DNA methylation and immune infiltration mediate the impact of tobacco exposure on pancreatic adenocarcinoma outcome: a high-dimensional mediation analysis

Florence Pittion[1], Elise Amblard[1], Emilie Devijver[1], Adeline Samson[2], Nelle Varoquaux[3], Magali Richard[1*]

1 Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000, Grenoble, France
2 Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
3 CNRS, UMR 5525, VetAgro Sup, Grenoble INP, TIMC, 38000 Grenoble, France

* magali.richard@univ-grenoble-alpes.fr

Abstract

Pancreatic ductal adenocarcinoma (PDAC) ranks among the most aggressive malignancies, characterized by exceptionally poor prognosis due to late diagnosis and therapeutic resistance. While tobacco exposure is an established risk factor for PDAC incidence and influences patient survival, the underlying molecular mechanisms mediating these effects remain poorly understood. We aimed to characterize the causal pathways linking tobacco exposure to PDAC patient survival, focusing on DNA methylation and tumor immune infiltration as potential mediators. For this purpose, we developed HDMAX2-surv, an extension of high-dimensional mediation analysis method HDMAX2, adapted to survival outcomes. Through extensive simulations across multiple scenarios, we demonstrated superior performance compared to existing methods in handling complex DNA methylation data. We integrated this approach with causal discovery frameworks and immune deconvolution algorithms to dissect multi-pathway mediation mechanisms. Using The Cancer Genome Atlas (TCGA) cohort of 112 PDAC patients, we identified 36 aggregated methylated regions (AMRs) with significant mediation effects, exhibiting both protective and deleterious impacts on survival. Our causal discovery analysis revealed complex relationships where tobacco effects operate through sequential pathways involving DNA methylation and immune infiltration changes. Serial mediation analysis identified four distinct pathways where tobacco-induced epigenetic modifications mediate survival effects through immune microenvironment alterations, involving the *ZBTB42*, *CLGN*, *RAD51AP2*, and *PRDM16* loci. Remarkably, these effects were undetectable through gene expression analysis alone, highlighting the unique value of methylation-based approaches. This study provides dual contributions: methodologically, through an integrated framework for high-dimensional survival mediation analysis with causal discovery; and biologically, by elucidating complex molecular mechanisms through which tobacco exposure influences PDAC outcomes. Our findings reveal the intricate interplay between environmental exposures, epigenetic modifications, and immune responses in cancer prognosis.

Introduction

Pancreatic adenocarcinoma (PDAC) represents approximately 90% of all pancreatic cancers and is associated with a high mortality rate, ranking as the third leading cause of cancer-related deaths in Europe [1]. PDAC outcomes exhibit significant variations based on gender and socioeconomic factors [2]. Among modifiable risk factors, smoking, alcohol consumption, and obesity have been consistently identified [1], highlighting the importance of considering tobacco exposure in disease prevention strategies. The relationship between smoking and PDAC development has been documented in numerous studies and comprehensively reviewed [1, 3]. A meta-analysis included 82 published epidemiological studies identified an 75% increased risk of developing pancreatic cancer in smokers compared to non-smokers [4], with an elevated risk that persists for at least 10 years after smoking cessation. Lynch et al. [5] observed a significantly elevated risk of developing pancreatic cancer (OR = 1.77) in current smokers versus never-smokers. Furthermore, data from the Panc4 cohort demonstrated a two-fold increased risk of developing pancreatic cancer among smokers, with the risk positively correlated with both the number of cigarettes smoked and the duration of exposure [6].

Beyond its role in PDAC development, exposure to tobacco also significantly impacts mortality and prognosis among patients diagnosed with pancreatic cancer. Ben et al. conducted a comprehensive meta-analysis of 20 studies and observed that current smokers (Hazard Ratio (HR) of 1.56) and former smokers (HR = 1.15) exhibited an elevated risk of mortality in patients diagnosed with pancreatic cancer, compared to never-smokers. Importantly, they determined that this elevated mortality risk among smoking patients persisted independently of other potential confounding factors, including alcohol use, BMI, and diabetes history [7]. Weissman et al. specifically examined the effect of smoking on PDAC prognosis across multiple studies [3]. Six of the eight reviewed studies identified a relationship between smoking and prognosis, with three demonstrating statistically significant effects. These findings led the authors to conclude that smoking should be considered as having important implications not only for the occurrence PDAC, but also for its prognosis. Further evidence comes from a meta-analysis by Lugo et al., which evaluated smoking's impact on pancreatic cancer risk. This study reported a 48% relative risk increase in people who had ever smoked compared to never-smokers, with current smokers showing a substantially higher excess risk (82%) compared to former smokers (17%). In particular, this risk decreased proportionally with the time elapsed since smoking cessation, eventually reaching levels comparable to non-smokers after approximately 20 years [8]. Taken together, these findings underscore the critical importance of investigating the mechanisms through which tobacco consumption influences PDAC prognosis, providing a rationale for our study.

Despite established associations between tobacco exposure and PDAC prognosis, the underlying molecular mechanisms are still poorly understood [9]. Mediation analysis offers a valuable approach to investigate the pathways through which exposure (tobacco) affects an outcome (survival). Specifically, high-dimensional mediation analysis enables the investigation of how epigenetic modifications, such as DNA methylation (DNAm), could mediate tobacco's effects on clinical outcomes. Multiple studies have shown that tobacco exposure significantly affects DNAm patterns across the genome [10]. Methods such as HDMAX2 [11, 12] have been developed to perform a high-dimensional mediation analysis in this context. HDMAX2 employs a two-step approach: first, identify potential mediators using latent factor models, and second, evaluate the mediated effect using the Sobel test and average causal mediation effect.

A critical objective in cancer research is understanding and predicting patient survival. To advance personalized medicine, oncologists need robust methods, such as

mediation analysis, to classify patients according to prognostic risk. High-dimensional mediation survival analysis presents multiple methodological challenges: selecting potential mediators (sure independant screening, false discovery rate control, regularization) [13], determining the appropriate survival model —whether non-parametric (Kaplan-Meier estimator, Nelson-Aalen estimator), parametric (Weibull, log-normal, exponential distributions), or semi-parametric (Cox proportional hazards models) [14, 15]— and integrating these approaches into a coherent statistical mediation framework (Sobel test [16], Baron and Kenny [17], boostraping , sensitivity analysis) [18, 19].

Several statistical methods have been proposed to address these challenges. HIMA combines sure independence screening (SIS) for the exposure–mediator model with minimax-concave-penalty (MCP) regularisation in the Cox outcome model to construct a sparse mediation set [20, 21]. MASH couples SIS and MCP selection and utilizes R^2-based measures to estimate mediation effects within the Cox regression model [22]. IUSMMT proposes a Cox model with random effects to identify mediators in high-dimensional settings [23]. CoxMKF uses aggregation of multiple knockoffs with Cox proportional hazards models to control the false discovery rate [24]. mediateR estimates high-dimensional mediation via dual ridge penalties and joint significance testing, and handles continuous, binary and survival outcomes [25]. Finally, SMAHP employs a multi-omic methodology capable of handling both high-dimensional mediators and exposures with FDR control [26]. Despite these advances, no consensus exists on the optimal methodological approach for high-dimensional mediation analysis in survival studies.

Causal mediation analysis is only meaningful within an established causal framework [27]. The integration of causal discovery with subsequent mediation analysis offers compelling advantages from both mechanistic and predictive perspectives. From a mechanistic standpoint, it enables understanding of the biological pathways responsible for observed phenotypes. From a predictive perspective, it provides a framework for survival prediction based on observational data while accounting for underlying causal relationships. In complex biological systems characterized by high-dimension, multicollinearity and heterogeneous underlying distributions, causality cannot be established through controlled experimental studies alone. Traditional epidemiological frameworks, such as the Bradford Hill criteria for establishing causality, become difficult to interpret in such high-dimensional molecular contexts [28, 29].

Numerous computational methods exist to infer causal relationships, broadly categorized into constraint-based approaches using (conditional) independence testing and score-based methods that optimize causal structure scores. Established algorithms such as the Peter-Clark (PC) algorithm, a constraint-based method, have demonstrated efficiency in bioinformatics applications [29]. However, these approaches often prove inadequate for high-dimensional data due to computational complexity and statistical power limitations [30]. To effectively identify and interpret causal mechanisms in high-dimensional biological data, there is a critical need for causal discovery algorithms ensuring sparsity, adding causal discovery structure for feature selection and integrating biological prior knowledge. Such hybrid frameworks are particularly crucial when investigating complex exposure-outcome relationships where multiple molecular pathways may operate simultaneously, as is the case with tobacco effects on cancer survival.

Indeed, beyond epigenetic alterations, emerging evidence suggests that immune infiltration within the tumor microenvironment may also mediate the relationship between tobacco exposure and PDAC outcomes. Weissman et al. proposed that tobacco's effects on PDAC operate through inflammation and immune system modulation pathways [3]. This hypothesis is supported by epidemiological observations

that certain immune-related factors, such as allergy history, confer protective effects against pancreatic cancer, presumably through enhanced immune surveillance mechanisms [1,31]. The role of immune infiltration in PDAC prognosis is complex and represents an active area of investigation in cancer immunology. Recent studies have demonstrated that immune composition profoundly affects PDAC patient outcomes, with distinct immune signatures associated with varying survival patterns [32]. These findings underscore the critical importance of incorporating immune infiltration parameters into comprehensive mediation frameworks when investigating tobacco effects on survival. The integration of both epigenetic and immunological mediators necessitates sophisticated multi-mediator causal analytical approaches capable of disentangling their individual and joint contributions to tobacco-related survival effects. Such comprehensive models may reveal previously unrecognized mechanistic pathways and provide deeper insights into the complex interplay between tobacco exposure, molecular alterations, and immune responses in PDAC pathogenesis.

In this work, we aimed to better characterize the relationship between tobacco exposure and PDAC prognosis. We developed a novel method for high-dimensional mediation analysis and causal discovery, adapted to survival outcomes. We validated it thoroughly using simulated data, and we then applied this methodology to a public TCGA dataset of PDAC patients to identify potential DNAm mediators in the causal pathway between tobacco exposure and clinical outcomes. To gain biological insights into the identified chromatin regions, we used deconvolution algorithms to quantify immune cell infiltration within each patient's tumor microenvironment. Then, we investigated the causal directionality between DNAm mediators and immune infiltration patterns, implementing multi-mediation analysis to elucidate the complete causal pathway linking tobacco exposure, DNAm alterations, immune infiltrate, and PDAC patient survival. Through this comprehensive approach, we identified several potential molecular mechanisms that may mediate the effect of tobacco exposure on PDAC prognosis, providing new insights into this clinically significant relationship.

Materials and methods

HDMAX2-surv method

Extension of HDMAX2 to HDMAX2-surv We extended the HDMAX2 method [11, 12] to a model that incorporates a censored survival outcome to assess indirect effects in the context of time-to-event data. The HDMAX2 method is based on a latent factor mixed regression model that allows the estimation of unobserved latent factors while performing a high-dimensional association analysis. HDMAX2 also implements a max-squared test to assess the statistical significance of potential mediators. The method then relies on the counterfactual framework with `mediation` R package [19] to estimate effects of the selected mediators. A key challenge in this study is the integration of a censored survival outcome within the mediation framework. We evaluated two modeling strategies: (i) `HDMAX2-surv.param` with an AFT (Accelareted Failure Time) parametric survival model in which the best-fitting distribution for the censored data was selected using the AIC (Akaike Information Criterion), and (ii) `HDMAX2-surv.aalen` with the Aalen additive-hazards model, a flexible non-parametric alternative for time-to-event analysis. See Supplementary Material section 1.1 for more details on the models.

Survival mediation methods comparison. Among high-dimensional mediation methods designed for survival outcomes, we retained HIMA [21] for benchmarking against our two methods. It is a publicly available pipeline that combines genome-wide screening with per-mediator inference. Although its reliance on a penalised Cox model was not itself a selection criteria, it provides a third survival framework, complementary to our AFT parametric and Aalen additive models, which lets us compare approaches under different modeling assumptions. The remaining methods were excluded for practical reasons: mediateR [25] and SMAHP [26], while advertised as high-dimensional, have been tested on moderate-sized data sets only and lack a variable selection module that scales to hundreds of thousands of CpGs. CoxMKF [24] is limited to knockoff-based selection and is unable to estimate mediation effects, so it cannot be compared directly with our methods. No `R` or `Python` code is provided by MASH. IUSMMT is designed for multiple DNA-methylation exposures and a single mediator. Finally, multimediate (survival) [33] computes mediation effects for a pre-specified panel of mediators and therefore cannot participate in the initial screening step. Some of these tools could be included in the second step (effect estimation) of our workflow, but their current implementations do not allow for side-by-side evaluation (see Supplementary Material Table 1). As results of this selection, we compared our two proposed methods `HDAMX2-surv.param`, `HDAMX2-surv.aalen` and the only external selected method `HIMA`.

Simulation studies

To evaluate HDMAX-surv performances and compare them with state-of-the-art method, we performed simulations. We simulated datasets mirroring our mediation setting with DNAm and a survival outcome.

Simulation framework. We simulated complete datasets (X, M, T, C) for N individuals and p CpG sites grouped into b correlated blocks (with within-block correlation fixed at 0.5). Each individual has K unobserved latent factors U. The variables were generated as follows:

- The **exposure** X is continuous and can be dichotomized as $X_{\text{bin}} = 1_{\{X \geq 0\}}$ in some scenarios.

- The **methylation matrix** M incorporates:
  - the influence of latent factors U,
  - the causal effect of the exposure on methylation ($X \to M$),
  - within-block correlation among probes,
  - additive Gaussian noise.
- The **survival time** T is drawn from an exponential model, with a hazard depending on both methylation and exposure ($M \to T$, $X \to T$).
- The **censoring time** C is sampled independently from a Gamma distribution and calibrated to yield approximately 20% censoring.

