Integrating genome and epigenome data to identify tissue-specific DNA methylation biomarkers for cancer risk

Yaohua Yang1†*, Yaxin Chen2†, Shuai Xu3, Xingyi Guo3, Guochong Jia3, Jie Ping3, Xiang Shu4, Tianying Zhao3, Fangcheng Yuan3, Gang Wang2, Yufang Xie2, Hang Ci2, Dan Liu2, Fei Ye5, Xiao-Ou Shu3, Wei Zheng3, Li Li6, Qiuyin Cai3*, Jirong Long3*

† These authors contribute equally to this work

*Corresponding authors contact information

Dr. Yaohua Yang, Center for Public Health Genomics, School of Medicine, University of Virginia. Address: 560 Ray C. Hunt Dr, Room 4408, Charlottesville, VA 22903. Email: vta8we@virginia.edu

Dr. Qiuyin Cai, Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center. Address: 1161 21st Avenue South, MCN B-2104, Nashville, TN, 37232. Email: qiuyin.cai@vanderbilt.edu

Dr. Jirong Long, Division of Epidemiology, Department of Medicine, Vanderbilt University Medical Center. Address: 2525 West End Avenue, Suite 800, Nashville, TN, 37203. Email: jirong.long@vumc.org

Affiliations:

1 Center for Public Health Genomics, Department of Public Health Sciences, UVA Comprehensive Cancer Center, School of Medicine, University of Virginia, Charlottesville, VA, USA

2 Frontiers Science Center for Disease-Related Molecular Network, Department of Respiratory and Critical Care Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan, China

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
3 Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA

4 Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

5 Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

6 Department of Family Medicine, UVA Comprehensive Cancer Center, School of Medicine, University of Virginia, Charlottesville, VA, USA
Abstract

The relationship between tissue-specific DNA methylation and cancer risk remains inadequately elucidated. Leveraging the Genotype-Tissue Expression (GTEx) consortium, we developed genetic models to predict DNA methylation at CpG sites (CpGs) across the genome for seven tissues and applied these models to genome-wide association study (GWAS) data of corresponding cancers, namely breast, colorectal, renal cell, lung, ovarian, prostate, and testicular germ cell cancers. At Bonferroni-corrected $P<0.05$, we identified 2,776 CpGs significantly associated with cancer risk, of which 92.7% (2,572) were specific to a particular cancer type. Notably, 57 CpGs within 35 putative novel loci retained significant associations with cancer risk after conditioning on proximal GWAS-identified signals. Further integrative multi-omics analyses revealed 791 CpG-gene-cancer trios, suggesting that DNA methylation at 248 distinct CpGs might influence cancer risk through regulating expression of 145 unique $cis$-genes. These findings substantially advance our understanding of the interplay between genetics, epigenetics, and gene expression in cancer etiology.
Genome-wide association studies (GWAS) have identified over 1,000 common variants associated with cancer risk\textsuperscript{1-7}. However, these variants are mainly situated in non-coding regions, posing challenges in identifying target genes and mechanisms\textsuperscript{1-7}. While expression quantitative trait loci (eQTL) studies have uncovered many genes associated with GWAS-identified variants, most of them were limited by the small proportion of gene expression variation captured by individual eQTL variants\textsuperscript{8}. Transcriptome-wide association studies (TWAS) addressed this by incorporating multiple \textit{cis}-variants to predict gene expression, unveiling hundreds of candidate cancer susceptibility genes\textsuperscript{9}. However, over half of GWAS loci lacked TWAS-identified genes, implying the possible existence of additional mechanisms contributing to the genetic susceptibility of cancer risk beyond \textit{cis}-gene expression regulation.

DNA methylation plays crucial roles in regulating gene expression regulation, maintaining genomic stability, and establishing of cell identity\textsuperscript{10}. Aberrant DNA methylation patterns, such as global hypomethylation and gene-specific hypermethylation, are hallmarks of cancers\textsuperscript{11}. In addition to environmental factors, DNA methylation is also shaped by genetics\textsuperscript{12}. In the largest GWAS of blood DNA methylation conducted to date (n=32,851), methylation QTLs (meQTLs) were identified for \textasciitilde45.2\% (190,102) of CpGs on the Illumina HumanMethylation450 BeadChip\textsuperscript{12}. A recent study employed the Illumina MethylationEPIC BeadChip to profile DNA methylation across nine distinct tissue types and discovered meQTLs for 37.9\% (286,152) of all investigated CpGs. Of these meQTLs, 37\% were detected among all tissues while 5\% were specific to a particular tissue type\textsuperscript{13}. In addition, a subset of these meQTLs were found to colocalize with GWAS-identified loci for various traits in biologically relevant tissues\textsuperscript{13}. However, most of these colocalizations did not involve eQTLs, suggesting that genetic effects on
trait variation in those loci were more likely to be mediated by DNA methylation rather than
gene expression. Therefore, dissecting tissue-specific genetically determined DNA methylation
holds promise in unravelling the genetic susceptibility to complex traits, including cancer
susceptibility.

We previously discovered 1,343 CpGs with genetically predicted DNA methylation levels in
blood associated with cancer risk. However, the lack of tissue DNA methylation data
hindered the evaluation of these findings in cancer-relevant tissues. The present study aimed to
identify tissue-specific DNA methylation biomarkers associated with cancer risk and decipher
the underlying mechanisms. Leveraging normal tissue DNA methylation data and paired genetic
data of cancer-free donors from the Gene-Tissue Expression (GTEx) consortium, we developed
statistical models for predicting DNA methylation at CpGs across the genome for seven tissue
types. These models were subsequently applied to cancer GWAS data to infer associations
between genetically predicted CpG methylation and the risk of breast, colorectal, renal cell, lung,
ovid, prostate, and testicular germ cell cancers, respectively. For identified cancer-associated-
CpGs, we employed integrative analyses of DNA methylomic, transcriptomic, genomic, and
cancer GWAS data to further explore whether they may affect cancer risk through modulating
the expression of nearby genes.

**Results**

**Tissue-specific DNA methylation prediction models**

The analytical framework of this study is illustrated in Figure 1. Processed DNA methylation
data, including beta mixture quantile (BMIQ)-normalized β values of 754,054 CpGs across 987
tissue samples from cancer-free GTEx subjects, were obtained from the Gene Expression Omnibus (GEO). After excluding breast tissue samples from males and samples from non-European descendants or without genetic data, 30 breast, 167 colon, 41 kidney, 163 lung, 118 ovary, 96 prostate, 44 testis, 42 whole blood, and 37 muscle tissue samples were retained. For each CpG in each tissue, a single-tissue model was built using elastic net and a joint-tissue models was established using multivariate-response penalized regression to borrow information from all the other tissues, respectively, and the one with higher prediction performance was kept. Of the 754,054 CpGs investigated, models for 510,444 (67.7%) exhibited reliable prediction performance of $R>0.1$ and $P<0.05$. Notably, 45% ($n=229,907$) of these models were highly tissue-specific, found exclusively in one tissue, while only 2% ($n=10,145$) were ubiquitous across all tissues. Specifically, we built 99,707 models for breast, 202,922 for colon, 138,849 for kidney, 204,339 for lung, 192,188 for ovary, 163,663 for prostate, and 128,843 for testis tissues, respectively.

