Leveraging methylation alterations to discover potential causal genes associated with the survival risk of cervical cancer in TCGA through a two-stage inference approach

Jinhui Zhang\(^1\)*, Ting Wang\(^1\)*, Xinghao Yu\(^1\), Shuiping Huang\(^1,2\), Huashuo Zhao\(^1\), Ping Zeng\(^1,2\)

1. Department of Epidemiology and Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu, 221004, China

2. Center for Medical Statistics and Data Analysis, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu, 221004, China

* The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

Correspondence Authors: Huashuo Zhao and Ping Zeng

Address: Department of Epidemiology and Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu, 221004, China.

Tel: +86 13305218786; E-mail: hszhao@xzhmu.edu.cn

Tel: +86 15996970535; FAX: +86 15996970535; E-mail: zpstat@xzhmu.edu.cn
ABSTRACT

Background: Multiple genes were previously identified to be associated with cervical cancer; however, the genetic architecture of cervical cancer remains unknown and many causal genes have yet been discovered.

Methods: To explore causal genes related to cervical cancer, a two-stage causal inference approach was proposed within the framework of Mendelian randomization, where the gene expression was treated as exposure, with methylations located within that gene serving as instrumental variables. Five prediction models were first utilized to characterize the relationship between the expression and methylations for each gene; then the methylation-regulated gene expression (MReX) was obtained and the association was evaluated via Cox mixed-effects model based on MReX. We further implemented the harmonic mean p-value (HMP) combination to take advantage of respective strengths of these prediction models while accounting for dependency among the p-values.

Results: A total of 14 causal genes were discovered to be associated with the survival risk of cervical cancer in TCGA when the five prediction models were separately employed. The total number of causal genes was brought to 23 when conducting HMP. Some of the newly discovered genes may be novel (e.g. YJEFN3, SPATA5L1, IMMP1L, C5orf55, PPIP5K2, ZNF330, CRYZL1, PPM1A, ESCO2, ZNF605, ZNF225, ZNF266, FICD and OSTC). Functional analyses showed these genes were enriched in tumor-associated pathways. Additionally, four genes (i.e. COL6A1, SYDE1, ESCO2 and GIPC1) were differentially expressed.

Conclusion: Overall, our study discovered promising candidate genes that are causally associated with the survival risk of cervical cancer and thus provided new insights into the genetic etiology of cervical cancer.

Keywords: cervical cancer, The Cancer Genome Atlas, causal gene, two-stage inference, instrumental variable, DNA methylation, gene expression, prediction model, harmonic mean p-value combination method, cox linear mixed-effects model
Background

Cervical cancer is a sexually transmitted disease, mostly caused by infection with human papillomavirus (HPV) \[1\]. In terms of cancer statistics in 2018, cervical cancer is the fourth most common malignancy and the fourth leading cause of cancer death among women worldwide, with an estimate of 570,000 cases and 311,000 deaths globally \[2\]. Moreover, cervical cancer is the second primary cause of cancer death in women aged 20 to 39 years \[3\]. Although great advanced have been achieved for cervical cancer, reliable diagnostic biomarkers and methods for early identification and screening remain lacking \[4, 5\]. In addition, despite the unitization of HPV vaccines for prevention and chemoradiotherapy as well as radical surgery offering satisfactory survival rate for early-stage cervical cancer patients, effective treatments for advanced patients are rarely available, especially in developing countries and regions \[5\].

Therefore, it is an urgent demand in clinical practice that valuable biomarkers should be well discerned and validated to signal the early stage or provide profile of cervical cancer progression \[6\]. As an effort to understand the genetic foundation of susceptibility to cervical cancer, in the past decade multiple genome-wide association studies (GWASs) were undertaken and discovered a group of cervical-cancer associated genetic variants; see Table S1 for details and see also \[5\] where a large number of associated germline genetic variants and genes were described for cervical cancer. These findings imply that the development of cervical cancer relies to a significant extend on inherited genetic components and genetic predisposing factors may affect the probability and persistence of, or sensitivity to HPV infection and the rate of tumor development as well as progression \[5\]. However, like many complex human traits and diseases, the genome-wide SNP-based heritability of cervical cancer estimated in GWAS is smaller than expected. For example, the heritability is 11.7% (se = 9.8%) in a Japanese population \[7\] and 24.0% (se =2.9%) in a Swedish population \[8\], both of which are lower than that observed in family studies \[9\]. The remaining missing heritability suggests that a large number of causal genes and genetic variants have yet been discovered and that continuous efforts to identify causative genes for cervical cancer are worthwhile \[5\].

As well demonstrated in many studies \[10-15\], mRNA-gene expression measured at
the transcript level influences the progression of complex diseases more directly than other omic measurements. However, the establishment of causal relationship between altered gene expressions and cervical cancer is not straightforward in observation studies due to unknown confounders and possible reverse causation. The latter is of particular concern because we cannot determine whether the regulated gene expressions are the causal factors or the consequence of the development or progression of cervical cancer due to considerably complicated biological network and interaction. Due to this reason, previous studies often aimed to examine association rather than causality between gene expression and cervical cancer.

In statistical genetics a powerful statistical tool to determine causal relationship and estimate causal effect of the exposure on the outcome in observational studies is Mendelian randomization (MR), which is built based on commonly used instrumental variable approaches developed in the field of causal inference [16-18]. Under some certain assumptions, the results of MR analysis are less susceptible to reverse causation and confounding factors [19]. One of key points in MR is to select valid instrumental variables for the exposure (i.e. expression level in our context). Biologically, methylation GpG sites of a specific gene within the unique function of transcript start site can down-regulate its expression level, and the deregulated expression can further influence the survival of cancer patients [10, 20-22], indicating that methylation alterations play a central role in cancers by regulating expression profile. This motivates us to propose a one-sample two-stage causal inference approach with methylations as instrumental variables of expression to detect causal genes for cervical cancer. This type of two-stage instrumental variable inference is widely employed in many research fields such as sociology, economics [23] and genetic medicine [24]. In addition, the utilization of methylations serving as instruments for causal inference is also commonly seen in recent genomic integrative analyses [25-29].