To define causal relationships, we randomly selected b blocks of 10 CpGs each, with $m_{\text{causal}} = b \times 10$ representing the total number of true simulated mediators. Within each block, the exposure-to-mediator coefficients α and mediator-to-survival coefficients β were drawn from normal distributions $\mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ and $\mathcal{N}(\mu_\beta, \sigma_\beta^2)$, respectively. To simulate realistic methylation patterns, all CpGs within a block shared the same effect sign to mimic hypo- or hyper-methylated regions.

An overlap proportion π_{overlap} controlled the degree to which the sets of CpGs with non-zero coefficients α_j and non-zero β_j coincided. Specifically, $(1 - \pi_{\text{overlap}})/\pi_{\text{overlap}} \times m_{\text{causal}}$ CpGs had non-zero exposure effects ($\alpha_j \neq 0$) but no survival effects ($\beta_j = 0$), and vice versa. The intersection of these two sets defined the true mediators, carrying both exposure-to-mediator and mediator-to-outcome effects. Finally, exposure X was simulated jointly with latent factors $\mathbf{U}_i$, with parameter ν_X governing their shared variability (see Supplementary Material for complete specifications).

For every setting in the parameter grid, we generated several replicate datasets, each initialized with a different seed to guarantee both variability and reproducibility. Full specifications of the equations, parameter grid, and underlying rationale appear in Supplementary Material, section 1.2.

Performances evaluation. We evaluated the performance of Step 1 (mediator selection) using the "precision score" defined as the count of True Positives (TP) among selected mediators, representing the number of correctly identified causal CpGs among the top-m_{causal} ranked candidates, where m_{causal} corresponds to the number of causal simulated CpGs. This metric directly assesses the method's ability to prioritize true mediators over non-mediators in high-dimensional settings. We assessed Step 2 performance using a "precision score" that requires both statistical significance and correct effect directionality. Specifically, a true prediction was validated when: (i) the Average Causal Mediation Effect (ACME) was correctly identified as statistically significant and (ii) the estimated effect sign matched the simulated ground truth. Among the top 100 CpGs selected in Step 1, we retained those with significant estimated effects using a Bonferroni-corrected threshold ($\alpha = 0.05/100 = 0.0005$) and verified whether their estimated effects were consistent with the simulated mediation effects. This stringent evaluation ensures that the methods not only detect mediation, but also provide accurate effect estimates.

Study population: the TCGA-PAAD cohort

The Cancer Genome Atlas (TCGA) is the result of a collaboration between the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) to generate molecular maps in major types and subtypes of cancer. It includes omic and

clinical data on enrolled cases provided by Biospecimen Core Repositories (BCRs) in schema-valid XML format. We downloaded the TCGA-PAAD study which is the PDAC atlas.

Molecular data. We used transcriptomic and methylation data of PDAC patients from the TCGA (n=184 patients with methylation data, n=178 for RNAseq) filtered as in [34]. The processing of the data was performed by the TCGA. For the methylation dataset, we also filtered out probes with NA values and probes located on sex chromosomes, retaining 371,033 features.

Exposure assessment. Among the tobacco-related variables available in the TCGA database, we used `exposures.tobacco_smoking_status`, which includes six categories: "Lifelong Non-Smoker", "Current Reformed Smoker for > 15 yrs", "Current Reformed Smoker for ≤ 15 yrs", "Current Smoker", "Current Reformed Smoker, Duration Not Specified", and "Unknown".

Patients with an "Unknown" smoking status ($n = 36$) were excluded from the analysis due to missing information. Patients categorized as "Current Reformed Smoker, Duration Not Specified" ($n = 8$) were excluded, as their smoking status could not be reliably assigned to either group.

We then created a binary smoking variable: "Lifelong Non-Smoker" and "Current Reformed Smoker for > 15 yrs" were classified as **non-smokers** (0), while "Current Reformed Smoker for ≤ 15 yrs" and "Current Smoker" were classified as **smokers** (1).

Follow-up. The follow-up variables are described as follows in the GDC Data Dictionary: "A visit by a patient or study participant to a medical professional. A clinical encounter that encompasses planned and unplanned trial interventions, procedures, and assessments that may be performed on a subject. A visit has a start and an end, each described with a rule. The process by which information about the health status of an individual is obtained before and after a study has officially closed; an activity that continues something that has already begun or that repeats something that has already been done." We used the following follow-up variables: (i) follow-up time (in days) and (ii) status (Censored = 0, event (death) = 1). We kept only common samples for which we have smoking, follow-up information, and non-zero survival time (n=112) (Table 1).

Co-variables. We also included co-variables in the models, such as age, gender, and grade of the tumor. In PDAC histologic grade is based on gland-forming differentiation and is reported as G1 well differentiated, G2 moderately differentiated, G3 poorly differentiated, and G4 undifferentiated/anaplastic.

Quantification of immune infiltration on the TCGA-PAAD cohort

Reference profiles. To perform supervised deconvolution, we used a dataset of bulk transcriptomic profiles from nine pure cell types found in PDAC: endothelial cells, neutrophils, macrophages, B cells, $CD4^+$ and $CD8^+$ cells, fibroblasts, and classical and basal-like tumour cells (GSE281204). We also downloaded two published single-cell datasets with labels to use as single-cell references during deconvolution. The first dataset [35] contained 57,530 cells with the following cell types labelled by the authors: fibroblasts, stellate cells (renamed fibroblasts), macrophages, endothelial cells, T and B cells, ductal cells type 2 (renamed tumor cells) and 1, endocrine cells, and acinar cells. The second dataset [36] contained 66,075 cells with the following cell types:

macrophages, NK cells, regulatory T cells, tumor cells, B cells, hepatocytes, pDCs and DCs, endothelial cells, plasma cells, and mesenchymal cells. We retained in both datasets, respectively 44,549 and 22,662 cells that match the cell types present in the bulk reference (endothelial cells, B and T cells, macrophages, fibroblasts, tumor cells for the Peng dataset [35], endothelial cells, B cells, macrophages, NK cells, regulatory T cells, pDCs and DCs and tumor cells for Raghavan [36]).

For each single-cell dataset, we sampled 5,000 cells randomly (to reduce computing time) while retaining the same number of cells from each patient (n=35 for Peng, n=2 for Raghavan), and used puriST, a classifier that separates tumor cells into basal or classical-like [37].

Deconvolution. We applied supervised deconvolution algorithms to retrieve cell type proportions (see Supplementary Figure 13). We deconvoluted the transcriptomic profiles of the samples used in the mediation analysis, using bulk or single-cell data as references (GSE281204 for the bulk reference, [35] and [36]) for the single-cell ones). We applied robust linear regression (RLR) with bulk references, and SCDC [38], InstaPrism [39], and MuSiC [40] with single-cell references. For RLR, we used the function `epidish` from the package EpiDISH with the "method" parameter set to "RPC." Of note, the references were normalized with DESeq2 beforehand [41]. For InstaPrism [39], we used the function `InstaPrism` from the eponymous package. We got two proportions matrices, one per single-cell reference. For SCDC [38], we used the function `SCDC_prop` to estimate the proportion matrix from a single reference. For MuSiC [40], we used the function `music_prop from` the package MuSiC and obtained a proportion matrix per single-cell reference. We finally computed a consensus prediction, based on the 7 predictions from the four methods described above. The consensus prediction was done based on the following procedure. First, all subsets of dendritic cells (DCs) and T cells (resp. plasmacytoid and XCR1$^+$ DCs for DCs and CD4$^+$, CD8$^+$ and regulatory T cells for T cells) were aggregated into a single type, respectively, called DCs and T cells. For the predicted proportion matrix, we sumed the proportions of all DCs (resp. T cells) subsets into a single type of DCs (resp. T cells) (A_{avg}). For the bulk reference matrix, we computed the average profile for DCs and T cells. For single-cell references, we first performed a pseudo-bulk before averaging the reference profiles of the DCs and T cell subsets (T_{avg}). Then for each prediction, we computed the Root Mean Square Error (RMSE) between the reconstruction $T_{\text{avg}} \times A_{\text{avg}}$ and the expression matrix. Finally, we computed the consensus proportion for each cell type. If a cell type is estimated by a single method out of the 7 described above, the consensus proportion is the prediction of this method. Otherwise, we performed a weighted average of all methods predicting this cell type, the weights being the normalized inverse of the RMSEs.

Execution of HDMAX2-surv on the TCGA-PAAD cohort

`HDMAX2-surv_step1`. We ran `HDMAX2-surv.aalen` with cohort, age at diagnosis, gender, and grade as observed confounding factors, and we estimate K unobserved latent factor. The optimal K had been estimated using Cattell's rule applied to the eigenvalues of principal component analysis (PCA) [42]. In the TCGA cohort, K was estimated to be 8 based on PCA of the residuals obtained after regressing methylation values in the adjustments factors (Supplementary Figure 9). The Latent Factor were estimated using a Latent Factor Mixed Model (LFMM) [43] :

$$M = T\alpha + \text{Imm}\,\beta^\top + \text{Cov}\,\gamma^\top + W + E \tag{1}$$

with M: the methylation matrix, T: a vector of tobacco exposure, Imm: the immune cell-type composition matrix (estimated by deconvolution), Cov: the matrix of

covariates (gender, age, tumor grade), $W = UV^\top$: the latent matrix, E: the residual noise. The K latent factors represent unobserved confounders, modeled through the matrix U, which is obtained via a singular value decomposition (SVD) of the matrix W, where V is the matrix of loadings. During the latent-factor estimation step, we incorporated immune-infiltration profiles, derived from the deconvolution pipeline (see section Quantification of immune infiltration on the TCGA-PAAD cohort), as fixed-effect covariates in the LFMM, thereby estimating latent factors conditional on these profiles. In subsequent analyses, we did not carry these immune-infiltration profiles forward: the high-dimensional mediation models were fitted without including them as covariates.

We then ran the function `run_AS_surv_param()` to evaluate the association between tobacco exposure, DNA methylation patterns, and survival outcomes, using the optimal number of latent factors K, previously determined, and the variables age at diagnosis, gender, and grade as observed adjustment factors.

AMR detection. We used a p-value aggregation approach to identify Aggregated Methylated Regions (AMR) of interest. We used `comb-p` [44, 45], a method that combines adjacent CpG p-values in sliding windows, as described in Jumentier et *al* [11] (see Supplementary Material Section 2). We defined significant regions as those containing at least two markers at a maximum distance of 1,000 bp and significant at the 5% threshold. The mean DNAm of CpGs located within each AMRs was calculated as summary metric for subsequent mediation analysis.

`HDMAX2-surv_step2`**.** In the second stage, we evaluated two scenarios. In Scenario A, we treated individual CpG sites as mediators and declared significance under false discovery rate control at 5% and 10%. In Scenario B, we first constructed AMRs and carry forward the 50 most significant AMRs from that step. We then applied the `estimate_effect_surv_param()` function to evaluate the indirect effects of AMRs in the pathway between tobacco exposure and survival, using the approach implemented in the `mediation` R package. We selected 37 significant AMRs with $P - value \leq 0.05$, which corresponds to adjusted $P - value \leq 0.065$ after Benjamini-Hochberg correction for multiple testing using the R function `p.adjust()` with `method = "BH"`.

Causal discovery framework for immune-mediated pathways

We developed a causal discovery framework to identify the relationships between tobacco exposure, DNA methylation, and immune infiltration in the survival of PDAC patients. We first verified that HDMAX2-surv latent factors were uncorrelated with tobacco exposure (all $P - value > 0.05$), allowing their exclusion from the causal graph. Using a three-step conditional independence testing procedure, we discriminated between two candidate causal graphs for each AMR-immune cell pair. Given the exploratory nature of the causal discovery framework, we used a liberal significance threshold ($P < 0.15$) for conditional independence tests to maximize the sensitivity in identifying potential causal relationships. The final causal inferences were interpreted within the context of biological plausibility and effect size magnitudes.

Variable definition and causal framework. We investigated the causal relationships between tobacco exposure and patient survival through immune-mediated pathways using a directed acyclic graph (DAG) framework. The analysis included six key variables: tobacco exposure status (T, binary: smoker vs. non-smoker), DNA methylation levels of significant AMRs (A, continuous), total immune infiltration proportion (I_{all}), B cell infiltration proportion (I_B, continuous), T cell infiltration

proportion (I_T, continuous), and survival outcome (S, continuous, measured as accelerated failure time).

Latent factor assessment. To ensure the validity of our causal framework and avoid spurious confounding relationships, we systematically evaluated the associations between HDMAX2-surv-derived latent factors (LFs, 8 components of the matrix U) and both immune infiltration variables and AMRs. This analysis aimed to identify latent factors that could represent unmeasured confounders in the causal pathways under investigation.

First, we tested the correlation between tobacco exposure and each of the eight LFs using Pearson correlation coefficients. All latent factors showed non-significant correlations with tobacco exposure (Pearson correlation P-values: 0.8216, 0.8814, 0.4849, 0.2818, 0.8056, 0.8428, 0.0891, and 0.3950 for LFs A to H, respectively; all $P - value > 0.01$,). Second, we tested the association of LFs with patient survival (Cox model P-values: 0.9403, 0.6609, 0.7011, 0.0619, 0.4923, 0.3478, 0.4997, and 0.6096 for LFs A to H, respectively; all $P - value > 0.01$.

Finally, we computed Pearson correlation coefficients between all latent factors ($n = 8$) and both immune infiltration variables ($n = 6$: total immune, B cells, T cells, DCs, NK cells, and macrophages) and AMRs ($n = 50$) (Supplementary Figure 14). Statistical significance was assessed using Pearson correlation tests ($P - value < 0.01$), with a Bonferroni correction for multiple testing. To identify latent factors that could act as confounders in AMR-immune relationships, we systematically examined all possible AMR-immune pairs ($n = 50 \times 6$ combinations). For each pair, we identified latent factors that simultaneously exhibited significant correlation with both the AMR and the immune variable. Such latent factors were flagged as potential confounders and accounted as covariates in the later models.