**Association between genetically predicted DNA methylation and cancer risk**

For each tissue type, prediction models were applied to GWAS data of the corresponding cancer using SPrediXcan to identify CpGs with genetically predicted DNA methylation levels significantly associated with cancer risk at Bonferroni-corrected $P<0.05$. These analyses involved GWAS data of seven cancers, including breast ($n=424,601$), colorectal ($n=254,791$), renal cell ($n=31,190$), lung ($n=85,716$), ovarian ($n=63,347$), prostate ($n=140,306$), and testicular germ cell ($n=28,135$) cancers. All subjects included in these GWAS were of European ancestry, except for those in colorectal cancer GWAS, of which 73% and 25% were of European and Asian (27%) ancestry, respectively. In total, 2,776 CpGs were found to be significantly
associated with the risk of at least one cancer (Figure 2; Supplementary Tables 1-7). A total of 95,438 genetic variants were included in prediction models of these 2,776 CpGs, with a median of 29 variants per model (interquartile range [IQR]: 17-46). Remarkably, 92.7% (2,572) of these 2,776 CpGs showed associations that were exclusively to a specific cancer type, including 371 (92.3% of 402) for breast cancer, 715 (95.6% of 748) for colorectal cancer, 17 (100%) for renal cell cancer, 386 (87.7% of 440) for lung cancer, 226 (89.3% of 253) for ovarian cancer, 721 (93.5% of 771) for prostate cancer, and 136 (93.8% of 145) for testicular germ cell cancer (Figure S1). Among these 2,776 CpGs, 2,572 are localized within 315 of the 690 (45.6%) cancer susceptibility loci identified by previous GWAS 1-7, while the remaining 204 CpGs at 42 loci are at least one mega base (Mb) away from any GWAS-identified cancer risk variants (Figure 2).

Upon adjusting for all independent GWAS signals in the nearest known loci, 130 of the 2,776 CpGs sustained a significant association with cancer risk at the same Bonferroni-corrected thresholds (Supplementary Tables 1-7). Among these 130 CpGs, 57 CpGs at 35 loci are >1Mb away from any GWAS-reported variants, including 15 CpGs at four loci for breast cancer, 35 CpGs at 27 loci for colorectal cancer, one CpG at one locus for lung cancer, three CpGs at two loci for prostate cancer, and three CpGs at one locus for testicular germ cell cancer (Table 1; Supplementary Tables 1-7), suggesting that these loci might be potential novel risk loci for cancers. The remaining 73 CpGs reside in 18 known risk loci, including eight CpGs at five loci for breast cancer, 34 CpGs at seven loci for colorectal cancer, and 31 CpGs at six loci for prostate cancer (Table 2; Supplementary Tables 1-7).
We compared the capability of approach we employed here with the TWAS approach in delineating associations at known cancer susceptibility loci identified by previous GWAS. Results of gene- and splicing-based TWAS were either acquired from recent studies or generated by SPrediXcan analyses using GTEx (v8)-based prediction models and cancer GWAS data. Despite the on average ~46% (9% to 85%) smaller sample size in prediction model development of our study, we uncovered significant associations in more known loci than TWAS (315 vs. 302), particularly for renal cell cancer (six vs. one) and prostate cancer (81 vs. 65). Noteworthy, in 27.9% of (88 out of 315) known loci which have CpGs significantly associated with cancer risk, TWAS was unable to identify any significant associations. These results emphasize the pronounced effectiveness of our approach in detecting association signals within GWAS-identified loci compared to TWAS.

The 2,776 cancer-associated-CpGs we identified showed significant enrichments in various genomic regulatory regions in relevant tissues and cell lines at a false discovery rate (FDR)<0.05 (Supplementary Figure 2). For instance, colorectal-cancer-associated CpGs were significantly enriched at active genomic regions that were featured by DNase I hypersensitive sites (DHS), chromatin states with enhancer and promoter signatures, and histone marks including H3K4me1, H3K4me3 and H3K36me3 in intestine tissues/cells. In contrast, lung-cancer-associated CpGs exhibited significant enrichments in inactive genomic regions that were characterized by Polycomb-repressed chromatin and the H3K27me3 mark in lung tissues/cells. For cancers with GWAS data stratified by histological types available, most CpGs showing significant associations with overall cancer risk were consistently found to be associated with the risk of most cancer subtypes. Nevertheless, significant cross-subtype heterogeneity was observed for
associations of 16, 3, and 8 CpGs with the risk of breast, lung, and ovarian cancers, respectively, at the Bonferroni-corrected thresholds utilized in main analyses (Supplementary Tables 1, 4, 5).

Except for ovarian cancer and testicular germ cell cancer, CpGs associated with any of the other five cancers were investigated for their differential methylation between tumor and adjacent normal tissues using data from The Cancer Genome Atlas (TCGA). Among these 2,378 CpGs, data on 1,198 (~50%) CpGs were available in TCGA, which is consistent with the coverage difference between the DNA methylome profiling arrays used by the present study (Illumina MethylaitonEPIC array) and TCGA (Humanmehylation450 BeadChip). Of these 1,198 CpGs, 300 (~25%) showed differential DNA methylation at $P<0.05$ with directions of effects consistent with those of CpG-cancer associations. This included 63 out of 188 CpGs for breast cancer, 88 out of 370 CpGs for colorectal cancer, two out of six CpGs for renal cell cancer, 68 out of 317 CpGs for lung cancer, and 79 out of 316 CpGs for prostate cancer (Supplementary Tables 8-12).

**DNA methylation influencing cancer risk through modulating cis-gene expression**

To search for potential target genes of cancer-associated CpGs, we performed expression quantitative trait methylation (eQTM) analyses using GTEx data of the corresponding tissue. At FDR<0.05, we identified 1,242 CpG-gene association pairs, including 245 (55 CpGs and 50 genes) in breast, 72 (49 CpGs and 58 genes) in colon, three (three CpGs and three genes) in kidney, 358 (87 CpGs and 51 genes) in lung, 503 (85 CpGs and 36 genes) in ovary, 48 (34 CpGs and 45 genes) in prostate, and 13 (10 CpGs and 12 genes) in testis. Genes involved in these CpG-gene pairs were then assessed for their impacts on the proliferation of relevant cancer cells using
Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-Cas9 essentiality screening data from DepMap. Out of 117 genes with data available in DepMap, 19 demonstrated essential roles in cell proliferation at a commonly used threshold of median CERES Score <-0.5 \(^{22}\), including *ELL*, *ADSL*, and *KANSL1* in breast cancer cells, *ZNF574* in colorectal cancer cells, *CCND1* for renal cell cancer cells, *DHX16*, *HYOU1*, *ABCF1*, *PPP1R10*, and *DDX39B* for lung cancer cells, *NSF* and *KANSL1* for ovarian cancer cells, *LSM2*, *CCT4*, *CTDP1*, *CPSF3*, *MRPL45*, *FDPS*, *PRRC2A*, and *NOLC1* in prostate cancer cells (Supplementary Figure 3).

