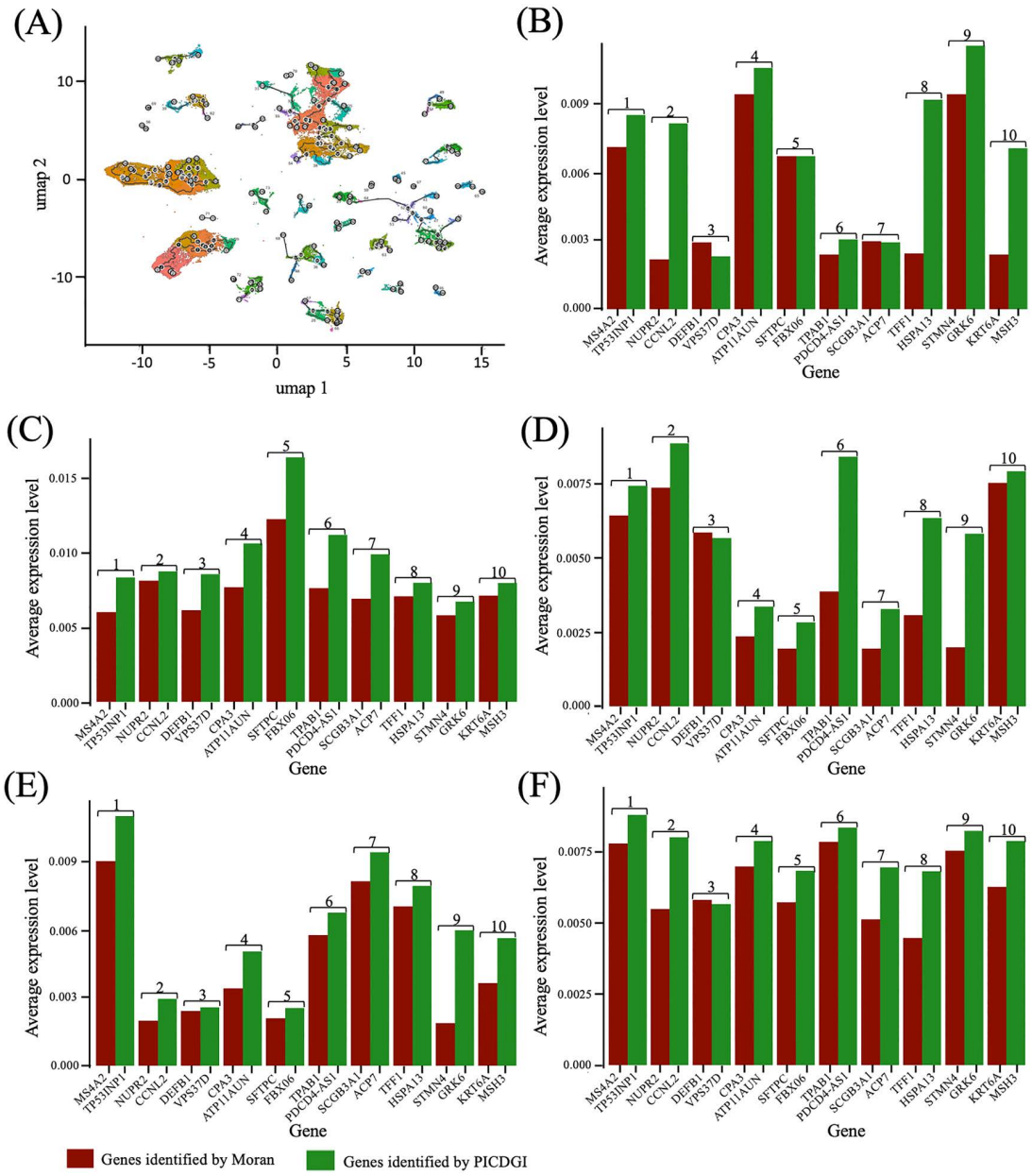

Fig 7. Comparison of the PICDG framework with the existing Moran's I test algorithm for predicting driver genes' inference in immune cells. The driver genes identified through Moran's I test display a lower average expression level compared to the expression level of driver genes presented by the PICDGI computational framework. The genes are ranked from the highest to the lowest immune-suppressive role (1 to 10): **(A)** Single-cell atlas map the trajectory and time values of cells progression; **(B)** Mast cell; **(C)** Natural Killer; **(D)** T cell; **(E)** B cell; **(F)** Dendritic cell.

https://doi.org/10.1371/journal.pcbi.1014143.g007

which ignores gene-gene interactions [76]. While Monocle3 excels at capturing dynamic expression patterns, PICDGI emphasizes gene interactions during progression, explaining the variation in results.

Although the overlap between the top-ranked genes from PICDGI and Monocle3 is limited, this divergence is expected due to the distinct methodological assumptions of each approach. Monocle3 primarily identifies genes whose expression changes significantly over pseudotime, independent of their regulatory influence on other genes. In contrast, PICDGI

ranks genes based on their conditional impact on the expression of other genes, modeled through a time-dependent interaction framework. This leads to the identification of genes with high regulatory importance, even if their individual expression dynamics are not prominent.

From a biological perspective, this difference underscores the complexity of cancer progression. Monocle3 captures responsive genes whose expression reflects temporal transitions or lineage states, while PICDGI is optimized to uncover putative causal regulators; genes that may drive these transitions through network-level influence. For instance, genes with relatively stable expression but central roles in oncogenic signaling may be highlighted by PICDGI but overlooked by Monocle3. In our study, PICDGI successfully identified TP53INP1, CA12, and LCNL1 as high-confidence cancer driver genes for patients 1, 2, and 3 respectively; each exhibiting the highest cancer driver coefficients within their individual profile. These genes are not merely transiently responsive but act as key regulators in tumor development. This demonstrates that, unlike Monocle3, our approach captures essential yet stable drivers of oncogenesis, offering more robust insights into patient-specific cancer mechanisms and paving the way for personalized therapeutic strategies. Therefore, the minimal overlap reflects complementary strengths of the methods in dissecting tumor progression from different angles.