Candidate causal graphs. Based on biological prior knowledge and preliminary association analyses, we defined two candidate causal graphs:

- **Graph 1**: $T \to A \to I \to S$ with additional direct effect $A \to S$
- **Graph 2**: $T \to A \to I \to S$ without direct effect $A \to S$

Systematic edge-removal procedure. We implemented a three-step procedure to identify the most plausible causal structure for each AMR-immune cell pair:

Step 1: Unconditional independence testing Starting from a fully connected graph, we tested unconditional independence between all pairs of variables using appropriate statistical models:

- Tobacco-Survival association: Cox proportional hazards model
- Tobacco-Immune associations: Linear regression models
- Tobacco-AMR associations: Linear regression models
- AMR-Survival associations: Cox proportional hazards models

All models were adjusted for potential confounders: age at diagnosis (continuous), sex (binary), and tumor grade (ordinal). Cox proportional hazards models identified immune variables significantly associated with survival: $P_{\text{B cells}} = 0.12$, $P_{\text{T cells}} = 0.14$, and $P_{\text{Tot.Imm.}} = 0.14$. Edges between unconditionally independent variables ($P - value > 0.15$) were removed from subsequent analyses. Additionally, we filtered out AMRs that did not show significant associations with either smoking (linear model,

$P - value > 0.15$) or survival (Cox model, $P - value > 0.15$), as these AMRs would not contribute meaningfully to the mediation pathways under investigation.

Step 2: conditional independence testing To discriminate between Graph 1 and Graph 2, we tested whether AMRs retained a direct causal effect on survival after conditioning on immune infiltration. For each AMR-immune pair, we fitted the linear model:

$$S = \beta_0 + \beta_1 I + \beta_2 A + \boldsymbol{\gamma}^T \mathbf{X} + \epsilon \tag{2}$$

where $\mathbf{X}$ represents the covariate vector (age, sex, grade), and $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

We retained only AMR-immune pairs where the immune variable showed significant association with survival ($P < 0.15$). The causal graph was determined based on the significance of the AMR coefficient (β_2):

- If β_2 remained significant ($P < 0.15$): Graph 1 (direct effect $A \to S$ preserved)
- If β_2 became non-significant ($P \geq 0.15$): Graph 2 (conditional independence $A \perp S \mid I$)

Step 3: edge orientation Causal directions were established using biological prior knowledge for most edges:

- $T \to A$: Tobacco exposure precedes and causally affects DNA methylation
- $A \to S$ and $I \to S$: Molecular changes affect survival outcomes

To orient the edge between AMRs and immune infiltration ($M \leftrightarrow I$), we tested whether AMRs act as colliders by fitting:

$$T = \alpha_0 + \alpha_1 I + \alpha_2 A + \boldsymbol{\delta}^T \mathbf{X} + \epsilon \tag{3}$$

The edge was oriented as $A \to I$ when the immune variable remained conditionally independent of tobacco given the AMR ($T \perp I \mid A$, i.e., α_1 non-significant at $P - value < 0.15$).

Serial mediation analysis

We analysed serial mediation between selected AMR pairs and immune-cell–type proportions by adapting the extended imputation approach of Zugna et al. [46] to accommodate time-to-event outcomes.

Guided by the DAG validated in the previous section, we decomposed the effect of tobacco on survival into direct, methylation-mediated, and immune-mediated components. We adopted an extended imputation strategy without inverse-probability weighting, which is equivalent to the sequential plug-in g-formula described by Daniel et al. [47] but avoids the variance inflation associated with mediator weights. We chose a non-weighted approach because we performed extensive causal discovery before running serial mediation analysis. Therefore, we made the assumption that we identified the entire set of possible causal chains, with no back-door path—i.e., no alternative route linking tobacco to the mediators through a shared confounder [48].

Adjustment for latent factors. To account for unmeasured confounding detected in the *latent-factor assessment* (Section Latent factor assessment), we refined the serial-mediation pipeline as follows: for each selected AMR–immune pair, latent factors that are significantly correlated with both the AMR and the immune profile are added as covariates in the immune-mediator regression (regression I). When no such latent factor is identified, regression I keeps its original specification.

Step 1: Outcome model. For each AMR/immune pair we fitted an accelerated-failure-time model

$$\log S = \gamma_0 + \gamma_T\, T + \gamma_A\, A + \gamma_I\, I + \boldsymbol{\gamma}_C^\top \mathbf{C} + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

selecting the baseline distribution (Weibull, Gaussian, log-normal, logistic, log-logistic, or exponential) by minimizing the AIC.

Step 2: Serial imputation. Each patient was copied four times to represent the counter-factual scenarios required by a two-step mediation chain. Counter-factual values of methylation (A) and immune infiltration (I) were imputed from linear models

$$A = \beta_0 + \beta_T\, T + \boldsymbol{\beta}_C^\top \mathbf{C} + \eta_A,$$

$$\text{Regression } I: \quad I = \alpha_0 + \alpha_T\, T + \alpha_A\, A + \begin{cases} \boldsymbol{\alpha}_C^\top \mathbf{C} + \eta_I, & \text{(no LF)} \\ \boldsymbol{\alpha}_U^\top \mathbf{U}^* + \boldsymbol{\alpha}_C^\top \mathbf{C} + \eta_I, & \text{(with LF)} \end{cases}$$

where $\mathbf{U}^*$ is the subset of latent factors simultaneously correlated with the AMR and the immune variable under study.

Step 3: Effect decomposition. The imputed survival times were regressed in the indicator variables (a_3, a_1, a_2) representing tobacco status at each stage of the counter-factual sequence, yielding the interventional analog components:

$$\text{IA–DE} = \theta_{a_3}, \qquad \text{IA–IE}_A = \theta_{a_1}, \qquad \text{IA–IE}_I = \theta_{a_2}, \qquad \text{IA–TE} = \theta_{a_3} + \theta_{a_1} + \theta_{a_2}.$$

These coefficients are interpreted as time-ratio differences on the AFT scale. The interventional analog effect answers how much longer patients would live if they all stopped smoking while keeping the methylation and immune profiles they naturally exhibit as non-smokers.

Uncertainty. Standard errors and 90 % percentile confidence intervals were computed from *1 000* bootstrap samples resampled at the patient level.

Statistical and bioinformatics analysis

Survival models. The association between smoking status and patient survival was evaluated using Cox proportional hazard models, using the `survival::coxph()` function of the 'survival' R package, with default parameters (Efron approximation to handle ties). The time at study entry was age at recruitment, while the exit time was age at death, loss, or end of follow-up, whichever came first. All Cox proportional hazard models were stratified by tumor grade, sex, and age at recruitment. Kaplan-Meier survival curves were generated using the `survminer::ggsurvplot()` function of the 'survminer' R package. Survival time was defined as the time from diagnosis to death or last follow-up.

Correlation analysis. Pearson correlation analysis were performed using the `cor.test` function with default parameters.

Bioinformatics analysis We used Illumina chip annotation file to match the CpGs in an AMR to genes.

LLM usage

Language corrections and stylistic improvements were assisted by large language models (LLMs).

Results

HDMAX2-surv: a robust method to perform high dimension mediation analysis for survival data

HDMAX2-surv is a high-dimensional mediation analysis method specifically designed for survival outcomes (detailed methodology in Materials and methods). Our framework follows the directed acyclic graph illustrated in Figure 1A, where we evaluate the causal effect of tobacco exposure (X) on survival outcome (Y) and quantify mediation through high-dimensional DNA methylation patterns (M). The method accounts for both observed and unobserved confounding factors through a latent factor approach. HDMAX2-surv extends the original HDMAX2 method [**?**, 11] by implementing a two-step approach: (i) feature selection of potential mediators using the max-squared ($\max^2$) test, and (ii) mediation analysis through a counterfactual framework to evaluate indirect effects. While the original HDMAX2 offers computational advantages—including rapid processing, accommodation of high-dimensional mediator sets, and incorporation of latent factors, its application was previously limited to continuous and binary outcomes. A critical challenge in survival mediation analysis is the violation of proportional hazards assumptions when mediator variables are incorporated into Cox models. Following recommendations from Lapointe-Shaw et al. [14], we implemented two alternative survival modeling approaches within the HDMAX2-surv framework (Figure 1B): (i) `HDMAX2-surv.aalen` incorporates Aalen additive hazards models into both HDMAX2-surv steps, building upon implementations by Domingo-Relloso et al. [33], and (ii) `HDMAX2-surv.param` implements parametric accelerated failure-time (AFT) models. The Aalen approach relaxes proportional hazards assumptions by allowing hazard ratios to vary over time through additive rather than multiplicative effects, whereas the AFT approach models survival time directly and completely avoid proportional hazards assumptions. We developed a comprehensive simulation framework to evaluate performance and enable comparison with existing approaches (Figure 1C). Our simulations generated realistic datasets with correlated DNAm blocks, binary exposure, and Cox-modeled survival outcomes, where consecutive CpG blocks served as causal mediators linking exposure to survival hazard. To avoid bias, survival data were simulated using Cox models that differed from the survival models implemented in HDMAX2-surv. We systematically evaluated three methods: `HDMAX2-surv.aalen`, `HDMAX2-surv.param`, and `HIMA` [21], a high-dimensional mediation method that performs feature selection using sure independence screening followed by debiased LASSO and joint significance testing. Our simulation campaign examined nine distinct analytical pipelines combining different selection (Step 1) and estimation procedures (Step 2), with systematic variation of effect sizes and sample sizes to assess method performance across realistic parameter ranges.

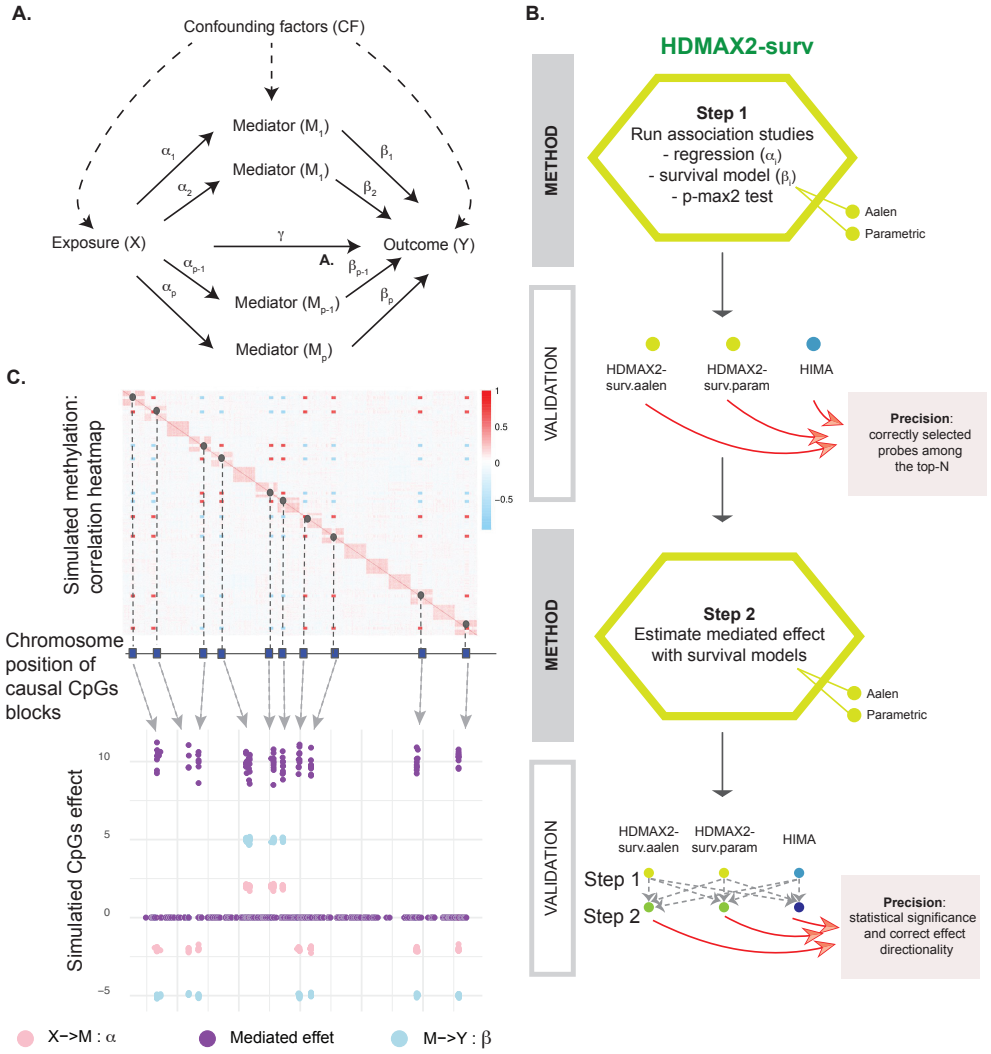

Fig 1. HDMAX2-surv: theoretical model, algorithm, and validation design. **(A)** Directed acyclic graph describing the high-dimensional mediation framework with an exposure X, p potential mediators M (DNAm CpG probes), a survival outcome Y, and confounding factors CF (observed and unobserved). **(B)** Schematic representation of the HDMAX2-surv method (Step 1: selection of candidate mediators; Step 2: mediation effect estimation), with a two-step evaluation workflow based on estimation precision using simulated data with known ground truth. `HDMAX2-surv.aalen` and `HDMAX2-surv.param` were compared to `HIMA`, a method designed for high-dimensional survival mediation analysis. **(C)** Schematic representation of the simulation framework. Heatmap shows pairwise correlations among simulated DNAm probes, illustrating the dependence structure imposed in the simulated data (blocks of correlated CpGs). We simulated genomic positions on a virtual chromosome. For selected blocks of causal CpGs (dark blue squares), corresponding to true mediators in our simulations, we assigned positive mediation effects with identical effect values for all CpGs within each block (purple). As we used Cox models to simulate survival outcomes, positive mediated effects are interpreted as deleterious for survival. This mediated effect represents the product of simulated α effects (exposure $\rightarrow$ mediator, pink) and β effects (mediator $\rightarrow$ outcome, blue).

