PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

IUSMMT: Survival mediation analysis of gene expression with multiple DNA methylation exposures and its application to cancers of TCGA

Zhonghe Shao[1‡], Ting Wang[1‡], Meng Zhang[1], Zhou Jiang[1], Shuiping Huang[1,2,3]*, Ping Zeng[1,2,3]*

1 Department of Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu, China, **2** Center for Medical Statistics and Data Analysis, Xuzhou Medical University, Xuzhou, Jiangsu, China, **3** Key Laboratory of Human Genetics and Environmental Medicine, Xuzhou Medical University, Xuzhou, Jiangsu, China

‡ These authors are co-first authors on this work.
* hsp@xzhmu.edu.cn (SH); zpstat@xzhmu.edu.cn (PZ)

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

Editor: Oscar Rueda, University of Cambridge, UNITED KINGDOM

Received: January 13, 2021

Accepted: July 6, 2021

Published: August 31, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1009250

Data Availability Statement: All files are available from the TCGA database (https://portal.gdc.cancer.gov/legacy-archive/).

Abstract

Effective and powerful survival mediation models are currently lacking. To partly fill such knowledge gap, we particularly focus on the mediation analysis that includes multiple DNA methylations acting as exposures, one gene expression as the mediator and one survival time as the outcome. We proposed IUSMMT (intersection-union survival mixture-adjusted mediation test) to effectively examine the existence of mediation effect by fitting an empirical three-component mixture null distribution. With extensive simulation studies, we demonstrated the advantage of IUSMMT over existing methods. We applied IUSMMT to ten TCGA cancers and identified multiple genes that exhibited mediating effects. We further revealed that most of the identified regions, in which genes behaved as active mediators, were cancer type-specific and exhibited a full mediation from DNA methylation CpG sites to the survival risk of various types of cancers. Overall, IUSMMT represents an effective and powerful alternative for survival mediation analysis; our results also provide new insights into the functional role of DNA methylation and gene expression in cancer progression/prognosis and demonstrate potential therapeutic targets for future clinical practice.

Author summary

DNA methylation has a causal effect on tumorigenesis and gene expression may be an important mediator of such influence. However, inferring the existence of mediation effect of gene expression is statistically challenging, especially when multiple and even high-dimensional DNA methylation exposures are collectively analyzed. To solve such challenge, we developed a new mediation approach called IUSMMT, in which mediation effects are determined by two separate tests: one for the association between methylations and the expression, the other for the association between the expression and the survival

Funding: The research of PZ was supported in part by the Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), the Natural Science Foundation of Jiangsu Province of China (BK20181472), the Chinese Postdoctoral Science Foundation (2018M630607 and 2019T120465), the QingLan Research Project of Jiangsu Province for Outstanding Young Teachers, the Six-Talent Peaks Project in Jiangsu Province of China (WSN-087), the Training Project for Youth Teams of Science and Technology Innovation at Xuzhou Medical University (TD202008), the Postdoctoral Science Foundation of Xuzhou Medical University, the National Natural Science Foundation of China (81402765), and the Statistical Science Research Project from National Bureau of Statistics of China (2014LY112). The research of SH was supported in part by the Social Development Project of Xuzhou City (KC19017). The research of TW was supported in part by the Social Development Project of Xuzhou City (KC20062).The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

outcome conditional on methylations. IUSMMT effectively combines the evidence of the two tests and infers the emergence of mediation effect by fitting an empirical three-component mixture null distribution. To evaluate the performance of IUSMMT, we conducted extensive simulation and analyzed ten TCGA cancers. Overall, we demonstrated the robustness, validity, and utility of IUSMMT under a wide variety of scenarios.

This is a *PLOS Computational Biology* Methods paper.

Introduction

Many recent studies have revealed that epigenetic abnormalities, particularly aberrant changes in methylation, exert an important causative effect on complex diseases [1,2]. However, the mechanisms regarding how alterations of DNA methylation influence diseases remain largely elusive. Biologically, it has been widely acknowledged that DNA methylation leads to a direct functional modification of the genome by regulating gene expression [3–13], which reversely affects various diseases [14–17]. Statistically, this motivates researchers to consider gene expression as a critical causal mediator of DNA methylation on the development of diseases. However, the principal genes that control the pathogenesis of diseases have not yet been identified and the underlying mechanisms of causal genes on diseases are also unclear.

With respect to statistics, mediation analysis, particularly popular in sociology, epidemiology, and psychology [18–22], offers a flexible means to interpret the interplay between DNA methylation, gene expression and diseases. The essential requirement of mediation analysis arises when researchers are interested in potential underlying mechanism between an exposure (e.g., DNA methylation) and an outcome (e.g., the survival time and status of cancer patients) and long to acquire an in-depth insight into such understanding. Formally, a mediating variable (or mediator) is defined as an intermediate variable (e.g., gene expression) in the causal sequence that relates the exposure to the outcome [18].

With datasets available from The Cancer Genome Atlas (TCGA) project as an illustrative example [23], in this study we aim to utilize appropriate mediation models to investigate the mechanism of gene expression on the signaling pathway from DNA methylation to the survival risk of cancers. In recent years, there have been a lot of methodology extensions of traditional linear mediation analysis to time-to-event outcome with censored data. Tein and MacKinnon (2003) [24] studied the estimation and inference of mediated effect for survival outcome under the context of the log-survival and log-hazard time models. Lange and Hansen (2011) [25] formulized the natural direct and indirect effects for time-to-event outcome using an additive hazard model within the counterfactual framework and illustrated its application by analyzing socioeconomic status, work environment, and long-term sickness absence. The natural direct and indirect effects were further described using a proportional hazards model with a rare outcome or an accelerated failure time model [26]. Wang and Zhang (2011) [27] proposed a Bayesian Tobit approach to examine the mediation effect for censored data and applied it to study whether verbal memory ability mediated the relationship between age and everyday functioning. More recently, Luo et al. (2020) [28] investigated survival mediation methods with high-dimensional mediators by borrowing the idea of screening-penalization procedure [29]. Although great advances have been achieved (see [30] for a comprehensive review of statistical methods for high-dimensional mediation analysis in the era of high-

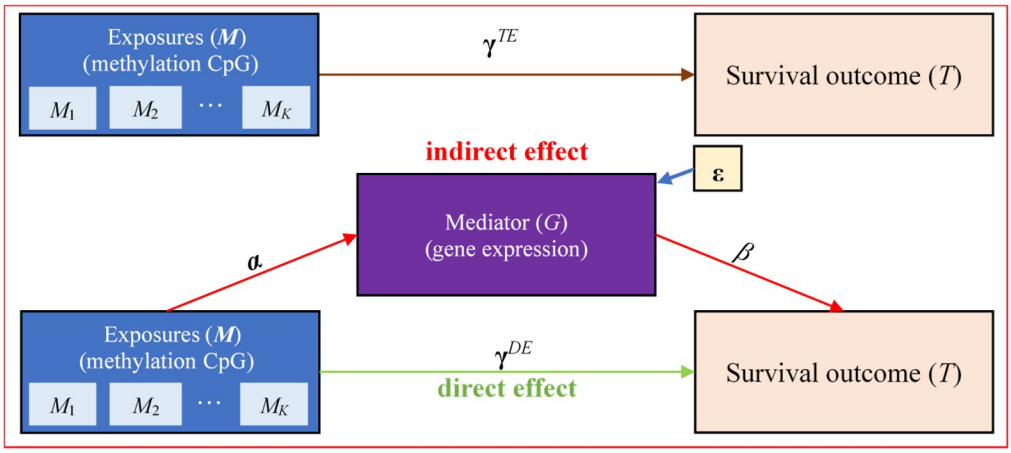

Fig 1. Statistical framework of the survival mediation analysis with multiple exposures, one mediator and one censored outcome. The mediation analysis involves in multiple DNA methylations (***M***) as exposures, one gene expression (*G*) as the mediator and one survival time as the outcome (*T*). There are two types of effects from ***M*** to *T*: the direct effect of ***M*** on *T* (i.e., γ^{DE}) and the indirect effect of ***M*** on *T* via the intermediate variable *G*. The indirect effect represents the amount of mediation coming from two sources: the effect from ***M*** to *G* (i.e., $\boldsymbol{\alpha}$) and the effect from *G* to *T* (i.e., β).

https://doi.org/10.1371/journal.pcbi.1009250.g001

throughput genomics), methods for survival mediation analysis are still yet to be developed. These prior methods often consider one or multiple mediators and cannot be readily applied to our setting where multiple exposures, sometimes high-dimensional, would be involved. Specially, we wish to implement a gene-centric mediation analysis to simultaneously model a group of DNA methylation CpG sites as exposures and explore the mediating role of gene expression in the influence of methylations on the survival risk of cancer patients (Fig 1).

To effectively examine mediation effect in survival studies with multiple exposures, we first employed the variance component-based score test to assess the association between a group of DNA methylation CpG sites and expression level for each gene under the framework of linear mixed-effects model (lmm) [31–34], and then tested for the effect of gene expression on the overall survival time while adjusting for the direct influence of these methylations within the framework of Cox linear mixed-effects model (coxlmm) [35–38]. This mediation analysis procedure is displayed in Fig 1, where the methylation-expression effect (i.e., $\boldsymbol{\alpha}$) and the expression-survival effect (i.e., β) can be causally interpreted if individual mediation models are correctly constructed and identifiability assumptions are satisfied [39–41]. Besides the explicit assumption of temporal ordering between methylation, gene expression and survival outcome, the additional assumptions for causal interpretation of these effects include: (i) the confounding between the methylations and the survival outcome is correctly controlled; (ii) the confounding between the gene expression and the survival outcome is correctly controlled; (iii) the confounding between the methylations and the gene expression is correctly controlled; and (iv) there should be no expression-survival confounders which are themselves affected by the methylations. The above assumptions are also known as sequential ignobility assumptions or no unmeasured confounding assumptions [20,42–45]. Note that, our mediation analysis method can still be employed for identifying potential methylation CpG sites or candidate genes for further exploration, even when the above causality conditions are not completely satisfied.

We subsequently determined the existence of mediation effect to detect candidate genes with potentially mediating roles by adherence to the principle of intersection-union test (IUT) [46–52], in which the maximum of the two *P*-values (denoted by $P_{\max}$) in the two tests above is

taken as the significance measurement. Although it is conceptually straightforward, this naïve IUT-based approach is oftentimes extremely conservative especially in large scale mediation effect tests due to its composite null nature [43,53], which can be equivalently expressed as a combination of three disjoint component null hypotheses. To correct the intrinsic conservativeness of IUT, we estimated the proportion for each component null hypothesis and constructed a novel null distribution for $P_{\max}$ by fitting a three-component mixture null distribution [54], which, in contrast to the naïve uniform null distribution of $P_{\max}$ assumed in previous literature [51,52,55], can lead to a desirable control of family-wise error rate (FWER) or false discovery rate (FDR). We thus refer to our proposed approach framework as the intersection-union survival mixture-adjusted mediation test (IUSMMT).

Finally, we applied IUSMMT to ten TCGA cancers and identified multiple genes with mediation effects. We revealed that, although DNA methylation CpG sites across the whole genome showed pleiotropic regularization on gene expressions of various cancers, most of the detected genetic regions, in which gene expressions played key roles as active mediators, were cancer type-specific and exhibited full mediations lying in the signaling pathway from DNA methylation CpG sites to the survival risk of cancers. Overall, IUSMMT represents an effective and powerful statistical tool for survival mediation analysis; our results also provide new insights into the functional role of DNA methylation and gene expression in cancer progression/prognosis and demonstrate potential therapeutic targets for future clinical practice.

Results

Overview of IUSMMT

We first offer an overview of the proposed IUSMMT method (Fig 2), with more details illustrated in the S1–S4 Texts and the section of Materials and Methods. IUSMMT is a gene-centric mediation method especially for survival data and dedicates to detect candidate genes with potential mediating effects standing on the signaling pathway from DNA methylation CpG sites to cancer survival. It proceeds in the following steps. First, IUSMMT calculates *P*-values of the methylation-expression effect (i.e., $P_{\boldsymbol{\alpha}}$) and the expression-survival effect (i.e., P_{β}) and takes them as input (Fig 2A); here $P_{\boldsymbol{\alpha}}$ is obtained via a variance component-based score test within lmm by assuming each of $\boldsymbol{a}$ following a mean-zero normal distribution with an unknown variance, and P_{β} is yielded through the Wald test within coxlmm. Second, to determine whether a gene of focus has a mediation effect, IUSMMT classifies the joint null hypothesis H_0: $\boldsymbol{\alpha\beta} = \mathbf{0}$ into three composite null sub-hypotheses and takes the maximum of $P_{\boldsymbol{\alpha}}$ and P_{β} as a significance measurement in terms of the IUT principle (Fig 2B). Finally, to enhance the statistical power, IUSMMT estimates the proportion for each component of the three null hypotheses (Fig 2C) and constructs a newly empirical exact null distribution by fitting a three-component mixture null distribution. Afterwards, an effective control of FWER or FDR is achieved on the basis of the estimated mixture null distribution. As shown below, the power of IUSMMT would improve and the bias in estimated mixture proportions would decrease with the increase of sample sizes.