Methodologically, our proposed approach follows the similar principle of prediXcan [30] that was developed recently to identify causal genes for complex diseases with genetic variants serving as instrumental variables in the framework of transcriptome-wide association studies (TWAS) [30-35]. Specifically, in our context we implement a relatively independent two-stage inference procedure (Figure 1): in
the first stage the weights of methylation alterations for each gene are estimated via genetic prediction models; in the second stage the methylation-regulated gene expression (MReX) is imputed with the corresponding predictive model and then the causal association between the gene and the survival risk of cervical cancer is examined based on MReX. More importantly, the two-stage based causal inference can be viewed as a special case of one-sample MR analysis from a statistical perspective [36]. Therefore, under the same conditions of MR our two-stage inference has the ability discovering putatively causal genes for cervical cancer. Furthermore, we consider five commonly used prediction models in the first stage of our two-stage inference procedure and exploit the harmonic mean $p$-values (HMP) method [37] — a novel combination strategy that is robust against high correlation [38, 39] — to take advantage of respective strengths of these models while accounting for dependency among the $p$-values of various models.

We finally apply the proposed approach to the cervical cancer dataset in The Cancer Genome Atlas (TCGA) program [40]. A total of 14 causal genes were discovered to be associated with the survival risk of cervical cancer when the five prediction models were separately implemented. The total number of causal genes was brought to 23 when conducting the combination test with HMP. Some of the newly discovered genes were reported in previous literature and differentially expressed between tumor and normal tissues. In addition, functional analyses showed that these genes were enriched in tumor-associated pathways.
Methods

TCGA cervical cancer data sets and quality control

Our analysis mainly relied on publicly available datasets of cervical cancer in TCGA [40]. From https://xenabrowser.net/hub/, we obtained clinical features on 317 samples, 20,530 RSEM normalized expressions on 308 samples and 485,577 DNA methylation alterations on 312 samples. To avoid racial heterogeneity, we kept 190 white cervical cancer patients with primary solid tumor after filtering out samples with too many missing values. The description of important characteristics of this cervical cancer dataset after filtering is given in Table 1. In our following analysis, we only considered protein-coding genes and determined whether DNA methylation alterations belonged to a given gene according to the TCGA annotation mapping file (i.e. illuminaMethyl450_hg19_GPL16304_TCGAlegacy). Then, each gene expression was quantile-transformed so that it followed a standard normal distribution and each methylation was standardized. Missing values were simply imputed with median. The flowchart for our study is shown in Figure 2.

Linear models predicting gene expression with DNA methylation alterations

Let $G$ be an $n$-vector of gene expression levels for the $i^{th}$ gene measured on $n$ individuals, $M$ be an $n \times p$ matrix for a group of DNA methylations that are located within this gene; note that $p$ varies gene by gene. We apply the following linear model to link $G$ and $M$

$$ G = Mw + \varepsilon, \varepsilon \sim N(0, \sigma_\varepsilon^2 I_n) $$

(1)

where $w$ is a $p$-vector for effect sizes of DNA methylations, $\varepsilon$ is an $n$-vector of residual errors following an independent and identical normal distribution with mean zero and variance $\sigma_\varepsilon^2$ and $I_n$ denotes the $n$-dimensional identical matrix. Because of the possible high-dimensional issue where the number of DNA methylations $p$ is larger than the sample size $n$ (see below), the commonly used least squares method is no longer applicable for estimating $w$. We instead employ several novel models which are specially designed for high-dimensional models, and particularly consider five regressions including linear mixed-effects model (LMM) [41-43], Bayesian sparse
linear mixed-effects model (BSLMM) [44], Latent Dirichlet Process Regression (DPR) [35] as well as Lasso [45] and elastic net (ENET) [46]. Among these methods, LMM, BSLMM and DPR explicitly incorporate all DNA methylations into the model by assuming diverse prior distributions for the effect sizes; while Lasso and ENET only include some most important DNA methylations with the way of regularization based on variable selection. The details of these models are described in [36]. We implement LMM and BSLMM with the GEMMA software (version 0.94), DPR with the DPR software [35], Lasso and ENET with the R glmnet package (version 2.0-18) [47]. Using these models, we can obtain the estimate of effect sizes of DNA methylations (denoted by $\hat{w}$) as well as the MReX level $\hat{G} = M\hat{w}$ for each gene.

Cox mixed-effects regression discovering methylation-regulated genes

We now investigate the association between the gene and the survival risk of cervical cancer using the Cox model [48]. Besides the direct gene effect based on MReX $\hat{G}$, we also incorporate the impact of DNA methylation alterations into the survival model to explain possible horizontal pleiotropy [49-54]

$$\frac{h(t \mid X, \hat{G}, M)}{h_0(t)} = \exp(Xa + \hat{G} \times b + Mc), \quad c \sim N(0, \sigma_c^2)$$

(2)

where $t$ is the observed survival time, $h_0(t)$ is an arbitrary baseline hazard function, $a = (a_1, a_2, \ldots, a_m)$ is an $m$-vector of effect sizes for available covariates $X$, $b$ is the effect size for the given gene and is of our primary interest, and $c = (c_1, c_2, \ldots, c_p)$ is a $p$-vector of effect sizes for DNA methylations. Because of the same reason of high-dimensional problem mentioned before, we assume $c$’s are random effects following a normal distribution with mean zero and variance $\sigma_c^2$, leading to the Cox linear mixed-effects regression model (denoted by coxlmm) [55]. When $c = 0$, or equivalently $\sigma_c^2 = 0$, coxlmm shown in (2) reduces into the general Cox model where only the influence of the methylation-driven gene exists. We fit coxlmm with the R coxme (version 2.2-10) package [56] via the Laplace approximation algorithm based on the second order Taylor series expansion [55]. The significance of MReX is examined through the Wald test ($H_0: b = 0$): $Z = \hat{b} / \sqrt{\text{var}(\hat{b})}$, where $\hat{b}$ is the
estimate of the effect size $b$, with $\text{var}(\hat{b})$ the variance of the estimate $\hat{b}$. The $p$-value of the Z statistic can be easily obtained because it asymptotically follows a standard normal distribution.

*Harmonic mean p-value method combining dependent p-values*

Because multiple prediction models are applied, for each gene we thus yield a set of $p$-value $p_k$ ($k = 1, 2, \ldots, K$; with $K$ the number of the prediction models) according to (2). Unfortunately, the simple and commonly used Fisher’s method for aggregating mutually independent multiple tests cannot be exploited due to highly positive correlation among individual tests since they are implemented for the same data set with the similar logic [57, 58]. Instead, we apply HMP to generate a pooled $p$ value across tests with various prediction models

$$
\begin{align*}
& P_{\text{HMP}} = \int_{1/T_{\text{HMP}}}^{\infty} f_x(x | \log T + 0.874, \frac{\pi}{2})dx \\
& T_{\text{HMP}} = (\sum_{k=1}^{K} \omega_k) / (\sum_{k=1}^{K} \omega_k / p_k)
\end{align*}
$$

(3)

where $\omega_k$ represents the non-negative weight for each $p_k$ with $\sum_{k=1}^{K} \omega_k = 1$ and $K = 5$ in our study and assume that $\omega_k$ is independent of $p_k$; $f_x$ denotes the Landau distribution probability density function. It has been theoretically demonstrated that the complicated dependency among $p$-values has little influence on the final pooled $p$-value in HMP [37], especially on exceedingly small $p$-values which are of particular interest in practice. We implement HMP with equal weights using the harmonicmeanp package (version 3.0) in R [59].