Evaluation of survival model performances

We evaluated the performance of the method using precision scores at both Step 1 and Step 2 of the framework. The results are presented for true positive count (correctly selected probes among the top-m_{causal}) at Step 1 (Figure 2A) and effect prediction accuracy (statistical significance and correct effect directionality) at Step 2 (Figure 2B). Our simulation campaign revealed consistent patterns across the parameter grid. Figure 2A and Supplementary Figure 3 show precision scores in the selection stage in five effect sizes (μ_α and μ_β: 0.05, 0.1, 0.5, 2, and 5), two sample sizes (150 or 300 individuals) and among 10 000 probes. As expected, performance of all methods increased with larger effect sizes and sample sizes. Across most scenarios, `HIMA` and `HDMAX2-surv.param` performed slightly better than `HDMAX2-surv.aalen`. At this stage, no clear advantage emerged between `HIMA` and `HDMAX2-surv.param`. We then evaluated precision scores at Step 2, which assesses whether selected CpGs both pass the significance threshold and show effect directions consistent with the simulated mediation effects (Figure 2B and Supplementary Figure 4). To perform fair and exhaustive comparison, we combined Step 2 methods with each Step 1 result, across the same parameter variations. We observed the same general trend with improved method performance as the effect size or the sample size increased. Notably, `HDMAX2-surv.param` Step 2 consistently outperformed other approaches, regardless of the Step 1 method with which it was combined. `HIMA` exhibited lower precision scores, likely due to its more conservative P-value outputs, which limit detection even after adjusting selection thresholds. Computational time analysis revealed that `HDMAX2-surv.aalen` was substantially slower than both `HDMAX2-surv.param` and `HIMA`, whose runtimes were roughly equivalent (Supplementary Figure 5).

As a control, we examined the behavior of each method in detecting false positives by running simulations with either $\alpha = 0$ (Supplementary Figure 3) or $\beta = 0$ (Supplementary Figure 4). When $\beta = 0$, `HIMA` displayed a high false positive rate, while smaller numbers of false positives were observed in `HDMAX2-surv` methods. These false positive rates are explained by the feature selection approach, which ranks potential mediators using scores that combine both pathway effects: $|\alpha_i \beta_i|$ for `HIMA` and $\max\{P_\alpha, P_\beta\}^2$ for `HDMAX2-surv`. Consequently, a strong effect on one pathway (α or β) can drive the ranking even when the other pathway has no effect (Supplementary Figure 8). Our observations indicate that HIMA is more susceptible to this issue than HDMAX2-surv, likely due to its multiplicative scoring approach that amplifies single-pathway effects.

Overall, `HDMAX2-surv.param` delivered the best performance: it matched `HIMA` in Step 1 selection and computational efficiency while providing substantially higher accuracy in effect estimation and lower false positives. Based on these simulation results, we selected `HDMAX2-surv.param` for downstream analysis of real data.

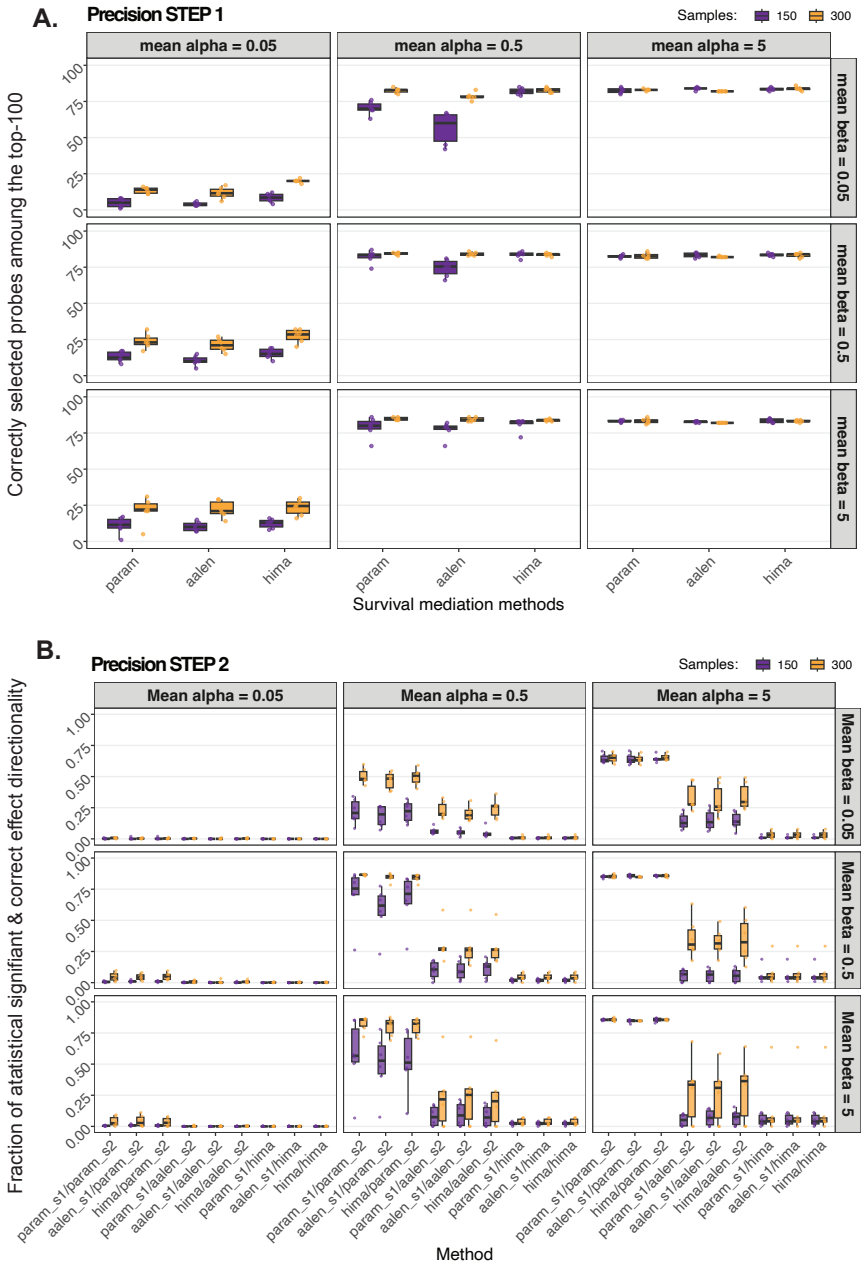

Fig 2. Performance evaluation of high-dimensional survival mediation methods. (A) Precision of methods in Step 1 (feature selection). True positive count among selected mediators, representing the number of correctly identified causal CpGs among the top-m_{causal} ranked by $|\alpha\beta|$ for HIMA and $\max\{P_\alpha, P_\beta\}^2$ for HDMAX2-surv. **(B)** Precision of methods in Step 2 (mediation analysis). Proportion of probes selected in Step 1 whose mediation effects in Step 2 exhibit both statistical significance and correct effect directionality. For both **(A)** and **(B)**, simulation parameters: $m_{\text{causal}} = 100$ causal probes among $p = 10\,000$ total CpGs. Exposure $\rightarrow$ mediator effects: $\mu_\alpha \in \{0.05, 0.5, 5\}$, $\sigma_\alpha = 0.1$. Mediator $\rightarrow$ outcome effects: $\mu_\beta \in \{0.05, 0.5, 5\}$, $\sigma_\beta = 0.1$. Overlap proportion: $\pi_{\text{overlap}} = 0.8$. Two sample sizes are shown ($n = 150$, purple; $n = 300$, orange) across 6 replications.

Variables	n (%)
Cohort	112 (100%)
Age	65 (Q1=57, Q3=74)
Sex	
Female	55 (49%)
Male	57(51%)
Tumor grade	
I	5 (4.4%)
II	50 (44.6%)
III	56 (50%)
IV	1 (0.8%)
Smoking	
Smoker (a and b)	37 (33%)
(a) Current Smoker	16 (14%)
(b) Reformed Smoker for $\leq$ 15 yrs)	21 (19%)
Non smoker (c and d)	75(67%)
(c) Reformed Smoker for $>$ 15 yrs	24 (21%)
(d) Lifelong Non-Smoker	51 (46%)
Deaths	59 (53%)

Table 1. Descriptive characteristics of the TCGA-PAAD cohort (n = 112). Age is reported as median with first and third quartiles (Q1–Q3). Sex, tumor grade, smoking status, and number of deaths are presented as counts and percentages.

Mediation of smoking status on PDAC survival

Although the relationship between tobacco exposure and PDAC incidence has been extensively characterized, the causal mechanisms linking smoking with the survival outcome in PDAC patients remain poorly understood. To address this knowledge gap, we investigated the causal relationship between tobacco exposure and PDAC patient survival, specifically examining how DNAm patterns mediate these effects. We used the comprehensive multi-omic dataset from The Cancer Genome Atlas (TCGA [49]), which provides integrated genomic, transcriptomic, and epigenetic profiles for PDAC patients along with detailed clinical characteristics, including smoking status, age, sex, tumor grade, and survival outcome (Table 1). Our analysis focused on a subset of 112 patients with confirmed PDAC diagnosis, as previously established in [34]. The composition of the cohort and clinical characteristics are summarized in Table 1 and illustrated in the Supplementary Figure 1. Detailed descriptions of clinical variables and inclusion criteria are provided in the Section Materials and methods.

We applied the `HDMAX2-surv.param` method to characterize the causal relationships between tobacco exposure, DNA methylation, and patient survival. The high-dimensional mediation analysis identified 10 mediating CpG sites with a false discovery rate (FDR) of 10% to ensure robust control of multiple testing correction in a total of 371,033 CpG sites analyzed (Figure 3A (Top) and Supplementary Figure 10). To enhance biological interpretability, we constructed Aggregated Mediator Regions (AMRs) by combining $P - \max^2$ values from individual CpGs identified in Step 1 of `HDMAX2-surv.param` (Figure 3A (Bottom)). From the top 50 candidate regions, we identified 36 AMRs with statistically significant Average Causal Mediation Effects

(ACMEs) (Figure 3B). The identified AMRs exhibited heterogeneous effects on patient survival. We identified 5 deleterious AMRs and 31 protective AMRs. Detailed characteristics of AMR with significant ACME are presented in Table 2, including AMR size, associated CpG sites, proximal genes, and number of pubmed hits associating the proximal gene and pancreatic cancer (Figure 3C). Since a specific parametric distribution must be specified in the AFT model for each AMR, ACME values are not strictly comparable across AMRs when different distributional assumptions (weibull, log-normal, log-logistic) are applied. Notably, some AMRs show substantial variability in effect size estimates, spanning several orders of magnitude with correspondingly wide confidence intervals. For example, AMR50 yields an ACME of -123.5 [95% CI: -285.1, -6.4]. Such wide intervals may reflect convergence issues in the bootstrap procedure and the limited sample size, underscoring the preliminary nature of these estimates. This variability highlights the exploratory scope of our analysis and the critical importance of experimental validation. While the direction of effects may be more reliable than their precise magnitudes, future studies with larger sample sizes will be required to refine these estimates and reduce uncertainty.

Among the significant AMRs, we identified four genes previously associated with PDAC occurrence or outcome, defined as having more than two prior citations in PubMed (Table 2): *MSX2*-AMR2, a member of the homeobox gene family, a target of the RAS signaling pathway, expressed in pancreatic cancer and related to chemoresistance [50, 51]; *KCNQ1*-AMR5, a potassium channel gene subject to tissue-specific imprinting and found to be differentially methylated in pancreatic cancer [52]; *PIWIL1*-AMR17, a member of the Argonaute protein family, implicated in pancreatic fibrosis and the regulation of pancreatic cancer [53, 54]; and *EPHA3*-AMR37, part of the ephrin receptor family, which contributes to glucose homeostasis in insulin-producing pancreatic cells [55](see Supplementary Figure 11).

These four AMRs demonstrated distinct mechanistic patterns, yet overall showed a protective effect on survival. AMR2 exhibited an ACME of 283.7 [95% CI: 30.6, 830.0] and AMR37 an ACME of 187.1 [95% CI: 1.9, 502.2]. In both cases, tobacco exposure globally increased methylation at the locus in smokers compared to non-smokers. This hypermethylation shifts the survival curve forward, thereby improving survival through a protective mediation pathway with synergistic positive effects. In contrast, AMR5 (123.6 [95% CI: 19.3, 299.3]) and AMR17 (87.7 [95% CI: 1.2,235.3]) showed the opposite trend: tobacco exposure decreased methylation, and methylation at these loci was negatively associated with survival. Consequently, hypomethylation conferred an overall protective effect, mediated through synergistic negative effects. However, not all effects were protective. In some cases, antagonistic patterns resulted in deleterious outcomes, such as AMR38, where tobacco-induced hypomethylation shifts the survival curve backward and reduced survival, or AMR25, where tobacco-induced hypermethylation had also a deleterious impact. These findings reveal complex, context-dependent relationships between tobacco exposure and survival outcomes, suggesting that DNA methylation alterations may both amplify and attenuate tobacco-related survival effects depending on the specific genomic locus. Notably, smoking had no significant impact on gene expression for these four genes (Supplementary Figure 12, Wilcoxon test at $\alpha = 5\%$), highlighting the unique value of methylation-based mediation approaches in high-dimensional settings. Beyond these established candidates, the biological mechanisms underlying the remaining AMRs represent promising avenues for future mechanistic investigation.