Results for simulation studies

Estimated null proportions

Following the statistical framework of survival mediation analysis shown in Fig 1, we generated expression and survival outcome based on a set of real methylation values of TCGA under various sample sizes and proportion parameters, with details of simulation described in the Materials and Methods section. We first present the results for simulation studies and

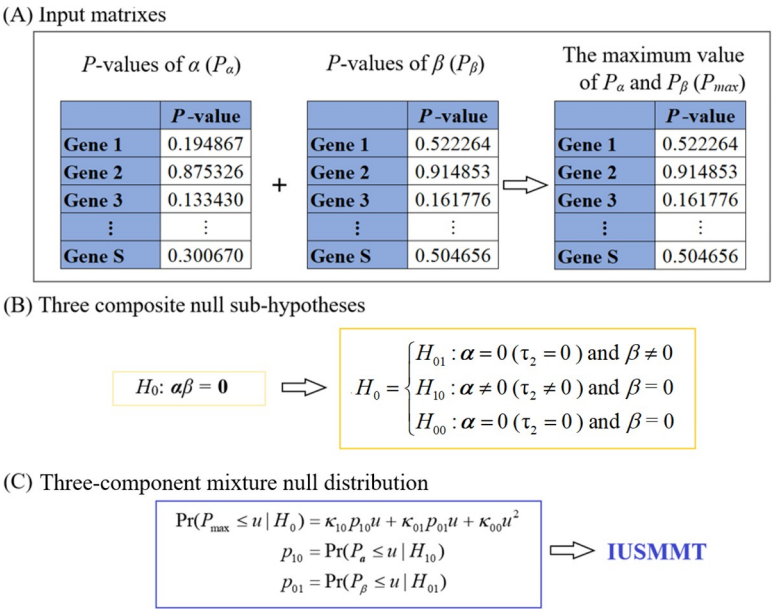

Fig 2. An overview of IUSMMT for examining the mediation effect in survival model. Here, $\boldsymbol{a} = (\alpha_1, \ldots, \alpha_K)$ is the vector of effect sizes of a set of DNA methylation CpG sites (i.e., the exposures) on the gene expression level (i.e., the mediator), with K the number of CpG sites within that gene; and β is the expression-survival effect; S indicates the total number of genes; n denotes the sample size. (A) IUSMMT first separately evaluates the significance of $\boldsymbol{a}$ and β, and calculates $P_{\boldsymbol{a}}$ and P_β; where $P_{\boldsymbol{a}}$ is obtained by a variance component-based score test within the linear mixed-effects model by assuming each of $\boldsymbol{a}$ following a mean-zero normal distribution with an unknown variance τ_2, while P_β is yielded through the Wald test within the Cox linear mixed-effects model. Then, IUSMMT takes the two P-values as input. (B) The hypothesis testing of mediation effect is to examine whether the product of $\boldsymbol{\alpha}$ and β is zero or not (i.e., H_0: $\boldsymbol{\alpha}\beta = \mathbf{0}$) and can be divided into three composite null sub-hypotheses. (C) In the three-component mixture null distribution, κ_{10} stands for the probability that the exposures are related to the mediator in the exposure-mediator model but the mediator is not associated with the survival outcome in the mediator-outcome model; κ_{01} stands for the probability that the exposures are not related to the mediator in the exposure-mediator model but the mediator is associated with the survival outcome in the mediator-outcome model; κ_{00} stands for the probability that the exposures are not related to the mediator in the exposure-mediator model and the mediator is not associated with the survival outcome in the mediator-outcome model. The definition of other notations used in B and C can be found in the Materials and Methods section.

https://doi.org/10.1371/journal.pcbi.1009250.g002

display estimated proportion parameters for the three-component mixture null distribution under various simulation scenarios in Tables 1 and S1. Totally, the estimates of these proportions show slight to moderate biases in the alternative scenarios but are relatively close to the true values in other cases. Nearly in all scenarios, κ_{00} is over-estimated especially when sample size is small. Specifically, in the sparse alternative cases, as can be expected, κ_{10} and κ_{01} are also over-estimated because they are non-negatively estimated but their true values are actually zero, which necessarily results in the underestimation for κ_{11}. In contrast, in the dense alternative cases, the opposite patterns are observed, with κ_{00} is over-estimated but κ_{01} is under-estimated, which is certainly a direct consequence of limited power when testing the effect of the exposure on the mediator (i.e., τ_2) and the effect of the mediator on the survival outcome (i.e., β) (S1 Fig). Particularly, in all the simulation scenarios, κ_{11} is always underestimated or approximately unbiased, implying that we can minimize the false discovery when examining the mediation effect. In addition, we find that the estimates of some proportion parameters (e.g., κ_{00}) have greater bias under the dense null compared to these under the sparse null. The reason is primarily due to relatively more sufficient information that can be available for estimating these non-zero proportion parameters under the sparse null compared to the dense null (e.g., 90% vs. 10%).

Table 1. Estimated and true proportion parameters (mean and standard deviation) in the three-component mixture null distribution under the five simulation scenarios with different sample sizes and numbers of mediation tests.

dense null	**$\kappa_{00} = 0.1$**	**$\kappa_{01} = 0.85$**	**$\kappa_{10} = 0.05$**	**$\kappa_{11} = 0$**
n = 250	0.441 (0.175)	0.536 (0.174)	0.022 (0.012)	0.001 (0.001)
n = 400	0.293 (0.155)	0.675 (0.154)	0.031 (0.012)	0.001 (0.001)
n = 548	0.215 (0.116)	0.750 (0.116)	0.034 (0.012)	0.000 (0.002)
sparse null	**$\kappa_{00} = 0.99$**	**$\kappa_{01} = 0.01$**	**$\kappa_{10} = 0$**	**$\kappa_{11} = 0$**
n = 250	0.977 (0.017)	0.022 (0.017)	0.000 (0.001)	0.001 (0.003)
n = 400	0.986 (0.011)	0.013 (0.011)	0.000 (0.001)	0.001 (0.001)
n = 548	0.977 (0.018)	0.023 (0.019)	0.000 (0.001)	0.000 (0.000)
complete null	**$\kappa_{00} = 1$**	**$\kappa_{01} = 0$**	**$\kappa_{10} = 0$**	**$\kappa_{11} = 0$**
n = 250	0.993 (0.011)	0.005 (0.010)	0.002 (0.006)	0.000 (0.000)
n = 400	0.998 (0.003)	0.001 (0.003)	0.001 (0.002)	0.000 (0.000)
n = 548	0.994 (0.013)	0.004 (0.012)	0.003 (0.008)	0.000 (0.000)
dense alternative	**$\kappa_{00} = 0.1$**	**$\kappa_{01} = 0.75$**	**$\kappa_{10} = 0.05$**	**$\kappa_{11} = 0.1$**
n = 250	0.394 (0.179)	0.537 (0.174)	0.070 (0.036)	0.000 (0.000)
n = 400	0.241 (0.145)	0.667 (0.143)	0.081 (0.035)	0.011 (0.022)
n = 548	0.172 (0.098)	0.725 (0.097)	0.077 (0.037)	0.027 (0.034)
sparse alternative	**$\kappa_{00} = 0.9$**	**$\kappa_{01} = 0$**	**$\kappa_{10} = 0$**	**$\kappa_{11} = 0.1$**
n = 250	0.900 (0.021)	0.055 (0.020)	0.033 (0.018)	0.013 (0.018)
n = 400	0.892 (0.017)	0.047 (0.026)	0.025 (0.018)	0.036 (0.028)
n = 548	0.870 (0.029)	0.061 (0.031)	0.040 (0.026)	0.028 (0.028)

Note: κ_{10} stands for the probability that the exposures are related to the mediator in the exposure-mediator model but the mediator is not associated with the survival outcome in the mediator-outcome model; κ_{01} stands for the probability that the exposures are not related to the mediator in the exposure-mediator model but the mediator is associated with the survival outcome in the mediator-outcome model; κ_{00} stands for the probability that the exposures are not related to the mediator in the exposure-mediator model and the mediator is not associated with the survival outcome in the mediator-outcome model; κ_{11} stands for the probability of the existence of mediation effects. The number of genes (i.e., the mediators) was set to 10^4. Note that, here we only present the estimates of one simulation scenario, with more results shown in S1 Table.

https://doi.org/10.1371/journal.pcbi.1009250.t001

Moreover, to evaluate the impact of various numbers of mediators on the estimators of proportions, we implemented an additional simulation with 10^3 genes (i.e., mediators) with $\kappa_{11} = 0.10$, $\kappa_{10} = 0.75$, $\kappa_{01} = 0.05$, and $\kappa_{00} = 0.10$ (a case that was very close to proportions obtained from our real applications). As can be anticipated, due to more information available (e.g., more genes), it turns out that increasing the number of mediators from 10^3 to 10^4 can generally improve the accuracy in estimating these proportion parameters (S2 Table).

Type I error and power. Based on these estimated proportion parameters, the mixture null distribution is constructed, and the mediation effect test is implemented. First, when assessing the performance of type I error control, we demonstrate that IUSMMT, which is based on the estimated mixture null distribution, can maintain the type I error correctly across these simulation scenarios (Figs 3 and S2); however, IUT, which utilizes the uniform distribution as its null distribution, is conservative. This conservativeness of IUT indicates it would be underpowered in the detection of significant mediation effect. In the power assessment, due to the adjustment of the mixture null distribution, IUSMMT is much more powerful compared to IUT (Figs 3 and S3). For instance, when n = 400 and $\beta = 0.05$, compared with IUT, IUSMMT has a 0.06, 0.12, 0.07 or 0.11 higher power under the sparse alternative and 0.25, 0.49, 0.45 or 0.30 higher power under the dense alternative when $\tau_2 = 0.02$, 0.04, 0.05 or 0.10, clearly indicating the benefit of estimating the empirical power function in the mixture null distribution. Moreover, it is clearly shown that IUSMMT has a much more pronounced

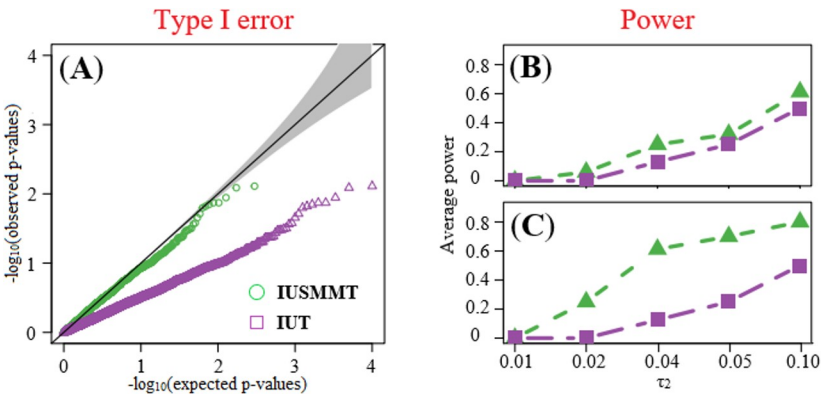

Fig 3. The QQ plot (A) and the average power curve (B-C) for IUSMMT and IUT. The QQ plot (A) is under the scenario of sparse nulls. The average power curve is calculated when the mediation strength parameter τ_2 increases under the sparse alternative (B) and the dense alternative (C). Here, n = 400, $\beta = 0.3$, and $\tau_2 = 0.01, 0.02, 0.04, 0.05$ or 0.10 at the x-axis. The magnitude of τ_2 quantifies the strength of association between DNA methylation CpG sites and gene expression. In B and C, the number of genes (i.e., mediators) was set to 10^4.

https://doi.org/10.1371/journal.pcbi.1009250.g003

advantage in power over IUT under the dense alternative (Figs 3 and S3). For example, IUSMMT is on average 29.8% more powerful than IUT under the dense alternative, while is on average only 7.3% more powerful under the sparse alternative.

DNA methylations are significantly associated with the survival risk of cancers

We now applied IUSMMT to ten types of cancers available from TCGA [23]; summary information of these cancers is shown in S3 and S4 Tables. We first examine the association between a group of DNA methylation CpG sites and the survival risk of cancers for each gene (i.e., H_0: $\boldsymbol{\gamma}^{TE} = 0$). A total of 57 (30 unique) regions of DNA methylation CpG sites are identified to be associated with the overall survival of these cancers (FDR < 0.05) except BLCA and HNSC (Table 2), including 8 methylation regions for BRCA, 7 for CESC, 7 for COAD, 17 for KIRP, 4 for LUAD, 1 for LUSC, 9 for SARC and 4 for STAD. Approximately 23.3 (= 7/30) of these regions of DNA methylation CpG sites exhibit pleiotropic effects (Fig 4A). For example, methylation CpG sites located in *ACAA2*, *NUSAP1*, *OGFOD1*, *PSMD5*, *SNRNP40* and *XRCC6* are simultaneously associated with five cancers including BRCA, CESC, COAD, KIRP and SARC; methylation CpG sites located in *USP37* are simultaneously related to four cancers (i.e., BRCA, CESC, COAD and SARC). Moreover, we recognize some cancer type-specific methylation CpG sites, including all the associated methylation sites identified for LUAD, STAD or LUSC, and partly for KIRP (11 out of 17), SARC (2 out of 9), or BRCA (1 out 8).

Identified associations between methylation and expression

We here evaluate the association between a group of DNA methylation CpG sites and gene expression (i.e., H_0: $\boldsymbol{\alpha} = 0$) through SKAT with the linear kernel. As a result, a particularly large number of associations are detected (FDR < 0.05) (Table 2), with the number of methylation-regulated genes ranging from 8,182 (62.8% = 8,182/13,029) for KIRP to 11,872 (89.5% = 11,872/13,270) for BRCA and an average of 79.3% across these cancers. There are approximately 98.2% methylation-regulated genes shared by at least two types of cancers, including 3,801 genes (e.g., *HIST1H4F*, *FBP1* and *CPEB4*) shared across all cancers.