We further investigated target genes of cancer-associated CpGs for their predicted expression in association with cancer risk using GTEx and cancer GWAS data, as well as for their differential expression between tumor and adjacent normal tissues using TCGA data. At FDR<0.05, an average of 88.2% (80.6% to 100.0%) of these genes showed a significant association or differential expression. By integrating findings from associations between CpGs and cancers, between CpGs and genes, and between genes and cancers, we revealed 791 CpG-gene-cancer trios. Within each trio, the relationships of CpG-cancer, CpG-gene, and gene-cancer showed consistent directions (Table 3; Supplementary Tables 13-19). Involved in these 791 trios were 248 unique CpGs, 145 distinct cis-genes of these CpGs, and seven cancers. Such trios provided evidence supporting the potential mechanism that these 248 CpG might influence cancer risk through regulating the expression of these 145 genes. For example, as shown in Table 3 and Figure 3, genetically predicted DNA methylation at cg22872885 was associated with decreased breast cancer risk, which was consistent with the lower methylation of this CpG in breast cancer tissues than in adjacent normal tissues. This association may be explained by the negative association between DNA methylation at cg22872885 and expression of the *ZMIZ1* gene, and the
positive association between genetically predicted expression of *ZMIZ1* and breast cancer risk. Consistently, the expression of *ZMIZ1* was significantly higher in breast cancer tissues than in adjacent normal tissues.

**Discussion**

In this comprehensive investigation of tissue-specific DNA methylation and cancer risk using genetic instruments, we identified 2,776 cancer-associated CpGs, with more than 92% being specific to a particular cancer type. Integrative analyses of multi-omics data unveiled 791 CpG-gene-cancer trios indicating that unique 248 CpGs may be associated with cancer risk via modulating the expression of 145 distinct *cis*-genes. These findings strengthen our understanding of the intricate interplays among genetics, epigenetics, and gene expression in cancer etiology.

Compared to a previous study that identified meQTLs for 286,152 CpGs across nine tissues, our study established prediction models for ~1.8 times more CpGs (n=510,444) spanning seven tissues. In addition, diverging from their observation that 5% of meQTLs were specific to a particular tissue and 37% were shared across all tissues, prediction models built in the present study showcased a distinctive pattern, with almost 45% of models exclusive to a specific tissue type and only 2% being detected among all tissues. These results indicate the increased sensitivity of our prediction models in capturing tissue-specific genetic determinants of DNA methylation. Notably, more than 92% of cancer-associated CpGs identified by our study were specific to a particular cancer, underlining the tissue-specific epigenetic mechanisms driving carcinogenesis. Nonetheless, it is worth noting that the relatively modest sample size of tissue
samples used for model construction and the varied samples sizes of cancer GWAS might also contributed to the tissue-specificity of models and cancer-associated-CpGs.

Our preceding studies using blood-based models identified multiple CpGs associated with the risk of four cancers \(^{14-17}\). In the present study based on tissue-specific models, we identified a comparable number of CpGs for breast, lung, and prostate cancer, but nearly three times more CpGs for ovarian cancer, than our previous studies. Nearly 80% of CpG-cancer associations identified for these four cancers in the current study were not found in our previous studies. This disparity might be attributed to the tissue-specific nature of these CpGs or their absence in the Illumina 450K array, which was used in our previous studies.

A recent study investigated genetically predicted colorectal tissue DNA methylation and colorectal cancer risk \(^{3}\). Our study of colorectal cancer utilized the same GWAS dataset, yet with profound improvements in prediction model development. First, we exclusively employed data of transverse colon tissues from cancer-free individuals, while the recent study involved various colorectal tissue types, including those adjacent to tumors of colorectal cancer patients. In addition, our study benefited from the inclusion of data from eight other distinct tissue types, enabling the development of colon-specific models while leveraging information from other tissues. In contrast, the recent study was confined to building single-tissue models. Finally, the current study, even with a more stringent threshold to select models (\(R>0.10\) and \(P<0.05\)), established models for almost 6.7 times more CpGs (202,922 vs. 30,385) than the recent study. As a result, our association analyses identified nearly 1.5 times as many colorectal-cancer-
associated CpGs (748 vs. 501) and replicated almost 20% of the CpGs reported by the recent study ³.

For ~12% of cancer-associated CpGs, we pinpointed potential target genes in corresponding tissues. This observation is in alignment with previous findings that a large proportion of meQTL-GWAS colocalizations lack eQTL involvement ¹³. These results suggest the potential presence of mechanisms that are alternative to gene expression regulation underlying most variant-cancer associations. Noteworthy, for 19 of these genes, we found evidence from CRISPR-Cas9 screening data supporting their essential roles in the proliferation of corresponding cancer cells, many of which have been implicated in cell proliferation, such as ADSL in breast cancer ²³ and FDPS in prostate cancer ²⁴. Further, 145 of these target genes were involved in 791 CpG-gene-cancer trios, implying the impacts of 248 CpG on cancer risk by regulating expression of these genes. Among them, nine genes were essential for cancer cell proliferation, including KANSL1 for breast cancer, CCND1 for renal cell cancer, DHX16, HYOU1 and ABCF1 for lung cancer, NSF for ovarian cancer, and CCT4, CTDP1, and CPSF3 for prostate cancer. Altogether, these findings demonstrated the capability of our study to identify functional genes that might be involved in putative genetic variants-DNA methylation-gene expression-cancer pathways.

Our study has notable strengths. First, we utilized Illumina MethylationEPIC BeadChip DNA methylation data from normal tissue samples of cancer-free individuals, enabling unbiased estimation of genetic determinants of DNA methylation at ~750,000 CpGs. In addition, both single- and joint-tissue prediction models were developed, considering both tissue-specific and
shared genetic determinants to improve prediction accuracy. Moreover, despite a smaller sample size involved in model development, our study successfully identified significant associations in a greater number of known GWAS loci compared to gene- and splicing-based TWAS. Further, the replication of a substantial proportion of associations using external data of tumor and adjacent normal tissue samples strengthened the validity of our findings. Finally, the discovery of CpG-gene-cancer trios provided mechanistic insights into the critical roles of epigenetics in the genetic etiology of cancer.