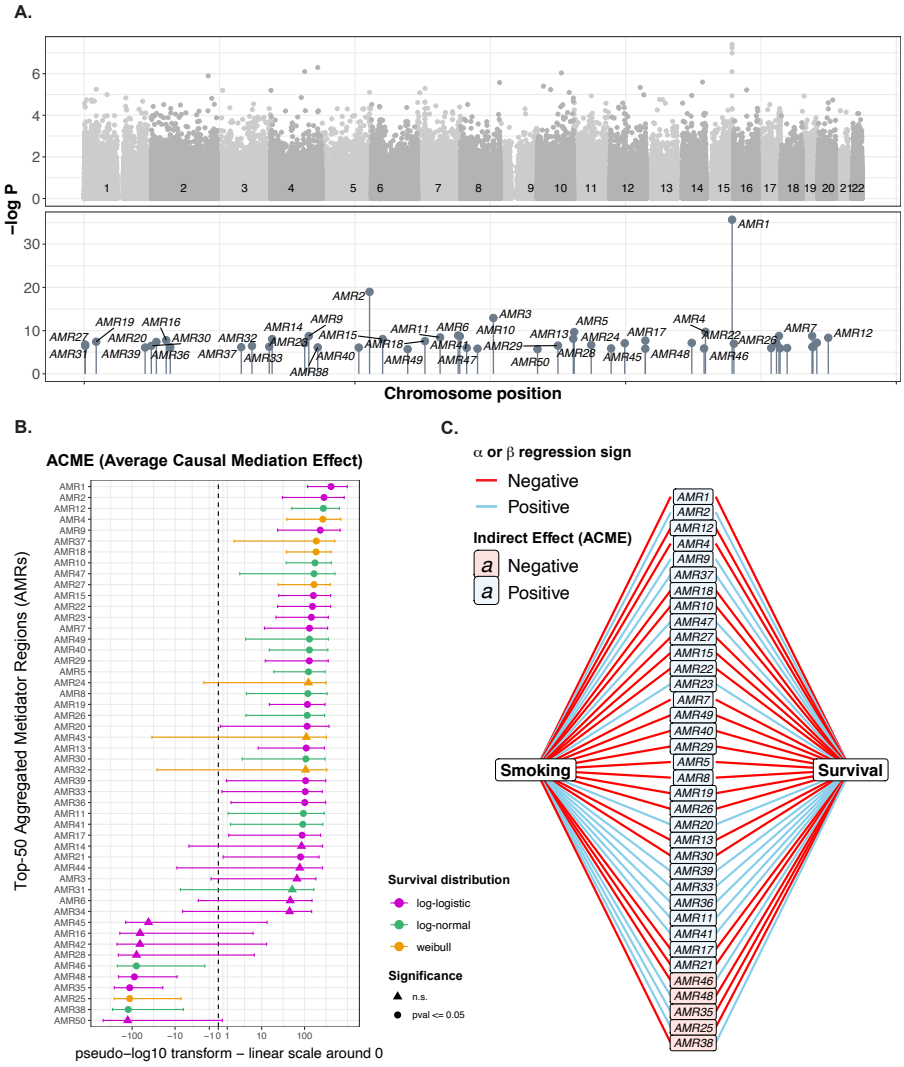

Fig 3. Mediation analysis of the effect of tobacco on survival. (A) Top: Manhattan plot of CpG $-\log_{10}(p)$ values from HDMAX2. Bottom: Manhattan plot of $-\log_{10}(p)$ values for candidate AMRs (37 hits). **(B)** Indirect effects (ACME) for each AMR; point= estimate, bar= 90 %CI. Data were transformed using the `transform_pseudo_log` of the ‘scales‘ R package (base = 10, sigma = 1), switching to a linear scale for absolute values below 1. Colors indicate the survival distribution retained in the AFT models, for each AMR. **(C)** Decomposition of significant ACME of smoking on survival: 31 AMRs show a positive indirect effect and 5 a negative effect. Boxes and segment colours reflect the sign of the ACME and indicates the direction of association (blue = protective, red = deleterious). *Note:* ACME, average causal mediation effect; AMR, aggregated mediator region; CI, confidence interval; HDMAX2, high-dimensional mediation analysis; AFT, Accelerated Failure Time.

AMR	Chromosome	Size (pb)	CpGs	Genes	ACME_CI	P-value	Pubmed Hits
AMR1	chr15	2906	cg00032912, cg00033666, cg00436496, cg02326806, cg05289920, cg07116712, cg07959469, cg09452082, cg10151901, cg10186131, cg17140992, cg17454263, cg18761422, cg20037575, cg22460896, cg22703162, cg22840361, cg24491784	RP11-262	410.6 [117.2,976.1]	0.000	0
AMR2	chr5	1321	cg01960016, cg02365742, cg06946371, cg12387713, cg15123984, cg17156565, cg20563910, cg25998402, cg26615160, cg26796283, cg27096144	MSX2	283.7 [30.6,830.0]	0.012	9
AMR4	chr14	683	cg02814525, cg09255505, cg14163691, cg14918743, cg19515081	ZBTB42	264.6 [38.8,687.4]	0.004	0
AMR5	chr11	670	cg00283576, cg01178971, cg03371125, cg04104132, cg04902871, cg10678459, cg11025829, cg11363972, cg15932136, cg19698309, cg19779211, cg20533553, cg23267890, cg23623667, cg24079038	KCNQ1	123.6 [19.3,299.3]	0.014	4
AMR7	chr17	299	cg10997810, cg15229994, cg16293484, cg19948701, cg20030711	SEPT9	130.1 [11.6,342.4]	0.028	1
AMR8	chr19	441	cg02587316, cg04332534, cg05020604, cg18630667, cg22517995, cg25397945	ZNF382	119.7 [4.2,334.8]	0.040	2, 0
AMR9	chr4	609	cg00169305, cg01725318, cg02651961, cg03168293, cg03575792, cg03710295, cg08129449, cg08821557, cg11539424, cg21246783, cg24695316, cg25323711	CLGN	231.7 [23.6,669.4]	0.020	0
AMR10	chr7	680	cg09016797, cg12712184, cg13104880, cg14635659, cg14938587, cg21428990		174.2 [37.4,416.3]	0.008	
AMR11	chr7	359	cg00702008, cg06129054, cg06994813, cg13909585, cg15478390, cg17137182, cg17611951, cg24399959	PCLO	93.7 [1.1,285.2]	0.046	1
AMR12	chr20	219	cg06198069, cg11252765, cg22582721	BLCAP, NNAT	272.3 [50.7,640.7]	0.006	0, 0
AMR13	chr10	879	cg02676262, cg05477444, cg15847996, cg24571871	INPP5A	109.3 [8.3,288.6]	0.024	0
AMR15	chr6	553	cg01350077, cg01883425, cg05345286, cg06688989, cg27200446	MDFI	160.3 [25.2,403.8]	0.008	2
AMR17	chr12	155	cg06365016, cg12020444, cg16244664, cg27261050	PIWIL1	87.7 [1.2,235.3]	0.044	4
AMR18	chr7	288	cg05767159, cg17400476	RP1-170O	185.2 [38.1,410.7]	0.002	0
AMR19	chr1	342	cg04486940, cg14210311, cg16681083, cg26233866	ERI3	116.7 [15.2,297.4]	0.016	0
AMR20	chr2	285	cg03002352, cg04804377, cg10167837, cg15124757, cg21642176, cg25565383, cg26119367	RAD51AP2	114.7 [0.2,368.8]	0.050	0
AMR21	chr19	227	cg03657031, cg06697294, cg11662224, cg21950287, cg23460210, cg27552287	PRKCG	80.7 [0.5,216.5]	0.046	1
AMR22	chr14	155	cg02054667, cg11051738		152.5 [23.7,402.9]	0.010	
AMR23	chr4	92	cg04373334, cg08575049	FAT4	143.9 [21.7,355.6]	0.006	1
AMR25	chr17	619	cg00117908, cg07004514, cg08901098, cg09980477, cg11102858, cg13173111, cg15253816, cg25456728	ARSG	-114.3 [-260.1,-7.3]	0.032	1
AMR26	chr16	286	cg02612270, cg06430465, cg10244976, cg13706784	LMF1	115.2 [4.2,290.4]	0.032	0
AMR27	chr1	227	cg01101873, cg06060874, cg27187555	PRDM16	165.7 [24.5,396.5]	0.010	1
AMR29	chr10	106	cg02987635, cg14207539, cg23264016	C10orf11	128.8 [12.2,354.3]	0.018	0
AMR30	chr1	181	cg07533148, cg20146541, cg20810478, cg23054189, cg26157385	TRIM58	107.2 [3.3,297.9]	0.034	1
AMR33	chr3	131	cg08002883, cg10511904, cg16376000, cg17393267	FGF12	103.6 [0.4,258.1]	0.048	0
AMR35	chr19	253	cg01174276, cg02671333, cg12029794, cg23854456	CTC-246B	-114.0 [-260.3,-19.4]	0.006	0, 0
AMR36	chr2	388	cg02304863, cg02518161, cg06870284, cg08194677, cg12784598, cg13165472, cg16675700, cg18564808	ANTXR1	101.0 [1.5,308.7]	0.046	0
AMR37	chr3	353	cg02517134, cg05676541, cg10315231, cg18055394, cg22619563	EPHA3	187.1 [1.9,502.2]	0.050	3
AMR38	chr4	332	cg10024909, cg15234400, cg16919569, cg21382382	HAND2-AS	-123.5 [-285.1,-6.4]	0.040	
AMR39	chr1	429	cg03199058, cg06802365, cg10174867, cg11804724, cg20492807, cg25379026	DNAH14	105.5 [0.9,309.3]	0.048	0
AMR40	chr5	151	cg04223420, cg07274716, cg22827250	PITX1	129.5 [15.2,347.1]	0.006	1
AMR41	chr8	381	cg03979241, cg04662594, cg14837598, cg16678564, cg26392737	DMTN	91.0 [1.4,264.6]	0.048	0
AMR46	chr14	210	cg02005147, cg02867079, cg03475665, cg06834875, cg06969206, cg09956302, cg12188268, cg26669257	HHIPL1	-79.7 [-220.5,-1.6]	0.042	0
AMR47	chr8	98	cg01906015, cg06410746, cg08855288, cg15858239	NKAIN3	165.9 [2.8,506.9]	0.038	0
AMR48	chr12	472	cg01802493, cg05863637, cg06153623, cg16236851, cg17226286	RP11-474	-88.6 [-207.5,-9.0]	0.036	0, 0
AMR49	chr6	111	cg17284168, cg18566177, cg21366688	SGK1	130.1 [4.1,356.1]	0.024	2

Table 2. Significant mediator-associated methylation regions (AMRs). Each AMR is defined by its chromosomal location, size (in base pairs), included CpG sites, annotated genes, the estimated Average Causal Mediation Effect (ACME) with 95% confidence interval, and the associated p-value. The "Pubmed Hits" column shows the number of PubMed articles retrieved with the query: *Humans[Mesh] AND (pancreatic cancer[MeSH]) AND GENE*. Empty cells indicate missing gene annotation or no publication found.

Causal discovery analysis reveals the immune role in Tobacco-Survival relationships

Previous studies indicate that immune infiltration might play role in mediating the effects of tobacco exposure on PDAC patient survival [3]. To investigate this hypothesis, we quantified immune cell infiltration proportions within the tumor microenvironment using computational deconvolution approaches. We developed a consensus deconvolution method that integrates both bulk and single-cell reference datasets to enhance accuracy and robustness. Our approach combines three well-validated methods previously benchmarked in comparative studies: (i) bulk reference-based: Robust Linear Regression (RLR) method [41]; and (ii) single-cell reference-based: Single-Cell Deconvolution (SCDC) [38], InstaPrism [39], and MuSiC algorithms [40]. This consensus approach allowed us to quantify the composition of tumor microenvironment across 9 distinct cellular components, including 5 immune cell types: B cells, T cells, dendritic cells (DCs), natural killer (NK) cells, and macrophages (Figure 4A).

Cox proportional hazards modeling revealed significant associations between several immune components and patient survival outcomes (P-value threshold = 0.15): (i) total immune infiltration (sum of all five immune components; HR = 0.85 per 1% increase in immune fraction [95% CI: 0.69, 1.056], P-value = 0.1442), (ii) B cells (HR = 0.48 per 1% increase in B cell fraction [95% CI: 0.19, 1.22], P-value = 0.1229), and (iii) T cells (HR = 0.58 per 1% increase in T cell fraction [95% CI: 0.28, 1.19], P-value = 0.1402) (Figure 4B). All models were adjusted for observed confounding factors: age at diagnosis, sex, and tumor grade. These findings suggest that the composition of immune microenvironment significantly influences PDAC patient prognosis, which warrants further investigation of potential mediation mechanisms.

To elucidate the role of immune infiltration in the causal pathway linking tobacco exposure to patient survival, we investigated whether immune cell proportions serve as additional mediators along with DNA methylation. Specifically, we focused on B cells and T cells given their significant associations with survival outcomes. We implemented a causal discovery framework based on conditional independence tests to identify causal relationships among the following variables: tobacco exposure (smoking, T), mediating AMRs (A), total immune infiltration (I_{all}), B cell infiltration proportion (I_B), T cell infiltration proportion (I_T), and survival (S) (Figure 4C). To distinguish spurious associations from genuine causal relationships and determine the directionality of the causal graph, we combined biological prior knowledge (established mechanistic understanding) with conditional independence testing (statistical tests to identify dependencies and independencies among variables under different conditioning sets).

To determine which directed acyclic graph (DAG) was supported for each AMR-immune pair (including significant immune variables: I_{all}, I_B, and I_T), we employed a systematic edge-removal approach, starting from a fully connected graph (Figure 4C). First, we removed edges based on unconditional independence tests (all adjusted for age at diagnosis, sex, and tumor grade): (i) smoking and survival : $T \perp S$ (survival cox model, $P - value = 0.80760$); (ii) smoking and total immune infiltration: $T \perp I_{\text{Tot.Imm.}}$ (linear regression, $P - value = 0.513$); (iii) Smoking and B cells infiltration: $T \perp I_{\text{Bcells}}$ (linear regression, $P - value = 0.954$); and (iv) smoking and Tcells infiltration: $T \perp I_{\text{DCs}}$ (linear regression, $P - value = 0.564$). Finally, we filtered out AMRs for which significant associations were not confirmed with either smoking (linear model at $P > 0.15$) or survival (Cox model at $P > 0.15$). HDMAX2-derived Latent factors, identified as potential confounders for these specific pairs of AMR-immune, were included as additional nodes in the causal graph (C_U, in addition to the observed confounders C_O) (see Materials and methods for more details). This systematic approach ensured that our causal inferences were not confounded by unmeasured factors captured by the latent factor model.