Table 2. Number of associated regions of DNA methylation CpG sites and genes identified in the survival mediation analysis.

cancers	γ^{TE}	α (%#)	β (%#)	γ^{DE} (%#)	Mediation effect test		
					IUT (%$)	IUSMMT (%$)	
BLCA	0	11,325 (86.29)	76 (0.58)	0 (0)	41 (0.31)	61 (0.46)	8 (13.1)
BRCA	8	11,872 (89.46)	56 (0.42)	8 (0.05)	30 (0.23)	46 (0.35)	14 (30.4)
CESC	7	10,059 (76.19)	9 (0.07)	8 (0.06)	1 (0.01)	1 (0.01)	0 (0)
COAD	7	8,963 (68.25)	0 (0)	7 (0.05)	0 (0)	0 (0)	0 (0)
HNSC	0	11,284 (84.73)	42 (0.32)	0 (0)	23 (0.17)	49 (0.37)	8 (16.7)
KIRP	17	8182 (62.80)	170 (1.30)	32 (0.22)	34 (0.26)	34 (0.26)	1 (2.9)
LUAD	4	10,342 (77.80)	17 (0.13)	1 (0.01)	176 (1.32)	188 (1.41)	39 (20.9)
LUSC	1	11,427 (84.73)	3 (0.02)	0 (0)	3 (0.02)	3 (0.02)	0 (0)
SARC	9	10,653 (81.73)	204 (1.57)	10 (0.07)	88 (0.68)	156 (1.2)	18 (11.5)
STAD	4	10,879 (80.84)	8 (0.06)	3 (0.02)	4 (0.03)	5 (0.04)	0 (0)
total	57	104,986	585	69	400	543	88 (16.3)
unique	30	13,801	570	41	392	529	
pleiotropy (%)	7 (23.3)	13,553 (98.2)	15 (2.6)	7 (17.1)	8 (2.0)	14 (2.6)	

Note

# denotes the proportion among the total genes under investigation

$ denotes the proportion among the genes that are associated with the survival risk of cancers; the proportion of pleiotropy is computed by the ratio between the number of associations with pleiotropic effects and the number of unique associations. The last second column shows the number and proportion of genes with mediating effects, and the last column shows the number and proportion of potential passenger methylation events among genes with mediating roles.

https://doi.org/10.1371/journal.pcbi.1009250.t002

We also performed the similar methylation-expression analysis in 193 normal tissues combined across the ten cancers and find that 43.9% (= 6075/13840) of genes are methylation-regulated (FDR < 0.05). This proportion is relatively smaller than that obtained from the tumor tissues, which may be a direct consequence of low power due to smaller sample size of normal tissues compared tumor tissues and the distinction mechanism in gene regulation between normal and tumor tissues. On average, 81.0% of methylation-regulated genes discovered in normal tissues are overlapped with these detected in tumor tissues across the cancers (S5A Fig). Interestingly, we observe that the effect of methylation on expression is much stronger in tumor tissues compared with that in the normal tissue. For example, the median value of τ_2, which can be employed to quantify the magnitude of the methylation effect on expression, is 3.7 times higher in the BRCA tissue than that in the normal tissue (S5B Fig).

Identified associations between expression and survival risk

We next explore the association between the expression level and the survival risk of cancers for each gene (i.e., H_0: $\beta = 0$) while adjusting for the direct effects of methylation alterations within the framework of coxlmm. The number of associated genes varies from zero for COAD to 204 for SARC (FDR < 0.05). Approximately 2.6% of these associated genes are simultaneously shared by at least two cancers (Fig 4B), while most of the associated genes (~97.4%) are cancer type-specific. In addition, a total of 69 (41 unique) regions of methylation CpG sites also exhibit direct influence (i.e., $\gamma^{DE} \neq 0$) on the overall survival of some cancers (except BLCA, HNSC and LUSC) (Table 2 and Fig 4C); and approximately 17.1% of significant regions of methylation CpG sites show direct pleiotropic effects. For example, the methylation regions located in *ACAA2*, *NUSAP1*, *OGFOD1*, *PSMD5*, *SNRNP40*, *USP37* and *XRCC6* are shared by five cancers (i.e., BRCA, CESC, COAD, KIRP and SARC). Moreover, we find that there are 12 methylation regions (i.e., located within *ACAA2*, *NUSAP1*, *OGFOD1*, *PSMD5*,

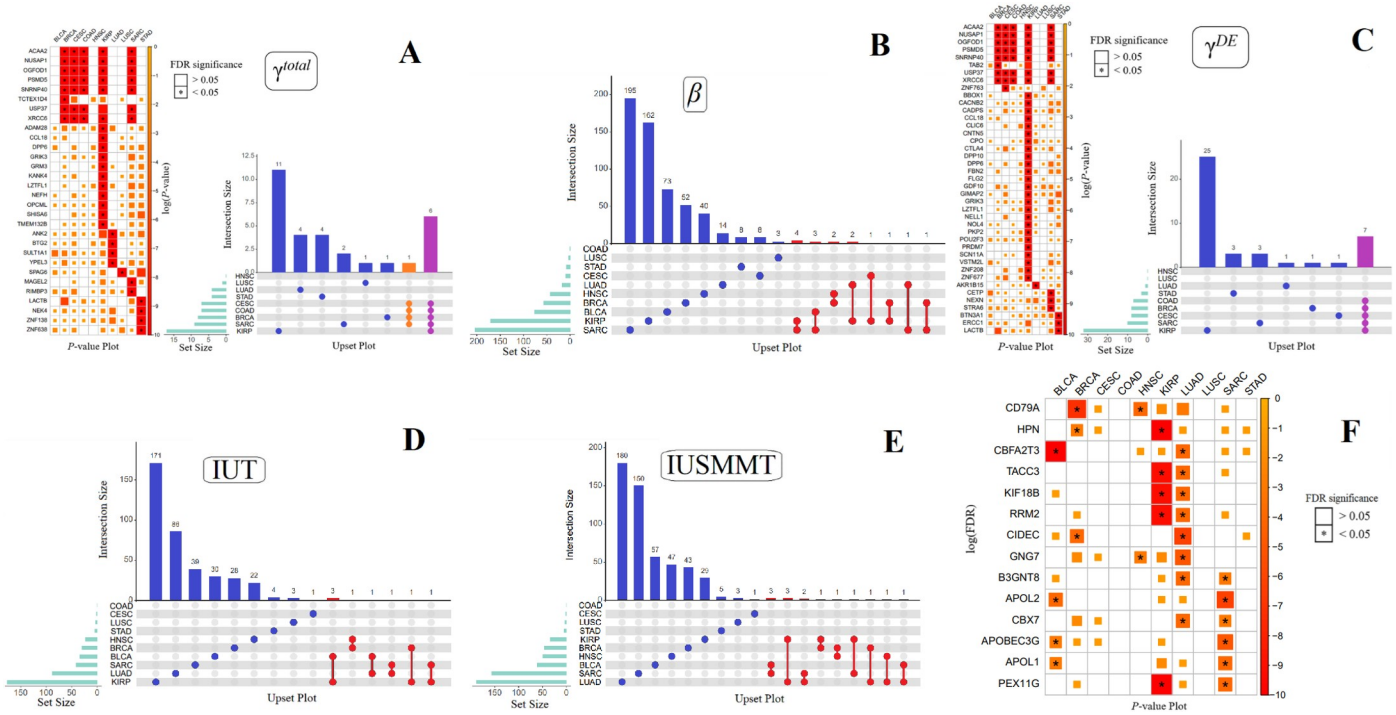

Fig 4. Upset plot and heatmap plot of the *P*-values (A-F). In the heatmaps of γ^{TE} (A) and γ^{DE} (C), the color of each box indicates the magnitude of the *P*-value. The number in the comment part represents the *P*-value processed by the negative logarithmic transformation. The darker the color, the smaller the *P*-value. In the Upset plots of γ^{TE}, β, γ^{DE}, IUT, and IUSMMT, each bar shows the number of shared genes. In these Upset plots, the blue part represents cancer type-specific genes, and the red, green, orange, and purple parts represent genes shared in two, three, four and five types of cancers. (F) The heatmap of the *P*-values of the 14 overlapped genes across all the cancers. The color of each box indicates the magnitude of the *P*-value. The number in the comment part represents the *P*-value processed by the negative log transformation, the darker the color, the smaller the *P*-value.

https://doi.org/10.1371/journal.pcbi.1009250.g004

SNRNP40, *USP37*, *XRCC6*, *CCL18*, *DPP6*, *GRIK3*, *LZTFL1*, and *LACTB*) that not only have non-zero total effect on the survival risk, but also have substantial direct influence on the survival risk of cancers.

Associated genes with mediating effects

We now utilize IUT and IUSMMT to assess whether the expression level of a gene has substantial mediating effect on the pathway from methylation CpG sites to the survival risk of cancers. The estimated proportion parameters for the three-component mixture null distribution for each cancer are shown in Table 3. These proportions also reflect the probability of the association between the methylations and the gene expression (i.e., κ_{10}), and between the gene expression and the survival time (i.e., κ_{01}), in line with the results described above. The QQ-plot of P_{max} for each cancer is demonstrated in S4 Fig. An observable upward deviation of P_{max} from the diagonal line implies the presence of mediating genes for most of these cancers except COAD. Specifically, IUSMMT detects 35.8% more genes with mediating effects compared to IUT, and all genes having mediation effects identified by IUT (a total of 400 genes) are also discovered by IUSMMT. Therefore, in the following we focus on the results of IUSMMT. As suggested by the association results described above and in terms of the principle of IUSMMT, except COAD (consistent with the pattern of the distribution of its P_{max} shown in S4 Fig), we identify multiple genes mediating the impact of methylation CpG sites on the survival risk of various cancers (Table 2 and Fig 4E). Specifically, there are a total of 543 (529 unique) significant methylation-expression mediation associations, with the number of genes exhibiting

Table 3. Estimated proportion parameters in the three-component mixture null distribution for 10 TCGA cancers.

cancers	κ_{10}	κ_{01}	κ_{00}	κ_{11}
BLCA	0.858	0.001	0.136	0.005
BRCA	0.891	0.001	0.105	0.004
CESC	0.761	0.000	0.238	0.001
COAD	0.683	0.000	0.317	0.000
HNSC	0.844	0.000	0.152	0.003
KIRP	0.621	0.006	0.366	0.007
LUAD	0.777	0.000	0.222	0.001
LUSC	0.847	0.000	0.153	0.000
SARC	0.803	0.002	0.181	0.014
STAD	0.808	0.000	0.191	0.000
average	0.789	0.001	0.206	0.004

Note: κ_{10} stands for the probability that the methylations are related to the gene expression in the exposure-mediator model but the gene expression is not associated with the survival outcome in the mediator-outcome model; κ_{01} stands for the probability that the methylations are not related to the gene expression in the exposure-mediator model but the gene expression is associated with the survival outcome in the mediator-outcome model; κ_{00} stands for the probability that the methylations are not related to the gene expression in the exposure-mediator model and the gene expression is not associated with the survival outcome in the mediator-outcome model.

https://doi.org/10.1371/journal.pcbi.1009250.t003

mediating effects ranging from 1 for CESC, to 156 for SARC and 188 for LUAD. It is very interesting that these mediating genes are more likely cancer type-specific as approximately 97.4% of genes are identified for a single cancer while only 2.6% (i.e., a total of 14 genes, including *CD79A*, *HPN*, *CBFA2T3*, *TACC3*, *KIF18B*, *RRM2*, *CIDEC*, *GNG7*, *B3GNT8*, *APOL2*, *CBX7*, *APOBEC3G*, *APOL1*, and *PEX11G*) are shared across distinct cancers (Fig 4F).

Functional role of selective DNA methylation and gene expression on TCGA cancers

It first needs to emphasize that among these methylation-expression mediation associations, nearly all are full mediations, with only one being partial mediation (i.e., *ZNF763* for CESC) (S6 Fig). *ZNF763* is a typical zinc finger protein containing Krüppel-associated box (KRAB), which is reportedly related to the development of cervical cancer. In addition, KRAB activates NF-κB by participating in the assembly of the IκB kinase (IKK) complex and promoting phosphorylation of IKK [56], while NF-κB is a multifunctional transcription factor and is related to the occurrence and development of cervical cancer [57]. Taken together, it is suggested that *ZNF763* may contribute to the occurrence of cervical cancer by enhancing NF-κB signaling and changing cell growth. In addition, among the 3,801 methylation-regulated genes that are shared by all analyzed cancers, 39 are histone genes. It has been revealed that histone gene loci are abnormally hypermethylated in a wide range of solid tumors [58,59]. For example, *HIST1H4F*, one of the identified histone genes with mediating effect, was abnormally hypermethylated in 17 types of cancers, acting as a potential universal-cancer-only methylation (UCOM) marker [59].

Prior studies also offered empirical evidence for these identified mediating genes, three of which are described in detail as follows, with all the mediating genes shown at https://github.com/biostatpzeng/IUSMMT. First, *TRIM26*, as a mediating gene associated with BLCA, was proved to play a key role in several types of cancers. For example, it was demonstrated that

TRIM26 had an oncogene impact on bladder cancer through regulating cell proliferation, migration and invasion via the Akt/GSK 3 β/β-catenin pathway [60], and methylation CpG sites had a significant effect on the regulation of *TRIM26* [61]. Second, *ZNF496*, as a mediating gene related to BRCA, was shown to significantly suppress ERα transactivation through over-expression, reduce the expression of estrogen receptor-alpha specific target genes, and inhibit the growth of breast cancer cells via ERα in an E2-dependent manner [62]. On the other hand, ERα is a key transcription factor involved in the proliferation and differentiation of mammary epithelia and has been demonstrated as an important predictor of breast cancer prognosis and a therapeutic target [63,64]. Moreover, the ZNF family, to which *ZNF496* belongs, has been shown to be mechanically linked to DNA methylation variability [65]. Third, *TMBIM6*, as a mediating gene discovered for HNSC, played an important role in the progression of laryngeal squamous cell carcinoma as a downstream target of *RBM15*-mediated N6-methyladenosine modification [66]. In addition, N6-methyladenosine methylation modification involved by *RBM15* regulates gene expression of *TMBIM*.

Enrichment pathway analysis for mediating genes

We performed the gene ontology (GO) and KEGG pathway enrichment analysis for all identified mediating genes using the DAVID database [67] and showed results in S7 Fig. In brief, the GO analysis demonstrates that the biological processes of these genes are concentrated in DNA template, immune response and cell-cell adhesion, the cellular components of these genes mainly involve cytoplasm, cytosol and perinuclear region of cytoplasm, and the molecular functions of these genes are primarily manifested in protein binding, cadherin binding and microtubule binding. The KEGG analysis of these genes also reveals multiple signaling pathways which are related to PI3K-Akt, Ras and histidine metabolism.

