Article

Genetically predicted levels of DNA methylation biomarkers and breast cancer risk: data from 228,951 women of European descent

Yaohua Yang[1], Lang Wu[1], Xiao-Ou Shu[1], Qiuyin Cai[1], Xiang Shu[1], Bingshan Li[2], Xingyi Guo[1], Fei Ye[3], Kyriaki Michailidou[4], Manjeet K. Bolla[4], Qin Wang[4], Joe Dennis[4], Irene L. Andrulis[5,6], Hermann Brenner[7,8,9], Georgia Chenevix-Trench[10], Daniele Campa[11], Jose E. Castelao[12], Manuela Gago-Dominguez[13,14], Thilo Dörk[15], Antoinette Hollestelle[16], Artitaya Lophatananon[17,18], Kenneth Muir[17,18], Susan L. Neuhausen[19], Håkan Olsson[20], Dale P. Sandler[21], Jacques Simard[22], Peter Kraft[23,24], Paul D. P. Pharoah[4], Douglas F. Easton[4], Wei Zheng[1], Jirong Long[1*]

[1] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA

[2] Department of Molecular Physiology & Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA

[3] Division of Cancer Biostatistics, Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, USA

[4] Center for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK

[5] Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada

© The Author(s) 2019. Published by Oxford University Press. All rights reserved. For permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

[6] Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada

[7] Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany

[8] German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany

[9] Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany

[10] Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia

[11] Department of Biology, University of Pisa, Pisa, Italy

[12] Oncology and Genetics Unit, Instituto de Investigacion Biomedica (IBI) Orense-Pontevedra-Vigo, Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain

[13] Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago De Compostela, Spain

[14] Moores Cancer Center, University of California San Diego, La Jolla, California, USA

[15] Gynaecology Research Unit, Hannover Medical School, Hannover, Germany

[16] Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands

[17] Division of Health Sciences, Warwick Medical School, Warwick University, Coventry, UK

[18] Institute of Population Health, University of Manchester, Manchester, UK

[19] Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, California, USA

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

[20] Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden

[21] Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina, USA

[22] Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, Québec, Canada

[23] Program in Genetic Epidemiology and Statistical Genetics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

[24] Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA

[*]**Corresponding Author:** Jirong Long, PhD, Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, 2525 West End Ave, Suite 800, Nashville, Tennessee, 37203, USA. Email: jirong.long@vanderbilt.edu

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Abstract

Background: DNA methylation plays a critical role in breast cancer development. Previous studies have identified DNA methylation marks in white blood cells as promising biomarkers for breast cancer. However, these studies were limited by low statistical power and potential biases. Utilizing a new methodology, we investigated DNA methylation marks for their associations with breast cancer risk.

Methods: Statistical models were built to predict levels of DNA methylation marks using genetic data and DNA methylation data from HumanMethylation450 BeadChip from the Framingham Heart Study (*N*=1,595). The prediction models were validated using data from the Women's Health Initiative (*N*=883). We applied these models to genome-wide association study (GWAS) data of 122,977 breast cancer cases and 105,974 controls to evaluate if the genetically predicted DNA methylation levels at CpGs are associated with breast cancer risk. All statistical tests were two-sided.

Results: Of the 62,938 CpG sites (CpGs) investigated, statistically significant associations with breast cancer risk were observed for 450 CpGs at a Bonferroni-corrected threshold of $P<7.94\times10^{-7}$, including 45 CpGs residing in 18 genomic regions which have not previously been associated with breast cancer risk. Of the remaining 405 CpGs located within 500 kilobase flaking regions of 70 GWAS-identified breast cancer risk variants, the associations for 11 CpGs were independent of GWAS-identified variants. Integrative analyses of genetic, DNA methylation and gene expression data found that 38 CpGs may affect breast cancer risk through regulating expression of 21 genes.

Conclusion: Our new methodology can identify novel DNA methylation biomarkers for breast cancer risk and can be applied to other diseases.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Breast cancer is the most common cancer for women in the United States as well as many countries around the world (1). DNA methylation plays critical roles in cancer development, including breast cancer (2).