3.6 External validation of PICDGI in an independent pediatric AML cohort

To assess whether PICDGI generalizes beyond lung adenocarcinoma and is not specific to a single tumor type or dataset, we next evaluated the framework in an independent pediatric acute myeloid leukemia (AML) single-cell RNA-seq cohort from Mumme *et al*.[77], which profiles bone marrow samples at Diagnosis (Dx), End-of-Induction chemotherapy (EOI), and Relapse. This dataset provides a stringent out-of-sample test because it spans longitudinal disease stages, resolves both malignant blasts and microenvironmental compartments, and includes curated AML blast signatures and relapse-associated biological programs.

Using a unified Seurat-based single-cell analysis pipeline and a comprehensive AML blast-centric annotation strategy, we resolved a continuum of leukemic myeloid states encompassing leukemia stem cell (LSC)-like cells, early myeloid progenitors, myeloblasts, cycling myeloid/monocytic/granulocytic populations, mature monocytic and granulocytic lineages, and inflammatory CXCL8^{+} states. These populations were defined based on established AML-associated genes including granule and protease markers (*MPO, ELANE, AZU1, CTSG, CTSD, PRTN3, LYZ*), inflammatory markers (*S100A8*/*A9*), transcriptional and developmental regulators (*RUNX1, HOXA9*), and stem/progenitor markers (*CD34, KIT*) [78–80]. In parallel, non-malignant immune populations (T, NK, immature B cells) and stromal compartments were cleanly separated using established lineage markers consistent with prior single-cell studies of hematopoitic differenciation and leukemia [81,82]. This standardized annotation enabled robust estimation of malignant cell fractions across cell types and stages.

Across Dx, EOI, and Relapse, myeloid leukemic populations consistently exhibited the highest malignant fractions, with early involvement at diagnosis, persistence following induction chemotherapy, and marked enrichment at relapse (S11 Text, S1 Fig A-C). In contrast, lymphoid and stromal populations remained largely non-malignant or showed only transient malignant signal, consistent with their roles as reactive or bystander compartments rather than leukemia-initiating lineages [80,83]. These longitudinal dynamics satisfy PICDGI's criteria for inferring a leukemia-originating lineage, namely: early presence, therapy resistance, and relapse enrichment, thereby identifying LSC-like, progenitor, and cycling myeloid populations as the dominant cancer-originating compartment in pediatric AML.

Leveraging these cell-state-resolved malignancy trajectories, PICDGI prioritizes candidate cancer driver genes by integrating longitudinal changes in malignant cell fractions with disease-transition aware driver coefficients (DrCoef). Across both Dx-EOI and Dx-Relapse transitions, the framework robustly recovered known pediatric AML-relevant drivers and facilitators (S11 Text, S2 Fig A-B), most notably *PRDM1* (*BLIMP1*), a tumor suppressor implicated in hematologic malignancies and leukemic differentiation control. In addition, PICDGI consistently prioritized genes involved in cellular stress responses and metabolic adaptation including *STIP1, GLO1, ISG15*, and *CD164* which have been linked to leukemic cell

survival, oxidative stress tolerance, immune signaling, and bone-marrow niche interactions, particularly in therapy-resistant AML contexts [84–86].

Beyond these established AML-associated genes, PICDGI nominated coherent sets of candidate drivers involved in RNA processing and splicing, protein homeostasis and proteasomal regulation, mitochondrial and metabolic pathways, and immune modulation. Notably, many of these candidates were consistently upregulated across malignant myeloid populations and disease stages, suggesting roles in leukemic maintenance, adaptation to cytotoxic therapy, or relapse-specific fitness rather than transient or lineage-restricted expression. Together, these results indicate that PICDGI captures both canonical AML driver biology and biologically plausible novel candidates, supporting its utility for systematic driver gene discovery in pediatric AML.

Overall, this external validation demonstrates that PICDGI generalizes across independent pediatric AML cohorts, robustly linking single-cell malignant cell dynamics to biologically meaningful driver gene prioritization. Full details of this external validation, including cell-type annotation, malignancy scoring, and cross-dataset driver inference, are provided in S11 Text.

4 Discussion

In this study, we introduced PICDGI, a variational Bayesian machine-learning framework for identifying cancer driver genes (CDGs) by modeling the dynamic impact of gene-gene interactions on cellular states during cancer progression. Leveraging single-cell RNA-seq data, PICDGI effectively distinguishes regulatory dynamics among malignant cells, immune cells co-opted by tumors, and their non-malignant counterparts. The model's predictions align with prior observations from in vitro, bulk and animal studies and novel CDGs predicted by PICDGI share strong functional similarities with known oncogenes (OGs) and tumor suppressor genes (TSGs), underscoring their biological plausibility and therapeutic relevance. Notably, several top-ranked novel CDGs also exhibited immunoregultory functions, highlighting their potential role in immune evasion and therapy resistance. By linking cellular origins of cancer to regulatory influence, PICDGI provides a framework for predicting drug response at subclonal resolution using tumor scRNA-seq profiles.

Unlike previous CDG discovery methods that rely on mutation recurrence or predefined gene sets, PICDGI explicitly models time varying gene-gene interactions, enabling the identification of rare, context-specific regulatory drivers. Compared to Monocle 3, which detects differentially expressed genes along pseudotime trajectories [76], without modeling gene dependencies [87], PICDGI ranks genes by their conditional influence on others over time. The limited overlap between the two methods reflects these fundamental differences. Notably, the CDGs identified by PICDGI are frequently supported by prior cancer literature [35,88] and exhibited consistently strong, progression-aligned expression dynamics. This reinforces the robustness of the framework in capturing tumorigenic pathways [89], making it a valuable tool for advancing cancer research and precision oncology [90].

PICDGI also reveals that top epithelial driver genes exhibit strong immunoregulatory influence (Fig 6B, S7 Text, S10B–S11B Figs), providing insight into mechanisms of immune evasion. Detecting these immunoregulatory CDGs enables the development of targeted therapies, to restore anti-tumor immune responses [91] and informs combination therapies to counteract immune evasion [92]. Furthermore, these genes may serve as predictive biomarkers for immunotherapy response [93], facilitating patient-specific treatment strategies [94].