*Functional analysis and differential expression analysis for newly identified associated genes*

Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analyses were conducted using the R clusterProfiler package (version 3.16.0) [60]. In addition, to further evaluate the expression profiles of these newly discovered genes, we performed differential expression analysis with 190 cervical tumors and
three normal tissues that were also available from TCGA. After normalization with
the trimmed mean of M values (TMM) method, differential expressed genes (DEGs)
were screened via the exact test based on quantile-adjusted conditional maximum
likelihood estimation [61, 62] implemented in the edgeR package (version 3.30.3) [63,
64]. Following previous work [65, 66], DEGs were defined if FDR < 0.05 and |log₂
FC (CT/CK)| ≥ 1.0.
Results

Cervical cancer datasets in TCGA and methylation-regulated genes

After quality control we reserved 485,577 DNA methylation GpG sites and 3 clinical covariates (i.e. age of onset, clinical stage, and tumor status) up to 190 cervical cancer patients of European ancestry. To avoid numerical instability, we focused on protein-coding genes which had at least ten methylations. We also first performed the LMM analysis [44, 67, 68] for each protein-coding gene based on its methylations and selected genes with the phenotypic variance explained by methylations larger than 1% (corresponding to a correlation coefficient of 10%). The remaining 12,623 genes are referred to as methylation-regulated genes and included in our subsequent analyses (Figure 2). The number of methylation GpG sites across genes ranges from 10 to 1,062, with the majority of analyzed genes (92.0% = 11,607/12,623) having methylations less than 50.

Identification of causal genes with Cox linear mixed-effects regression

We employed the coxlmm [55] with various prediction models to examine the relationship between MReX and the survival risk of cervical cancer patients while adjusting for the direct effect of methylations and the confounding effect of clinical covariates. First, we observe that these prediction models display various performance across genes (Figure 3A). Specifically, some prediction models have higher prediction accuracy for some genes but behave less satisfactorily for others. For example, in terms of $R^2$, BSLMM behaves well for 38.3% genes ($= 4,834/12,623$), while Lasso, ENET, LMM and DPR have higher $R^2$ for 26.82% ($= 3,386/12,623$), 14.79% ($= 1,867/12,623$), 10.92% ($= 1,378/12,623$) and 9.17% ($= 1,158/12,623$) genes, respectively. As expected, the resulting $p$-values of these prediction methods in coxlmm are highly correlated (Figure 3B). For example, the Pearson's correlation of the $p$-values (in a scale of -log10) ranges from 0.63 between DPR-coxlmm and Lasso-coxlmm to 0.96 between LMM-coxlmm and BSLMM-coxlmm.

Based on the results of coxlmm, a total of 14 unique associated genes (false discovery rate [FDR] < 0.05) are identified (Table 2). Specifically, we detect three associated genes with DPR-coxlmm, ten associated genes via Lasso-coxlmm and eight
associated genes through ENET-coxlmm, but do not discover any associated genes using LMM-coxlmm or BSLMM-coxlmm (Figure 4). Among these, six genes (i.e. YJEFN3, SPATA5L1, C5orf55, PPIP5K2, ESCO2 and ZNF225) are simultaneously found by Lasso-coxlmm and ENET-coxlmm, while only one gene (i.e. VPS4B) is simultaneously discovered by DPR-coxlmm and ENET-coxlmm (Figure 3C).

Among these associated genes we find PCM1 (FDR\textsubscript{ENET} = 0.032), classified to the cell cycle control network, was previously discovered to be associated with the early stage of cervical cancer [69]. SPR (FDR\textsubscript{Lasso} = 0.003) is located within 1 Mb genetic region of previous GWAS-identified gene ALMS1 [7]. In addition, VPS4B (FDR\textsubscript{DPR} = 0.024 and FDR\textsubscript{ENET} = 0.031) is a subtype of VPS4 which is the component of the ESCRT machinery and plays an essential role in HPV infectious entry and capsid disassembly [70]. The remaining ten genes (i.e. YJEFN3, SPATA5L1, C5orf55, PPM1A, IMMP1L, ZNF330, PPIP5K2, ESCO2, FICD and ZNF225) are not directly reported to be related to the survival risk of cervical cancer in previous literature. However, for these genes we find suggestive indirect evidence that may support their association with cervical cancer. Specifically, for example, YJEFN3 is a member of the human YJEFN domain containing protein family strongly expressing in Leydig cell tumors and in the fibromas and participates in cholesterol processing and steroid hormone metabolism [71]. SPATA5L1 might play a key role in inhibits ATP Hydrolysis and four-way junction helicase activity and further influences DNA replication and pathogenesis [72, 73]. Smac/DIABLO was expressed de novo in certain subset of cervical tumors [74], while mature Smac/DIABLO was produced on the mitochondrial inner membrane via IMMP1L [75]. PPIP5 kinases (e.g. PPIP5K2) mediate PP-IPs binding, activate casein kinase 2 (CK2) and promote the phosphorylation of the TTT complex, which stimulates DNA-PK/ATM to activate p53 on the cancer cellular [76, 77]. There exists evidence that miR-135b leads to cervical cancer cell transformation [78] and down-regulated miR-135b expression could inhibit the proliferation and invasion of tumor cells by up-regulating PPM1A [79].

\textit{HMP to combine different p-values}

As mentioned before, because the p-values obtained from coxlmm with diverse prediction models are highly dependent (Figure 3B), we effectively apply HMP to
combine the five $p$ values and generate an overall significance for each gene (Figure 2 and Table 2). Nine associated genes are additionally discovered (Figure 3C), including CRYZL1, ZNF605, ZNF266, SNAI1, OSTC, FAM73A, COL6A1, GIPC1 and DCTPP1. Among these genes, five (i.e. SNAI1, COL6A1, GIPC1, DCTPP1 and FAM73A) were identified in prior work and SYDE1 locates within 1Mb generic region of GIPC1 (Table S1).