Several limitations should be noted. First, the sample size for prediction model development, despite being one of the largest for many tissue types, remained relatively small compared to the extensive number of examined CpGs. Enlarging the sample sizes, particularly for tissues with limited samples, would enhance model precision and possibly unveil additional significant associations. Second, the GWAS data of colorectal cancer included individuals of both European and East Asian ancestries. Genetic differences between populations might affect the performance of certain models, particularly in regions with ancestral variations, which might consequently impact the detection of significant associations. Finally, the differential methylation analyses using TCGA data were limited by the small sample size and the potential differences in DNA methylation profiles between normal tissues adjacent to tumors and those obtained from cancer-free subjects. Future studies employing normal tissue samples from cancer-free individuals, coupled with functional experiments, are needed to further corroborate our findings.

In summary, we identified more than 2,500 CpGs showing tissue-specific associations with cancer risk, nearly 250 of which may influence cancer risk by regulating neighbor gene
expression. Our findings emphasize the effectiveness of multi-omics integration in cancer biomarker discovery and enhance our comprehension of the critical role of genetics and epigenetics in cancer etiology.

**Online Methods**

**Data Acquisition**

Whole-genome sequencing (WGS) data of blood samples and Illumina MethylationEPIC BeadChip DNA methylation data of normal tissue samples from GTEx (v8) were used as references to build DNA methylation prediction models $^8,_{13}$. Genotype and phenotype data were downloaded from the database of Genotype and Phenotype (dbGaP) (phs000424.v8.p2). Normalized DNA methylation data of nine tissue types, including breast, colon, kidney, lung, ovary, prostate, testis, whole blood, and muscle, was obtained from GEO (GSE213478). Detailed information on sample preparation, sequencing, and data processing were described elsewhere $^8,_{13}$. Briefly, WGS libraries built from blood DNA samples from 838 donors were sequenced on the Illumina HiSeq X or Hiseq 2000 platform at the Broad Institute with a median coverage of ~32X. Genotype data of European ancestry were extracted and non-palindromic variants with missing data <5%, minor allele frequency (MAF) >5%, and Hardy-Weinberg equilibrium (HWE) $P > 10^{-4}$ were retained for subsequent analyses. Epigenome-wide DNA methylation profiling was performed using the Illumina MethylationEPIC BeadChip based on 1,000 tissue DNA samples across nine unique tissue types obtained from 424 subjects. The R package *ChAMP* (v.2.8.6) $^{25}$ was utilized to process raw data to exclude low-quality samples and CpGs and estimate DNA methylation $\beta$ values $^{13}$. After background correction using the single sample normal-exponential out-of-band (ssnoob) method implemented in the R package *minfi* (v.1.36.0) $^{26}$, $\beta$ values were
normalized using the BMIQ method. Finally, DNA methylation data of 754,119 CpGs among 738 samples and genotype data of ~5.1 (IQR: 4.6-5.6) variants from 317 subjects were included in prediction model development.

Summary statistics of GWAS data for breast, colorectal, renal cell, lung, ovarian, prostate, and testicular germ cell cancers were acquired from different sources. Except for colorectal cancer, data of which were from a meta-analysis of GWAS among European and Asian descendants, data for all the other cancers were from GWAS among European descendants. Breast cancer data came from a meta-analysis of the Breast Cancer Association Consortium and UK Biobank, including 133,511 cases and 291,090 controls. Colorectal cancer data was accessed from GWAS catalog (GCST90129505), including 100,204 CRC cases and 154,587 controls, comprising 78,473 cases and 107,143 controls of European ancestry from Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the Colon Cancer Transdisciplinary Study (CORECT) and the Colon Cancer Family Registry (CCFR), and 21,731 and 47,444 of Asian Ancestry from the Asia Colorectal Cancer Consortium (ACCC). Renal cell cancer data was retrieved from dbGaP (phs001736.v2.p1), including 10,784 cases and 20,406 controls from six datasets, two from International Agency for Research on Cancer (IARC), two from National Cancer Institute (NCI), one from the University of Texas MD Anderson Cancer, and one from the Institute of Cancer Research, UK. Lung cancer data was downloaded from GWAS catalog (GCST004746), including 29,266 cases and 56,450 controls from the Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium (TRICL-ILCCO) and the Lung Cancer Cohort Consortium (LC3). Ovarian cancer data was obtained from the Ovarian Cancer Association Consortium, including 22,406 cases and 40,941
controls. Prostate cancer data was accessed from the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL) consortium, including 79,194 cases and 61,112 controls. Testicular germ cell cancer data was retrieved from dbGaP (phs001349), including 10,156 cases and 17,979 from The Testicular Cancer Consortium.

**DNA Methylation Prediction Model Development**

For each tissue, BMIQ-normalized DNA methylation β values were inverse-normalized within each CpG and regressed on covariates to get residuals. These covariates included top five genetic principle components (PCs), Probabilistic Estimation of Expression Residuals (PEER) factors (n=5 for breast, kidney, testis, muscle, and blood; n=20 for colon, lung, ovary, and prostate), sex (only for colon, kidney, lung, muscle, blood), and indicators for WGS sequencing platform (HiSeq X or HiSeq 2000) and library construction protocol indicator (PCR based or PCR-free). For each CpG, a single-tissue prediction model was built using genetic variants within its 500Kb flanking region to predict its inverse-normalized methylation residuals by fitting an elastic net model (α=0.5). In addition, a cross-tissue prediction model borrowing information from DNA methylation data of all the other eight tissues from the same set of donors was established using the multivariate-response penalized regression method implemented in Unified Test for MOlecular SignaTures (UTMOST). For both strategies, five-fold cross-validation was performed to evaluate prediction performance and satisfactory models were identified at R>0.1, 10% correlation between predicted and measured DNA methylation levels, and P<0.05. CpGs with at least one model meeting these criteria were considered in downstream association analyses. If both models are qualified, the model with higher R value was used in downstream association analyses.
Association Analyses between Genetically Predicted DNA Methylation and Cancer Risk

SPrediXcan\textsuperscript{20} was utilized to assess associations between genetically predicted DNA methylation level at CpGs and cancer risk. The association Z score was calculated following the below formula, in which \( W_{SNP-m} \) represents the weight of variant \( S \) on DNA methylation levels at CpG \( m \), \( \hat{\sigma}_S \) and \( \hat{\sigma}_m \) represents estimated variances of variant \( S \) and CpG \( m \), and \( \hat{\beta}_S \) and \( se(\hat{\beta}_S) \) represents effect size and standard error of the association between variant \( S \) and cancer risk, respectively.

\[
Z_m = \sum_{s \in Model_m} W_{SNP-m} \frac{\hat{\sigma}_S}{\hat{\sigma}_m} \frac{\hat{\beta}_S}{se(\hat{\beta}_S)}
\]

Association analyses and Bonferroni-correction were conducted for each cancer type separately and significant associations were identified at Bonferroni-corrected \( P<0.05 \), corresponding to \( 5.01 \times 10^{-7} \) (0.05/99,707) for breast cancer, \( 2.46 \times 10^{-7} \) (0.05/202,922) for colorectal cancer, \( 3.60 \times 10^{-7} \) (0.05/138,849) for renal cell cancer, \( 2.45 \times 10^{-7} \) (0.05/204,339) for lung cancer, \( 2.60 \times 10^{-7} \) (0.05/192,188) for ovarian cancer, \( 3.05 \times 10^{-7} \) (0.05/163,663) for prostate cancer, and \( 3.88 \times 10^{-7} \) (0.05/128,843) for testicular germ cell cancer.