Across LUAD patients, PICDGI identified approximately 30 top-ranked CDGs per patient, of which 38% were previously unreported (underlined genes in S9 Text). Several of these drivers merit particular attention. For example, JPH1 (junctophilin-1), a key structural protein forming junctional complexes between the plasma membrane and the sarco/endoplasmic reticulum [95], was predicted by PICDGI as a CDG in patient 1. Previously hypothesized to be a disease-modifier gene in individuals with Charcot-Marie-Tooth disease type 2K (*CMT2K*) [96], JPH1 was predicted by PICDGI as a CDG in patient 1. Similarly, *CHEK1*, originally identified as a regulator of the G2/M checkpoint and DNA repair [97], was identified as a CDG and may serve as a prognostic biomarker for LUAD [98].

Importantly, PICDGI generalized beyond LUAD. In an independent pediatric acute myeloid leukemia (AML) scRNA-seq cohort, by using PICDGI on the same hyperparameters, we recovered known relapse-associated regulators and prioritized biologically coherent programs linked to leukemic persistence, immune modulation, and metabolic adaptation. The framework identified malignant myeloid populations as the leukemia-originating compartment and highlighted context-dependent facilitators such as *PRDM1, STIP1, GLO1, ISG15*, and *CD164*, all of which have established relevance to hematologic malignancy biology and therapy resistance. This external validation demonstrates that PICDGI captures conserved principles of cancer progression across distinct tumor types and disease contexts.

Despite these advances, limitations remain. Not all driver genes act independently. Many function within coordinated modules or pathways that collectively drive oncogenesis [24]. Currently, PICDGI does not fully capture pathway-level contributions or cell-cell communication networks. Future work will extend PICDGI to model cohesive gene networks with high interaction densities and incorporate cell-cell communication [99], better reflecting multicellular dynamics underlying cancer evolution.

In summary, PICDGI combines variational Bayesian inference with dynamic modeling of gene-gene interaction to identify functional, regulatory CDGs from scRNA-seq data. By revealing patient-specific driver profiles and generalizing across both solid and hematologic cancers, PICDGI advances precision oncology and provides a principled foundation for studying tumor progression, immune evasion, and therapy resistance at single-cell resolution.

Supporting information

S1 Text. Parameters of ARMA model.
(DOCX)

S2 Text. Optimization of the hurst parameter H in covariance computation.
(DOCX)

S3 Text. Variational bayesian inference.
(DOCX)

S4 Text. Incorporating gene-gene interactions to enhance cancer driver gene prediction.
(DOCX)

S1 Fig. Comparison of predicted versus observed gene expression levels. We compare predicted versus observed gene expression levels using a baseline model that assumes gene independence and an interaction-aware model that incorporates gene-gene interactions.
(TIFF)

S2 Fig. Evaluation of prediction accuracy using two metrics. We evaluate prediction accuracy using two metrics: Mean Squared Error (MSE), which quantifies the average squared difference between predicted and observed gene expression levels, and Negative Log Posterior (NLP), which reflects how well the model explains the observed data under the posterior distribution. Lower values in both metrics indicate improved model performance.
(TIFF)

S5 Text. Conceptual illustration of the driver coefficient (DrCoef).
(DOCX)

S1 Table. Posterior mean, posterior variability, and effect interpretation for representative genes.
(DOCX)

S2 Table. Ranking of gene-level driver coefficients and their biological interpretation.
(DOCX)

S1 Fig. Posterior distribution panels of gene-specific regulatory effects (Toy example). We illustrate posterior distributions of gene-specific regulatory effects using four toy genes. For each gene, we model the effect parameter β_g as a normal distribution with mean μ_g and standard deviationσ_g. We show how the posterior mean and uncertainty (reflected by curve width) influence the resulting driver coefficient. This example demonstrates how we use both effect magnitude and certainty to rank genes within the PICDGI framework.
(TIFF)

S2 Fig. Ranking of genes by driver coefficient (DrCoef). We illustrate how DrCoef ranks the four toy genes by combining effect size and uncertainty. We show that G1 receives the highest DrCoef because it has both a strong effect and low variance, while G3 ranks above G2 due to its more precise estimate despite a smaller effect. G4, with no effect, appropriately receives a DrCoef of zero. This example demonstrates how we use DrCoef to prioritize genes based on both magnitude and confidence of their inferred effects.
(TIFF)

S6 Text. Comparing PICDGI features against existing cancer driver-discovery and dynamic network-inference frameworks.
(DOCX)

S1 Table. Comparative summary of PICDGI and Existing cancer driver-discovery and network-inference method families.
(DOCX)

S7 Text. Single-cell RNA-Seq data acquisition, preprocessing and analysis across cancer progression stages for individual patients.
(DOCX)

S1 Fig. Single-cell RNA-seq data visualization for patient1 at early stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 1, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S2 Fig. Single-cell RNA-Seq data visualization for patient1 for patient1 at mid stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 1, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S3 Fig. Single-cell RNA-seq data visualization for patient1 at late stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 1, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S4 Fig. Single-Cell RNA-seq data visualization for patient2 at early stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 2, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S5 Fig. Single-Cell RNA-seq data visualization for patient2 at mid stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 2, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S6 Fig. Single-Cell RNA-seq data visualization for patient2 at late stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 2, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S7 Fig. Single-Cell RNA-seq data visualization for patient3 at early stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 3, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S8 Fig. Single-Cell RNA-seq cell type visualization for patient3 at mid stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 3, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S9 Fig. Single-Cell RNA-seq data visualization for patient3 at late stage. We show tSNE plots of marker gene expression across major cell lineages in early-stage Patient 3, highlighting lineage-specific markers for related immune and non-immune cells.
(TIFF)