From there, we identified two possible candidate causal graphs: Graph 1, where AMRs have causal effects on both immune content and survival; and Graph 2, where AMRs have causal effects only on immune infiltration. Then we run conditional independence tests for graph discrimination (Figure 4D). To distinguish between Graph 1 and Graph 2 for each AMR-immune pair, we tested the presence of a direct edge from AMR to survival using the linear model (1) (see the Materials and methods section). We retained only pairs in which the immune variable was significant ($P - value < 0.15$). If both M and I remained significant, we preserved the edge $A \rightarrow S$ (Graph 1). When A lost significance after controlling for the immune variable ($A \perp S \mid I$), we removed the edge $A \rightarrow S$ (Graph 2). Finally, causal directions from smoking to AMRs and from AMRs/immune variables to survival were established based on biological prior knowledge. To orient the edge between AMRs and immune variables, we tested whether AMRs act as colliders using the linear model (2) (see Section Materials and methods). We retained pairs where the immune variable remained conditionally independent of tobacco given the AMR ($T \perp I \mid A$, i.e., α_1 not significant at $P - value < 0.15$), supporting the direction $A \rightarrow I$ in both graphs. This integrated approach enables robust causal inference while accounting for the complex, multi-pathway nature of smoking's effects on PDAC prognosis. In total, 29 AMR-immune pairs supported Graph 1 and 2 pairs supported Graph 2, likely reflecting insufficient statistical power to detect real Graph 1 associations.

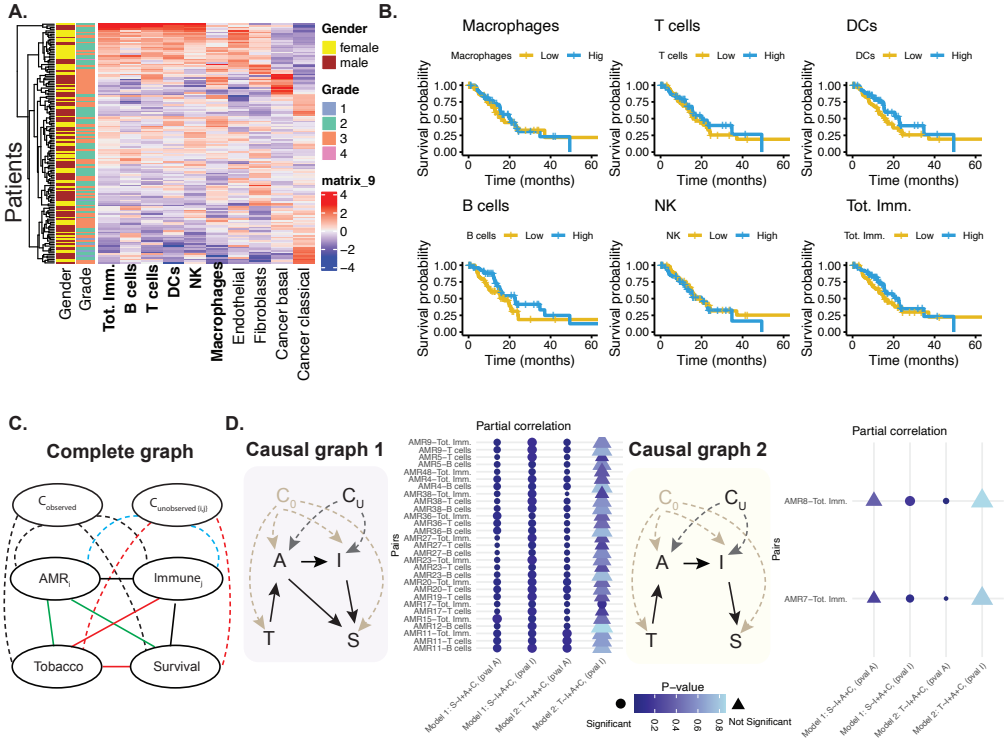

Fig 4. Immune composition, DNA methylation, and survival. (A) Heatmap shows standardized (Z-scored) deconvolution estimates from the consensus method for selected cell types in the TCGA-PAAD dataset. Neutrophils were excluded from visualization because their estimated proportions were consistently zero across all samples. Rows represent TCGA samples, hierarchically clustered based on cell-type composition. Immune cell types are indicated in bold. *Tot. Imm.* corresponds to total immune content, computed as the sum of all individual immune cell-type estimates. **(B)** Kaplan–Meier survival compare overall survival between patients with low and high levels of the immune cell type. Immune infiltration was dichotomized based on median values for each cell-type. High expression is shown in blue, low expression in orange. **(C)** Complete directed acyclic graph (DAG) framework. Observed confounders (age, sex, and tumor grade) are denoted C_O, and latent factors (LFs) representing potential unobserved confounders are denoted C_U. Red edges indicate relationships removed by unconditional independence testing: $T \perp S$, $T \perp I_j$ where $j \in \{\text{Tot.Imm.}, \text{B cells}, \text{T cells}\}$, $T \perp C_U$ (tobacco independent of latent factors) and $S \perp C_U$ (survival independent of latent factors). Green edges show relationships retained after testing: $T \not\perp \text{AMR}_i$ and $\text{AMR}_i \not\perp S$ for $i \in \{1, 2, \ldots, 36\}$. Blue edges represent latent factor confounding relationships that are specific to individual AMR-immune pair combinations. **(D)** Causal discovery results showing two candidate DAG structures. Graph 1 represents pathways where AMRs have effects on both immune infiltration and survival. Graph 2 shows pathways where AMRs affect survival solely through immune infiltration. Heatmaps display partial correlation analysis results, where each point represents an AMR-immune pair and indicates the significance of associations in conditional independence models used for causal graph discrimination. Color indicates the p-value (darker = more significant), and shape distinguishes statistically significant results (pval ≤ 0.15). Only significant immune–AMR pairs meeting selection criteria (DAG1 or DAG2) are shown. See Methods for model details.

Serial mediation analysis of AMR-Immune pathways

Next, we sought to identify whether indirect effects exist between tobacco exposure and PDAC patient survival through identified mediator chains (AMR-immune infiltration pairs identified by causal discovery), focusing our work on pairs supported by Graph 1 (see Figure 4D). In the context of multiple mediators, it is possible to decompose the total effect by considering potential mediators simultaneously, with several decomposition approaches available [47] . We adopted a counterfactual approach, adapting a recent R implementation proposed by Zugna et al. [46] to parametric survival models. We selected the "Extended imputation-based approach, estimation procedure and bootstrap without interaction between A and M1 and M2 for conditional effects" because it is the only implementation that allows decomposition of effects mediated by the first mediator (AMR, M_1) and the second mediator (immune infiltration, M_2). This sequential method first estimates the effect mediated by M_1, then estimates the joint mediated effect through both M_1 and M_2. The partial indirect effect via M_2 without passing through M_1 is then estimated using the extended imputation approach (Figure 5A). For each pair of AMR–immune, latent factors simultaneously correlated with both AMR and immune variable (Section Latent factor assessment) were included as additional covariates in the immune-mediator regression; when no latent factor met this criterion the original specification (inclusion of observed confounder age, gender and grade) was retained. We applied this method to all pairs identified as following the sequential pathway $T \to M_1 \to M_2 \to S$ in our causal discovery framework (Figure 5B-C). As expected, we observed no indirect effects via M_2 alone, since immune infiltration was not directly dependent on tobacco exposure in our dataset.

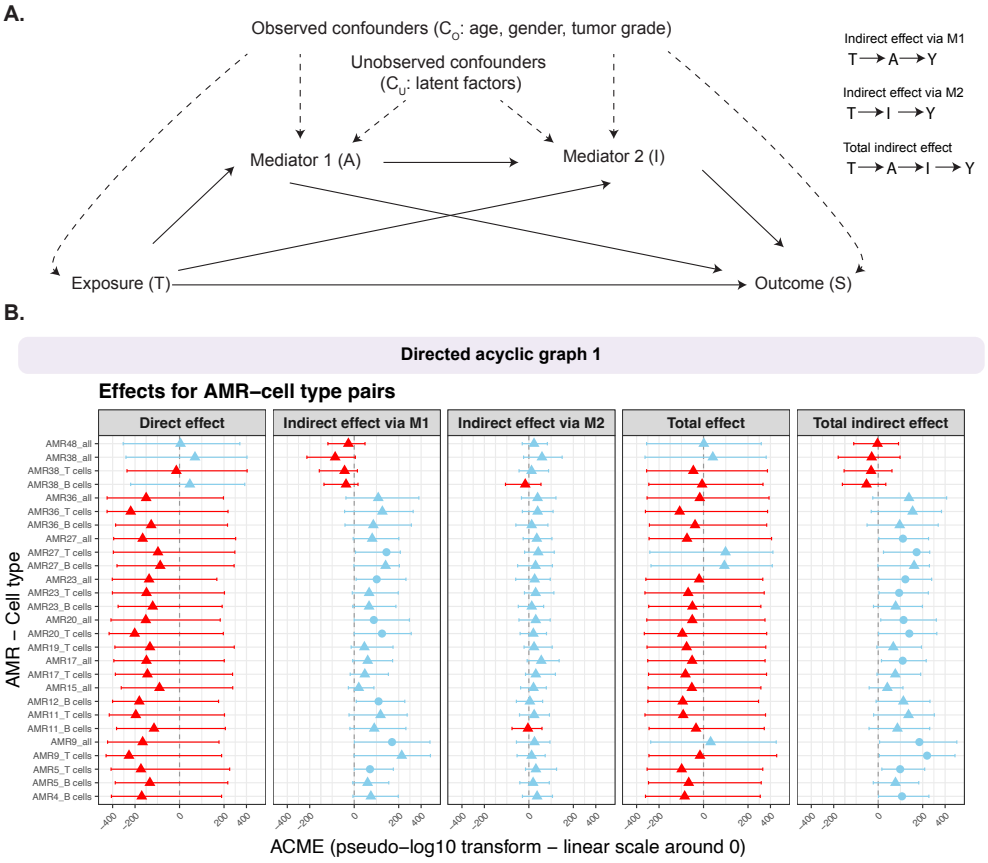

Fig 5. Estimates of total, direct, and indirect effects of smoking on PDAC patient survival through DNA methylation and immune infiltration. (A). DAG representing the studied structure. (B) Estimated effects for AMR-Immune pairs identified in DAG1. Circles represent significant effect, Triangle non significant effect, Colours represent blue for protective effects and red for deleterious effect. Each line represents 90 % percentile confidence intervals.

Characterization of protective and deleterious mediation pathways

We applied serial mediation analysis to the 29 pairs retained from the causal discovery step. Effect estimation failed for two pairs (AMR11–Total Immune and AMR4–B cells), leaving 27 pairs for testing. Among these, eleven showed statistically significant total indirect effects, including seven AMRs of interest (AMR4, AMR5, AMR9, AMR17, AMR20, AMR23, and AMR27). As *KCNQ1*-AMR5 and *PIWIL1*-AMR17 are already known candidates, we focused on the four AMRs containing more than two CpGs as particularly relevant examples (Figure 5B and Supplementary Figure 15). In subsequent analyses, we used refined smoking categories: never-smoker, former smoker ($>$ 15 years), former smoker ($\leq$ 15 years), and current smoker. Although more informative, these subcategories were not used in the mediation analysis to preserve statistical power by pooling into larger smoker vs. non-smoker groups. Remarkably, for all four selected AMRs (*ZBTB42*-AMR4, *CLGN*-AMR9, *RAD51AP2*-AMR20, and *PRDM16*-AMR27) we observed an indirect protective effect of tobacco exposure on survival (Figure 5B).

For *ZBTB42*-AMR4, the effect was mediated jointly by B cells. AMR4 spans the coding region of *ZBTB42*, a zinc finger protein family member not previously linked to

PDAC. Tobacco exposure induced slight demethylation of this constitutively methylated locus, particularly in current smokers (Figure 6A). Interestingly, *ZBTB42* has recently been identified as a prognostic factor associated with immune cell infiltration in glioma [56]. For *CLGN*-AMR9, the effect was mediated jointly by T cells. AMR9 lies in a CpG-dense promoter region of Calmegin, a testis-specific endoplasmic reticulum chaperone. While the locus is generally unmethylated, partial methylation was observed in some smoking patients (Figure 6B). Although not previously linked to PDAC, Calmegin has been proposed as a predictor of clear cell renal carcinoma [57] and is differentially methylated in metabolic dysfunction–associated steatotic liver disease [58]. For *RAD51AP2*-AMR20 and *PRDM16*-AMR27, tobacco exposure also demonstrated protective indirect effects, mediated either by the AMR alone or jointly with T-cell infiltration (Figure 6C–D). AMR20 is located in the promoter of RAD51AP2, which is constitutively unmethylated but shows partial methylation in smokers. RAD51AP2 interacts with RAD51, a key DNA repair factor. While RAD51AP2 has been implicated in cholangiocarcinoma progression [59], its role in PDAC has not, to our knowledge, been previously reported. Finally, AMR27 is located in the gene body of *PRDM16*, a zinc finger transcription factor known for recurrent translocations in acute myeloid leukemia [60], but with no previously reported association with PDAC. The locus is normally hemi-methylated, yet displays a bias toward demethylation in current smokers. Importantly, across all four genes, smoking status was not associated with significant changes in gene expression (Wilcoxon rank-sum test; Figure 6). These findings demonstrate the potential of integrating high-dimensional mediation analysis with causal discovery frameworks to uncover complex, multi-pathway mechanisms linking environmental exposures to cancer outcomes. Notably, no specific trend emerged for immune infiltration, consistent with the complexity and incomplete understanding of its role in PDAC [61]. Taken together, our results highlight the intricate interplay between tobacco-induced epigenetic modifications and immune microenvironment alterations in determining PDAC patient prognosis.