For each CpG-cancer association, we evaluated its independence from its nearest GWAS risk signals. Briefly, for each CpG, we identified variants that were independently associated with cancer risk at \( P<5 \times 10^{-8} \) in its nearest GWAS-reported cancer susceptibility locus through a stepwise model selection procedure implemented in GCTA-COJO (v1.91.3 beta)\textsuperscript{28}. Then for each variant included in the prediction model of this CpG, its association with cancer risk conditioning on all variants identified in the first step was evaluated using GCTA-COJO. Finally, SPrediXcan analysis was conducted using the summary statistics generated in the second step.
and the Bonferroni-corrected thresholds used in main analyses were applied to determine significance.

For CpGs significantly associated with cancer risk, we first carried out eFORGE (v2.0) analyses with default settings to evaluate their overlap with DHS, chromatin states, and histone marks, respectively, in relevant tissues and cell lines using data from the Roadmap Epigenomics Project. Then for those associated with breast, renal cell, lung, and ovarian cancer, stratification analyses by cancer subtypes were performed and for each cancer, the heterogeneity in associations between subtypes was examined via Cochrane’s Q test and significant heterogeneity was detected at the Bonferroni-corrected thresholds used in main analyses. Further, except for ovarian and testicular germ cell cancer, CpGs associated with other cancer were interrogated for their differential methylation between tumor and adjacent normal tissues using Illumina HumanMethylation450 BeadChip DNA methylation data from TCGA. DNA methylation β values of 485,577 CpGs and patient information were obtained from the National Cancer Institute (NCI) Genomic Data Commons Data Portal. For each cancer, data from subtypes were combined to improve statistical power and subjects of non-European ancestry and CpGs with any missing data were excluded. Then, β values were quantile-normalized within-sample and then inverse-normalized within-CpG. For each CpG, differential methylation analysis was performed by fitting a linear mixed-effects model implemented in the R package nlme (v 3.1.140) with tissue type (tumor/adjacent normal) modeled as a random effect, adjusting for age, sample type indicator (FFPE or not), and top three DNA methylation PCs. For analyses of colorectal, renal cell, and lung cancers, sex and cancer subtype were additionally adjusted. For lung cancer, smoking status (current/former/never) and pack-year of smoking were further
adjusted. Numbers of tumor and adjacent normal tissue samples included in the final analyses were 791 vs. 97 for breast, 404 vs. 45 for colorectal, 597 vs. 205 for renal cell, 839 vs. 74 for lung, and 502 vs. 50 for prostate cancer. CpGs showing a differential expression at $P<0.05$ with directions of effect sizes consistent with Z scores of their associations with cancer risk in the main analyses were considered validated.

**Identifying CpG-gene-cancer Trios**

For each cancer-associated CpG, eQTM analyses were conducted to search for potential target genes in its 500Kb flanking region using data of corresponding tissue. DNA methylation data involved in prediction model development in the present study and gene expression data downloaded from GTExPortal were utilized for these analyses. Subjects of non-European descent were also involved to improve statistical power. In total, 34 breast, 75 colon, 131 lung, 112 ovary, 44 prostate, and 25 testis tissue samples with both DNA methylation and gene expression data available were involved in this analysis. For each tissue type, genes with $\geq6$ read and $>0.1$ Transcript Per Million (TPM) were retained and TPM values were quantile-normalized within samples and then inverse-normalized within genes. Then for each CpG-gene pair, a linear regression model was fitted with inverse-normalized DNA methylation values as the exposure and inverse-normalized gene expression values as the outcome. Five DNA methylation PEERs and five gene expression PEERs were additionally adjusted. Finally, FDR correction was applied to the nominal $P$ values and significant associations were identified at FDR$<0.05$. Such analyses could not be performed for kidney tissue due to the extremely small sample size ($n=5$). To address this, results of CpG-gene associations based on data of 414 normal non-neoplastic kidney tissue samples were accessed from a previous study.31
For genes significantly associated with cancer-associated CpGs, we first evaluated their effects on essentiality for proliferation of corresponding cancer cells using CRISPR-Cas9 screening data. CERES values of these genes in cells relevant to breast (n=48), colon and rectum (n=57), kidney (n=32), lung (n=114), ovary (n=57), and prostate (n=10) were obtained from the DepMap Public 23Q2 release. For a particular cancer, genes with a median CERES value <-0.5 across all cells were considered essential for proliferation. Next, these genes were investigated for their genetically predicted expression in association with cancer risk. Single- and cross-tissue gene expression prediction models developed using GTEx (v8) data via elastic net and Joint-Tissue Imputation (JTI) approaches were acquired from PredictDB and Zenodo, respectively. For each of the genes that had at least one model with $R^2>0.01$ and $P<0.05$, only the model with higher $R^2$ value was used in association analyses with cancer risk using SPrediXcan. Further, we examined the differential expression of these genes between tumor and normal tissues. Results from the Gene Expression Profiling Interactive Analysis (GEPIA2) web server were used. Data curation and analyses by the GEPIA2 team are described in detail elsewhere. Briefly, for each of 33 cancer types, raw RNA-seq data from TCGA and GTEx were processed using a uniform pipeline and differential expression analyses between TCGA tumor tissues and TCGA adjacent normal tissues plus GTEx normal tissues were conducted using the R limma package. For both SPrediXcan and differential expression analyses, FDR-correction was applied for each cancer separately and FDR<0.05 was used to determine significance.

Finally, to identify CpG-gene-cancer trios supporting DNA methylation of cancer-associated CpGs influencing cancer risk by modulating neighboring gene expression, we integrated results
from CpG-cancer associations, CpG-gene associations, and gene-cancer relationships to evaluate
the consistency of association directions. For gene-cancer relationships, we counted genes either
consistently associated with cancer risk or consistently differentially expressed between tumor
and normal tissues.

Declaration of Interest
The authors declare no competing interests.

Data availability
Genotype, DNA methylation, and gene expression data of GTEx participants were obtained from
dbGaP (phs000424.v8.p2), GEO (GSE213478), and GTExPortal
(https://www.gtexportal.org/home/), respectively. DNA methylation data of TCGA participants
was acquired from NCI Genomic Data Commons Data Portal (https://portal.gdc.cancer.gov/).
GTEx v8-based gene expression and splicing prediction models were downloaded from
PredictDB Data Repository (https://predictdb.org/). Differential gene expression data was
accessed from GEPIA2 (http://gepia2.cancer-pku.cn/#index).

Code availability
All codes that could be used to replicate our findings, along with all DNA methylation prediction
models developed in this study are available at Zenodo (https://zenodo.org/deposit/8226213).