S10 Fig. Cancer driver genes with the highest driver coefficients for Patient 2. We illustrate epithelial cell genes with the highest driver coefficients for Patient 2, including a barplot highlighting known oncogenes and tumor suppressor genes, a heatmap showing gene driver inference across immune cell types from Patient 2's single-cell data, and boxplots comparing transcription factor activity and expression between normal and cancer epithelial cells, revealing significant discordance in some cases supported by statistical tests.
(TIFF)

S11 Fig. Cancer driver genes with the highest driver coefficients for Patient 3. We illustrate epithelial cell genes with the highest driver coefficients for Patient 3, including a barplot highlighting known oncogenes and tumor suppressor genes, a heatmap showing gene driver inference across immune cell types from Patient 3's single-cell data, and boxplots comparing transcription factor activity and expression between normal and cancer epithelial cells, revealing significant discordance in some cases supported by statistical tests.
(TIFF)

S12 Fig. Comparison of PICDGI and Moran's I test for driver gene prediction in immune cells for patient 2. We compare the PICDGI framework with Moran's I test for predicting driver genes in immune cells of Patient 2, finding that driver genes identified by Moran's I have lower average expression levels than those from PICDGI, with genes ranked by immune-suppressive role across various cell types including mast cells, natural killer cells, T cells, B cells, and dendritic cells, alongside a single-cell atlas mapping cell progression and pseudo-time values.
(TIFF)

S13 Fig. Comparison of PICDGI and Moran's I test for driver gene prediction in immune cells for patient 3. We compare the PICDGI framework with Moran's I test for predicting driver genes in immune cells of Patient 3, finding that

driver genes identified by Moran's I have lower average expression levels than those from PICDGI, with genes ranked by immune-suppressive role across various cell types including mast cells, natural killer cells, T cells, B cells, and dendritic cells, alongside a single-cell atlas mapping cell progression and pseudo-time values.
(TIFF)

S8 Text. Ambient RNA correction and validation of progenitor-cell identification.
(DOCX)

S9 Text. Top 30 cancer driver genes identified by PICDGI across all three patients.
(DOCX)

S10 Text. Pathway enrichment analysis of PICDGI-predicted driver genes across patients.
(DOCX)

S1 Fig. Pathway enrichment for patient 1 driver and modulator genes. We performed pathway enrichment analysis for patient 1 and found that driver and modulator genes were enriched in cholesterol homeostasis and several malignancy-associated programs, including cell-cycle regulation, microtubule and centrosome processes, autophagy, oxidative stress responses, and pseudopodium activity.
(TIFF)

S2 Fig. Pathway enrichment for patient 2 driver and modulator genes. We performed pathway enrichment analysis for patient 2 and found that driver and modulator genes collectively activate core cancer programs, including cell-cycle regulation, spindle and checkpoint control, and major signaling pathways such as PDGF/RAF/PKC, mTORC1, hypoxia response, and apoptosis. We also observed enrichment for GO biological processes related to mitotic mechanics, metabolic and redox remodeling, and cellular regeneration and localization. These results highlight the coordinated functional roles of the predicted driver genes.
(TIFF)

S3 Fig. Pathway enrichment for patient 3 driver and modulator genes. We performed pathway enrichment analysis for patient 3's driver and modulator genes and found that these genes cluster into coherent biological programs. We observed significant enrichment of autophagy-related pathways, including PI3KC3 complex I/II and mTORC1 signaling. We also identified GO processes involving regulation of lipid kinase activity, cytoplasmic translation, glycolysis and related metabolic pathways, NADH regeneration, cellular responses to acidic pH, and angiogenesis. Together, these results indicate that patient 3's key regulators participate in coordinated PI3K-mTOR/autophagy signaling and metabolic stress–adaptation programs.
(TIFF)

S11 Text. External validation of PICDGI using an independent pediatric AML scRNA-seq cohort and cross-dataset driver inference analysis.
(DOCX)

S1 Fig. Cellular composition and cancer cell fraction dynamics across disease stages. This figure summarizes how PICDGI integrates single-cell cellular composition, malignant cell fraction dynamics, and gene-level ranking to identify leukemia-originating populations and candidate driver genes. After uniform preprocessing and annotation, we recover diverse cellular landscapes dominated by myeloid leukemic states alongside non-malignant immune populations. By aggregating cell-level malignancy scores by lineage and stage, we show that myeloid populations, particularly LSC-like, progenitor, cycling, and granulocytic states, exhibit the highest and most persistent malignant fractions from diagnosis through end-of-induction and relapse, whereas lymphoid compartments remain largely non-malignant. Finally, PICDGI

leverages these dynamics to prioritize driver genes, recovering known pediatric AML-associated drivers and nominating additional biologically plausible candidates, thereby linking cellular evolution to gene-level driver inference.
(TIFF)

S2 Fig. Differential gene ranking across disease transitions with known cancer drivers highlighted. This figure summarizes PICDGI-based prioritization of candidate cancer driver genes across disease transitions in pediatric AML. For both Dx–EOI and Dx–Relapse comparisons, genes are ranked by their driver coefficient (DrCoef), capturing consistent, malignant cell–associated expression changes across cell states. Known cancer drivers or AML-relevant genes are marked, demonstrating that PICDGI recovers established biology (e.g., PRDM1, GLO1, ISG15, STIP1, CD164) while also nominating additional, functionally coherent candidates involved in metabolism, stress response, RNA processing, and protein homeostasis. The similarity in coefficient magnitudes across top-ranked genes indicates a robust and non-noise–driven driver signal, with relapse-associated rankings highlighting genes linked to leukemic persistence and adaptation.
(TIFF)

Acknowledgments

We are grateful to Dr. Siddharth Rawat for thoughtfully reviewing an earlier version of this manuscript and offering valuable feedback.

Author contributions

Conceptualization: Komlan Atitey, Benedict Anchang.

Data curation: Komlan Atitey.

Formal analysis: Komlan Atitey, Benedict Anchang.

Software: Komlan Atitey.

Supervision: Benedict Anchang.

Validation: Komlan Atitey.

Writing – original draft: Komlan Atitey.

Writing – review & editing: Benedict Anchang.