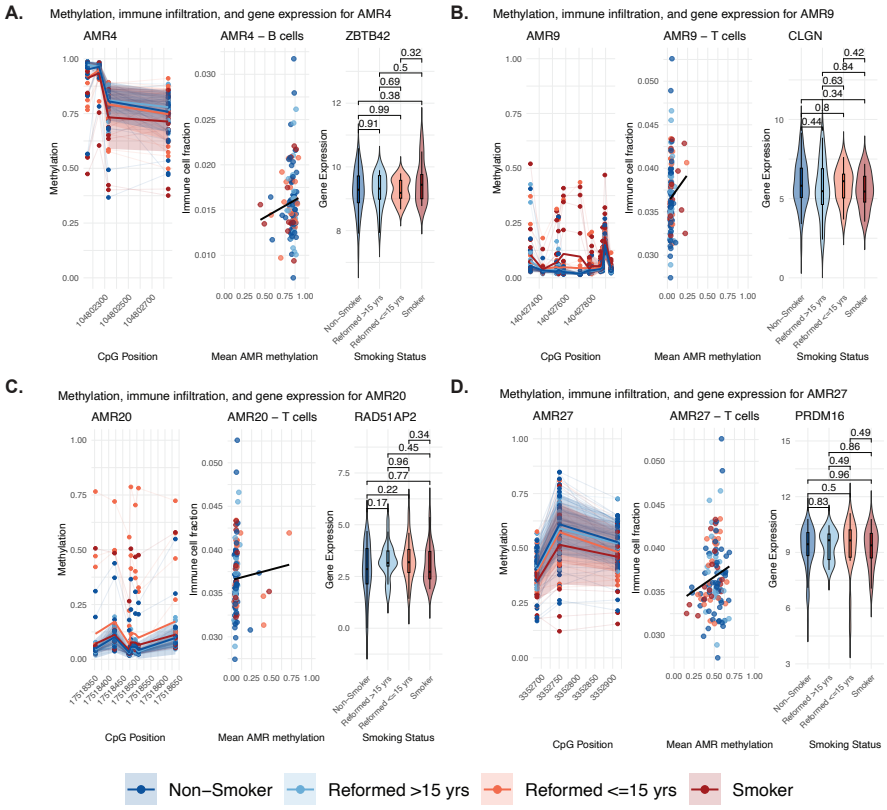

Fig 6. DNA methylation, immune infiltration, and gene expression associated with selected AMRs. A. *ZBTB42*-AMR4, **B.** *CLGN*-AMR9, **C.** *RAD51AP2*-AMR20, and **D.** *PRDM16*-AMR27. For each AMR, three panels are shown: (i) CpG site–specific methylation patterns stratified by smoking status, with individual patient trajectories (light lines), group means (solid lines), and standard deviation intervals (shaded ribbons); (ii) correlation between mean AMR methylation and immune cell fraction for the relevant cell type; and (iii) gene expression levels stratified by smoking categories; Violin plots show the distribution; boxplots indicate the median and interquartile range. P-values are from pair-wise Wilcoxon rank-sum tests. Colors are consistent across panels: red for smokers and blue for non-smokers.

Discussion

In this study, we developed a novel high-dimensional mediation analysis method for survival outcomes and applied it to investigate the role of DNA methylation in mediating tobacco effects on PDAC patient survival. By integrating computational immune deconvolution with causal discovery frameworks, we identified complex mediation pathways that involve both epigenetic modifications and immune microenvironment alterations.

Our HDMAX2-surv method addresses a critical gap in high-dimensional mediation analysis for survival data. Through extensive simulations, we demonstrated superior performance compared to existing approaches, particularly in handling the scale and complexity of epigenome-wide DNAm data. The parametric survival modeling approach offers computational advantages while maintaining robust statistical inference, although it remains dependent on the correct model specification. Additional simulations exploring diverse effect size distributions, inter-probe correlations, and DNAm noise structures would further strengthen its evaluation. Our findings also show that current methods claiming to handle high-dimensional mediation may not remain robust in genuinely large-scale settings, highlighting the need for methodological advances. This is particularly important for DNAm mediation analysis, which is increasingly applied in studies of chronic disease and environmental exposures [62]. Beyond DNAm, the proposed framework can be adapted to other exposure–outcome contexts where epigenetic mechanisms are implicated. Estimating total mediated effects in survival analysis remains a complex challenge, particularly when mediators exert opposing influences, as illustrated in our tobacco–survival analysis. Developing robust strategies for such settings should be a priority for future methodological research.

We also developed an ad hoc framework to identify causal relationships and perform serial mediation analysis between AMRs and immune infiltration. Several limitations of this approach warrant consideration. First, accurate estimation of direct and indirect effects in serial mediation relies on the assumption of no unmeasured confounding across the exposure–mediator–outcome relationships [46]. Although our causal discovery framework incorporated measured confounders (age, sex, tumor grade) and verified the independence of latent factors from tobacco exposure, the potential influence of unmeasured confounders—such as genetic predisposition, environmental exposures, or lifestyle factors—remains a limitation that could bias our causal effect estimates. Second, our framework assumes linear relationships between intermediary molecular variables, which may not capture the full complexity of biological interactions [63]. Nonlinear mediation effects, such as threshold effects or interaction-dependent relationships, could represent important biological mechanisms that warrant investigation through more flexible modeling approaches. Third, aspects of our variable construction approach also require careful interpretation. While our consensus-based immune deconvolution method represents an advance, it remains a proxy measure, and the estimated cell-type proportions are inherently dependent on the quality of reference datasets. Although the consensus approach limits estimation errors, future methodological developments in the field will likely improve deconvolution accuracy, warranting revalidation of our findings. Finally, the aggregated nature of tobacco exposure variables in the TCGA dataset may mask important dose-response relationships or timing effects. Future studies with more granular exposure information will be critical to refine these analyses.

Given the limited cohort size, we faced inherent challenges regarding statistical power and false discovery control. Additionally, our analytical pipeline involved sequential testing across multiple stages: high-dimensional CpG screening, AMR construction, mediation testing, causal discovery, and serial mediation analysis. This multi-stage approach likely inflated the risk of false discoveries, necessitating careful

threshold selection and interpretation. We adopted a conceptual framework focused on identifying candidate regions for subsequent experimental validation rather than providing an exhaustive catalog of all mediation relationships. Consequently, we employed varying significance thresholds across different analytical stages, each justified by specific methodological considerations. In our simulation framework, we applied conservative Bonferroni correction to evaluate algorithm performance under stringent conditions. Simulations used 10,000 probes (approximately one chromosome), whereas real data contained 370,000 probes. Under real-world conditions, FDR control at 10% identified only 10 significant CpGs, illustrating our statistical power limitations. To address high-dimensional challenges, we employed a spatial aggregation strategy using `comb-p` [45] to combine adjacent CpG P-values, then selected the top 50 candidate regions. This approach prioritized candidate identification over exhaustive system characterization. We then implemented a two-tiered approach for causal discovery. First, we used conservative Bonferroni correction for latent factor identification to minimize model complexity and preserve statistical power. Subsequently, we employed a more liberal $P - value$ threshold (0.15) for the remaining causal discovery steps to maximize sensitivity for potential associations, anticipating that only a subset of AMR-immune pairs would advance to serial mediation validation. Final candidate selection relied on 90% confidence intervals, identifying nine significant AMR-immune pairs. This multi-stage strategy demonstrates how to navigate constrained analytical environments with limited statistical power. While our approach addressed power limitations pragmatically, several caveats remain. Significance thresholds retain some arbitrariness, and pairs classified as Graph 2 might actually represent Graph 1 relationships undetected due to insufficient power. However, the biological plausibility of our identified candidates provides encouraging evidence for our approach's validity and suggests that statistical power limitations can be overcome to yield biologically meaningful results. These findings require experimental validation to confirm the proposed mechanistic pathways.

Our analysis revealed four key loci: *ZBTB42*, *CLGN*, *RAD51AP2*, and *PRDM16*, where DNA methylation mediates tobacco effects on survival through immune infiltration changes. These findings are particularly compelling because they demonstrate detectable effects through DNAm analysis that were undetectable through gene expression profiling alone. This underscores the unique value of DNAm mediation analysis in uncovering regulatory mechanisms that may be missed by transcriptomic approaches. The identification of protective and deleterious mediation pathways illustrates the complex biological reality, where tobacco exposure can exert opposing effects depending on the genomic context. This complexity may explain the absence of a statistically significant total effect of tobacco on survival in our cohort, as direct and mediated effects, or effects from multiple mediators, may compensate for each other [64]. Moreover, we observed complex effects of immune infiltration, consistent with findings reported in the literature. For instance, chronic pancreatitis influences PDAC outcomes, probably through remodeling of the immune microenvironment, but the mechanisms remain complex and not fully understood [61]. Such scenarios highlight the importance of mediation analysis in dissecting underlying biological mechanisms even when overall associations appear non-significant.

Mediation analysis, while classically used in epidemiology, remains underutilized in oncology research. PDAC is known to be refractory to immunotherapy, but the underlying mechanisms have yet to be fully characterized [65]. Our work opens new perspectives on the mechanisms underlying tobacco effects on survival, highlighting specific AMR-immune cell combinations that warrant experimental validation. The identified loci represent potential targets for diagnostic, prognostic, and therapeutic development. Furthermore, our findings suggest that epigenetic-immune interactions

may serve as actionable biomarkers for personalized treatment strategies in tobacco-exposed PDAC patients.

Conclusion

In conclusion, this study demonstrates the power of integrating high-dimensional mediation analysis with causal discovery approaches to uncover complex biological mechanisms. Our findings provide new insights into tobacco-related carcinogenesis and establish a methodological framework for future investigations of epigenetic-immune interactions in cancer progression and response to treatment.

Competing interests

Authors declare that they have no competing interests.

Funding

This work was funded by the project THEMA, from "Appel à projets IRGA 2021-2022" form the University Grenoble Alpes and by the French Agency for National Research (CauseHet // ANR-22-CE45-0030). Finally, it has also been carried out with financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm (ACACIA project AAP-MIC-2021).

Authors' contributions

Conceptualization and methodology: F.P., E.A., E.D, A.S, N.V. and M.R. . Project administration: M.R. Data analysis: F.P., E.A, N.V. and M.R, Writing: F.P., E.A. and M.R. All authors have reviewed the manuscript and agreed to its publication.

Acknowledgments

We thank Yuna Blum for critical reading of the manuscript. We thank the GRICAD UGA mesocenter for computing resources. We thank Hugo Barbot and Olivier François for insightful scientific discussions.

Availability of data and materials

The code for data analysis and figure generation is available on GitHub (https://github.com/bcm-uga/tims-pdac).

References

1. Afghani E, Klein AP. Pancreatic Adenocarcinoma: Trends in Epidemiology, Risk Factors, and Outcomes. Hematology/Oncology Clinics of North America. 2022;36(5):879–895. doi:10.1016/j.hoc.2022.07.002.

2. Pastrana Del Valle J, Mahvi DA, Fairweather M, Wang J, Clancy TE, Ashley SW, et al. Associations of gender, race, and ethnicity with disparities in

short-term adverse outcomes after pancreatic resection for cancer. Journal of Surgical Oncology. 2022;125(4):646–657. doi:10.1002/jso.26748.

3. Weissman S, Takakura K, Eibl G, Pandol SJ, Saruta M. The Diverse Involvement of Cigarette Smoking in Pancreatic Cancer Development and Prognosis. Pancreas. 2020;49(5):612. doi:10.1097/MPA.0000000000001550.

4. Iodice S, Gandini S, Maisonneuve P, Lowenfels AB. Tobacco and the risk of pancreatic cancer: a review and meta-analysis. Langenbeck's Archives of Surgery. 2008;393(4):535–545. doi:10.1007/s00423-007-0266-2.

5. Lynch SM, Vrieling A, Lubin JH, Kraft P, Mendelsohn JB, Hartge P, et al. Cigarette Smoking and Pancreatic Cancer: A Pooled Analysis From the Pancreatic Cancer Cohort Consortium. American Journal of Epidemiology. 2009;170(4):403–413. doi:10.1093/aje/kwp134.

6. Bosetti C, Lucenteforte E, Silverman DT, Petersen G, Bracci PM, Ji BT, et al. Cigarette smoking and pancreatic cancer: an analysis from the International Pancreatic Cancer Case-Control Consortium (Panc4). Annals of Oncology: Official Journal of the European Society for Medical Oncology. 2012;23(7):1880–1888. doi:10.1093/annonc/mdr541.

7. Ben QW, Liu J, Sun YW, Wang LF, Zou DW, Yuan YZ. Cigarette Smoking and Mortality in Patients With Pancreatic Cancer: A Systematic Review and Meta-analysis. Pancreas. 2019;48(8):985. doi:10.1097/MPA.0000000000001383.

8. Lugo A, Peveri G, Bosetti C, Bagnardi V, Crippa A, Orsini N, et al. Strong excess risk of pancreatic cancer for low frequency and duration of cigarette smoking: A comprehensive review and meta-analysis. European Journal of Cancer. 2018;104:117–126. doi:10.1016/j.ejca.2018.09.007.

9. Momi N, Kaur S, Ponnusamy MP, Kumar S, Wittel UA, Batra SK. Interplay between Smoking-induced Genotoxicity and Altered Signaling in Pancreatic Carcinogenesis. Carcinogenesis. 2012;33(9):1617–1628. doi:10.1093/carcin/bgs186.

10. Rousseaux S, Seyve E, Chuffart F, Bourova-Flin E, Benmerad M, Charles MA, et al. Immediate and durable effects of maternal tobacco consumption alter placental DNA methylation in enhancer and imprinted gene-containing regions. BMC Medicine. 2020;18(1):1–20. doi:10.1186/s12916-020-01736-1.