Specifically, it is shown that SNAI1, along with ZEB1, regulated the epithelial-mesenchymal transition and was then involved in the metastasis of cervical cancer [80]. The up-regulated COL6A1 expression in the tissues of cervical cancer was related to poor clinical prognosis and treated as an important biomarker of cervical cancer progression [81]. The down-regulation of GIPC1 in cervical cancer with HPV-18 infection can lead to the resistance to cytostatic transforming growth factor $\beta$ signaling through TGF$\beta$R3 destabilization [82]. In addition, DCTPP1 was found to be differentially expressed in normal and cancerous tissues and it was significantly accumulated in the nucleus of cervical carcinoma, implying the important role of DCTPP1 under malignant pathology [83]. Family with sequence similarity 73, member A (FAM73A) is the down-regulated gene of DNA from exfoliated cervical cells in terms of the HPV-16 variant analysis [84, 85]. CRYZL1 contains an reduced nicotinamide adenine dinucleotide (phosphate) ($\text{NAD(P)H}$) binding site which is involved in cellular metabolism, while cervical lesions are associated with cellular metabolic abnormalities [86]. It is previously found that the members of the ZNF family interact with nucleic acids, proteins and small molecules and are involved in a variety of crucial molecular processes in cervical tumor cells at replication, transcriptional and translational levels. Thus, ZNF605 and ZNF266 may be potentially targetable [87-89]. OSTC can regulate gamma-secretase [90] while this secretase affects the ability of HPV pseudo-viruses infection both in human HaCat cells and mouse cells [91].

In summary, compared with the tests via individual prediction methods, it is demonstrated that HMP greatly improves statistical power by combining dependent tests and thus identifies more prognosis-associated genes for cervical cancer. Totally, 23 genes are discovered to be related to the survival risk of cervical cancer, among which 14 genes are likely newly novel genes (i.e. YJEFN3, SPATA5L1, IMMP1L, ...
C5orf55, PPIP5K2, ZNF330, CRYZL1, PPM1A, ESCO2, ZNF605, ZNF225, ZNF266, 
FICD and OSTC).

Identification of DEGs, GO and KEGG pathway annotation

In terms of the differential expression analysis, four DEGs are detected among the 23 
new genes identified above (Figure 5A). In particular, COL6A1 and SYDE1 are 
up-regulated genes, while ESCO2 and GIPC1 are down-regulated genes (Figure 5B). 
To explore the potential functions of these genes that may be associated with the 
tumorigenesis and development of cervical cancer, we performed functional 
enrichment analysis with GO and KEGG using the R package clusterProfiler (version 
3.16.0) [60]. The top 5 significantly enriched GO terms of three parts and two KEGG 
pathways identified are shown in Figure 5C.

The GO biological process (BP) terms are remarkably enriched in polyol metabolic 
process, regulation of biosynthetic process and signaling pathway, chondrocyte 
differentiation. For the GO cellular component (CC) terms, the target genes are 
concentrated in the midbody, pericentriolar material, and so on. The molecular 
function (MF) category was focused on NADP binding, platelet-derived growth factor 
binding (Table S2). The KEGG enrichment analysis indicates that these genes are 
remarkably enriched in tumor-associated pathways, including protein export ($P = 
0.028$) and folate biosynthesis ($P = 0.032$) (Figure 5C). The combined action of folate 
biosynthesis and graft-versus-host disease were demonstrated to be significantly 
associated with cervical cancer in suit: HLA-DPB1 [92]. The up-regulated 
differentially expressed genes are mostly associated with cartilage morphogenesis 
(ontology: BP), collagen trimer (ontology: CC) and extracellular matrix structural 
constituent conferring tensile strength (ontology: MF). The down-regulated 
differentially expressed genes are mostly associated with organic hydroxy compound 
biosynthetic process (ontology: BP) and organic hydroxy compound metabolic 
process (ontology: BP), dendritic shaft (ontology: CC) (Table S2). The functional 
enrichment results suggest that these newly discovered causal genes may participate 
in oncogenicity and tumor progression in cervical cancer through regulating relevant 
biological processes and critical pathways.
Discussion

Given the severe health threat among women and little knowledge of genetic basis for cervical cancer, persistent work should be done to discover genes that are causally related to cervical cancer [5]. The present study is one of such efforts with the aim to detect newly causal genes for cervical cancer through integrative genomic methods. The two-stage inference analysis pipeline applied in this work can be considered as a gene-centered integration approach by aggregating omics datasets and clinical information. With the growing high-throughput omics datasets in cancer research over recent years [40], it is well recognized that the utilization of only one single level of genomic measurements is insufficient to completely untangle the etiology of cancer prognosis [13, 40]. Based on the omics datasets of TCGA measured from multiple platforms, we treated the gene expression as the exposure and the survival time as the outcome to explore causal genes of cervical cancer within the framework of two-stage MR study to avoid the reverse causation.

For each gene, under the biologically plausible assumption that the methylation CpG sites in the gene can regulate gene expression [10, 20-22], we employed these methylations as instrumental variables. Note that, DNA methylations can also influence the survival through alternative mechanisms different from the pathway by the regulation of gene expression; such a phenomenon is referred to as horizontal pleiotropy in MR [49-54] and violates the necessary assumptions of instrumental variable causal inference. To guard against such violation, we also included the direct effect of methylations when inferring the causal relationship between the gene and the survival of cervical cancer.

One critical step in our two-stage inference is to evaluate the effect relationship between a group of DNA methylation CpG sites and the expression level for each gene. The power of the subsequent association performed in cox1mm would greatly depend on how well the prediction model utilized can capture the underlying genetic architecture of the transcriptome [30-35], which can differ in the numbers, effect sizes and effect directions of causal methylation alterations in diverse genes. Therefore, a powerful two-stage inference approach should in the first stage choose a prediction model whose prior effect distribution closely matches the true effect distribution so
that it can approximate well the genetic architecture of the gene [35, 36, 44]. For example, if DNA methylation alterations have effect sizes following a normal distribution, then LMM-cox1mm would be more powerful; on the other hand, if only a very small fraction of DNA methylation alterations may be predictive for the gene expression, then the test with sparse prediction models (e.g. Lasso-cox1mm and ENET-cox1mm) would be superior. Due to unknown true association patterns, there is no uniformly most powerful test. As a result, the two-stage association test may perform well for one gene, but not necessarily for another.