DNA methylation of several genes in white blood cells had been associated with breast cancer risk, however inconsistent results showed (3-7). Most of these studies had a retrospective design. Two prospective studies found that overall DNA hypo-methylation in white blood cells was associated with increased breast cancer risk (8, 9). In addition, a panel of 250 CpGs in white blood cell DNA was identified to be predictive of breast cancer risk (10). However, none of these CpGs were consistently observed in a later study (9). These studies were limited by small sample size, lack of replication and/or reverse causation. Furthermore, the repeatability of DNA methylation measurements at some CpGs using the HumanMethylation450 BeadChip was not optimal (11), which may have contributed to the inconsistency across studies.

A recent study indicated the epigenetic supersimilarity of monozygotic twin pairs (12). More recently, 24 heritable CpGs were associated with breast cancer risk (13). Multiple genetic variants had been identified as DNA methylation quantitative trait loci (meQTL) (14-16), suggesting that DNA methylation at some CpGs are genetically determined and thus can be predicted using genetic variants. Studies using *cis* (500 kilobase [Kb] flanking regions) meQTL single nucleotide polymorphisms (SNPs) had discovered novel CpGs for diseases (17, 18). However, the proportion of variance explained by a single meQTL SNP for most CpGs is typically small. Herein, we proposed a new methodology to build statistical models to predict DNA methylation in white blood cells via multiple SNPs in a reference dataset and then apply the models to large genome-wide association study (GWAS) datasets to evaluate genetically predicted DNA methylation in association with disease risk. We tested this methodology by

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

investigating the association of genetically predicted DNA methylation with breast cancer risk using data from 122,977 breast cancer cases and 105,974 controls.

Methods

Building DNA methylation prediction models and evaluating prediction performance

Figure 1 presents the overall workflow. Genetic and DNA methylation data of white blood cell samples from 1,595 unrelated European participants included in the Framingham Heart Study (FHS) were obtained from dbGaP (phs000342 and phs000724). The HumanMethylation450 BeadChip DNA methylation data were QC-ed and normalized by using the "minfi" package (19) and were then regressed on age, sex, cell type composition and top ten principal components to get residuals. Genotyping was conducted using the Affymetrix 500K mapping array and data were imputed to the 1000 Genomes Phase I Version 3. SNPs with imputation quality ≥0.80 and minor allele frequency ≥0.05 were used. For each CpG, we built a statistical model using allelic dosage data of *cis* SNPs to predict DNA methylation residuals, following the elastic net method (α=0.50) with ten-fold cross-validation (20) (**Supplementary Methods**). Genetic data and white blood cell DNA methylation data from the HumanMethylation450 BeadChip of 883 unrelated European women included in the Women's Health Initiative (WHI) from dbGaP (phs000315, phs000675 and phs001335) were used for the external validation of models. Genotyping was conducted using the HumanOmni1-Quad_v1-0_B and the HumanOmniExpress array and data were imputed to the 1000 Genomes Phase I Version 3. DNA methylation and genotyping data were processed following a similar procedure used for the FHS data. For each CpG, the predicted DNA methylation level was estimated using its prediction model and the correlation between predicted and measured DNA methylation level was evaluated using Spearman's

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

correlation. In total, prediction models for 63,000 CpGs built by using FHS data were externally validated by WHI data (**Supplementary Methods**).

Statistical analyses

We used MetaXcan (20, 21) to investigate genetically predicted DNA methylation in association with breast cancer risk by applying prediction models to summary statistics of breast cancer GWAS from the Breast Cancer Association Consortium (BCAC) (22), including 122,977 cases and 105,974 controls of European descent. The BCAC includes three datasets, i.e. 46,785 cases and 42,892 controls genotyped on the iCOGS, 61,282 cases and 45,494 controls on the OncoArray, and 14,910 cases and 17,588 controls on varied GWAS arrays (22). Genotyping data were imputed to the 1000 Genomes Phase I Version 3. Among the 751,157 SNPs included in the predicting levels for 63,000 CpGs, summary statistics for associations between SNPs and breast cancer risk in the BCAC were available for 750,914 (>99.9%) SNPs, corresponding to 62,938 CpGs, which were included in the final analyses. This study was approved by the BCAC Data Access Coordination Committee. The association Z score was estimated using the following formula:

$$Z_m \approx \sum_{s \in \text{Model}_m} w_{sm} \frac{\hat{\sigma}_s}{\hat{\sigma}_m} \frac{\hat{\beta}_s}{\text{se}(\hat{\beta}_s)}$$