11. Jumentier B, Barrot CC, Estavoyer M, Tost J, Heude B, Fran , et al. High-Dimensional Mediation Analysis: A New Method Applied to Maternal Smoking, Placental DNA Methylation, and Birth Outcomes. Environmental Health Perspectives;131(4):047011. doi:10.1289/EHP11559.

12. Pittion F, Jumentier B, Nakamura A, Lepeule J, François O, Richard M. High-dimensional mediation analysis: Unraveling pathways linking external exposures to health outcomes. Peer Community in Genomics. 2025;1:100416. doi:10.24072/pci.genomics.100416.

13. Blum MGB, Valeri L, François O, Cadiou S, Siroux V, Lepeule J, et al. Challenges Raised by Mediation Analysis in a High-Dimension Setting. Environmental Health Perspectives. 2020;128(5):55001. doi:10.1289/EHP6240.

14. Lapointe-Shaw L, Bouck Z, Howell NA, Lange T, Orchanian-Cheff A, Austin PC, et al. Mediation analysis with a time-to-event outcome: a review of use and reporting in healthcare research. BMC medical research methodology. 2018;18(1):118. doi:10.1186/s12874-018-0578-7.

15. Lin SH, Young JG, Logan R, VanderWeele TJ. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. Statistics in medicine. 2017;36(26):4153–4166. doi:10.1002/sim.7426.

16. Sobel ME. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociological Methodology. 1982;13:290–312. doi:10.2307/270723.

17. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology. 1986;51(6):1173–1182. doi:10.1037//0022-3514.51.6.1173.

18. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychological Methods. 2010;15(4):309–334. doi:10.1037/a0020761.

19. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. **mediation** : *R* Package for Causal Mediation Analysis. Journal of Statistical Software. 2014;59(5). doi:10.18637/jss.v059.i05.

20. Luo C, Fa B, Yan Y, Wang Y, Zhou Y, Zhang Y, et al. High-dimensional mediation analysis in survival models. PLOS Computational Biology. 2020;16(4):e1007768. doi:10.1371/journal.pcbi.1007768.

21. Zhang H, Zheng Y, Hou L, Zheng C, Liu L. Mediation analysis for survival data with high-dimensional mediators. Bioinformatics (Oxford, England). 2021;37(21):3815–3821. doi:10.1093/bioinformatics/btab564.

22. Chi S, Flowers CR, Li Z, Huang X, Wei P. MASH: Mediation analysis of survival outcome and high-dimensional omics mediators with application to complex diseases. The Annals of Applied Statistics. 2024;18(2):1360–1377. doi:10.1214/23-AOAS1838.

23. Shao Z, Wang T, Zhang M, Jiang Z, Huang S, Zeng P. IUSMMT: Survival mediation analysis of gene expression with multiple DNA methylation exposures and its application to cancers of TCGA. PLOS Computational Biology. 2021;17(8):e1009250. doi:10.1371/journal.pcbi.1009250.

24. Tian P, Yao M, Huang T, Liu Z. CoxMKF: a knockoff filter for high-dimensional mediation analysis with a survival outcome in epigenetic studies. Bioinformatics. 2022;38(23):5229–5235. doi:10.1093/bioinformatics/btac687.

25. Huang L, Long JP, Irajizad E, Doecke JD, Do KA, Ha MJ. A unified mediation analysis framework for integrative cancer proteogenomics with clinical outcomes. Bioinformatics. 2023;39(1):btad023. doi:10.1093/bioinformatics/btad023.

26. Ahn S, Fu W, van Gerwen M, Liu L, Li Z. A Multi-Omics Framework for Survival Mediation Analysis of High-Dimensional Proteogenomic Data. ArXiv. 2025; p. arXiv:2503.08606v1.

27. VanderWeele TJ. Explanation in causal inference: developments in mediation and interaction. International Journal of Epidemiology. 2016;45(6):1904–1908. doi:10.1093/ije/dyw277.

28. Höfler M. The Bradford Hill considerations on causality: a counterfactual perspective. Emerging Themes in Epidemiology. 2005;2:11. doi:10.1186/1742-7622-2-11.

29. Ribeiro AH, Crnkovic M, Pereira JL, Fisberg RM, Sarti FM, Rogero MM, et al. AnchorFCI: harnessing genetic anchors for enhanced causal discovery of cardiometabolic disease pathways. Frontiers in Genetics. 2024;15. doi:10.3389/fgene.2024.1436947.

30. Glymour C, Zhang K, Spirtes P. Review of Causal Discovery Methods Based on Graphical Models. Frontiers in Genetics. 2019;10:524. doi:10.3389/fgene.2019.00524.

31. Gomez-Rubio P, Zock JP, Rava M, Marquez M, Sharp L, Hidalgo M, et al. Reduced risk of pancreatic cancer associated with asthma and nasal allergies. Gut. 2017;66(2):314–322. doi:10.1136/gutjnl-2015-310442.

32. Sivakumar S, Jainarayanan A, Arbe-Barnes E, Sharma PK, Leathlobhair MN, Amin S, et al. Distinct immune cell infiltration patterns in pancreatic ductal adenocarcinoma (PDAC) exhibit divergent immune cell selection and immunosuppressive mechanisms. Nature Communications. 2025;16(1):1397. doi:10.1038/s41467-024-55424-2.

33. Domingo-Relloso A, Jerolon A, Tellez-Plaza M, Bermudez JD. Causal mediation for uncausally related mediators in the context of survival analysis. medRxiv: The Preprint Server for Health Sciences. 2024; p. 2024.02.16.24302923. doi:10.1101/2024.02.16.24302923.

34. Nicolle R, Raffenne J, Paradis V, Couvelard A, de Reynies A, Blum Y, et al. Prognostic Biomarkers in Pancreatic Cancer: Avoiding Errata When Using the TCGA Dataset. Cancers. 2019;11(1):126. doi:10.3390/cancers11010126.

35. Peng J, Sun BF, Chen CY, Zhou JY, Chen YS, Chen H, et al. Single-cell RNA-seq highlights intra-tumoral heterogeneity and malignant progression in pancreatic ductal adenocarcinoma. Cell Research. 2019;29(9):725–738. doi:10.1038/s41422-019-0195-y.

36. Raghavan S, Winter PS, Navia AW, Williams HL, DenAdel A, Lowder KE, et al. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. Cell. 2021;184(25):6119–6137.e26. doi:10.1016/j.cell.2021.11.017.

37. Rashid NU, Peng XL, Jin C, Moffitt RA, Volmar KE, Belt BA, et al. Purity Independent Subtyping of Tumors (PurIST), a clinically robust, single sample classifier for tumor subtyping in pancreatic cancer. Clinical cancer research : an official journal of the American Association for Cancer Research. 2020;26(1):82–92. doi:10.1158/1078-0432.CCR-19-1467.

38. Dong M, Thennavan A, Urrutia E, Li Y, Perou CM, Zou F, et al. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Briefings in Bioinformatics. 2021;22(1):416–427. doi:10.1093/bib/bbz166.

39. Hu M, Chikina M. InstaPrism: an R package for fast implementation of BayesPrism. Bioinformatics. 2024;40(7):btae440. doi:10.1093/bioinformatics/btae440.

40. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature Communications. 2019;10(1):380. doi:10.1038/s41467-018-08023-x.

41. Zheng SC, Breeze CE, Beck S, Teschendorff AE. Identification of differentially methylated cell types in epigenome-wide association studies. Nature Methods. 2018;15(12):1059–1066. doi:10.1038/s41592-018-0213-x.

42. Cattell RB. The Scree Test For The Number Of Factors. Multivariate Behavioral Research. 1966;1(2):245–276. doi:10.1207/s15327906mbr0102$_1$0.

43. Caye K, Jumentier B, Lepeule J, François O. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular Biology and Evolution. 2019;36(4):852–860. doi:10.1093/molbev/msz008.

44. Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Research. 2016;44(3):e20. doi:10.1093/nar/gkv907.

45. Pedersen BS, Schwartz DA, Yang IV, Kechris KJ. Comb-p: software for combining, analyzing, grouping and correcting spatially correlated P-values. Bioinformatics. 2012;28(22):2986–2988. doi:10.1093/bioinformatics/bts545.

46. Zugna D, Popovic M, Fasanelli F, Heude B, Scelo G, Richiardi L. Applied causal inference methods for sequential mediators. BMC Medical Research Methodology. 2022;22(1):301. doi:10.1186/s12874-022-01764-w.

47. Daniel RM, De Stavola BL, Cousens SN, Vansteelandt S. Causal Mediation Analysis with Multiple Mediators. Biometrics. 2015;71(1):1–14. doi:10.1111/biom.12248.

48. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. Epidemiology (Cambridge, Mass). 2017;28(2):258–265. doi:10.1097/EDE.0000000000000596.

49. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics. 2013;45(10):1113–1120. doi:10.1038/ng.2764.

50. Hamada S, Satoh K, Hirota M, Kanno A, Umino J, Ito H, et al. The homeobox gene MSX2 determines chemosensitivity of pancreatic cancer cells via the regulation of transporter gene ABCG2. Journal of Cellular Physiology. 2012;227(2):729–738. doi:10.1002/jcp.22781.

51. Duarte-Medrano G, Lopez-Méndez I, Ramírez-Luna M Valdovinos-Andraca F, Cruz-Martínez R, Medina-Vera I, et al. Analysis of circulating blood and tissue biopsy PDX1 and MSX2 gene expression in patients with pancreatic cancer: A case-control experimental study. Medicine. 2019;98(26):e15954. doi:10.1097/MD.0000000000015954.

52. Wang H, Yin F, Yuan F, Men Y, Deng M, Liu Y, et al. Pancreatic cancer differential methylation atlas in blood, peri-carcinomatous and diseased tissue. Translational Cancer Research. 2020;9(2). doi:10.21037/tcr.2019.11.26.

53. Xue R, Zhou J, Wu J, Meng Q, Gong J, Shen L. P-element-Induced Wimpy-Testis-Like Protein 1 Regulates the Activation of Pancreatic Stellate Cells Through the PI3K/AKT/mTOR Signaling Pathway. Digestive Diseases and Sciences. 2023;68(4):1339–1350. doi:10.1007/s10620-022-07605-6.

54. Xie J, Xing S, Shen BY, Chen HT, Sun B, Wang ZT, et al. PIWIL1 interacting RNA piR-017061 inhibits pancreatic cancer growth via regulating EFNA5. Human Cell. 2021;34(2):550–563. doi:10.1007/s13577-020-00463-2.

55. Volta F, Scerbo MJ, Seelig A, Wagner R, O'Brien N, Gerst F, et al. Glucose homeostasis is regulated by pancreatic -cell cilia via endosomal EphA-processing. Nature Communications. 2019;10:5686. doi:10.1038/s41467-019-12953-5.

56. Li Y, Zhu Y, Chen L, Xia S, Adegboro AA, Wanggou S, et al. Transcription factor ZBTB42 is a novel prognostic factor associated with immune cell infiltration in glioma. Frontiers in Pharmacology. 2023;14:1102277. doi:10.3389/fphar.2023.1102277.

57. Cai M, Yang Q, He J, Wu H, Li Z, Fang Z, et al. Identification and Validation of a Necroptosis-Related Prognostic Signature for Kidney Renal Clear Cell Carcinoma. Stem Cells International. 2023;2023:8446765. doi:10.1155/2023/8446765.

58. Li J, Liu X, Tran TT, Lee M, Tsai RYL. DNA Methylation and Target Gene Expression in Fatty Liver Progression From Simple Steatosis to Advanced Fibrosis. Liver International: Official Journal of the International Association for the Study of the Liver. 2025;45(3):e70040. doi:10.1111/liv.70040.

59. Yao Q, Chen W, Gao F, Wu Y, Zhou L, Xu H, et al. Characteristic Analysis of Featured Genes Associated with Cholangiocarcinoma Progression. Biomedicines. 2023;11(3):847. doi:10.3390/biomedicines11030847.

60. Nishikata I, Sasaki H, Iga M, Tateno Y, Imayoshi S, Asou N, et al. A novel EVI1 gene family, MEL1, lacking a PR domain (MEL1S) is expressed mainly in t(1;3)(p36;q21)-positive AML and blocks G-CSF-induced myeloid differentiation. Blood. 2003;102(9):3323–3332. doi:10.1182/blood-2002-12-3944.

61. Mahadevan KK, Dyevoich AM, Chen Y, Li B, Sugimoto H, Sockwell AM, et al. Type I conventional dendritic cells facilitate immunotherapy in pancreatic cancer. Science. 2024;384(6703):eadh4567. doi:10.1126/science.adh4567.

62. Fujii R, Sato S, Tsuboi Y, Cardenas A, Suzuki K. DNA methylation as a mediator of associations between the environment and chronic diseases: A scoping review on application of mediation analysis. Epigenetics. 2022;17(7):759–785. doi:10.1080/15592294.2021.1959736.

63. Cai Q, Fu Y, Lyu C, Wang Z, Rao S, Alvarez JA, et al. A new framework for exploratory network mediator analysis in omics data. Genome Research. 2024;34(4):642–654. doi:10.1101/gr.278684.123.

64. O'Rourke HP, MacKinnon DP. Reasons for Testing Mediation in the Absence of an Intervention Effect: A Research Imperative in Prevention and Intervention Research. Journal of Studies on Alcohol and Drugs. 2018;79(2):171–181. doi:10.15288/jsad.2018.79.171.

65. Cappellesso F, Orban MP, Shirgaonkar N, Berardi E, Serneels J, Neveu MA, et al. Targeting the bicarbonate transporter SLC4A4 overcomes immunosuppression and immunotherapy resistance in pancreatic cancer. Nature Cancer. 2022;3(12):1464–1483. doi:10.1038/s43018-022-00470-2.