To leverage the advantage of distinct prediction models to improve power, instead of selecting an optimal prediction model, in the present study we considered a wide range of prediction models in our two-stage inference procedure. It can be imaged that the resulting $p$ values would be highly correlated because they are generated with the same data set following the similar logic (Figure 2). The correlation structure of these $p$-values also depends on the true architecture of gene expression, which however is rarely known in advance and is likely to vary from one gene to another across the genome. Therefore, it is desirable to construct an omnibus test that integrates the advantage of multiple prediction approaches and is robust against distinct transcriptomic architectures. To achieve this aim, we exploited HMP [37] combining these correlated $p$-values and integrating individual strengths of various tests. As illustrated in our empirical application, HMP achieves relatively higher power since it aggregates genetic association information across different tests.

Compared to previous similar methods, the proposed two-stage inference approach differs in three aspects. First, unlike prediXcan [30] we constructed the two-stage inference procedure in one sample, leading to the so-called one-sample two-stage regression [24]. Second, multiple competing prediction models rather a single model were utilized and combined with HMP which was $p$-value calibrated [37]. Thus, our strategy often has higher power compared to the test with single prediction model. Third, due to widespread pleiotropic effects in omics [49-54], we also considered the direct influence of methylations. Therefore, our results would be robust against the bias of pleiotropy of instrumental variables that are commonly encountered in MR.

However, the present study is not without limitation. First, the methylation-regulated
genes were analyzed only in TCGA; no external relevant expression profiles were applied for further validation. Second, we only employed methylations as instrumental variables, other omic measurements that regulate gene expression (e.g. genetic variants [43, 93, 94]) can be also simultaneously incorporated to further improve power. Third, we only utilized local methylation GpG sites of a gene as candidate instruments. It is not known whether the power can be further enhanced if the global methylation GpG sites are exploited. Fourth, the present study assumed a linear relationship for each methylations-expression pair. While a linear relationship can be methodologically interpreted as a first-order approximation to non-linear relationship [44], modeling a linear relationship may be suboptimal and suffer from power loss if the true relationship is non-linear. Fifth, due to the complicated standard error structures for those prediction models, in terms of the assumption of no measurement error (NOME) [51], we did not incorporate the uncertainty in the estimated effect sizes of methylations into our two-stage approach, although such uncertainty may be important in integrative genomic causal inference [95, 96]. Actually, we note that many previous two-stage MR studies or TWAS approaches followed this NOME principle [30, 31, 35, 51].

**Conclusion**

In summary, using the proposed two-stage causal inference approach within the framework of MR analysis, we discovered a total of 14 causal genes which were associated with the survival risk of cervical cancer patients when separately applying five commonly used prediction models. The number of causal genes was brought to 23 when employing the combination method of HMP. Some may be newly novel genes (i.e. YJEFN3, SPATA5L1, IMMP1L, C5orf55, PPIP5K2, ZNF330, CRYZL1, PPM1A, ESCO2, ZNF605, ZNF225, ZNF266, FICD and OSTC), and some of those newly discovered genes were reported in previous literature and differentially expressed between tumor and normal tissues. In addition, functional analyses showed that those genes were enriched in tumor-associated pathways. Our findings provide new insights into the genetic etiology of cervical cancer and suggest possibly potential therapeutic targets for cervical cancer in the future.
Supplementary Material

Supplementary material accompanies this paper at BMC Cancer.

Additional file 1: Table S1. Genes reported to be associated with risk of cervical cancer through GWAS. Abbreviation: PMID, PubMed ID; OR, Odd Ratio; CI, Confidence Interval.

Additional file 1: Table S2. Functional term enrichment analysis by casual genes. Abbreviation: GOID, gene ontology id; BP, biological function; CC, cellular component; MF, molecular function
ABBREVIATIONS

BP: biological process; BSLMM: Bayesian sparse linear mixed model; CoxLmm: Cox linear mixed-effects model; CC: cellular component; DEG: differential expressed genes; DPR: Latent Dirichlet Process Regression; EBI: the European Bioinformatics Institute; ENET: elastic net; FDR: false discovery rate; GO: Gene Ontology; GWAS: genome-wide association study; HMP: Harmonic mean p-value method; HPV: human papillomavirus; KGEE: Kyoto Encyclopedia of Genes and Genomes; LMM: Linear mixed model; MF: molecular function; MR: Mendelian randomization; TCGA: The Cancer Genome Atlas; TWAS: transcriptome-wide association study

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request or in https://xenabrowser.net/hub/.

Competing interests

The authors declare that they have no competing interests

Funding

This study was partly supported by Youth Foundation of Humanity and Social Science funded by Ministry of Education of China (18YJC910002), Natural Science Foundation of Jiangsu Province (BK20181472), China Postdoctoral Science Foundation (2018M630607 and 2019T120465), Postdoctoral Science Foundation of Xuzhou Medical University, QingLan Research Project of Jiangsu Province for Outstanding Young Teachers, Six-Talent Peaks Project in Jiangsu Province of China (WSN-087), Social Development Project of Xuzhou City (KC19017), National Natural Science Foundation of China (81402765), Statistical Science Research Project from National Bureau of Statistics of China (2014LY112) and Training Project for Youth Science and Technology Innovation Team at Xuzhou Medical University.
Authors’ contributions

P.Z. and S.Z. conceived the idea for the study; P.Z., T.W., X.Y. and J.Z. obtained the data and performed the data analyses; P.Z., S.H., T.W. and J.Z. interpreted the results of the data analyses. P.Z., T.W. and J.Z. wrote the manuscript with the participation of all authors.

Acknowledgements

The TCGA data was publicly available from https://xenabrowser.net/. The data analyses in the present study were supported by the high-performance computing cluster at Xuzhou Medical University.

Author details

1 Department of Epidemiology and Biostatistics, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu, 221004, China  
2 Center for Medical Statistics and Data Analysis, School of Public Health, Xuzhou Medical University, Xuzhou, Jiangsu, 221004, China *The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint first authors.