In the formula, w_{sm} is the weight of SNP s on CpG m. $\hat{\sigma}_s$ and $\hat{\sigma}_m$ are the estimated variances of SNP s and CpG m respectively. $\hat{\beta}_s$ and $\text{se}(\hat{\beta}_s)$ are the effect size and standard error of SNP s on breast cancer risk respectively (**Supplementary Methods**). Association *P* values were also calculated by MetaXcan and a Bonferroni-corrected threshold of $P<7.94\times10^{-7}$ (0.05/62,938) was used to determine statistically significant associations of genetically predicted levels of CpGs with breast cancer risk. All statistical tests were two-sided.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

We then conducted GCTA-COJO (23) and MetaXcan (21) analyses to assess whether associations of predicted DNA methylation with breast cancer risk were independent of GWAS-identified breast cancer susceptibility variants. We also performed stratification analyses by datasets within the BCAC, i.e. iCOGS, OncoArray and GWAS, and by estrogen (ER) status. Heterogeneity across BCAC datasets, and between ER status was estimated by Cochran's Q test (**Supplementary Methods**).

Functional analyses

Functional annotation of CpGs were conducted using ANNOVAR (24). The enrichments of breast-cancer-associated CpGs in putative functional elements, including DNase I hypersensitive sites and genomic regions overlapping with histone modification marks, e.g. H3K27me3, H3K36me3 and H3K4me3, were evaluated by eFORGE (25) v1.2 using data from the Roadmap Epigenomics Project (26) (**Supplementary Methods**).

Identifying consistent directions of associations across DNA methylation, gene expression and breast cancer risk

For breast-cancer-associated CpGs, we investigated DNA methylation of them in correlation with expression of their nearby genes annotated by ANNOVAR, using data of 1,367 unrelated European participants from the FHS (**Supplementary Methods**). For genes with expression levels statistically significantly correlated with DNA methylation levels at these CpGs, we built genetic models to predict their expression levels using data from 6,124 different tissue samples of 369 participants of European ancestry from the Genotype-Tissue Expression (GTEx) (27). The models were applied to the BCAC data to estimate associations between genetically

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

predicted expression of these genes and breast cancer risk, utilizing MetaXcan (20, 21) (**Supplementary Methods**). For both analyses, we used false discovery rate (FDR)<0.05 to determine statistically significant correlations/associations. To elucidate putative pathways through which DNA methylation impacts breast cancer risk, association results across DNA methylation, gene expression and breast cancer risk were integrated to assess the consistency of association directions (**Supplementary Methods**).

Comparison between prediction model approach and single meQTL SNP approach

Of the 62,938 CpGs investigated, meQTLs had been identified for 24,845 CpGs (15). We compared the prediction performance of these 24,845 CpGs via prediction models or single meQTL SNPs. We also investigated these 24,845 CpGs for their DNA methylation levels, predicted via single meQTL SNPs, in association with breast cancer risk using the BCAC data, following the inverse-variance weighted method (28) (**Supplementary Methods**). The association results were compared with those from the prediction model approach.

Results

DNA methylation prediction models

A flow diagram describing the number of CpGs and SNPs during each analysis step is shown in **Supplementary Figure 1**. Genetic and white blood cell DNA methylation data from FHS were used to build DNA methylation prediction models. In total, 473,865 autosomal CpGs were assayed and 370,785 were retained after QC. Statistical models were established to predict DNA methylation levels for 223,959 CpGs, 61,219 of which were within CpG islands. Of these 223,959 CpGs, the predicted and measured DNA methylation levels are positively correlated

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

with a correlation coefficient of at least 0.10, i.e. $R_{FHS} \geq 0.10$ and $R_{FHS}^2 \geq 0.01$ for 81,361 CpGs. To validate these 81,361 models, we applied them to the WHI data and calculated the squared values of correlation coefficients between predicted and measured DNA methylation levels, i.e. R_{WHI}^2. For these 81,361 models, a high correlation of R_{FHS}^2 and R_{WHI}^2, was observed with a Pearson correlation r=0.95 (**Figure 2**), indicating that CpGs predicted well in FHS were also predicted well in WHI. Notably, 70,269 of the 81,361 CpGs showed a $R_{WHI}^2 \geq 0.01$. For 7,269 of these 70,269 CpGs, their corresponding probes on the HumanMethylation450 BeadChip overlap with genetic polymorphisms (based on dbSNP Build 151). Such overlapping could potentially affect the estimation of their DNA methylation levels (15), hence these 7,269 CpGs were excluded. In total, 63,000 CpGs were included in downstream analyses.