References

1. Dakal TC, Dhabhai B, Pant A, Moar K, Chaudhary K, Yadav V, et al. Oncogenes and tumor suppressor genes: functions and roles in cancers. MedComm (2020). 2024;5(6):e582. https://doi.org/10.1002/mco2.582 PMID: 38827026
2. Vogt PK. Cancer genes. West J Med. 1993;158(3):273–8. PMID: 8460509
3. Prindiville SA, Mandrekar SJ, Meropol NJ, Denicoff A, Grad O, Hautala JA, et al. Streamlining the conduct of cancer clinical trials: new standard data collection practices for National Cancer Institute late-phase clinical studies. J Natl Cancer Inst. 2025;117(3):396–401. https://doi.org/10.1093/jnci/djae239 PMID: 39325869
4. Torkamani A, Schork NJ. Identification of rare cancer driver mutations by network reconstruction. Genome Res. 2009;19(9):1570–8. https://doi.org/10.1101/gr.092833.109 PMID: 19574499
5. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214–8. https://doi.org/10.1038/nature12213 PMID: 23770567
6. Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. Nucleic Acids Res. 2012;40(21):e169. https://doi.org/10.1093/nar/gks743 PMID: 22904074
7. 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 Biol. 2016;17(1):128. https://doi.org/10.1186/s13059-016-0994-0 PMID: 27311963
8. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238–44. https://doi.org/10.1093/bioinformatics/btt395 PMID: 23884480

9. Tang Y-Y, Wei P-J, Zhao J-P, Xia J, Cao R-F, Zheng C-H. Identification of driver genes based on gene mutational effects and network centrality. BMC Bioinformatics. 2021;22(Suppl 3):457. https://doi.org/10.1186/s12859-021-04377-0 PMID: 34560840
10. Li G, Hu Z, Luo X, Liu J, Wu J, Peng W, et al. Identification of cancer driver genes based on hierarchical weak consensus model. Health Inf Sci Syst. 2024;12(1):21. https://doi.org/10.1007/s13755-024-00279-6 PMID: 38464463
11. Du X-W, Li G, Liu J, Zhang C-Y, Liu Q, Wang H, et al. Comprehensive analysis of the cancer driver genes in breast cancer demonstrates their roles in cancer prognosis and tumor microenvironment. World J Surg Oncol. 2021;19(1):273. https://doi.org/10.1186/s12957-021-02387-z PMID: 34507558
12. Li X, Xu J, Li J, Gu J, Shang X. Towards simplified graph neural networks for identifying cancer driver genes in heterophilic networks. Brief Bioinform. 2024;26(1):bbae691. https://doi.org/10.1093/bib/bbae691 PMID: 39751645
13. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, et al. RNA velocity of single cells. Nature. 2018;560(7719):494–8. https://doi.org/10.1038/s41586-018-0414-6 PMID: 30089906
14. Bergen V, Lange M, Peidli S, Wolf FA, Theis FJ. Generalizing RNA velocity to transient cell states through dynamical modeling. Nat Biotechnol. 2020;38(12):1408–14. https://doi.org/10.1038/s41587-020-0591-3 PMID: 32747759
15. Schiebinger G, Shu J, Tabaka M, Cleary B, Subramanian V, Solomon A, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. Cell. 2019;176(4):928-943.e22. https://doi.org/10.1016/j.cell.2019.01.006 PMID: 30712874
16. Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35(12):2159–61. https://doi.org/10.1093/bioinformatics/bty916 PMID: 30445495
17. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, et al. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017;33(15):2314–21. https://doi.org/10.1093/bioinformatics/btx194 PMID: 28379368
18. Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, Saeys Y, et al. Trajectory-based differential expression analysis for single-cell sequencing data. Nat Commun. 2020;11(1):1201. https://doi.org/10.1038/s41467-020-14766-3 PMID: 32139671
19. Song D, Li JJ. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. Genome Biol. 2021;22(1):124. https://doi.org/10.1186/s13059-021-02341-y PMID: 33926517
20. Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. Proc Natl Acad Sci U S A. 2016;113(50):14330–5. https://doi.org/10.1073/pnas.1616440113 PMID: 27911828
21. Hou JP, Ma J. DawnRank: discovering personalized driver genes in cancer. Genome Med. 2014;6(7):56. https://doi.org/10.1186/s13073-014-0056-8 PMID: 25177370
22. Guo W-F, Zhang S-W, Zeng T, Li Y, Gao J, Chen L. A novel network control model for identifying personalized driver genes in cancer. PLoS Comput Biol. 2019;15(11):e1007520. https://doi.org/10.1371/journal.pcbi.1007520 PMID: 31765387
23. Guo W-F, Zhang S-W, Liu L-L, Liu F, Shi Q-Q, Zhang L, et al. Discovering personalized driver mutation profiles of single samples in cancer by network control strategy. Bioinformatics. 2018;34(11):1893–903. https://doi.org/10.1093/bioinformatics/bty006 PMID: 29329368
24. Pham VVH, Liu L, Bracken C, Goodall G, Li J, Le TD. Computational methods for cancer driver discovery: a survey. Theranostics. 2021;11(11):5553–68. https://doi.org/10.7150/thno.52670 PMID: 33859763
25. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. Mol Syst Biol. 2013;9:637. https://doi.org/10.1038/msb.2012.68 PMID: 23340843
26. Shi P, Han J, Zhang Y, Li G, Zhou X. IMI-driver: integrating multi-level gene networks and multi-omics for cancer driver gene identification. PLoS Comput Biol. 2024;20(8):e1012389. https://doi.org/10.1371/journal.pcbi.1012389 PMID: 39186807
27. Huang M, Ma J, An G, Ye X. Unravelling cancer subtype-specific driver genes in single-cell transcriptomics data with CSDGI. PLoS Comput Biol. 2023;19(12):e1011450. https://doi.org/10.1371/journal.pcbi.1011450 PMID: 38096269
28. Raimondi D, Passemiers A, Fariselli P, Moreau Y. Current cancer driver variant predictors learn to recognize driver genes instead of functional variants. BMC Biol. 2021;19(1):3. https://doi.org/10.1186/s12915-020-00930-0 PMID: 33441128
29. Mbemi A, Khanna S, Njiki S, Yedjou CG, Tchounwou PB. Impact of Gene-Environment Interactions on Cancer Development. Int J Environ Res Public Health. 2020;17(21):8089. https://doi.org/10.3390/ijerph17218089 PMID: 33153024
30. Chappell MA, Groves AR, Whitcher B, Woolrich MW. Variational bayesian inference for a nonlinear forward model. IEEE Trans Signal Process. 2009;57(1):223–36. https://doi.org/10.1109/tsp.2008.2005752
31. Atitey K, Loskot P, Mihaylova L. Variational Bayesian inference of hidden stochastic processes with unknown parameters. 2019. https://arxiv.org/abs/1911.00757
32. Atitey K. DEGBOE: discrete time evolution modeling of gene mutation through bayesian inference using qualitative observation of mutation events. J Biomed Inform. 2022;134:104197. https://doi.org/10.1016/j.jbi.2022.104197 PMID: 36084801
33. Kim N, Kim HK, Lee K, Hong Y, Cho JH, Choi JW, et al. Single-cell RNA sequencing demonstrates the molecular and cellular reprogramming of metastatic lung adenocarcinoma. Nat Commun. 2020;11(1):2285. https://doi.org/10.1038/s41467-020-16164-1 PMID: 32385277
34. Rade M, Grieb N, Weiss R, Sia J, Fischer L, Born P, et al. Single-cell multiomic dissection of response and resistance to chimeric antigen receptor T cells against BCMA in relapsed multiple myeloma. Nat Cancer. 2024;5(9):1318–33. https://doi.org/10.1038/s43018-024-00763-8 PMID: 38641734