References

1. Šarenac T, Mikov M: Cervical Cancer, Different Treatments and Importance of Bile Acids as Therapeutic Agents in This Disease. Frontiers in pharmacology 2019, 10:484.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A: Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 2018, 68(6):394-424.
3. Siegel RL, Miller KD, Jemal A: Cancer statistics, 2019. CA Cancer J Clin 2019, 69(1):7-34.
4. Nakamura K, Sawada K, Yoshimura A, Kinose Y, Nakatsuoka E, Kimura T: Clinical relevance of circulating cell-free microRNAs in ovarian cancer. Mol Cancer 2016, 15(1):48.
5. Chen D, Gyllensten U: Lessons and implications from association studies and post-GWAS analyses of cervical cancer. Trends Genet 2015, 31(1):41-54.
6. Nahand JS, Taghizadeh-Boroujeni S, Karimzadeh M, Borran S, Pourhanifeh MH, Moghoofei M, Bokharaei-Salim F, Karampoor S, Jafari A, Asemi Z et al: microRNAs: New prognostic, diagnostic, and therapeutic biomarkers in cervical cancer. *J Cell Physiol* 2019, **234**(10):17064-17099.

7. Masuda T, Low S-K, Akiyama M, Hirata M, Ueda Y, Matsuda K, Kimura T, Murakami Y, Kubo M, Kamatani Y et al: GWAS of five gynecologic diseases and cross-trait analysis in Japanese. *Eur J Hum Genet* 2020, **28**(1):95-107.

8. Chen D, Cui T, Ek WE, Liu H, Wang H, Gyllensten U: Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability. *Carcinogenesis* 2015, **36**(9):992-998.

9. Magnusson PKE, Lichtenstein P, Gyllensten UB: Heritability of cervical tumours. *Int J Cancer* 2000, **88**(5):698-701.

10. 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):149-159.

11. Zhao H, Ljungberg B, Grankvist K, Rasmuson T, Tibshirani R, Brooks JD: Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS medicine* 2005, **3**(1):e13.

12. Kim Y, Kang YS, Seok J: GAIT: gene expression Analysis for Interval Time. *Bioinformatics* 2018, **34**(13):2305-2307.

13. Zhao Q, Shi X, Xie Y, Huang J, Shia B, Ma S: Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Briefings in Bioinformatics* 2014, **16**(2):291-303.

14. Zhu B, Song N, Shen R, Arora A, Machiela MJ, Song L, Landi MT, Ghosh D, Chatterjee N, Baladandayuthapani V: Integrating clinical and multiple omics data for prognostic assessment across human cancers. *Scientific reports* 2017, **7**(1):16954.

15. 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 (in press)* 2020.

16. Angrist JD, Imbens GW, Rubin DB: Identification of Causal Effects Using Instrumental Variables. *J Am Stat Assoc* 1996, **91**(434):444-455.

17. Greenland S: An introduction to instrumental variables for epidemiologists. *Int J Epidemiol* 2000, **29**:722-729.

18. Sheehan NA, Didelez V, Burton PR, Tobin MD: Mendelian randomisation and causal inference in observational epidemiology. *PLoS Med* 2008, **5**(8):e177.

19. Davey Smith G, Ebrahim S: ‘Mendelian randomization’: can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol* 2003, **32**(1):1-22.
20. 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):2613-2625.

21. Fabiani E, Leone G, Giachelia M, D’Alo F, Greco M, Criscuolo M, Guidi F, Rutella S, Hohaus S, Voso MT: 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):2275-2284.

22. 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.

23. Angrist JD, Keueger AB: Does Compulsory School Attendance Affect Schooling and Earnings? The Quarterly Journal of Economics 1991, 106(4):979-1014.

24. Xue H, Pan W, for the Alzheimer's Disease Neuroimaging I: Some statistical consideration in transcriptome-wide association studies. Genet Epidemiol 2020, 44(3):221-232.

25. Qi T, Wu Y, Zeng J, Zhang F, Xue A, Jiang L, Zhu Z, Kemper K, Yengo L, Zheng Z et al: Identifying gene targets for brain-related traits using transcriptomic and methylocim data from blood. Nat Commun 2018, 9(1):2282-2282.

26. Hannon E, Gorrie-Stone TJ, Smart MC, Burrage J, Hughes A, Bao Y, Kumari M, Schalkwyk LC, Mill J: Leveraging DNA-Methylation Quantitative-Trait Loci to Characterize the Relationship between Methylomic Variation, Gene Expression, and Complex Traits. Am J Hum Genet 2018, 103(5):654-665.

27. Liu L, Zeng P, Yang S, Yuan Z: Leveraging methylation to identify the potential causal genes associated with survival in lung adenocarcinoma and lung squamous cell carcinoma. Oncology Letters 2020, 20(1):193-200.

28. Yu H, Cheng W, Zhang X, Wang X, Yue W: Integration analysis of methylation quantitative trait loci and GWAS identify three schizophrenia risk variants. Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology 2020, 45(7):1179-1187.

29. Wu Y, Zeng J, Zhang F, Zhu Z, Qi T, Zheng Z, Lloyd-Jones LR, Marioni RE, Martin NG, Montgomery GW et al: Integrative analysis of omics summary data reveals putative mechanisms underlying complex traits. Nat Commun 2018, 9(1):918.

30. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, Eyler AE, Denny JC, Consortium GT, Nicolae DL et al: A gene-based association method for mapping traits using reference transcriptome data. Nature Genetics 2015, 47(9):1091-1098.
Geus EJC, Boomsma DI, Wright FA et al: Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 2016, 48(3):245-252.

Hu Y, Li M, Lu Q, Weng H, Wang J, Zekavat SM, Yu Z, Li B, Gu J, Muchnik S et al: A statistical framework for cross-tissue transcriptome-wide association analysis. Nat Genet 2019, 51(3):568-576.

Wainberg M, Sinnott-Armstrong N, Mancuso N, Barbeira AN, Knowles DA, Golan D, Ermel R, Ruusalepp A, Quertermous T, Hao K et al: Opportunities and challenges for transcriptome-wide association studies. Nat Genet 2019, 51(4):592-599.

Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK: Integrating predicted transcriptome from multiple tissues improves association detection. PLoS Genet 2019, 15(1):e1007889.

Zeng P, Zhou X: Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nature Communications 2017, 8(1):456.

Zhu H, Zhou X: Transcriptome-wide association studies: a view from Mendelian randomization. Quantitative Biology 2020.

Wilson D: The harmonic mean p-value for combining dependent tests. Proceedings of the National Academy of Sciences 2019, 116(4):1195-1200.

Held L: On the Bayesian interpretation of the harmonic mean p-value. Proceedings of the National Academy of Sciences 2019, 116(13):5855-5856.

Wilson DJ: Reply to Held: When is a harmonic mean p-value a Bayes factor? Proceedings of the National Academy of Sciences 2019, 116(13):5857-5858.

Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, Shen R, Taylor AM, Cherniack AD, Thorsson V 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.e296.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW et al: Common SNPs explain a large proportion of the heritability for human height. Nature Genetics 2010, 42(7):565-569.