Direction of mediation effects

We further examined the direction of the effect of methylation CpG sites on expression and the effect of expression on the survival risk for these mediating genes (Table 4). It is found that the expression level of most of the genes (68.7% = 373/543) were inhibited by methylation alterations, while the expression level of some genes (~31.3%) was also upregulated by

Table 4. Direction of the effect of methylations on expression and the effect of expression on the survival risk of cancers for identified genes with mediating influence.

cancers	*N*	Direction (α & β)			
		+ +	- -	- +	+ -
BLCA	61	12	24	12	13
BRCA	46	2	28	4	12
CESC	1	1	0	0	0
HNSC	49	2	18	3	26
KIRP	34	5	11	16	2
LUAD	188	19	70	59	40
LUSC	3	2	0	1	0
SARC	156	8	59	64	25
STAD	5	1	1	3	0
total	543	52	211	162	118

Note: *N* denotes the number of genes with mediating effect; + or—refers to the positive or negative direction in the effect of methylations on expression (i.e., α) or in the effect of expression on the survival risk of cancers (β).

https://doi.org/10.1371/journal.pcbi.1009250.t004

methylation alterations, in line with prior observations of the dual function of methylations on gene expression [68, 69]. The downregulated expressions may lead to higher (43.4% = 162/373) or lower (56.6% = 211/373) survival risk of cancers, and the upregulated expressions can also likely result in higher (30.6% = 52/170) or lower (68.8% = 118/170) survival risk of cancers. Totally, 48.4% of (= 263/543) genes have the methylation-expression effect and the expression-survival effect in the same direction and 51.6% (= 280/543) of genes have opposite directions.

Distinguish passenger methylation events from mediation methylation events

Finally, we highlight that the effect pathway in our mediation analysis depends on a critical assumption that DNA methylation CpG sites regulate gene expression rather than conversely. However, such assumption may not fully hold as recent methylation studies in large cancer datasets have revealed that a bulk of methylation alterations observed in cancers might be a consequence of global epigenetic remodeling mechanism including (i) global loss via replication related errors; and (ii) CpG island/shore methylation gain associated with tumor proliferation [70,71]. Under this case, altered expression of a handful of key genes would have a reverse influence on DNA methylation CpG sites over the cancer epigenome. Namely, the expression of a gene can affect the survival risk of cancers while impacting DNA methylations, leading to the biological consequence that changed DNA methylations are just passenger events.

Therefore, distinguishing passenger methylation events from mediation methylation events that are likely to have a direct or indirect (via expression mediation) influence on the survival risk of cancers is necessary for the biological interpretation of results in our mediation analysis. In terms of the result of multivariate variance-component score test conducted via MultiSKAT [72], we find that approximately 16.3% (= 88/543) of the discovered mediation methylation events might be passenger methylation events across these cancers after the adjustment with Bonferroni's method ($P < 0.05$) (Table 2).

Discussions

In the present study we have proposed a novel statistical approach, called IUSMMT, for examining high-dimensional mediation effects in survival models with multiple exposures and one mediator. We have applied IUSMMT to ten TCGA cancers to identify genes that likely exhibited mediation effects of gene expression on the signaling pathway from DNA methylation CpG sites to the survival risk of cancers and observed some interesting results which provide supplementary insights on the biological mechanism from DNA methylation to gene expression and to survival outcome. First, only a few of DNA methylation CpG sites showed a total impact on the survival risk of cancers, although we cannot completely rule out the possibility of low statistical power due to small sample sizes and high censoring rates of these analyzed cancers. Second, we found that DNA methylation CpG sites can influence expression level directly for a wide range of genes and that most of these methylation-regulated genes were shared across distinct cancers, authenticating the prior finding that DNA methylation is a pervasive epigenomic mechanism of gene regulation [73,74].

Third, many genes were related to the survival risk of cancers, whereas most of them (~84.6%) were cancer type-specific, with only a few simultaneously shared between various cancers. As a result, although a wealth of genes mediated the influence of DNA methylation CpG sites on the overall survival time of cancers, the majority of the mediations associations were also cancer type-specific. This specificity largely reflects the inter-tumor heterogeneity of

the transcriptomic influence on cancers and the tissue distinction in cancer prognosis [75], implying that mediated pathways of the epigenomic impact vary cancer by cancer although DNA methylation pervasively regulates gene expression. It also has important implication on targeted therapies of precision cancers medicine by intervening DNA methylations [76–78].

Fourth, besides the indirect impact mediated by genes, multiple regions of DNA methylation CpG sites also have direct effects on these cancers. Fifth, we found that almost all the associations were full mediations, with only few partial mediations. Besides the truly biological mechanism, other possible explanations also include low power in examining the direct effects (i.e., γ^{DE}) of DNA methylation CpG sites for these TCGA cancers. Note that, potentially due to the lack of coverage of distal regulatory elements in the TCGA 450K chip, our analysis cannot include distal enhancer when defining methylation loci for every gene although we can consider methylation alterations around the promoter. It has been shown that the distal enhancers of CpG sites are unmethylated in normal cells but often gain methylations in cancer cells and tissues [79,80]. With the growing increase in the use of whole genome bisulfite whole genome sequencing (WGBS) and Illumina's newly released methylated EPIC BeadChip [81], including methylation loci around the enhancer would become feasible and important in epigenetic mediation analysis from both the biological and statistical perspectives. For instance, a mediation event, which was identified as partial mediation using only promoter CpG sites, may be tagged as fully mediated if enhancers/transcription factor binding would be considered as well. This is certainly an interesting topic for future work. It also needs to emphasize that our analysis cannot directly indicate these mediating genes include in proliferation or cell cycle related genes. Further experimental studies are warranted to elucidate the biological function of these genes.

In addition, although many studies have demonstrated that the variability in DNA methylation level can be largely attributable to diverse cell types and that cellular composition can be an important factor for elucidating biological processes [82] (e.g., the ratio of infiltrating neutrophils to plasma cells has important prognostic significance in cancer survival [83]), in the present analysis we followed prior work (e.g., [43,84]) and did not consider cell types as covariates due to two main reasons listed below. First, because cell types are often inferred directly from gene expressions or methylations [85–87], we did not incorporate them in the mediation model to avoid using data twice. Second, to our knowledge, adjusting cell types is primarily for DNA in the blood sample, and is very rare in the tumor tissue. Nevertheless, we recognize the critical role of cell types in the high-dimensional mediation analysis [88] and would pursue this issue in our future work.

It needs to highlight that IUSMMT adopted the principle of intersection-union test and is conceptually straightforward [47,49,50]. Like many previous mediation analyses, we implemented IUSMMT in two stages. Specifically, in the first step, the influence of DNA methylation CpG sites on the gene expression was examined and in the second step the association between the expression levels was investigated. Note that, although there are two null hypotheses involved in IUSMMT, no adjustment for multiplicity is needed because the overall null hypothesis of no mediation effect is rejected if and only if each of individual null hypotheses (rather than any of individual null hypotheses) is rejected. It is evident that the power of IUSMMT depends on the individual power in each stage.

Our analysis relies on an implicit assumption that DNA methylation CpG sites can regulate gene expression [89–92]. This assumption is largely supported by prior finding that epigenetic modification is closely relevant to the gene expression level which in turn has a direct function consequence on complex human diseases including malignant tumors [93,94]. Previous studies also showed that genes often exhibited cancer type-specific methylation changes and contributed to a higher incidence in cancer patients [95]. In addition, methylation CpG sites not

only affect the survival through the expression of its located genes but also work via a variety of other mechanisms, including splice variants and enhancer regions [96–98]. Recent work discovered that the methylation profiles of patients of four cancer subtypes played an important role in regulating gene expression during many biological processes and identified some functional genes with different methylation status in different subtypes [99].

However, the regulation assumption from DNA methylations to gene expression may not fully hold as revealed in terms of recent studies which found that methylation alterations can be reversely regulated by gene expression under the global epigenetic remodeling mechanism [70,71]. As a consequence, altered DNA methylation may simply represent a passenger event rather than a mediation event. To distinguish passenger methylation events from mediation methylation events, using the reverse regression analysis we found statistically supportive evidence that a small fraction of mediation associations may be passenger methylation events. However, this reverse analysis strategy is an *ad-hoc* approach which has no clear theoretical basis and which only considered the *cis* influence of gene expression regulation on methylation. It also demonstrates an important limitation of mediation analysis; namely, although it is useful for uncovering evidence of causal association, mediation analysis *per se* can be inadequate for completely determining the direction of the identified causal relationship. Therefore, addressing passenger methylation event in our epigenetic mediation analysis comprehensively is a promising but challenging direction for future investigation.

We finally highlight that there is still some room for further enhancement of the power in the first stage where the relationship between a set of DNA methylation CpG sites and gene expression level was examined with the variance-component test with a linear kernel function in the present study. As well-documented in previous studies, other more complicated kernels or a composite optimal kernel may have the potential to improve the power [100–106]. In addition, herein we implicitly suppose that all DNA methylation CpG sites had an influence on the gene expression, which may not hold in the real-life dataset. Instead, there may be only a few DNA methylation CpG sites regulating the expression level of a gene. Under this circumstance, a sparse relationship between DNA methylation CpG sites and the expression level should be modeled [51,107]. Furthermore, IUSMMT may be sub-optimal if the effect sizes of DNA methylation CpG sites located within the gene have the same direction (e.g., all are positive or negative). In this scenario, the burden test with a weighted score across these methylation CpG sites often leads to more powerful mediation approaches [33,101,108]. However, it seems very challenging to select a test that is consistently optimal across various settings because of the true relationship is unknown. Alternatively, it is hence desirable to construct a feasible omnibus mediation effect test that can aggregate all these strengths stated above. Nevertheless, in the present study we offer a very general framework for the assessment of mediation effect and we leave these issues mentioned above as an important and promising research avenue in our further work.

Materials and methods

Modeling framework for survival mediation analysis with multiple exposures

IUSMMT is a gene-centric high-dimensional survival mediation approach and considers one gene at each time. Assume there are multiple exposures ($\boldsymbol{M}$; e.g., a set of DNA methylation values for CpG sites located within a given gene), one mediator (G; e.g., expression level of that gene) and one survival outcome (T; including the survival time t and the survival status d) for n individuals. In the conventional mediation analysis [18–22], the

impact of $\boldsymbol{M}$ on T stands for the direct effect, and the influence of $\boldsymbol{M}$ on G and subsequently G on T for the indirect effect. Our objective is to evaluate the causal effect of $\boldsymbol{M}$ on T that is mediated via G. If $\boldsymbol{M}$ affects T only through G (i.e., $\boldsymbol{\gamma}^{\mathrm{DE}} = 0$), it is called full mediation; otherwise, it is referred to as partial mediation if $\boldsymbol{\gamma}^{\mathrm{DE}} \neq 0$ [109]. For a gene under investigation, the associations between DNA methylations, expression, or clinical covariates ($\boldsymbol{X}$) and the survival outcome can be determined within the framework of mediation analysis through the following procedures.

Step 1: Cox linear mixed-effects model testing for the total effect of methylations on the survival outcome

Following previous work [28,38], we first fit an exposure-outcome Cox model (i.e., methylation-survival model) to examine the association between a group of DNA methylation CpG sites ($\boldsymbol{M}$) of a gene of focus and the survival outcome (T) (i.e., $\boldsymbol{M} \rightarrow T$) while adjusting for the impact of existing covariates ($\boldsymbol{X}$)

$$\log(h(t|\boldsymbol{M},\boldsymbol{X})/h_0(t)) = \sum_{k=1}^{K} M_k \gamma_k^{TE} + \boldsymbol{X}\boldsymbol{w}_1 = \boldsymbol{M}\boldsymbol{\gamma}^{TE} + \boldsymbol{X}\boldsymbol{w}_1 \quad (1)$$

where $h(t|\boldsymbol{M},\boldsymbol{X})$ is the hazard risk at time t given $\boldsymbol{M}$ and $\boldsymbol{X}$; $\boldsymbol{\gamma}^{TE} = (\gamma_1^{TE}, \ldots, \gamma_K^{TE})$ is the vector of total effect sizes of exposures on the survival outcome; $\boldsymbol{w}_1 = (w_{11},\ldots,w_{1L})$ denotes the vector of effect sizes of L covariates in the exposure-outcome model; $h_0(t)$ is an arbitrary baseline hazard function, and K is the number of methylations. Nota bene, K varies gene by gene. In terms of the weighted residual method proposed in [110], we find that the proportional hazards assumption required by the utilization of the Cox model (here only covariates are considered; see below) is satisfied across all the ten TCGA cancers ($P>0.05$) (S5 Table).

To assess the relationship between $\boldsymbol{M}$ and T (H_0: $\boldsymbol{\gamma}^{TE} = 0$), one may treat $\boldsymbol{\gamma}^{TE}$ to be fixed effects and apply the classical score test (or multivariate Wald test) for hypothesis testing. However, as the number of methylations (i.e., K) in a gene might be very large up to several hundred and highly correlated with each other (S8 Fig), the fixed-effects test methods would have a large degree of freedom and are thus underpowered [31–33,111]. Alternatively, we assume $\gamma_k^{TE}(k = 1, \ldots, K)$ follows a normal distribution $N(0, \tau_1)$ in terms of previous relevant studies [32–34], leading to the so-called coxlmm [35,38]. Based on this effect assumption, we estimate and test the direct path within the framework of kernel-machine (KM) based Cox model relying on another equivalent null hypothesis H_0: $\tau_1 = 0$ via the coxKM package [112].

Traditionally, the presence of a significant total effect (i.e., $\boldsymbol{\gamma}^{TE} \neq 0$) is the prerequisite for the subsequent mediation test [22]. However, in practice it is not uncommon for the situation where the total effect is nonsignificant (i.e., $\boldsymbol{\gamma}^{TE} = 0$) but a substantial mediation effect remains [113]. Therefore, we always further investigate the existence of mediation effect irrespective of the significance or insignificance of $\boldsymbol{\gamma}^{TE}$.