35. Lusby R, Demirdizen E, Inayatullah M, Kundu P, Maiques O, Zhang Z, et al. Pan-cancer drivers of metastasis. Mol Cancer. 2025;24(1):2. https://doi.org/10.1186/s12943-024-02182-w PMID: 39748426
36. Wamsley JJ. The roles of NF-κB, Activin, and sphingosine-1-phosphate in promoting non-small cell lung cancer-initiating cell phenotypes. University of Virginia. 2013.
37. Atitey K, Hughes CE, Fusco JC. Physics-informed AI with chemical master equation dynamics for driver-gene subclone detection and risk labeling. Comput Struct Biotechnol J. 2025;27:4566–85. https://doi.org/10.1016/j.csbj.2025.10.046 PMID: 41234486
38. Anderson NM, Simon MC. The tumor microenvironment. Curr Biol. 2020;30(16):R921–5. https://doi.org/10.1016/j.cub.2020.06.081 PMID: 32810447
39. Visvader JE. Cells of origin in cancer. Nature. 2011;469(7330):314–22. https://doi.org/10.1038/nature09781 PMID: 21248838
40. Maman S, Witz IP. A history of exploring cancer in context. Nat Rev Cancer. 2018;18(6):359–76. https://doi.org/10.1038/s41568-018-0006-7 PMID: 29700396
41. Ma N, Whitt W. Efficient simulation of non-Poisson non-stationary point processes to study queueing approximations. Statist Probabil Lett. 2016;109:202–7. https://doi.org/10.1016/j.spl.2015.11.018
42. Dingli D, Pacheco JM. Stochastic dynamics and the evolution of mutations in stem cells. BMC Biol. 2011;9:41. https://doi.org/10.1186/1741-7007-9-41 PMID: 21649942
43. Grenier Y. Time-dependent ARMA modeling of nonstationary signals. IEEE Trans Acoust Speech Signal Process. 1983;31(4):899–911. https://doi.org/10.1109/tassp.1983.1164152
44. Chou-Chen SW, Morettin PA. Indirect inference for locally stationary ARMA processes with stable innovations. J Stat Comput Simulat. 2020;90(17):3106–34. https://doi.org/10.1080/00949655.2020.1797030
45. Løvsletten O. Consistency of detrended fluctuation analysis. Phys Rev E. 2017;96(1–1):012141. https://doi.org/10.1103/PhysRevE.96.012141 PMID: 29347071
46. Corona-Ruiz M, Hernandez-Cabrera F, Cantú-González JR, González-Amezcua O, Javier Almaguer F. A stochastic phylogenetic algorithm for mitochondrial DNA analysis. Front Genet. 2019;10:66. https://doi.org/10.3389/fgene.2019.00066 PMID: 30906309
47. Chiang J-Y, Huang J-W, Lin L-Y, Chang C-H, Chu F-Y, Lin Y-H, et al. Detrended fluctuation analysis of heart rate dynamics is an important prognostic factor in patients with end-stage renal disease receiving peritoneal dialysis. PLoS One. 2016;11(2):e0147282. https://doi.org/10.1371/journal.pone.0147282 PMID: 26828209
48. Zunino L, Pérez DG, Kowalski A, Martín MT, Garavaglia M, Plastino A, et al. Fractional brownian motion, fractional gaussian noise, and Tsallis permutation entropy. Physica A: Stat Mech Appl. 2008;387(24):6057–68. https://doi.org/10.1016/j.physa.2008.07.004
49. Hoffman MD, Blei DM, Wang C, Paisley J. Stochastic variational inference. J Machine Learn Res. 2013.
50. Yoo CW, Nam B-H, Kim J-Y, Shin H-J, Lim H, Lee S, et al. Carbonic anhydrase XII expression is associated with histologic grade of cervical cancer and superior radiotherapy outcome. Radiat Oncol. 2010;5:101. https://doi.org/10.1186/1748-717X-5-101 PMID: 21040567
51. Vrettas MD, Opper M, Cornford D. Variational mean-field algorithm for efficient inference in large systems of stochastic differential equations. Phys Rev E Stat Nonlin Soft Matter Phys. 2015;91(1):012148. https://doi.org/10.1103/PhysRevE.91.012148 PMID: 25679611
52. Chib S. Markov chain monte carlo methods: computation and inference. In: Handbook of econometrics. Elsevier; 2001. 3569–649. https://doi.org/10.1016/s1573-4412(01)05010-3
53. Meredith M, Kruschke J. HDInterval: highest (posterior) density intervals. CRAN: Contributed Packages; 2016.
54. Titsias M, Lawrence ND. Bayesian Gaussian process latent variable model. Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings; 2010.
55. Gargiulo G, Serresi M, Marine J-C. Cell states in cancer: drivers, passengers, and trailers. Cancer Discov. 2024;14(4):610–4. https://doi.org/10.1158/2159-8290.CD-23-1510 PMID: 38571419
56. Olaniru OE, Kadolsky U, Kannambath S, Vaikkinen H, Fung K, Dhami P, et al. Single-cell transcriptomic and spatial landscapes of the developing human pancreas. Cell Metab. 2023;35(1):184-199.e5. https://doi.org/10.1016/j.cmet.2022.11.009 PMID: 36513063
57. H. Kelejian H, Prucha IR. On the asymptotic distribution of the Moran I test statistic with applications. J Economet. 2001;104(2):219–57. https://doi.org/10.1016/s0304-4076(01)00064-1
58. Cao J, Spielmann M, Qiu X, Huang X, Ibrahim DM, Hill AJ, et al. The single-cell transcriptional landscape of mammalian organogenesis. Nature. 2019;566(7745):496–502. https://doi.org/10.1038/s41586-019-0969-x PMID: 30787437
59. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015;33(5):495–502. https://doi.org/10.1038/nbt.3192 PMID: 25867923
60. Gires O, Pan M, Schinke H, Canis M, Baeuerle PA. Expression and function of epithelial cell adhesion molecule EpCAM: where are we after 40 years?. Cancer and Metastasis Reviews. 2020;39(3):969–87.
61. Rao CG, Chianese D, Doyle GV, Miller MC, Russell T, Sanders RA Jr, et al. Expression of epithelial cell adhesion molecule in carcinoma cells present in blood and primary and metastatic tumors. Int J Oncol. 2005;27(1):49–57. https://doi.org/10.3892/ijo.27.1.49 PMID: 15942643