Makowsky R, Pajewski NM, Klimentidis YC, Vazquez AI, Duarte CW, Allison DB, de Los Campos G: Beyond Missing Heritability: Prediction of Complex Traits. Plos Genet 2011, 7(4):e1002051.

Zeng P, Zhou X, Huang S: Prediction of gene expression with cis-SNPs using mixed models and regularization methods. BMC Genomics 2017, 18:368.

Zhou X, Carbonetto P, Stephens M: Polygenic modeling with bayesian sparse linear mixed models. Plos Genet 2013, 9(2):e1003264.
Tibshirani R: Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 1996, 58(1):267-288.

Zou H, Hastie T: Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, 67(2):301-320.

Friedman J, Hastie T, Tibshirani R: Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 2010, 33(1):1-22.

Cox DR: Regression Models and Life-Tables. *Journal of the royal statistical society Series B (Methodological)* 1972, 34(2):187-220.

Verbanck M, Chen C-Y, Neale B, Do R: Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet* 2018, 50(5):693-698.

Bowden J, Davey Smith G, Burgess S: Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* 2015, 44(2):512-525.

Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR: Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I² statistic. *Int J Epidemiol* 2016, 45(6):1961-1974.

Burgess S, Thompson SG: Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol* 2017, 32(5):377-389.

Slob EA, Groenen PJ, Thurik AR, Rietveld CA: A note on the use of Egger regression in Mendelian randomization studies. *Int J Epidemiol* 2017:dyx191.

Barfield R, Feng H, Gusev A, Wu L, Zheng W, Pasaniuc B, Kraft P: Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genet Epidemiol* 2018, 42(5):418-433.

Therneau TM, Grambsch PM, Pankratz VS: Penalized survival models and frailty. *Journal of computational and graphical statistics* 2003, 12(1):156-175.

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

Fisher RA: Statistical Methods for Research Workers, 5th Edn. Biological monographs and manuals. Edinburgh: Oliver and Boyd Ltd; 1934.

Rice K: A Decision-Theoretic Formulation of Fisher’s Approach to Testing. *Am Stat* 2010, 64(4):345-349.

Wilson D: Harmonic Mean p-Values and Model Averaging by Mean Maximum Likelihood. R package version 3.0.
Yu G, Wang LG, Han Y, He QY: clusterProfiler: an R package for comparing biological themes among gene clusters. Omics: a journal of integrative biology 2012, 16(5):284-287.

Robinson MD, Smyth GK: Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 2008, 9(2):321-332.

Li CI, Su PF, Shyr Y: Sample size calculation based on exact test for assessing differential expression analysis in RNA-seq data. BMC bioinformatics 2013, 14:357.

McCarthy DJ, Chen Y, Smyth GK: Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res 2012, 40(10):4288-4297.

Chai H, Zhou X, Cui Z, Rao J, Hu Z, Lu Y, Zhao H, Yang Y: Integrating multi-omics data with deep learning for predicting cancer prognosis. bioRxiv 2019:807214.

Deng S, Ma J, Zhang L, Chen F, Sang Z, Jia Z, Ma L: De novo transcriptome sequencing and gene expression profiling of Magnolia wufengensis in response to cold stress. BMC Plant Biology 2019, 19(1):321.

Visscher PM, Hill WG, Wray NR: Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 2008, 9(4):255-266.

Yang J, Manolio TA, Pasquale LR, Boerwinkle E, Caporaso N, Cunningham JM, de Andrade M, Feenstra B, Feingold E, Hayes MG et al: Genome partitioning of genetic variation for complex traits using common SNPs. Nature Genetics 2011, 43(6):519-525.

Güzel C, Govorukhina NI, Wisman GBA, Stingl C, Dekker LJM, Klip HG, Hollema H, Guryev V, Horvatovich PL, van der Zee AGJ et al: Proteomic alterations in early stage cervical cancer. Oncotarget 2018, 9(26):18128-18147.

Broniarczyk J, Pim D, Massimi P, Bergant M, Gozdzicka-Jozefiak A, Crump C, Banks L: The VPS4 component of the ESCRT machinery plays an essential role in HPV infectious entry and capsid disassembly. Sci Rep 2017, 7:45159.

Rudolph C, Siguener A, Hartmann A, Orso E, Bals-Pratsch M, Gronwald W, Seifert B, Kalbitzer HR, Verdorfer I, Luetjens CM et al: ApoA-I-binding protein (AI-BP) and its homologues hYjeF_N2 and hYjeF_N3 comprise the YjeF_N domain protein family in humans with a role in spermiogenesis and oogenesis. Hormone and metabolic research = Hormons- und Stoffwechselforschung = Hormones et metabolisme 2007, 39(5):322-335.
Tanaka AJ, Cho MT, Millan F, Juusola J, Retterer K, Joshi C, Niyazov D, Garnica A, Gratz E, Deardorff M et al: Mutations in SPATA5 Are Associated with Microcephaly, Intellectual Disability, Seizures, and Hearing Loss. *American journal of human genetics* 2015, *97*(3):457-464.

White PW, Faucher AM, Massariol MJ, Welchner E, Rancourt J, Cartier M, Archambault J: Biphenylsulfonacetic acid inhibitors of the human papillomavirus type 6 E1 helicase inhibit ATP hydrolysis by an allosteric mechanism involving tyrosine 486. *Antimicrobial agents and chemotherapy* 2005, *49*(12):4834-4842.

Martinez-Ruiz G, Maldonado V, Ceballos-Cancino G, Grajeda JP, Melendez-Zajgla J: Role of Smac/DIABLO in cancer progression. *Journal of experimental & clinical cancer research : CR* 2008, *27*(1):48.

Lee S, Kim MG, Ahn H, Kim S: Inositol Pyrophosphates: Signaling Molecules with Pleiotropic Actions in Mammals. *Molecules (Basel, Switzerland)* 2020, *25*(9).

Fridy PC, Otto JC, Dollins DE, York JD: Cloning and characterization of two human VIP1-like inositol hexakisphosphate and diphosphoinositol pentakisphosphate kinases. *The Journal of biological chemistry* 2007, *282*(42):30754-30762.

Leung CO, Deng W, Ye TM, Ngan HY, Tsao SW, Cheung AN, Pang RT, Yeung WS: miR-135a leads to cervical cancer cell transformation through regulation of β-catenin via a SIAH1-dependent ubiquitin proteosomal pathway. *Carcinogenesis* 2014, *35*(9):1931-1940.