Step 2: Linear mixed-effects model testing for the effect of methylations on gene expression

Next, in the exposure-mediator model (i.e., methylation-expression model) we evaluate the association path from methylation CpG sites to expression while controlling the influences of other covariates (i.e., $\boldsymbol{M} \rightarrow$ G). With the similar reason described in the first step, like $\boldsymbol{\gamma}^{TE}$ we

also suppose the effects of methylations $\boldsymbol{\alpha}$ have a normal distribution with variance τ_2

$$\begin{aligned} G &= \sum_{k=1}^{K} M_k \alpha_k + \boldsymbol{X}\boldsymbol{w}_2 + \boldsymbol{\varepsilon} = \boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{X}\boldsymbol{w}_2 + \boldsymbol{\varepsilon} \\ \alpha_k &\sim N(0, \tau_2) \\ \boldsymbol{\varepsilon} &\sim N(0, \sigma_\varepsilon^2) \end{aligned} \tag{2}$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is the vector of effect sizes of exposures on the mediator; $\boldsymbol{w}_2 = (w_{21}, \ldots, w_{2L})$ denotes the vector of the effect sizes of covariates in the exposure-mediator model; and $\boldsymbol{\varepsilon}$ is a mean-zero normal residual with variance σ_ε^2. We examine the null hypothesis H_0: $\boldsymbol{\alpha} = 0$ (or equivalently $\tau_2 = 0$) by utilizing the variance component-based score test within the framework of lmm [31,32,114–116] albeit the likelihood ratio-based test is also applicable [33,117,118]

$$Q = \sum_{k=1}^{K} \left\{ \sum_{i=1}^{n} \left[M_{ik} (G_i - \boldsymbol{X}_i \hat{\boldsymbol{w}}_2) \right]^2 \right\} \tag{3}$$

where $\hat{\boldsymbol{w}}_2$ is the estimate of $\boldsymbol{w}_2$ under the null model (i.e., $G = \boldsymbol{X}\boldsymbol{w}_2 + \boldsymbol{\varepsilon}$). Under H_0, the score statistic Q follows a mixture of chi-square distribution and the P-value is approximately obtained by Davies' method [32,119]. We implement this test via the SKAT package [32].

Step 3: Cox linear mixed-effects model testing for the effect of gene expression on the survival outcome

In the third step under the mediator-outcome model (i.e., expression-survival model), we examine the path from gene expression (G) to the survival outcome (T) conditional on methylation CpG sites (i.e., $G \rightarrow T$) through coxlmm

$$\log(h(t|\boldsymbol{M}, G, \boldsymbol{X})/h_0(t)) = \sum_{k=1}^{K} \boldsymbol{M}_k \gamma_k^{DE} + G\beta + \boldsymbol{X}\boldsymbol{w}_3 = \boldsymbol{M}\boldsymbol{\gamma}^{DE} + G\beta + \boldsymbol{X}\boldsymbol{w}_3 \tag{4}$$

where $\boldsymbol{\gamma}^{TE} = (\gamma_1^{TE}, \ldots, \gamma_K^{TE})$ is the vector of the direct effects of methylations, β is the effect size of gene expression, and $\boldsymbol{w}_3 = (w_{31}, \ldots, w_{3L})$ denotes the vector of the effect sizes of covariates in the mediator-outcome model. In the same principle, we suppose $\gamma_k^{DE} (k = 1, \ldots, K)$ follows a normal distribution $N(0, \tau_3)$. Herein, we are interested in testing H_0: $\beta = 0$. We estimate β and implement a Wald test via the coxme package [35,120].

Based on these models described above, the average nature direct effect and nature indirect effect can then be derived (S1 Text) [26,121]; the modeling assumptions required for estimation identifiability and causal interpretation of these effects are also given (S1 Text).

Intersection-union survival mixture-adjusted mediation test (IUSMMT)

Finally, to verify whether a given gene has mediation effect on the path from methylation CpG sites to the survival risk, we test the joint null hypothesis in which both effects have to be zero: H_0: $\boldsymbol{\alpha} = \beta = 0$ (or equivalently, H_0: $\tau_2 = \beta = 0$). As mentioned before, we need to implement the similar hypothesis testing across the whole genome. Therefore, this is a high-dimensional problem of mediation effect test. That is, we have H_{0j} ($j = 1, \ldots, S$) for all S genes. For simplicity, in the following we here ignore the subscript j. The individual mediation effect test is also equivalent to the hypothesis testing whether the mediation effect exists or not (H_0: $\boldsymbol{\alpha}\beta = \mathbf{0}$ versus H_1: $\boldsymbol{\alpha}\beta \neq \mathbf{0}$) in the absence of interactions between the exposures and the mediator and

can be divided into three composite null sub-hypotheses

$$H_0 = \begin{cases} H_{10} : \boldsymbol{\alpha} \neq 0 (\tau_2 \neq 0) \text{ and } \beta = 0 \\ H_{01} : \boldsymbol{\alpha} = 0 (\tau_2 = 0) \text{ and } \beta \neq 0 \\ H_{00} : \boldsymbol{\alpha} = 0 (\tau_2 = 0) \text{ and } \beta = 0 \end{cases} \quad (5)$$

The hypothesis in (5) can be formulated within the framework of the intersection-union test (IUT) (S2 Text) [46–50]

$$H_0 = H_{10} \cup H_{01} \cup H_{00} \text{ versus } H_1 = H_{10}^c \cap H_{01}^c \cap H_{00}^c \quad (6)$$

with A denoting the complement of set A.

To conduct IUT, we additionally define $H_0^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} = 0$ versus $H_1^{\boldsymbol{\alpha}} : \boldsymbol{\alpha} \neq 0$ and $H_0^{\boldsymbol{\beta}} : \boldsymbol{\beta} = 0$ versus $H_1^{\boldsymbol{\beta}} : \boldsymbol{\beta} \neq 0$, and suppose the P-value obtained from $H_0^{\boldsymbol{\alpha}}$ is $P_{\boldsymbol{\alpha}}$ and the P-value obtained from H_0^{β} is P_{β}. To ensure the desirable type I error control of IUT, we exploit $P_{\max} = \max(P_{\boldsymbol{\alpha}}, P_{\beta})$ to evaluate the overall significance of the mediation effect. IUT is advantageous in that the resulting mediation effect test has conceptual simplicity and feasibility of maneuver, and $P_{\max}$ *per se* can be employed as the P-value for assessing the significance of mediation [51,52,55,122]. Obviously, when rejecting H_0 if $P_{\max}$ is less than the given significance level of α, IUT is indeed a level-α test, representing that the type I error of IUT is guaranteed at most α once the rejection decision for H_0 is made [47,49].

However, IUT is oftentimes extremely conservative especially when both $H_0^{\boldsymbol{\alpha}}$ and H_0^{β} (i.e., H_{00}) hold [43,53], which is particularly true in genome-wide mediation studies where most of molecular markers such as gene expression or DNA methylation may be not expected to be related to the outcome of interest. To our knowledge, the commonly-used Sobel test [123,124], which is also often overly conservative [53,125], cannot be directly applicable for examining gene-centric high-dimensional mediation effects. Alternatively, to efficiently correct this conservativeness of IUT, we estimate the proportion for each component of the three null hypotheses and construct a new empirical null distribution for $P_{\max}$ by fitting a three-component mixture null distribution (S3 Text) [54]

$$\begin{aligned} \Pr(P_{\max} \leq u | H_0) &= \kappa_{10} p_{10} u + \kappa_{01} p_{01} u + \kappa_{00} u^2 \\ p_{10} &= \Pr(P_a \leq u | H_{10}) \\ p_{01} &= \Pr(P_\beta \leq u | H_{01}) \end{aligned} \quad (7)$$

where κ_{10}, κ_{01} and κ_{00} are the proportions corresponding to the three null hypotheses illustrated in (5), u is a given threshold value for the significance evaluation, while p_{10} and p_{01} are actually the power of rejecting $\boldsymbol{a} = 0$ under H_{10} or $\beta = 0$ under H_{01} respectively; and can be estimated via the Grenander method [126]. The estimation of these proportion parameters required in (7) can be easily implemented with methods that were well-established in the FDR literature (S4 Text) [29,127–135]. Once the estimates of these proportions are obtained, the estimated mixture null distribution for $P_{\max}$ can be built to control FWER or FDR. A comprehensive theoretical derivation with regards to the control of FWER or FDR can be conferred in [54].

To distinguish from the naïve IUT-based mediation test which takes $P_{\max}$ as the statistic and the uniform as the null distribution, we refer to the proposed approach as IUSMMT (intersection-union survival mixture-adjusted mediation test) which exploits the estimated mixture as the null distribution. IUSMMT is freely available at https://github.com/biostatpzeng/IUSMMT.

Simulation studies for type I error control and power evaluation

To evaluate the performance of IUSMMT, we undertake extensive simulations to investigate the type I error control and power. To simulate the truth in practice, we generated the gene expression level (G) and the survival outcome with 50 methylation CpG sites ($\boldsymbol{M}$) of *B3GALT4* on 548 BRCA individuals in TCGA [23]. Of note, the selection of this gene and this cancer was primarily because the number of methylation CpG sites and the sample size satisfied our simulation settings. Specifically, we first simulated G via a linear mixed-effect model with the number of methylation CpG sites varying in terms of a uniform distribution (ranging from 10 to 30, with an average of 20). Two covariates x_1 (based on the age of the BRCA patients) and x_2 (based on the cancer stage of the BRCA patients) ($\boldsymbol{X}$) were also included, both having an effect size of 0.5 in the exposure-mediator model

$$\begin{aligned} G &= \sum_{k=1}^{K} M_k \alpha_k + 0.5x_1 + 0.5x_2 + \boldsymbol{\varepsilon} = \boldsymbol{M\alpha} + 0.5x_1 + 0.5x_2 + \boldsymbol{\varepsilon} \\ \alpha_k &\sim N(0, \tau_2) \\ \boldsymbol{\varepsilon}_1 &\sim N(0, 1) \end{aligned} \tag{8}$$

Thereafter, we employed the inverse probability method to create the survival time (i.e. t) in terms of the Weibull distribution with a fixed shape parameter $\lambda = 1$ and a fixed scale parameter $\rho = 0.01$ [136]. The location parameter of the Weibull distribution was determined by $\boldsymbol{M}$, G and $\boldsymbol{X}$ in the mediator-outcome model

$$\begin{aligned} \log t &= \frac{1}{\rho}\log(-\log(u)/(\lambda\exp(\eta))) \\ \eta &= \sum_{k=1}^{K} \boldsymbol{M}_k \gamma_k^{DE} + G\beta + 0.5x_1 + 0.5x_2 = \boldsymbol{M\gamma}^{DE} + G\beta + 0.5x_1 + 0.5x_2 \\ u &\sim U(0, 1) \\ \boldsymbol{\gamma}^{DE} &\sim N(0, \tau_3) \end{aligned} \tag{9}$$

where u was a 0–1 uniform variable and $\tau_3 = 0.02$. The censored rate was fixed to be 50% in a random manner (the high censored rate corresponded to the similar situation observed in TCGA cancer dataset; see below).

To balance the statistical power, we set $\tau_2 = 0$, 0.01, 0.02, 0.04, 0.05 or 0.10 and $\beta = 0$, 0.15, 0.2, 0.25 or 0.30, and considered various configurations of the two parameters. For each configuration, we set the sample size n = 250, 400 or 548, and generated 10^4 datasets with $\boldsymbol{M}$, G, X, and T; that is, we had $S = 10^4$ genes (i.e., mediators). Because there were too many possibilities for the effect sizes and the mixture proportions, in the present study we were primarily interested in several cases that matched closely to our context application. Following the previous study [54], five scenarios for proportion parameters were designed (S1 Table), including the dense null under which κ_{11} was set to zero but κ_{10} or (and) κ_{10} had a small value, the sparse null under whichκ_{11} was set to zero, but in contrast to the sparse nulls, κ_{00} had a small value, and both κ_{10} and κ_{10} had a relatively large value, the complete null under whichκ_{00} was set to one, the sparse alternative under whichκ_{11} was not equal to zero, κ_{00} had a large value, and κ_{10} and κ_{10} were set to zero, and the dense alternative under whichκ_{11} was not equal to zero, but in contrast to the sparse alternative. The control of type I error was assessed with QQ plot and the power was calculated by the proportion of true positive discoveries among the true mediation effects. We repeated the simulations 100 times and took the average across the replications.

Application to ten TCGA cancer datasets

We finally applied IUSMMT to ten cancer datasets (S3 Table) publicly available from TCGA [23]. For these cancers, we downloaded their clinical information, RNAseq expressions as well as DNA methylations from UCSC Xena (https://xenabrowser.net/). DNA methylation levels were determined by Illumina Infinium HumanMethylation 450K platform and gene expression profiles were measured via the Illumina HiSeq 2000 RNA Sequencing platform. For methylation measurements, we removed non-CpG sites (i.e., these probes with *ch* labels) and CpG sites located on sex chromosomes; we also excluded cross-reactive probes as suggested in [137]. The beta values of methylation levels were logit-transformed to obtain the M-value for each CpG locus as the M-value is not bounded between zero and one and is thus more valid for our subsequent statistical analysis [138]. For each cancer, we focused on patients of self-report European ancestry while excluding patients whose tissues were formalin-fixed and paraffin-embedded. For gene expression measurements, we only considered protein coding genes and removed genes with over 50% zero expressions and variances smaller than 20% quantile of expression values. Then, we yielded methylation-expression pair for each gene in terms of the position annotation file provided by UCSC Xena. In each gene pair, we standardized both DNA methylation and expression level so that methylation and expression have a mean zero and variance one. According to TCGA gene annotation mapping file, we defined in our analysis methylations as those located within the entire gene body and a 500 bps upstream of the transcription start site (TSS) so that the promoter can be included. The distance of 500 bps is to some extent arbitrary and experience-based. In fact, the optimal extension distance upstream of TSS is not clear and various choices were applied in prior literature [139]. To examine the robustness of various extension distances, we performed a sensitivity analysis by incorporating methylation CpG sites within distinct extended regions, and found that the extension of the distance upstream of TSS in defining methylation loci seems to be very robust and has little influence on our final identification of mediating genes for methylation levels measured with the 450K platform (S9 Fig).