Acknowledgments
This analysis includes data generated by the GTEx consortium and the TCGA National Cancer Institute Program. These data were obtained from dbGaP (phs000424.v8.p2), GEO (GSE213478), and NCI Genomic Data Commons Data Portal, respectively. Y.Y. is partially supported by the NCI grant R00CA248822. This work was also supported in part by NCI grants R01CA249863 (MPIs: C.Q. and L.J.) and R01CA247987 (MPIs: L.J. and Y.F.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Data analyses were conducted on the Rivanna High-Performance Computing (PHC) system at the University of Virginia and the Advanced Computing Center for Research and Education (ACCRE) HPC system at the Vanderbilt University.

**Contributions**

Y.Y., L.J., and C.Q. conceived of and directed the study. Y.Y. and C.Y. performed all bioinformatics and statistical analyses, prepared tables, figures, and supplements, and wrote the manuscript, under the supervision of L.J., and C.Q. L.J., C.Q., Z.W., S.XO., Y.F., and L.L. provided valuable advice on analytical methods, results interpretation, and manuscript writing, and critically revised the original version of manuscript. X.S., W.G., X.Y., C.H., L.D. intently checked analyses pipelines and codes and finalized tables, figures, and supplements. All authors contributed essential feedback throughout the study, and reviewed and approved the final manuscript.
References


Figure Legends

Figure 1. Overall workflow and resources of the present study. A, the analysis pipeline. WGS, whole genome sequencing; GTEx, Gene-Tissue Expression; QC, quality control; eQTM, expression quantitative trait methylation; TCGA, The Cancer Genome Atlas. B, tissue samples used in DNA methylation prediction model development and cancer GWAS data used in association analyses.

Figure 2. Manhattan plots showing associations between genetically predicted DNA methylation at CpGs and cancer risk. A, breast cancer. B, colorectal cancer. C, renal cell cancer. D, lung cancer. E, ovarian cancer. F, prostate cancer. G, testicular germ cell cancer. The dashed red line in each plot denotes a threshold of Bonferroni-corrected $P<0.05$. Cytoband information instead of CpGs were annotated because of the large number of cancer-associated-CpGs. Potential novel loci are highlighted in red.

Figure 3. Examples of CpG-gene-cancer trios suggest DNA methylation affecting cancer risk by modulating nearby gene expression. Red arrows, lines, and blocks denote positive associations, while green ones denote negative associations. In boxplots, red and green boxes represent data of tumor and adjacent normal tissues, respectively. DNAm, DNA methylation; GEx, gene expression; OR, odds ratio; CI, confidence interval; TCGA, The Cancer Genome Atlas; BRCA, breast invasive carcinoma; COAD, colon adenocarcinoma; READ, rectum adenocarcinoma; LUAD, lung adenocarcinoma, LUSC, lung squamous cell carcinoma; OV, ovarian serous cystadenocarcinoma. A, DNA methylation at cg22872885 may decrease breast cancer risk by suppressing the expression of ZMIZ1. B, DNA methylation at cg07007382 may
decrease colorectal cancer risk by suppressing the expression of *HLA-DQA2*. C, DNA methylation at cg09476067 may decrease lung cancer risk by promoting the expression of *TRIM39*. D, DNA methylation at cg17117718 may increase ovarian cancer risk by suppressing the expression of *LRCC37A4P*.
### Tables

**Table 1. DNA methylation marks associated with cancer risk identified in genomic regions not yet reported for cancer risk**

<table>
<thead>
<tr>
<th>Cpg</th>
<th>Chromatin classification</th>
<th>Nearest risk variants</th>
<th>Distance (Mb)</th>
<th>( P^a ) adjusted for risk variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Breast cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg26218269</td>
<td>4q31.3</td>
<td>MAB21L2; LRBA</td>
<td>rs6828523</td>
<td>24.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5'UTR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg03611078</td>
<td>6p25.2</td>
<td>C6orf201</td>
<td>rs11242765</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5'UTR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg03637179</td>
<td>7q22.1</td>
<td>TAF6</td>
<td>rs79518236</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5'UTR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg22652561</td>
<td>17p12</td>
<td>MAP2K4</td>
<td>rs78378222</td>
<td>4.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSS200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg18755616</td>
<td>1q21.3</td>
<td>KCNN3</td>
<td>rs5028523</td>
<td>18.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg15503006</td>
<td>1q25.2</td>
<td>FAM20B</td>
<td>rs10911251</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSS200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg06965496</td>
<td>3p14.3</td>
<td>FLNB</td>
<td>rs9831861</td>
<td>5.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg24683684</td>
<td>3q13.12</td>
<td>BBX</td>
<td>rs12635946</td>
<td>5.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg10945313</td>
<td>4q25</td>
<td>ANK2</td>
<td>rs238976</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5'UTR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg03172688</td>
<td>5q31.3</td>
<td>PCDH1</td>
<td>rs647161</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg24248978</td>
<td>6p22.1</td>
<td>ZNF165</td>
<td>rs1476570</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5'UTR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg25824330</td>
<td>6p21.1</td>
<td>MRPS18A</td>
<td>rs5793940</td>
<td>1.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg06766817</td>
<td>6q25.3</td>
<td>SYNJ2</td>
<td>rs151127921</td>
<td>24.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg04095069</td>
<td>7q36.1</td>
<td>NUB1</td>
<td>rs7316193</td>
<td>20.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg05877768</td>
<td>10p11.21</td>
<td>PASSD3</td>
<td>rs1773860</td>
<td>5.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg00171421</td>
<td>10q22.3</td>
<td>KCNMA1</td>
<td>rs704017</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exonic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg12338214</td>
<td>10q26.3</td>
<td>BNP3</td>
<td>rs11916721</td>
<td>19.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg16984897</td>
<td>11p15.4</td>
<td>STIM1</td>
<td>rs4450168</td>
<td>6.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSS1500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg07438103</td>
<td>11p11.2</td>
<td>C11orf49</td>
<td>rs174537</td>
<td>14.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg05190884</td>
<td>12q24.31</td>
<td>ZNF664-RPLNA</td>
<td>rs73208120</td>
<td>6.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg24776343</td>
<td>13q12.12</td>
<td>TNNFSF19</td>
<td>rs11964464</td>
<td>3.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg24467989</td>
<td>13q14.11</td>
<td>ELF1</td>
<td>rs7336670</td>
<td>4.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg13259103</td>
<td>14q31.3</td>
<td>GALC</td>
<td>rs61975764</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg24797896</td>
<td>15q24.2</td>
<td>C15orf59</td>
<td>rs8031386</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg27634195</td>
<td>16p13.3</td>
<td>UNKL</td>
<td>rs9929218</td>
<td>67.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg09232225</td>
<td>16q22.1</td>
<td>CES4A</td>
<td>rs9929218</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSS200</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg04290162</td>
<td>16q23.3</td>
<td>MPHOSPH6</td>
<td>rs9930005</td>
<td>2.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSS1500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg04996089</td>
<td>16q24.3</td>
<td>PABPN1; CBFA2T3</td>
<td>rs2696839</td>
<td>2.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intergenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg13876325</td>
<td>17p11.2</td>
<td>PIgL</td>
<td>rs1078643</td>
<td>5.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg15210992</td>
<td>17q21.31</td>
<td>AOC3</td>
<td>rs983318</td>
<td>29.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TSS1500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg11650372</td>
<td>21q22.11</td>
<td>ITSN1</td>
<td>rs9983528</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intronic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lung cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg23119604</td>
<td>6p22.1</td>
<td>HLA-L</td>
<td>rs9267123</td>
<td>1.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exonic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prostate cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg21411781</td>
<td>3q22.1</td>
<td>CPNE4; ACPP</td>
<td>rs3500612</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intergenic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg10658021</td>
<td>6p21.33</td>
<td>TNF</td>
<td>rs9275160</td>
<td>1.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Exonic</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testicular germ cell cancer</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cg18042004</td>
<td>22q13.2</td>
<td>POLR3H</td>
<td>rs739525</td>
<td>20.6</td>
</tr>
</tbody>
</table>