Gao J, Zhang L, Liu Z, Yao S, Gao S, Wang L: Effect of miR-135b inhibitor on biological characteristics of osteosarcoma cells through up-regulating PPM1A. *International journal of clinical and experimental pathology* 2019, *12*(3):689-699.

Chen Z, Li S, Huang K, Zhang Q, Wang J, Li X, Hu T, Wang S, Yang R, Jia Y et al: The nuclear protein expression levels of SNAI1 and ZEB1 are involved in the progression and lymph node metastasis of cervical cancer via the epithelial-mesenchymal transition pathway. *Human pathology* 2013, *44*(10):2097-2105.

Hou T, Tong C, Kazobinka G, Zhang W, Huang X, Huang Y, Zhang Y: Expression of COL6A1 predicts prognosis in cervical cancer patients. *Am J Transl Res* 2016, *8*(6):2838-2844.

Katoh M: Functional proteomics, human genetics and cancer biology of GIPC family members. *Experimental & molecular medicine* 2013, *45*(6):e26.
26. Y: dCTP pyrophosphohydrase exhibits nucleic accumulation in multiple carcinomas. *European journal of histochemistry : EJH* 2013, 57(3):e29.

27. Green ES: *Analysis of HPV16 Variants in the Carolina Women’s Care Study and a Comparison of Gene Expression Profiles of Exfoliated Cervical Cells From Women Who Either Clear or Do Not Clear an HPV16 Infection*. (Doctoral dissertation). 2019.

28. Meng L, Dian F, Yinan X, Hao Z, Dingyue Z: *Anticancer Effect of Natural Product Sulforaphane by Targeting MAPK Signal through miRNA-1247-3p in Human Cervical Cancer Cells*. *Biointerface Research in Applied Chemistry* 2020, 11(1):7943-7972.

29. Wang X, Wang Y, Zhang Z, Huang M, Fei Y, Ma J, Mi L: *Discriminating different grades of cervical intraepithelial neoplasia based on label-free phasor fluorescence lifetime imaging microscopy*. *Biomedical optics express* 2020, 11(4):1977-1990.

30. Das P, Bansal A, Rao SN, Deodhar K, Mahantshetty U, Shrivastava SK, Sivaraman K, Mulherkar R: *Somatic Variations in Cervical Cancers in Indian Patients*. *PLoS One* 2016, 11(11):e0165878.

31. Li P, Guo H, Zhou G, Shi H, Li Z, Guan X, Deng Z, Li S, Zhou S, Wang Y et al: *Increased ZNF84 expression in cervical cancer*. *Archives of gynecology and obstetrics* 2018, 297(6):1525-1532.

32. Network CGAR: *Integrated genomic and molecular characterization of cervical cancer*. *Nature* 2017, 543(7645):378-384.

33. Wilson CM, Magnaudieix A, Yardin C, Terro F: *DC2 and keratinocyte-associated protein 2 (KCP2), subunits of the oligosaccharyltransferase complex, are regulators of the gamma-secretase-directed processing of amyloid precursor protein (APP)*. *The Journal of biological chemistry* 2011, 286(36):31080-31091.

34. Huang HS, Buck CB, Lambert PF: *Inhibition of gamma secretase blocks HPV infection*. *Virology* 2010, 407(2):391-396.

35. Ivansson EL, Juko-Pecirep I, Erlich HA, Gyllensten UB: *Pathway-based analysis of genetic susceptibility to cervical cancer in situ: HLA-DPB1 affects risk in Swedish women*. *Genes and immunity* 2011, 12(8):605-614.

36. Manor O, Segal E: *Robust Prediction of Expression Differences among Human Individuals Using Only Genotype Information*. *Plos Genet* 2013, 9(3):e1003396.

37. Manor O, Segal E: *GenoExp: a web tool for predicting gene expression levels from single nucleotide polymorphisms*. *Bioinformatics* 2015, 31(11):1848-1850.

38. Yuan Z, Zhu H, Zeng P, Yang S, Sun S, Yang C, Liu J, Zhou X: *Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies*. *Nat Commun* 2020, 11(1):3861.

26/29
Yeung K-F, Yang Y, Yang C, Liu J: CoMM: A Collaborative Mixed Model That Integrates GWAS and eQTL Data Sets to Investigate the Genetic Architecture of Complex Traits. Bioinformatics and Biology Insights 2019, 13:1177932219881435.
Figure legends

Figure 1. Schematic framework of our proposed two-stage causal inference approach. Top: estimate the weight of each methylation site based on the methylation-expression pair of a given gene with various prediction models; Bottom: evaluate the association between methylation-regulated gene expression (MReX) and the survival of cervical cancer using the Cox linear mixed-effects model and then discover causal genes for cervical cancer in TCGA.

Figure 2. Flowchart for the present study with datasets of cervical cancer available from TCGA. (1) Various levels of raw datasets were included for cervical cancer; we conducted a series of quality control for those raw datasets; (2) gene expressions predicted with methylations were generated with diverse prediction models, the Cox linear mixed-effects model was applied to identify methylation-driven genes based on predicted expression levels; we aggregated the p values of genes from different prediction models through a p-values combination manner to find significant genes that were related to the survival of cervical cancer. Finally, we further implemented functional and differential expression analyses for newly identified associated genes.

Figure 3. (A) The number of prediction models that have maximum $R^2$ across all the genes analyzed when predicting expression level with using methylations. (B) Pearson's correlation of the p values (in a scale of -log10) obtained via in the Cox linear mixed-effects model with five different prediction models. In the plot the intensity of the color and the size of the circle represent the magnitude of the correlation. (C) UpSet plot to illustrate the intersection of associated genes identified by tests with five prediction models. LMM: Linear mixed model; BSLMM: Bayesian sparse linear mixed model; DPR: Latent Dirichlet Process Regression; ENET: elastic net; HMP: harmonic mean p-value combination method.

Figure 4. Manhattan plot showing the significance of all genes. Each plot is in a -log10 (false discovery rate [FDR]) scale. Genes with -log10 FDR > 1.3 (i.e. FDR < 0.05) are highlighted. DPR: Latent Dirichlet Process Regression; ENET: elastic net; HMP: harmonic mean p-value method
Figure 5. (A) Heatmap of expression levels for these 23 newly identified causal genes of cervical cancer. (B) Heatmap for differentially expressed genes. (C) Gene ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analyses for the 23 genes. Count Number denotes the number of genes related to the enriched GO or KEGG pathway.