Associations of genetically predicted DNA methylation with breast cancer risk

We conducted MetaXcan analyses to estimate genetically predicted DNA methylation of the 63,000 CpGs in association with breast cancer risk. Among 751,157 SNPs included in prediction models of these 63,000 CpGs, data were available for 750,914 (>99.9%) SNPs in BCAC, corresponding to 62,938 CpGs. For most of these CpGs, a substantial majority of SNPs were used in association analyses. The Manhattan plot for the associations results is shown in **Supplementary Figure 2**. Of the 62,938 CpGs, statistically significant associations were observed for 450 at $P<7.94\times10^{-7}$, a Bonferroni-corrected threshold (**Tables 1-2 and Supplementary Table 1**). For these 450 CpGs, 12,286 SNPs were included in their prediction models, with the average of 27 SNPs for each CpG. Of the 12,286 SNPs, genotypes of 10,099 (82.2%) were associated with DNA methylation levels of these CpGs at $P<0.05$. Among these 450 CpGs, 45 reside in 18 genomic regions that are 500Kb away from GWAS-identified breast

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

cancer risk variants. After adjusting for proximally located GWAS-identified risk variants, statistically significant associations ($P<7.94\times10^{-7}$) retained for 17 CpGs within 10 genomic loci (**Table 1**). Among the remaining 405 CpGs located within 500Kb flaking regions of 70 GWAS-identified breast cancer risk variants, statistically significant associations ($P<7.94\times10^{-7}$) remained for 11 CpGs within seven genomic regions after adjusting for corresponding GWAS-identified risk variants (**Table 2**). The predicted DNA methylation levels of these 450 CpGs could explain 2.2% of the familial relative risk of breast cancer (**Supplementary Methods**).

For these 450 CpGs, stratified analyses by ER status were conducted to evaluate the heterogeneity between ER-positive and ER-native diseases. Most CpGs were associated with risks of both (**Supplementary Tables 2-3**), nevertheless, 39 and eight CpGs were respectively more statistically significantly associated with ER-positive and ER-negative disease with Cochran's $P<1.11\times10^{-4}$ (0.05/450) (**Supplementary Table 2**). All 450 CpGs showed consistent associations across three BCAC subsets (**Supplementary Table 4**).

To explore potential regulatory functions of the 450 CpGs, eFORGE was used to estimate the enrichments of them in putative functional genomic regions. These 450 CpGs were enriched in genomic regions harboring H3K4me1 marks, indicative of enhancers, in human mammary epithelial cells as well as in 36 out of 38 cell types and tissues assayed in the Roadmap Epigenomics Project (26) (**Supplementary Figure 3**). Compared with all the 62,938 CpGs, these 450 CpGs were statistically significantly enriched in ncRNA_exonic regions with hypergeometric distribution test $P<5.55\times10^{-3}$ (0.05/9) (**Supplementary Table 5**). In addition, of these 450 CpGs, 36, 37 and seven were respectively within or close to (10Kb) metastable epialleles identified by three recent studies (29-31).

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

To determine whether these 450 CpGs are soma-wide, we re-built prediction models without adjusting for cell type composition. Totally, models for 411 CpGs were established with $R_{FHS}^2 \geq 0.01$. For these 411 CpGs, a high correlation between R^2 based on models with or without cell type composition adjustment was observed (Spearman correlation r=0.95, **Supplementary Figure 4A**). All these 411 CpGs were associated with breast cancer risk at $P<7.57\times10^{-5}$ and the association Z scores were highly correlated with those based on prediction models with cell type composition adjustment (Spearman correlation r=0.98, **Supplementary Figure 4B**).

DNA methylation impacting breast cancer risk through regulating gene expression

We investigated whether DNA methylation of the 450 CpGs could influence flanking gene expression using the FHS data. Among 342 CpGs and 158 genes with DNA methylation and gene expression data, statistically significant correlations were observed for 100 CpG-gene pairs, including 100 CpGs and 62 genes, at FDR<0.05 (**Supplementary Table 6**). In total, 60 of these 100 statistically significant correlations were negative. Especially, for 22 CpGs that reside in promoter regions, DNA methylation at 20 CpGs were negatively correlated with gene expression. We evaluated the associations between genetically predicted expression of these 62 genes and breast cancer risk using the GTEx and BCAC data. Gene expression prediction models were established for 45 genes, 32 of which were statistically significantly associated with breast cancer risk at FDR<0.05 (**Supplementary Table 7**).