62. Young MD, Behjati S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. Gigascience. 2020;9(12):giaa151. https://doi.org/10.1093/gigascience/giaa151 PMID: 33367645
63. Crick FH, Barnett L, Brenner S, Watts-Tobin RJ. General nature of the genetic code for proteins. Nature. 1961;192:1227–32. https://doi.org/10.1038/1921227a0 PMID: 13882203
64. Cavenee WK, White RL. The genetic basis of cancer. Sci Am. 1995;272(3):72–9. https://doi.org/10.1038/scientificamerican0395-72 PMID: 7871410
65. Seillier M, Peuget S, Gayet O, Gauthier C, N'Guessan P, Monte M, et al. TP53INP1, a tumor suppressor, interacts with LC3 and ATG8-family proteins through the LC3-interacting region (LIR) and promotes autophagy-dependent cell death. Cell Death Differ. 2012;19(9):1525–35. https://doi.org/10.1038/cdd.2012.30 PMID: 22421968
66. von Neubeck B, Gondi G, Riganti C, Pan C, Parra Damas A, Scherb H, et al. An inhibitory antibody targeting carbonic anhydrase XII abrogates chemoresistance and significantly reduces lung metastases in an orthotopic breast cancer model in vivo. Int J Cancer. 2018;143(8):2065–75. https://doi.org/10.1002/ijc.31607 PMID: 29786141
67. Li Y, Xiao X, Li J, Han Y, Cheng C, Fernandes GF, et al. Lung cancer in ever- and never-smokers: findings from multi-population GWAS studies. Cancer Epidemiol Biomarkers Prev. 2024;33(3):389–99. https://doi.org/10.1158/1055-9965.EPI-23-0613 PMID: 38180474
68. Roshan-Zamir M, Khademolhosseini A, Rajalingam K, Ghaderi A, Rajalingam R. The genomic landscape of the immune system in lung cancer: present insights and continuing investigations. Front Genet. 2024;15:1414487. https://doi.org/10.3389/fgene.2024.1414487 PMID: 38983267
69. Zhang B, Leung PC, Cho WCS, Wong CK, Wang D. Targeting PI3K signaling in lung cancer: advances, challenges and therapeutic opportunities. J Transl Med. 2025;23(1):1–12.
70. Marusyk A, Polyak K. Tumor heterogeneity: causes and consequences. Biochim Biophys Acta. 2010;1805(1):105–17. https://doi.org/10.1016/j.bbcan.2009.11.002 PMID: 19931353
71. Corbet C, Feron O. Tumour acidosis: from the passenger to the driver's seat. Nat Rev Cancer. 2017;17(10):577–93. https://doi.org/10.1038/nrc.2017.77 PMID: 28912578
72. Vallejos CA, Risso D, Scialdone A, Dudoit S, Marioni JC. Normalizing single-cell RNA sequencing data: challenges and opportunities. Nat Methods. 2017;14(6):565–71. https://doi.org/10.1038/nmeth.4292 PMID: 28504683
73. Svensson V, Natarajan KN, Ly L-H, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14(4):381–7. https://doi.org/10.1038/nmeth.4220 PMID: 28263961
74. Luo W, Lin GN, Song W, Zhang Y, Lai H, Zhang M, et al. Single-cell spatial transcriptomic analysis reveals common and divergent features of developing postnatal granule cerebellar cells and medulloblastoma. BMC Biol. 2021;19(1):135. https://doi.org/10.1186/s12915-021-01071-8 PMID: 34210306
75. Atitey K, Motsinger-Reif AA, Anchang B. Model-based evaluation of spatiotemporal data reduction methods with unknown ground truth through optimal visualization and interpretability metrics. Brief Bioinform. 2023;25(1):bbad455. https://doi.org/10.1093/bib/bbad455 PMID: 38113074
76. Pham D, Tan X, Balderson B, Xu J, Grice LF, Yoon S, et al. Robust mapping of spatiotemporal trajectories and cell-cell interactions in healthy and diseased tissues. Nat Commun. 2023;14(1):7739. https://doi.org/10.1038/s41467-023-43120-6 PMID: 38007580
77. Mumme HL, Huang C, Ohlstrom D, Bakhtiari M, Raikar SS, DeRyckere D, et al. Identification of leukemia-enriched signature through the development of a comprehensive pediatric single-cell atlas. Nat Commun. 2025;16(1):4114. https://doi.org/10.1038/s41467-025-59362-5 PMID: 40316535
78. van Galen P, Hovestadt V, Wadsworth Ii MH, Hughes TK, Griffin GK, Battaglia S, et al. Single-Cell RNA-Seq reveals AML hierarchies relevant to disease progression and immunity. Cell. 2019;176(6):1265-1281.e24. https://doi.org/10.1016/j.cell.2019.01.031 PMID: 30827681
79. Zeng AGX, Bansal S, Jin L, Mitchell A, Chen WC, Abbas HA, et al. A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. Nat Med. 2022;28(6):1212–23. https://doi.org/10.1038/s41591-022-01819-x PMID: 35618837
80. Shlush LI, Mitchell A, Heisler L, Abelson S, Ng SWK, Trotman-Grant A, et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. Nature. 2017;547(7661):104–8. https://doi.org/10.1038/nature22993 PMID: 28658204
81. Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573-3587.e29. https://doi.org/10.1016/j.cell.2021.04.048 PMID: 34062119
82. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM 3rd, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-1902.e21. https://doi.org/10.1016/j.cell.2019.05.031 PMID: 31178118
83. Mumme H, Thomas BE, Bhasin SS, Krishnan U, Dwivedi B, Perumalla P, et al. Single-cell analysis reveals altered tumor microenvironments of relapse- and remission-associated pediatric acute myeloid leukemia. Nat Commun. 2023;14(1):6209. https://doi.org/10.1038/s41467-023-41994-0 PMID: 37798266
84. Mandelbaum J, Bhagat G, Tang H, Mo T, Brahmachary M, Shen Q, et al. BLIMP1 is a tumor suppressor gene frequently disrupted in activated B cell-like diffuse large B cell lymphoma. Cancer Cell. 2010;18(6):568–79. https://doi.org/10.1016/j.ccr.2010.10.030 PMID: 21156281
85. Thornalley P. Glyoxalase I–structure, function and a critical role in the enzymatic defence against glycation. Biochem Soc Transact. 2003;31(6):1343–138.
86. Desai SD. ISG15: a double edged sword in cancer. Oncoimmunology. 2015;4(12):e1052935. https://doi.org/10.1080/2162402X.2015.1052935 PMID: 26587329

87. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):1–14. https://doi.org/10.1038/s12276-018-0071-8 PMID: 30089861

88. Wang P-W, Su Y-H, Chou P-H, Huang M-Y, Chen T-W. Survival-related genes are diversified across cancers but generally enriched in cancer hallmark pathways. BMC Genomics. 2022;22(Suppl 5):918. https://doi.org/10.1186/s12864-022-08581-x PMID: 35508961

89. Ricciuti B, Arbour KC, Lin JJ, Vajdi A, Vokes N, Hong L, et al. Diminished efficacy of programmed death-(Ligand)1 inhibition in STK11- and KEAP1-mutant lung adenocarcinoma is affected by KRAS mutation status. J Thorac Oncol. 2022;17(3):399–410. https://doi.org/10.1016/j.jtho.2021.10.013 PMID: 34740862

90. Xu J, Zhang H, Yang L. Rab3B proteins: cellular functions, regulatory mechanisms, and potential as a cancer therapy target. Cell Biochem Biophys. 2025;83(1):263–77. https://doi.org/10.1007/s12013-024-01549-6 PMID: 39320613

91. Tie Y, Tang F, Wei Y-Q, Wei X-W. Immunosuppressive cells in cancer: mechanisms and potential therapeutic targets. J Hematol Oncol. 2022;15(1):61. https://doi.org/10.1186/s13045-022-01282-8 PMID: 35585567

92. Tay RE, Richardson EK, Toh HC. Revisiting the role of CD4+ T cells in cancer immunotherapy-new insights into old paradigms. Cancer Gene Ther. 2021;28(1–2):5–17. https://doi.org/10.1038/s41417-020-0183-x PMID: 32457487

93. McKean WB, Moser JC, Rimm D, Hu-Lieskovan S. Biomarkers in precision cancer immunotherapy: promise and challenges. Am Soc Clin Oncol Educ Book. 2020;40:e275–91. https://doi.org/10.1200/EDBK_280571 PMID: 32453632

94. Wang D-R, Wu X-L, Sun Y-L. Therapeutic targets and biomarkers of tumor immunotherapy: response versus non-response. Signal Transduct Target Ther. 2022;7(1):331. https://doi.org/10.1038/s41392-022-01136-2 PMID: 36123348

95. Pla-Martín D, Calpena E, Lupo V, Márquez C, Rivas E, Sivera R, et al. Junctophilin-1 is a modifier gene of GDAP1-related Charcot-Marie-Tooth disease. Hum Mol Genet. 2015;24(1):213–29. https://doi.org/10.1093/hmg/ddu440 PMID: 25168384

96. Lehnart SE, Wehrens XHT. The role of junctophilin proteins in cellular function. Physiol Rev. 2022;102(3):1211–61. https://doi.org/10.1152/physrev.00024.2021 PMID: 35001666

97. Neizer-Ashun F, Bhattacharya R. Reality CHEK: understanding the biology and clinical potential of CHK1. Cancer letters. 2021;497:202–11.

98. Tan Z, Chen M, Wang Y, Peng F, Zhu X, Li X, et al. CHEK1: a hub gene related to poor prognosis for lung adenocarcinoma. Biomark Med. 2022;16(2):83–100. https://doi.org/10.2217/bmm-2021-0919 PMID: 34882011

99. Egbon OA, Hickey JW, Anchang B. Fusion of spatiotemporal and network models to prioritize multiscale effects in single-cell perturbations. Brief Bioinform. 2025;26(3):bbaf277. https://doi.org/10.1093/bib/bbaf277 PMID: 40545244