\( P^a \) adjusted for risk variants

\( OR \), odds ratio per standard deviation (SD) increase in genetically predicted DNA methylation level; CI, confidence interval; Mb, megabase; TSS, transcription start site; UTR, untranslated region.

\( ^a \) Association \( Z \) scores, 95% CIs, and \( P \) values were estimated using SPrediXcan. All statistical tests were two-sided.

\( ^b \) Coefficients of correlation between predicted and measured DNA methylation levels.
<table>
<thead>
<tr>
<th>CpG</th>
<th>Cytopband</th>
<th>Closest gene</th>
<th>Classification</th>
<th>Z score a</th>
<th>OR (95% CI) a</th>
<th>P a</th>
<th>R b</th>
<th>Closest risk variants</th>
<th>Distance (Mb)</th>
<th>P a adjusted for risk variants</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg23178958</td>
<td>2p13.3</td>
<td>ANTXR1</td>
<td>Intronic</td>
<td>-5.56</td>
<td>0.92 (0.89-0.94)</td>
<td>2.77×10⁻⁸</td>
<td>0.38</td>
<td>rs4602255</td>
<td>0.005</td>
<td>1.08×10⁻⁷</td>
</tr>
<tr>
<td>cg20242213</td>
<td>5p12</td>
<td>MRPS30-DT</td>
<td>Intronic</td>
<td>7.97</td>
<td>1.07 (1.05-1.09)</td>
<td>1.64×10⁻¹⁵</td>
<td>0.49</td>
<td>rs10941679</td>
<td>0.068</td>
<td>8.74×10⁻⁸</td>
</tr>
<tr>
<td>cg01444058</td>
<td>6q25.1</td>
<td>LOC102723831; AP12</td>
<td>Intergenic</td>
<td>-7.35</td>
<td>0.87 (0.83-0.90)</td>
<td>2.04×10⁻¹³</td>
<td>0.55</td>
<td>rs3757322</td>
<td>0.383</td>
<td>1.31×10⁻¹⁰</td>
</tr>
<tr>
<td>cg08353955</td>
<td>11q13.3</td>
<td>LOC102724265; LINC01488</td>
<td>Intergenic</td>
<td>-20.70</td>
<td>0.66 (0.64-0.69)</td>
<td>3.48×10⁻⁹⁵</td>
<td>0.64</td>
<td>rs554219</td>
<td>0.042</td>
<td>3.82×10⁻¹¹</td>
</tr>
<tr>
<td>cg02633117</td>
<td>19p13.11</td>
<td>MYO9B</td>
<td>Intronic</td>
<td>-5.04</td>
<td>0.90 (0.87-0.94)</td>
<td>4.64×10⁻⁷</td>
<td>0.44</td>
<td>rs67397200</td>
<td>0.214</td>
<td>4.15×10⁻⁸</td>
</tr>
<tr>
<td>cg10474949</td>
<td>6p22.1</td>
<td>ZFP57; HLA-F</td>
<td>Intergenic</td>
<td>7.10</td>
<td>1.72 (1.48-2.00)</td>
<td>1.25×10⁻¹²</td>
<td>0.26</td>
<td>rs1476570</td>
<td>0.132</td>
<td>1.51×10⁻⁹</td>
</tr>
<tr>
<td>cg09037630</td>
<td>6p21.33</td>
<td>LOC100287329</td>
<td>Exonic</td>
<td>-6.98</td>
<td>0.71 (0.65-0.78)</td>
<td>3.01×10⁻¹²</td>
<td>0.24</td>
<td>rs3830041</td>
<td>0.663</td>
<td>9.46×10⁻¹²</td>
</tr>
<tr>
<td>cg04105091</td>
<td>6p21.32</td>
<td>LOC100507547</td>
<td>Exonic</td>
<td>-6.95</td>
<td>0.86 (0.82-0.89)</td>
<td>3.63×10⁻¹²</td>
<td>0.21</td>
<td>rs3830041</td>
<td>0.070</td>
<td>1.20×10⁻⁷</td>
</tr>
<tr>
<td>cg12444411</td>
<td>7p22.2</td>
<td>GNA12</td>
<td>5’UTR</td>
<td>5.49</td>
<td>1.07 (1.05-1.10)</td>
<td>3.94×10⁻⁸</td>
<td>0.58</td>
<td>rs11812197</td>
<td>0.061</td>
<td>6.66×10⁻⁸</td>
</tr>
<tr>
<td>cg14334258</td>
<td>9q31.2</td>
<td>KLF4; ACTL7B</td>
<td>Intergenic</td>
<td>-7.69</td>
<td>0.91 (0.89-0.94)</td>
<td>1.51×10⁻¹⁴</td>
<td>0.31</td>
<td>rs10978941</td>
<td>0.222</td>
<td>2.10×10⁻¹¹</td>
</tr>
<tr>
<td>cg14447205</td>
<td>9q34.2</td>
<td>GBGT1</td>
<td>Intronic</td>
<td>13.15</td>
<td>1.60 (1.49-1.71)</td>
<td>1.59×10⁻³⁹</td>
<td>0.19</td>
<td>rs7038489a</td>
<td>0.643</td>
<td>6.95×10⁻¹⁰</td>
</tr>
<tr>
<td>cg05475172</td>
<td>22q13.31</td>
<td>WNT7B</td>
<td>Intron</td>
<td>-5.85</td>
<td>0.94 (0.93-0.96)</td>
<td>5.05×10⁻⁹</td>
<td>0.65</td>
<td>rs736037</td>
<td>0.646</td>
<td>1.08×10⁻⁸</td>
</tr>
<tr>
<td>cg11886554</td>
<td>3q26.2</td>
<td>SKIL</td>
<td>Intronic</td>
<td>5.12</td>
<td>1.10 (1.06-1.13)</td>
<td>3.01×10⁻⁷</td>
<td>0.67</td>
<td>rs78416326</td>
<td>0.002</td>
<td>2.96×10⁻⁸</td>
</tr>
<tr>
<td>cg26668675</td>
<td>6p21.33</td>
<td>PSORSIC3</td>
<td>Intronic</td>
<td>7.28</td>
<td>1.28 (1.19-1.36)</td>
<td>3.37×10⁻¹³</td>
<td>0.41</td>
<td>rs62407547</td>
<td>0.932</td>
<td>1.69×10⁻⁹</td>
</tr>
<tr>
<td>cg06131755</td>
<td>6q25.3</td>
<td>SOD2-OT1</td>
<td>Exonic</td>
<td>-5.82</td>
<td>0.90 (0.87-0.93)</td>
<td>5.89×10⁻⁹</td>
<td>0.50</td>
<td>rs963800</td>
<td>0.032</td>
<td>9.36×10⁻⁸</td>
</tr>
<tr>
<td>cg20253542</td>
<td>7p15.2</td>
<td>EVX1; HIBADH</td>
<td>Intergenic</td>
<td>-9.14</td>
<td>0.85 (0.82-0.88)</td>
<td>6.40×10⁻¹⁰</td>
<td>0.54</td>
<td>rs6956484</td>
<td>0.004</td>
<td>2.54×10⁻⁷</td>
</tr>
<tr>
<td>cg04726730</td>
<td>11p15.5</td>
<td>TNNI2; LSP1</td>
<td>Intergenic</td>
<td>7.54</td>
<td>1.89 (1.60-2.24)</td>
<td>4.53×10⁻¹⁴</td>
<td>0.26</td>
<td>rs1881502</td>
<td>0.361</td>
<td>1.35×10⁻¹⁴</td>
</tr>
<tr>
<td>cg16166568</td>
<td>11q13.3</td>
<td>LOC338694; MYEOV</td>
<td>Intergenic</td>
<td>-19.88</td>
<td>0.79 (0.77-0.80)</td>
<td>6.39×10⁻⁸⁸</td>
<td>0.77</td>
<td>rs11825796</td>
<td>0.006</td>
<td>4.62×10⁻¹⁶</td>
</tr>
</tbody>
</table>