To explore whether DNA methylation at CpGs could impact breast cancer risk through regulating gene expression, we integrated all association results and identified consistent directions of associations across 38 CpGs, 21 genes and breast cancer risk (**Table 3 and Supplementary Table 8**). Among these 38 CpGs, five reside in genomic regions not previously

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

reported for breast cancer risk via GWAS, including cg22221025 and cg26668989 in *GSTM4*, cg16647868 and cg19040266 in *SLC22A5*, and cg25839482 in *IMP3*. Except for *LRRC25*, the associations between predicted expression of the other 20 genes and breast cancer risk attenuated substantially upon adjusting for SNPs included in prediction models of their corresponding CpGs (**Table 3 and Supplementary Table 8**). The associations of these 20 genes with breast cancer risk may be modulated by DNA methylation.

Comparison of the genetic prediction model approach with the single meQTL SNP approach

To evaluate the performance of DNA methylation prediction improved by prediction model approach, we compared the R^2 from prediction models with those from meQTLs. Among the 24,845 CpGs having both models and meQTLs, prediction performances of models (R_{FHS}^2) were statistically significantly higher than those of single meQTL SNPs (R_{meQTL}^2) (**Figure 3**). Especially, 21,874 CpGs (84.1%) were predicted better ($R_{FHS}^2 > R_{meQTL}^2$) using models.

To determine whether our prediction model approach could identify more breast-cancer-associated CpGs than the meQTL approach, for the 24,845 CpGs having both prediction models and meQTLs, we investigated their DNA methylation levels, predicted by single meQTL SNPs, in association with breast cancer risk. For these 24,845 CpGs, a strong correlation (Pearson correlation r=0.88) between Z scores from prediction model and single meQTL SNP approaches. The *P* values from prediction model approach were lower than those from single meQTL SNP approach (**Supplementary Figure 5**). Of the 450 breast-cancer-associated-CpGs, meQTLs were identified for only 162 CpGs and 128 reached $P<2.01\times10^{-6}$ (Bonferroni-correction; 0.05/24,845)

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

based on single meQTL SNP approach (**Supplementary Table 9**). Therefore, only 128 (28.4%) of the 450 breast-cancer-associated-CpGs could be identified using single meQTL SNP approach.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Discussion

Using breast cancer as an example, we tested a novel methodology to identify CpGs associated with disease risk. We identified 450 CpGs that were statistically significantly associated with breast cancer risk. Of these, 38 CpGs may affect breast cancer risk through regulating expression of 21 genes. We demonstrate that our methodology is successful in identifying novel DNA methylation biomarkers for disease risk. Our findings provide substantial new information regarding the mechanistic relationship of genetics, epigenetics and gene expression and their associations with breast cancer risk.

Traditional epidemiologic studies of DNA methylation were limited by small sample size, confounders and reverse causation. Our study focused on genetically determined DNA methylation, which is immune from reverse causation and confounders. Compared with the approach using single meQTL SNPs as genetic instruments, our prediction model approach statistically significantly improved prediction performances. More importantly, over half of the breast-cancer-associated CpGs would be missed using single meQTL SNP approach. Although we focused on breast cancer, this methodology can be applied to other diseases.

We observed consistent directions of associations across 38 CpGs, 21 genes and breast cancer risk. Among them, five CpGs in three genes, *GSTM4*, *SLC22A5* and *IMP3*, are within genomic regions which had not been associated with breast cancer risk via GWAS. *GSTM4* over-expression could help to maintain a reduced state of cytochrome *c*, which contributes to methotrexate resistance in breast cancer cells (32). A mutation in *SLC22A5* was reported to

enhance cancer cell metastasis in breast tissues (33). The over-expression of *IMP3* was observed in *BRCA*- mutated invasive breast carcinomas (34). Three CpGs in *CD160* were associated with increased breast cancer risk by the down-regulation of *CD160* expression. This gene was suggested to have anti-cancer activity (35). Another three CpGs were located in *MAPT*, which were associated with breast cancer metastasis (36). In ER-negative breast cancers, the knockdown of a natural antisense of *MAPT*, *MAPT-AS1*, resulted in inhibited cancer cell proliferation (37). Recently, a CpG in *GREB1*, cg18584561, was associated with breast cancer risk (13). In the present study this CpG was removed because of low quality, hence a comparison couldn't be made. Another study (38) reported that a rare variant *BRCA1* c.-107A>T could silence *BRCA1* and increase breast cancer risk via DNA methylation. However, this variant is very rare and not included in data of FHS, WHI and BCAC, hence we could not investigate this variant for its association with either *BRCA1* methylation or breast cancer risk.