In addition, some clinical covariates available with only a few missing values, such as gender (coded as 0 or 1), age (treated as continuous variable), stage (coded from 1 to 5 and treated as continuous variable), estrogen receptor status (ER) (coded as 0 or 1) and progesterone receptor status (PR) (coded as 0 or 1), were also considered. We had to ignore other common covariates (e.g., alcohol consumption) that had too many missing values although they may be also greatly important. For the remaining datasets, the missing values were simply imputed by the mean if any. Notably, some of these covariates were cancer type-specific (e.g., ER and PR). Following previous studies [38,140–143], we employed the overall survival time and the corresponding survival status as the outcome as there was minimal ambiguity in defining an overall survival event [144]. In brief, overall survival in TCGA was the duration from the diagnosis of cancer to the death of patients. The employed TCGA cancer datasets and data process are summarized in S2 and S4 Tables.

Identify passenger methylation event via MultiSKAT

To distinguish passenger methylation events from mediation methylation events that are likely to have a direct or indirect (via expression mediation) influence on the survival risk of cancers, we here assume the gene expression may affect a set of DNA methylations and perform a reverse analysis by regressing methylation measurements on expression for all the identified mediating genes based on a multivariate model

$$\boldsymbol{M} = \sum_{k=1}^{K} G\tilde{\beta}_k + \boldsymbol{X}\tilde{\mathbf{W}}_2 + \tilde{\boldsymbol{\varepsilon}} = G\tilde{\boldsymbol{\beta}} + \boldsymbol{X}\tilde{\mathbf{W}}_2 + \tilde{\boldsymbol{\varepsilon}} \tag{10}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_K)$ denotes the vector of effect sizes of gene expression on DNA methylation sites, $\tilde{\mathbf{W}}_2$ is the matrix of effect sizes for covariates, and $\tilde{\boldsymbol{\varepsilon}}$ is the matrix of residual error terms. We further suppose that each $\tilde{\beta}_k$ ($k = 1, \ldots, K$) follows a normal distribution with mean zero and variance τ. Then, our objective is to test for the null hypothesis H_0: $\tilde{\boldsymbol{\beta}} = 0$, which is equivalent to examining H_0: $\tau = 0$. The corresponding variance component score test statistic is given as

$$Q_M = \{G^T(\boldsymbol{M} - \hat{\boldsymbol{\mu}})\hat{\mathbf{V}}^{-1}\}^T\{G^T(\boldsymbol{M} - \hat{\boldsymbol{\mu}})\hat{\mathbf{V}}^{-1}\} \quad (11)$$

where $\hat{\boldsymbol{\mu}}$ and $\hat{\mathbf{V}}$ are the estimated mean and covariance of $\boldsymbol{M}$ under the null hypothesis. Then, the test statistic Q_M follows a mixture of chi-square distribution and this test can be implemented via MultiSKAT [72].

Supporting information

S1 Fig. Estimated power for α and β. These powers are estimated under six alternative simulation scenarios from A to I with various mixture proportions. The graph in the left column is the power of β, and the x-axis is the value of $\beta = 0.15, 0.20, 0.25$ or 0.30; The graph in the left column is the power of α, and the x-axis is the value of $\tau_2 = 0.01, 0.02, 0.04, 0.05$ or 0.10. These powers are estimated by the average across the 100 replications.
(TIF)

S2 Fig. QQ plot for IUSMMT and IUT under various the scenarios of nulls and sample sizes. Here (A), (B) and (C) represent the sample size n = 250, 400 and 548, respectively.
(TIF)

S3 Fig. Estimated power for IUSMMT and the IUT method. Here, $\tau_2 = 0.01, 0.02, 0.04, 0.05$ or 0.10 at the x-axis, $\beta = 0.15, 0.20, 0.25$ or 0.30 on the top. These powers are estimated for the nine alternative simulation scenarios from A to I with various values for the mixture proportions by the average across the 100 replications. (A), (B) and (C) represent the sample size n = 250, 400 and 548, respectively.
(TIF)

S4 Fig. QQ plot for the mediation effect test statistic $P_{\max}$ for 10 TCGA cancers.
(TIF)

S5 Fig. (A) **Proportion of overlapped methylation-regulated genes discovered in various cancer tumor tissues and these discovered in normal tissue.** (B) **Estimated values of τ_2 in the BRCA tissue and in the normal tissue.** Here, τ_2 can be employed to quantify the magnitude of the methylation effect on expression, the sample size for the normal tissue is combined and analyzed across all the 10 cancers.
(TIF)

S6 Fig. Partial mediation framework of *ZNF763* for CESC. Please refer to Fig 1 for the interpretation of these parameters shown herein.
(TIF)

S7 Fig. Results of KEGG pathway enrichment analysis and the significant terms identified by GO enrichment analysis for the genes.
(TIF)

S8 Fig. Number of methylation markers belonging to each gene across the while genome in TCGA cancer in terms of the annotation mapping file provided by UCSC Xena.
(TIF)

S9 Fig. Correlation between *P*-values (in–log10 scale) of the methylation-expression association for all genes with meditation effects with various distances before the transcription start site (TSS) so that the promoter can be included. Here, various distances before the TSS were extended, ranging from 500bp to 5000bp with an increment of 500bp. For each extension, the methylations within that extended region and gene body were included to examine their relationship with gene expression using the variance-component score test. The *P*-values calculated with methylations within a 500bp upstream of the TSS were treated as the reference.
(TIF)

S1 Table. Estimated and true proportion parameters (mean and standard deviation) in the three-component mixture null distribution under the five simulation scenarios and different sample sizes and different numbers of mediators.
(DOCX)

S2 Table. Comparison of the effect of different numbers of mediators on the bias of the estimators for these proportion parameters.
(DOCX)

S3 Table. Data process of the ten TCGA cancers used in our mediation analysis.
(DOCX)

S4 Table. Basic characteristics of the ten TCGA cancer datasets.
(DOCX)

S5 Table. Results for testing for the proportional hazards assumption in the used Cox model.
(DOCX)

S1 Text. Effects in survival mediation analysis with multiple exposures, and identifiability assumptions.
(DOCX)

S2 Text. Intersection-union test.
(DOCX)

S3 Text. Three-component mixture null distribution.
(DOCX)

S4 Text. Estimation of proportion parameters.
(DOCX)

Acknowledgments

We would like to express our gratitude to TCGA for making genomic datasets publicly available and are indebted to all the investigators and participants contributed to this project. Data analyses and simulations in the present study were carried out with the high-performance computing cluster that was supported by the special central finance project of local universities for Xuzhou Medical University.

Author Contributions

Conceptualization: Ping Zeng.

Data curation: Zhonghe Shao, Ting Wang.

Formal analysis: Zhonghe Shao, Ting Wang.

Funding acquisition: Ping Zeng.

Methodology: Zhonghe Shao.

Project administration: Ping Zeng.

Resources: Zhonghe Shao.

Software: Zhonghe Shao.

Supervision: Shuiping Huang.

Validation: Meng Zhang, Zhou Jiang.

Visualization: Zhonghe Shao, Meng Zhang.

Writing – original draft: Zhonghe Shao.

Writing – review & editing: Zhonghe Shao, Ping Zeng.

References

1. Xiong J, Li Y, Huang K, Lu M, Shi H, Ma L, et al. Association between DAPK1 promoter methylation and cervical cancer: a meta-analysis. PLoS One. 2014; 9(9):e107272–e. https://doi.org/10.1371/journal.pone.0107272 PMID: 25268905.
2. Baylin SB. The cancer epigenome: its origins, contributions to tumorigenesis, and translational implications. Proc Am Thorac Soc. 2012; 9(2):64–5. https://doi.org/10.1513/pats.201201-001MS PMID: 22550245.
3. Momparler RL, Bovenzi V. DNA methylation and cancer. J Cell Physiol. 2000; 183(2):145–54. https://doi.org/10.1002/(SICI)1097-4652(200005)183:2<145::AID-JCP1>3.0.CO;2-V PMID: 10737890
4. Busslinger M, Hurst J, Flavell RA. DNA methylation and the regulation of globin gene expression. Cell. 1983; 34(1):197–206. https://doi.org/10.1016/0092-8674(83)90150-2 PMID: 6883509.
5. Deaton AM, Bird A. CpG islands and the regulation of transcription. Genes Dev. 2011; 25(10):1010–22. https://doi.org/10.1101/gad.2037511 PMID: 21576262.
6. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology: official publication of the American College of Neuropsychopharmacology. 2013; 38(1):23–38. https://doi.org/10.1038/npp.2012.112 PMID: 22781841.
7. Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. Genome Biol. 2014; 15(2):R37. https://doi.org/10.1186/gb-2014-15-2-r37 PMID: 24555846
8. Huang WY, Hsu SD, Huang HY, Sun YM, Chou CH, Weng SL, et al. MethHC: a database of DNA methylation and gene expression in human cancer. Nucleic Acids Res. 2015; 43(Database issue): D856–61. Epub 2014/11/16. https://doi.org/10.1093/nar/gku1151 PMID: 25398901; PubMed Central PMCID: PMC4383953.
9. Schübeler D. Function and information content of DNA methylation. Nature. 2015; 517(7534):321–6. https://doi.org/10.1038/nature14192 PMID: 25592537
10. Das PM, Singal R. DNA Methylation and Cancer. J Clin Oncol. 2004; 22(22):4632–42. https://doi.org/10.1200/JCO.2004.07.151 PMID: 15542813.
11. Blake LE, Roux J, Hernando-Herraez I, Banovich NE, Perez RG, Hsiao CJ, et al. A comparison of gene expression and DNA methylation patterns across tissues and species. Genome Res. 2020; 30(2):250–62. https://doi.org/10.1101/gr.254904.119 PMID: 31953346
12. Taylor DL, Jackson AU, Narisu N, Hemani G, Erdos MR, Chines PS, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. Proceedings of the National Academy of Sciences. 2019; 116(22):10883–8. https://doi.org/10.1073/pnas.1814263116 PMID: 31076557

13. Gong J, Wan H, Mei S, Ruan H, Zhang Z, Liu C, et al. Pancan-meQTL: a database to systematically evaluate the effects of genetic variants on methylation in human cancer. Nucleic Acids Res. 2019; 47 (D1):D1066–D72. https://doi.org/10.1093/nar/gky814 PMID: 30203047.

14. Dermitzakis ET. From gene expression to disease risk. Nat Genet. 2008; 40(5):492–3. https://doi.org/10.1038/ng0508-492 PMID: 18443581

15. GTEx Consortium. Genetic effects on gene expression across human tissues. Nature. 2017; 550 (7675):204–13. https://doi.org/10.1038/nature24277 http://www.nature.com/nature/journal/v550/n7675/abs/nature24277.html#supplementary-information. PMID: 29022597

16. GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015; 348(6235):648–60. https://doi.org/10.1126/science.1262110 PMID: 25954001

17. Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, et al. Genetics of gene expression and its effect on disease. Nature. 2008; 452(7186):423–8. https://doi.org/10.1038/nature06758 PMID: 18344981

18. MacKinnon DP. Introduction to statistical mediation analysis: Routledge; 2008.

19. VanderWeele T. Explanation in causal inference: methods for mediation and interaction: Oxford University Press; 2015.

20. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010; 15(4):309. https://doi.org/10.1037/a0020761 PMID: 20954780

21. Hicks R, Tingley D. Causal mediation analysis. The Stata Journal. 2011; 11(4):605.

22. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986; 51(6):1173–82. https://doi.org/10.1037//0022-3514.51.6.1173 PMID: 3806354

23. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018; 173(2):291–304.e6. https://doi.org/10.1016/j.cell.2018.03.022 PMID: 29625048.

24. Tein J-Y, MacKinnon DP, editors. Estimating Mediated Effects with Survival Data. New Developments in Psychometrics; 2003 2003//; Tokyo: Springer Japan.

25. Lange T, Hansen JV. Direct and Indirect Effects in a Survival Context. Epidemiology. 2011; 22 (4):575–81. https://doi.org/10.1097/EDE.0b013e31821c680c 00001648-201107000-00024. PMID: 21552129

26. VanderWeele TJ. Causal mediation analysis with survival data. Epidemiology. 2011; 22(4):582. https://doi.org/10.1097/EDE.0b013e31821db37e PMID: 21642779

27. Wang L, Zhang Z. Estimating and Testing Mediation Effects with Censored Data. Structural Equation Modeling: A Multidisciplinary Journal. 2011; 18(1):18–34. https://doi.org/10.1080/10705511.2011.534324

28. Luo C, Fa B, Yan Y, Wang Y, Zhou Y, Zhang Y, et al. High-dimensional mediation analysis in survival models. PLoS Comput Biol. 2020; 16(4):e1007768. https://doi.org/10.1371/journal.pcbi.1007768 PMID: 32302299

29. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics. 2016; 32(20):3150–4. https://doi.org/10.1093/bioinformatics/btw351 PMID: 27357171

30. Zeng P, Shao Z, Zhou X. Statistical methods for mediation analysis in the era of high-throughput genomics: current successes and future challenges. Computational and Structural Biotechnology Journal. 2021;in press. https://doi.org/10.1016/j.csbj.2021.05.042 PMID: 34141140.

31. Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, et al. Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. Am J Hum Genet. 2010; 86(6):929–42. https://doi.org/10.1016/j.ajhg.2010.05.002 PMID: 20560208

32. Wu Michael C, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011; 89(1):82–93. https://doi.org/10.1016/j.ajhg.2011.05.029 PMID: 21737059

33. Zeng P, Zhao Y, Liu J, Liu L, Zhang L, Wang T, et al. Likelihood Ratio Tests in Rare Variant Detection for Continuous Phenotypes. Ann Hum Genet. 2014; 78(5):320–32. https://doi.org/10.1111/ahg.12071 PMID: 25117149

34. Zeng P, Zhao Y, Li H, Wang T, Chen F. Permutation-based variance component test in generalized linear mixed model with application to multilocus genetic association study. BMC Med Res Methodol. 2015; 15:37. https://doi.org/10.1186/s12874-015-0030-1 PMID: 25897803

35. Therneau TM, Grambsch PM, Pankratz VS. Penalized survival models and frailty. J Comput Graph Statist. 2003; 12:156–75.

36. Ripatti S, Palmgren J. Estimation of Multivariate Frailty Models Using Penalized Partial Likelihood. Biometrics. 2000; 56(4):1016–22. https://doi.org/10.1111/j.0006-341x.2000.01016.x PMID: 11129456

37. Therneau TM, Grambsch PM. Modelling Survival Data: Extending the Cox Model. New York: Springer; 2000.

38. Yu X, Wang T, Huang S, Zeng P. How can gene expression information improve prognostic prediction in TCGA cancers: an empirical comparison study on regularization and mixed-effect survival models. Frontiers in Genetics. 2020; 11(920). https://doi.org/10.3389/fgene.2020.00920 PMID: 32973875

39. VanderWeele TJ. Mediation analysis: a practitioner's guide. Annu Rev Public Health. 2016; 37:17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402 PMID: 26653405

40. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992:143–55. https://doi.org/10.1097/00001648-199203000-00013 PMID: 1576220

41. MacKinnon DP, Fairchild AJ, Fritz MS. Mediation analysis. Annu Rev Psychol. 2007; 58:593–614. https://doi.org/10.1146/annurev.psych.58.110405.085542 PMID: 16968208

42. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. Stat Sci. 2010; 25(1):51–71.

43. Huang Y-T. Genome-wide analyses of sparse mediation effects under composite null hypotheses. The Annals of Applied Statistics. 2019; 13(1):60–84. https://doi.org/10.1214/18-aoas1181

44. Pearl J. The causal mediation formula—a guide to the assessment of pathways and mechanisms. Prev Sci. 2012; 13(4):426–36. https://doi.org/10.1007/s11121-011-0270-1 PMID: 22419385

45. Pearl J. Interpretation and identification of causal mediation. Psychol Methods. 2014; 19(4):459–81. https://doi.org/10.1037/a0036434 PMID: 24885338.

46. Berger RL. Uniformly More Powerful Tests for Hypotheses concerning Linear Inequalities and Normal Means. J Am Stat Assoc. 1989; 84(405):192–9. https://doi.org/10.1080/01621459.1989.10478755

47. Berger RL. Multiparameter Hypothesis Testing and Acceptance Sampling. Technometrics. 1982; 24 (4):295–300. https://doi.org/10.1080/00401706.1982.10487790

48. Sen PK, Tsai M-T. Two-Stage Likelihood Ratio and Union–Intersection Tests for One-Sided Alternatives Multivariate Mean with Nuisance Dispersion Matrix. J Multivariate Anal. 1999; 68(2):264–82. https://doi.org/10.1006/jmva.1998.1791.

49. Berger RL, Hsu JC. Bioequivalence trials, intersection-union tests and equivalence confidence sets. Stat Sci. 1996; 11(4):283–319. https://doi.org/10.1214/ss/1032280304

50. Berger RL. Likelihood Ratio Tests and Intersection-Union Tests. In: Panchapakesan S, Balakrishnan N, editors. Advances in Statistical Decision Theory and Applications. Boston, MA: Birkhäuser Boston; 1997. p. 225–37.

51. Gao Y, Yang H, Fang R, Zhang Y, Goode EL, Cui Y. Testing Mediation Effects in High-Dimensional Epigenetic Studies. Frontiers in Genetics. 2019; 10:1195. https://doi.org/10.3389/fgene.2019.01195 PMID: 31824577.

52. Zhong W, Spracklen CN, Mohlke KL, Zheng X, Fine J, Li Y. SMUT: Multi-SNP mediation intersection-union test. Bioinformatics. 2019; 35(22):4724–9. https://doi.org/10.1093/bioinformatics/btz285 PMID: 31099385

53. Barfield R, Shen J, Just AC, Vokonas PS, Schwartz J, Baccarelli AA, et al. Testing for the indirect effect under the null for genome-wide mediation analyses. Genet Epidemiol. 2017; 41(8):824–33. https://doi.org/10.1002/gepi.22084 PMID: 29082545

54. Dai JY, Stanford JL, LeBlanc M. A Multiple-Testing Procedure for High-Dimensional Mediation Hypotheses. J Am Stat Assoc. 2020:1–16. https://doi.org/10.1080/01621459.2020.1765785

55. Zhong W, Darville T, Zheng X, Fine J, Li Y. Generalized Multi-SNP Mediation Intersection-Union Test. bioRxiv. 2019:780767. https://doi.org/10.1101/780767

56. Wang W, Guo M, Hu L, Cai J, Zeng Y, Luo J, et al. The zinc finger protein ZNF268 is overexpressed in human cervical cancer and contributes to tumorigenesis via enhancing NF-κB signaling. J Biol Chem. 2012; 287(51):42856–66. Epub 2012/10/22. https://doi.org/10.1074/jbc.M112.399923 PMID: 23091055.

57. Nair A, Venkatraman M, Maliekal TT, Nair B, Karunagaran D. NF-kappaB is constitutively activated in high-grade squamous intraepithelial lesions and squamous cell carcinomas of the human uterine cervix. Oncogene. 2003; 22(1):50–8. Epub 2003/01/16. https://doi.org/10.1038/sj.onc.1206043 PMID: 12527907.

58. Fritz AJ, Ghule PN, Boyd JR, Tye CE, Page NA, Hong D, et al. Intranuclear and higher-order chromatin organization of the major histone gene cluster in breast cancer. Journal of Cellular Physiology. 2018; 233(2):1278–90. https://doi.org/10.1002/jcp.25996 PMID: 28504305

59. Dong S, Li W, Wang L, Hu J, Song Y, Zhang B, et al. Histone-Related Genes Are Hypermethylated in Lung Cancer and Hypermethylated HIST1H4F Could Serve as a Pan-Cancer Biomarker. Cancer Research. 2019; 79(24):6101. https://doi.org/10.1158/0008-5472.CAN-19-1019 PMID: 31575549

60. Xie X, Li H, Pan J, Han X. Knockdown of TRIM26 inhibits the proliferation, migration and invasion of bladder cancer cells through the Akt/GSK3β/β-catenin pathway. Chemico-Biological Interactions. 2021; 337:109366. https://doi.org/10.1016/j.cbi.2021.109366 PMID: 33549581

61. Zhang X, Pei L, Li R, Zhang W, Yang H, Li Y, et al. Spina bifida in fetus is associated with an altered pattern of DNA methylation in placenta. J Hum Genet. 2015; 60(10):605–11. Epub 2015/07/17. https://doi.org/10.1038/jhg.2015.80 PMID: 26178427.

62. Wang J, Zhang X, Ling J, Wang Y, Xu X, Liu Y, et al. KRAB-containing zinc finger protein ZNF496 inhibits breast cancer cell proliferation by selectively repressing ERα activity. Biochim Biophys Acta Gene Regul Mech. 2018. Epub 2018/07/18. https://doi.org/10.1016/j.bbagrm.2018.07.003 PMID: 30012466.

63. Burns KA, Korach KS. Estrogen receptors and human disease: an update. Archives of Toxicology. 2012; 86(10):1491–504. https://doi.org/10.1007/s00204-012-0868-5 PMID: 22648069

64. Arnal J-F, Lenfant F, Metivier R, Flouriot G, Henrion D, Adlanmerini M, et al. Membrane and Nuclear Estrogen Receptor Alpha Actions: From Tissue Specificity to Medical Implications. Physiological Reviews. 2017; 97(3):1045–87. https://doi.org/10.1152/physrev.00024.2016 PMID: 28539435

65. Bertozzi TM, Elmer JL, Macfarlan TS, Ferguson-Smith AC. KRAB zinc finger protein diversification drives mammalian interindividual methylation variability. Proceedings of the National Academy of Sciences of the United States of America. 2020; 117(49):31290–300. Epub 2020/11/25. https://doi.org/10.1073/pnas.2017053117 PMID: 33239447.

66. Wang X, Tian L, Li Y, Wang J, Yan B, Yang L, et al. RBM15 facilitates laryngeal squamous cell carcinoma progression by regulating TMBIM6 stability through IGF2BP3 dependent. J Exp Clin Cancer Res. 2021; 40(1):80–. https://doi.org/10.1186/s13046-021-01871-4 PMID: 33637103.

67. The Medicaid Outcomes Distributed Research N. Use of Medications for Treatment of Opioid Use Disorder Among US Medicaid Enrollees in 11 States, 2014–2018. JAMA. 2021; 326(2):154–64. https://doi.org/10.1001/jama.2021.7374 PMID: 34255008

68. Jones PA. The DNA methylation paradox. Trends Genet. 1999; 15(1):34–7. https://doi.org/10.1016/s0168-9525(98)01636-9 PMID: 10087932.

69. Kass S, Landsberger N, Wolffe A. DNA methylation directs a time-dependent repression of transcription initiation. Curr Biol. 1997; 7:157–65. https://doi.org/10.1016/s0960-9822(97)70086-1 PMID: 9395433

70. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018; 50(4):591–602. Epub 2018/04/04. https://doi.org/10.1038/s41588-018-0073-4 PMID: 29610480; PubMed Central PMCID: PMC5893360.

71. Meir Z, Mukamel Z, Chomsky E, Lifshitz A, Tanay A. Single-cell analysis of clonal maintenance of transcriptional and epigenetic states in cancer cells. Nat Genet. 2020; 52(7):709–18. Epub 2020/07/01. https://doi.org/10.1038/s41588-020-0645-y PMID: 32601473; PubMed Central PMCID: PMC7610382.

72. Dutta D, Scott L, Boehnke M, Lee S. Multi-SKAT: General framework to test for rare-variant association with multiple phenotypes. Genet Epidemiol. 2019; 43(1). https://doi.org/10.1002/gepi.22156 PMID: 30298564.

73. Jones PA, Baylin SB. The fundamental role of epigenetic events in cancer. Nat Rev Genet. 2002; 3(6):415–28. Epub 2002/06/04. https://doi.org/10.1038/nrg816 PMID: 12042769.

74. Feinberg AP, Koldobskiy MA, Göndör A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet. 2016; 17(5):284–99. Epub 2016/03/15. https://doi.org/10.1038/nrg.2016.13 PMID: 26972587; PubMed Central PMCID: PMC4888057.

75. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. Mol Cancer. 2017; 16(1):41–. https://doi.org/10.1186/s12943-017-0600-4 PMID: 28209166.

76. Cheng Y, He C, Wang M, Ma X, Mo F, Yang S, et al. Targeting epigenetic regulators for cancer therapy: mechanisms and advances in clinical trials. Signal Transduction and Targeted Therapy. 2019; 4(1):62. https://doi.org/10.1038/s41392-019-0095-0 PMID: 31871779

77. Issa J-PJ. DNA methylation as a therapeutic target in cancer. Clinical cancer research: an official journal of the American Association for Cancer Research. 2007; 13(6):1634–7. https://doi.org/10.1158/1078-0432.CCR-06-2076 PMID: 17363514.

78. Li J, Su X, Dai L, Chen N, Fang C, Dong Z, et al. Temporal DNA methylation pattern and targeted therapy in colitis-associated cancer. Carcinogenesis. 2019; 41(2):235–44. https://doi.org/10.1093/carcin/bgz199 PMID: 31802101

79. Pidsley R, Zotenko E, Peters TJ, Lawrence MG, Risbridger GP, Molloy P, et al. Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. Genome Biol. 2016; 17(1):208. Epub 2016/10/09. https://doi.org/10.1186/s13059-016-1066-1 PMID: 27717381; PubMed Central PMCID: PMC5055731.

80. Akhtar-Zaidi B, Cowper-Sal-lari R, Corradin O, Saiakhova A, Bartels CF, Balasubramanian D, et al. Epigenomic enhancer profiling defines a signature of colon cancer. Science. 2012; 336(6082):736–9. Epub 2012/04/14. https://doi.org/10.1126/science.1217277 PMID: 22499810; PubMed Central PMCID: PMC3711120.

81. Stirzaker C, Taberlay PC, Statham AL, Clark SJ. Mining cancer methylomes: prospects and challenges. Trends Genet. 2014; 30(2):75–84. Epub 2013/12/26. https://doi.org/10.1016/j.tig.2013.11.004 PMID: 24368016.

82. Jaffe AE, Irizarry RA. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol. 2014; 15(2):R31. Epub 2014/02/06. https://doi.org/10.1186/gb-2014-15-2-r31 PMID: 24495553; PubMed Central PMCID: PMC4053810.

83. Gentles AJ, Newman AM, Liu CL, Bratman SV, Feng W, Kim D, et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. Nat Med. 2015; 21(8):938–45. Epub 2015/07/21. https://doi.org/10.1038/nm.3909 PMID: 26193342; PubMed Central PMCID: PMC4852857.

84. Huang YT. Variance component tests of multivariate mediation effects under composite null hypotheses. Biometrics. 2019; 75(4):1191–204. Epub 2019/04/23. https://doi.org/10.1111/biom.13073 PMID: 31009061.

85. Arneson D, Yang X, Wang K. MethylResolver—a method for deconvoluting bulk DNA methylation profiles into known and unknown cell contents. Communications Biology. 2020; 3(1):422. https://doi.org/10.1038/s42003-020-01146-2 PMID: 32747663

86. Li T, Fu J, Zeng Z, Cohen D, Li J, Chen Q, et al. TIMER2.0 for analysis of tumor-infiltrating immune cells. Nucleic Acids Res. 2020; 48(W1):W509–W14. https://doi.org/10.1093/nar/gkaa407 PMID: 32442275.