OR, odds ratio per standard deviation (SD) increase in genetically predicted DNA methylation level; CI, confidence interval; Mb, megabase; UTR, untranslated region.

a Association Z scores, ORs, 95% CIs, and P values were estimated using SPrediXcan. All statistical tests were two-sided.

b Correlation coefficients of predicted and measured DNA methylation levels.
Table 3. Consistent directions of associations across DNA methylation, gene expression, and cancer risk

<table>
<thead>
<tr>
<th>CpG</th>
<th>Chr</th>
<th>Position (HG19)</th>
<th>Gene</th>
<th>Distance (Kb)</th>
<th>Cytoband</th>
<th>CpG-cancer P-value</th>
<th>CpG-gene P-value</th>
<th>Gene-cancer P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>cg08381504</td>
<td>3</td>
<td>46792347</td>
<td>CGR8</td>
<td>847.7</td>
<td>3p21.31</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>cg07546779</td>
<td>8</td>
<td>29495175</td>
<td>LEPRTOLI</td>
<td>-457.7</td>
<td>8q12</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>cg22872288</td>
<td>10</td>
<td>80615577</td>
<td>LEML1</td>
<td>-213.0</td>
<td>10q22.3</td>
<td>-</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>cg01741372</td>
<td>11</td>
<td>783889</td>
<td>AP006621.51</td>
<td>+</td>
<td>11p15.5</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>cg20393308</td>
<td>17</td>
<td>43728044</td>
<td>KANSL1-ASI</td>
<td>-542.9</td>
<td>17q21.31</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>cg02100011</td>
<td>22</td>
<td>38635773</td>
<td>MIFN</td>
<td>+</td>
<td>22q11.1</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

**Breast cancer**

- cg08381504
- cg07546779
- cg22872288
- cg01741372
- cg20393308
- cg02100011

**Colorectal cancer**

- cg12658694
- cg19445579
- cg01329289
- cg18892128
- cg07007382
- cg22421503
- cg12444411
- cg16459265
- cg12169691
- cg09595922
- cg08956474
- cg13238214
- cg03894242
- cg01493522
- cg24467893
- cg20070332
- cg12934611
- cg21136508
- cg08728796
- cg19133199
- cg11506090

**Renal cell cancer**

- cg13524857
- cg06511366

**Lung cancer**

- cg25506641
- cg16898333
- cg03432955
- cg09476067
- cg01347674
- cg18825676
- cg13835168

**Ovarian cancer**

- cg26754761
- cg17177718

**Prostate cancer**

- cg09539077
- cg17397364

**Testicular germ cell cancer**

- cg15120454
- cg19626747
- cg22340370

Dr. [Author's Name] and [Affiliation] have declared competing interests.

Chr: chromosome; Kb: kilobase; Dir: direction of association or differential expression.

*P* values were calculated using SPrEdiXcan. All statistical tests were two-sided.

1. *P* values were calculated using SPrEdiXcan and *R* language packages for data from GTEX and GEPIA2, respectively. For genes showing in both GTEX and GEPIA2 data, *P* values from GEPIA2 data are presented. *P* values from GTEX data are highlighted in bold. All statistical tests were two-sided.
Genotype data in blood
WGS (GTEx)

DNA methylome data in tissues
Methylation EPIC BeadChip (GTEx)

Single-/cross-tissue models
Elastic net/UTMOST

~5 million SNPs

99,709-204,339 models

754,054 CpGs

0-16 heterogeneous associations

17-741 CpGs associated with cancer risk

SPrediXcan

3-503 CpG-gene pairs

Differential methylation sites analyses (TCGA)

2-79 CpGs consistently differentially methylated

3-46 genes associated with cancer risk or differentially expressed

3-274 consistent DNA methylation-gene-cancer associations

Breast
GWAS summary: 29,266 cases and 56,450 controls

Kidney
GWAS summary: 10,784 cases and 20,406 controls

Prostate
GWAS summary: 79,194 cases and 61,112 controls

Testis
GWAS summary: 10,156 cases and 179,683 controls

Lung
GWAS summary: 29,266 cases and 56,450 controls

Kidney
GWAS summary: 10,784 cases and 20,406 controls

Prostate
GWAS summary: 79,194 cases and 61,112 controls

Testis
GWAS summary: 10,156 cases and 179,683 controls

Colon
GWAS summary: 100,204 cases and 154,587 controls

Ovary
GWAS summary: 100,204 cases and 154,587 controls

(n=163)

(n=41)

(n=96)

(n=44)

(n=118)

(n=30)

(n=167)

(n=118)

(n=30)

(n=167)

(n=118)

(n=30)

(n=167)

(n=118)

(n=30)

(n=167)

(n=118)