87. Jaakkola MK, Elo LL. Computational deconvolution to estimate cell type-specific gene expression from bulk data. NAR Genom Bioinform. 2021; 3(1):lqaa110. Epub 2021/02/13. https://doi.org/10.1093/nargab/lqaa110 PMID: 33575652; PubMed Central PMCID: PMC7803005.

88. Luo X, Schwartz J, Baccarelli A, Liu Z. Testing cell-type-specific mediation effects in genome-wide epigenetic studies. Brief Bioinform. 2021; 22(3). Epub 2020/07/08. https://doi.org/10.1093/bib/bbaa131 PMID: 32632436; PubMed Central PMCID: PMC8138838.

89. Glinsky GV. Integration of HapMap-Based SNP Pattern Analysis and Gene Expression Profiling Reveals Common SNP Profiles for Cancer Therapy Outcome Predictor Genes*. Cell Cycle. 2006; 5(22):2613–25. https://doi.org/10.4161/cc.5.22.3498 PMID: 17172834

90. Fabiani E, Leone G, Giachelia M, D'Alo F, Greco M, Criscuolo M, et al. Analysis of genome-wide methylation and gene expression induced by 5-aza-2'-deoxycytidine identifies BCL2L10 as a frequent methylation target in acute myeloid leukemia. Leuk Lymphoma. 2010; 51(12):2275–84. https://doi.org/10.3109/10428194.2010.528093 PMID: 21077739.

91. Wang W, Baladandayuthapani V, Morris JS, Broom BM, Manyam G, Do K-A. iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data. Bioinformatics. 2013; 29(2):149–59. https://doi.org/10.1093/bioinformatics/bts655 PMID: 23142963

92. de Tayrac M, Lê S, Aubry M, Mosser J, Husson F. Simultaneous analysis of distinct Omics data sets with integration of biological knowledge: Multiple Factor Analysis approach. BMC Genomics. 2009; 10(1):32. https://doi.org/10.1186/1471-2164-10-32 PMID: 19154582

93. Bernstein BE, Meissner A, Lander ES. The mammalian epigenome. Cell. 2007; 128(4):669–81. Epub 2007/02/27. https://doi.org/10.1016/j.cell.2007.01.033 PMID: 17320505.

94. Liang G, Weisenberger DJ. DNA methylation aberrancies as a guide for surveillance and treatment of human cancers. Epigenetics. 2017; 12(6):416–32. Epub 2017/03/30. https://doi.org/10.1080/15592294.2017.1311434 PMID: 28358281.

95. Tsai H, Baylin S. Cancer epigenetics: linking basic biology to clinical medicine. Cell research. 2011; 21(3):502–17. https://doi.org/10.1038/cr.2011.24 PMID: 21321605.

96. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009; 1(2):239–59. https://doi.org/10.2217/epi.09.33 PMID: 20495664.

97. Lopez-Serra P, Esteller M. DNA methylation-associated silencing of tumor-suppressor microRNAs in cancer. Oncogene. 2012; 31(13):1609–22. https://doi.org/10.1038/onc.2011.354 PMID: 21860412

98. Wan J, Oliver VF, Wang G, Zhu H, Zack DJ, Merbs SL, et al. Characterization of tissue-specific differential DNA methylation suggests distinct modes of positive and negative gene expression regulation. BMC Genomics. 2015; 16(1):49. https://doi.org/10.1186/s12864-015-1271-4 PMID: 25652663

99. Chen L, T Z, X P, YH Z, T H, YD C. Identifying Methylation Pattern and Genes Associated with Breast Cancer Subtypes. International journal of molecular sciences. 2019; 20(17). https://doi.org/10.3390/ijms20174269 PMID: 31480430.

100. Fang R, Yang H, Gao Y, Cao H, Goode EL, Cui Y. Gene-based mediation analysis in epigenetic studies. Brief Bioinform. 2020. https://doi.org/10.1093/bib/bbaa113 PMID: 32608480

101. Wu MC, Maity A, Lee S, Simmons EM, Harmon QE, Lin X, et al. Kernel Machine SNP-Set Testing Under Multiple Candidate Kernels. Genet Epidemiol. 2013; 37(3):267–75. https://doi.org/10.1002/gepi.21715 PMID: 23471868

102. Urrutia E, Lee S, Maity A, Zhao N, Shen J, Li Y, et al. Rare variant testing across methods and thresholds using the multi-kernel sequence kernel association test (MK-SKAT). Stat Interface. 2015; 8 (4):495–505. https://doi.org/10.4310/SII.2015.v8.n4.a8 PMC4698916. PMID: 26740853

103. Wang X, Xing EP, Schaid DJ. Kernel methods for large-scale genomic data analysis. Brief Bioinform. 2014; 16(2):183–92. https://doi.org/10.1093/bib/bbu024 PMID: 25053743

104. Yang H, Cao H, He T, Wang T, Cui Y. Multilevel heterogeneous omics data integration with kernel fusion. Brief Bioinform. 2020; 21(1):156–70. https://doi.org/10.1093/bib/bby115 PMID: 30496340

105. Yang H, Li S, Cao H, Zhang C, Cui Y. Predicting disease trait with genomic data: a composite kernel approach. Brief Bioinform. 2016; 18(4):591–601. https://doi.org/10.1093/bib/bbw043 PMID: 27255915

106. He T, Li S, Zhong P-S, Cui Y. An optimal kernel-based U-statistic method for quantitative gene-set association analysis. Genet Epidemiol. 2019; 43(2):137–49. https://doi.org/10.1002/gepi.22170 PMID: 30456931

107. Song Y, Zhou X, Zhang M, Zhao W, Liu Y, Kardia SLR, et al. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. Biometrics. 2020; 76(3):700–10. https://doi.org/10.1111/biom.13189 PMID: 31733066

108. Ionita-Laza I, Lee S, Makarov V, Buxbaum Joseph D, Lin X. Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants. Am J Hum Genet. 2013; 92(6):841–53. https://doi.org/10.1016/j.ajhg.2013.04.015 PMID: 23684009

109. VanderWeele TJ. Introduction to Statistical Mediation Analysis by MACKINNON, D. P. Biometrics. 2009; 65(3):998–1000. https://doi.org/10.1111/j.1541-0420.2009.01315_12.x

110. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika. 1994; 81(3):515–26. https://doi.org/10.1093/biomet/81.3.515

111. Qu L, Guennel T, Marshall Scott L. Linear score tests for variance components in linear mixed models and applications to genetic association studies. Biometrics. 2013; 69(4):883–92. https://doi.org/10.1111/biom.12095 PMID: 24328714

112. Lin X, Cai T, Wu MC, Zhou Q, Liu G, Christiani DC, et al. Kernel machine SNP-set analysis for censored survival outcomes in genome-wide association studies. Genet Epidemiol. 2011; 35(7):620–31. Epub 2011/08/04. https://doi.org/10.1002/gepi.20610 PMID: 21818772.

113. Duan Z, Qiao Y, Lu J, Lu H, Zhang W, Yan F, et al. HUPAN: a pan-genome analysis pipeline for human genomes. Genome Biol. 2019; 20(1):149. https://doi.org/10.1186/s13059-019-1751-y PMID: 31366358

114. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004; 20(1):93–9. https://doi.org/10.1093/bioinformatics/btg382 PMID: 14693814

115. Goeman JJ, Van De Geer SA, Van Houwelingen HC. Testing against a high dimensional alternative. J R Stat Soc Ser B. 2006; 68(3):477–93. https://doi.org/10.1111/j.1467-9868.2006.00551.x

116. Kwee LC, Liu D, Lin X, Ghosh D, Epstein MP. A Powerful and Flexible Multilocus Association Test for Quantitative Traits. Am J Hum Genet. 2008; 82(2):386–97. https://doi.org/10.1016/j.ajhg.2007.10.010 PMID: 18252219

117. Crainiceanu CM, Ruppert D. Likelihood ratio tests in linear mixed models with one variance component. J R Stat Soc Ser B. 2004; 66(1):165–85. https://doi.org/10.1111/j.1467-9868.2004.00438.x

118. Zeng P, Zhao Y, Zhang L, Huang S, Chen F. Rare Variants Detection with Kernel Machine Learning Based on Likelihood Ratio Test. PLoS ONE. 2014; 9(3):e93355. https://doi.org/10.1371/journal.pone.0093355 PMID: 24675868

119. Davies RB. Algorithm AS 155: The Distribution of a Linear Combination of chi-2 Random Variables. Journal of the Royal Statistical Society: Series C (Applied Statistics). 1980; 29(3):323–33. https://doi.org/10.2307/2346911

120. Therneau TM. coxme: Mixed Effects Cox Models. R package version 2.2–14. https://CRAN.R-project.org/package=coxme. 2019.

121. Smith AA, Huang Y-T, Eliot M, Houseman EA, Marsit CJ, Wiencke JK, et al. A novel approach to the discovery of survival biomarkers in glioblastoma using a joint analysis of DNA methylation and gene expression. Epigenetics. 2014; 9(6):873–83. Epub 2014/03/26. https://doi.org/10.4161/epi.28571 PMID: 24670968.

122. Teumer A, Chaker L, Groeneweg S, Li Y, Di Munno C, Barbieri C, et al. Genome-wide analyses identify a role for SLC17A4 and AADAT in thyroid hormone regulation. Nat Commun. 2018; 9(1):4455. https://doi.org/10.1038/s41467-018-06356-1 PMID: 30367059

123. Sobel ME. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. Sociological Methodology. 1982; 13:290–312.

124. Sobel ME. Some New Results on Indirect Effects and Their Standard Errors in Covariance Structure. Sociological Methodology. 1986:159–86.

125. Mackinnon DP, Warsi G, Dwyer JH. A Simulation Study of Mediated Effect Measures. Multivariate Behavioral Research. 1995; 30(1):41–. https://doi.org/10.1207/s15327906mbr3001_3 PMID: 20157641.

126. Langaas M, Lindqvist BH, Ferkingstad E. Estimating the proportion of true null hypotheses, with application to DNA microarray data. J R Stat Soc Ser B. 2005; 67(4):555–72. https://doi.org/10.1111/j.1467-9868.2005.00515.x

127. Jin J, Cai TT. Estimating the Null and the Proportion of Nonnull Effects in Large-Scale Multiple Comparisons. J Am Stat Assoc. 2007; 102(478):495–506. https://doi.org/10.1198/016214507000000167

128. Jiang H, Doerge RW. Estimating the proportion of true null hypotheses for multiple comparisons. Cancer Inform. 2008; 6:25–32. PMID: 19259400

129. Huang Y-T, Pan W-C. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. Biometrics. 2016; 72(2):402–13. https://doi.org/10.1111/biom.12421 PMID: 26414245

130. Efron B. Size, power and false discovery rates. Ann Stat. 2007; 35(4):1351–77.

131. Storey J. The positive false discovery rate: a Bayesian interpretation and the q-value. Ann Stat. 2003; 31:2013–35. https://doi.org/10.1214/aos/1074290335

132. Efron B, Zhang NR. False discovery rates and copy number variation. Biometrika. 2011; 98(2):251–71. https://doi.org/10.1093/biomet/asr018

133. Storey J, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci USA. 2003; 100:9440–5. https://doi.org/10.1073/pnas.1530509100 PMID: 12883005

134. Storey J. Journal of the Royal Statistical Society: Series B (Statistical Methodology). JR Stat Soc B. 2002; 64:479–98. https://doi.org/10.1111/1467-9868.00346

135. Efron B, Tibshirani R. Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002; 23(1):70–86. https://doi.org/10.1002/gepi.1124 PMID: 12112249

136. Bender R, Augustin T, Blettner M. Generating survival times to simulate Cox proportional hazards models. Statistics in Medicine. 2005; 24(11):1713–23. https://doi.org/10.1002/sim.2059 PMID: 15724232

137. Chen Y, Lemire M, Choufani S, Butcher D, Grafodatskaya D, Zanke B, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. Epigenetics. 2013; 8:203–9. https://doi.org/10.4161/epi.23470 PMID: 23314698

138. Du P, Zhang X, Huang C-C, Jafari N, Kibbe WA, Hou L, et al. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics. 2010; 11(1):587. https://doi.org/10.1186/1471-2105-11-587 PMID: 21118553

139. Zhang J, Lu H, Zhang S, Wang T, Zhao H, Guan F, et al. How can gene expression information improve prognostic prediction in TCGA cancers: an empirical comparison study on regularization and mixed-effect survival models. Frontiers in Genetics. 2021; https://doi.org/10.3389/fgene.2020.00920 PMID: 34149809

140. Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S. Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. Brief Bioinform. 2014:bbu003. https://doi.org/10.1093/bib/bbu003 PMID: 24632304

141. Shen S, Wang G, Shi Q, Zhang R, Zhao Y, Wei Y, et al. Seven-CpG-based prognostic signature coupled with gene expression predicts survival of oral squamous cell carcinoma. Clinical Epigenetics. 2017; 9(1):88. https://doi.org/10.1186/s13148-017-0392-9 PMID: 28852427

142. Wei Y, Liang J, Zhang R, Guo Y, Shen S, Su L, et al. Epigenetic modifications in lysine demethylases associate with survival of early-stage NSCLC. Clinical Epigenetics. 2018; 10:41. https://doi.org/10.1186/s13148-018-0474-3 PMID: 29619118.

143. Shen S, Zhang R, Guo Y, Loehrer E, Wei Y, Zhu Y, et al. A multi-omic study reveals BTG2 as a reliable prognostic marker for early-stage non-small cell lung cancer. Mol Oncol. 2018; 12(6):913–24. https://doi.org/10.1002/1878-0261.12204 PMID: 29656435.

144. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018; 173 (2):400–16.e11. https://doi.org/10.1016/j.cell.2018.02.052 PMID: 29625055