One limitation of our study is that prediction models were built using data from white blood cells, not breast tissues. However, it is unfeasible to obtain breast tissues from healthy individuals. Although in the Cancer Genome Atlas, both genotype and DNA methylation were profiled for histologically normal tissue samples from 115 breast cancer patients of European descent (39), the DNA methylation profiles of "histologically normal" tissue samples from breast cancer patients may differ from those of tissue samples from healthy women. Multiple studies have suggested that meQTLs could be consistently detected across different tissues (40-43), indicating that genetically determined DNA methylation at many CpGs have cross-tissue consistency. Therefore, the genetic models we built using data from white blood cells should capture DNA methylation of many CpGs in breast tissues.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

The present study has multiple strengths. Our novel methodology overcomes limitations of traditional epidemiological studies and is more accurate and powerful than studies based on single meQTL approach. A large number of samples in the reference dataset were used for model building and the performances of models were excellent as demonstrated by external validation. The BCAC GWAS data with to-date the largest sample size provided strong statistical power to identify associations between CpGs and breast cancer risk. By integrating multi-omics data, we found consistent evidence to support that DNA methylation may impact breast cancer risk through regulating gene expression.

In summary, using a novel methodology, we identified multiple CpGs statistically significantly associated with breast cancer risk and proposed that several CpGs may affect breast cancer risk through regulating gene expression. Our study demonstrates the utility of integrative analyses of multi-omics data in identifying novel biomarkers for risk of developing breast cancer and provides new insights into the etiology of this malignancy.

Funding

This project was supported in part by grants R01CA158473 and R01CA148677 from the U.S. National Institutes of Health as well as funds from the Anne Potter Wilson endowment. This project was also supported by development funds from the Department of Medicine at Vanderbilt University Medical Center. Artitaya Lophatananon and Kenneth Muir were partly funded through the ICEP program which is supported by CRUK (C18281/A19169). Genotyping of the OncoArray was principally funded by three sources: the PERSPECTIVE project, funded from the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de l'Économie, de la Science et de l'Innovation du Québec through

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Genome Québec, and the Quebec Breast Cancer Foundation; the National Cancer Institute at the National Institutes of Health Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project (National Institutes of Health grants U19 CA148065 and X01HG007492); and Cancer Research UK (C1287/A10118 and. C1287/A16563). BCAC is funded by Cancer Research UK [C1287/A16563], by the European Community's Seventh Framework Programme under grant agreement 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec – grant # PSR-SIIRI-701. Combining the GWAS data was supported in part by The National Institutes of Health Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE, part of the GAME-ON initiative).

Notes

The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication. The authors have no conflicts of interest to disclose.

The authors thank Jing He, Wanqing Wen of the Vanderbilt Epidemiology Center and Ran Tao of the Department of Biostatistics, Vanderbilt University Medical Center for their help with the data analysis of this study. The authors also would like to thank all of the individuals

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

who participated in the parent studies and all the researchers, clinicians, technicians and administrative staff for their contributions. The data of FHS Offspring Cohort, WHI and GTEx used in this study are publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap; dbGaP Study Accession: phs000342 and phs000724 for FHS, phs000315, phs000675 and phs001335 for WHI, and phs000424.v6.p1 for GTEx). Most of the BCAC data used in this study are or will be publicly available via dbGaP. Data from some BCAC studies are not publicly available due to restraints imposed by the ethics committees of individual studies; requests for further data can be made to the BCAC (http://bcac.ccge.medschl.cam.ac.uk/) Data Access Coordination Committee. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University.

References

1. DeSantis CE, Fedewa SA, Goding Sauer A, *et al.* Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. CA: a cancer journal for clinicians 2016;66(1):31-42.
2. Sarkar S, Horn G, Moulton K, *et al.* Cancer development, progression, and therapy: an epigenetic overview. International journal of molecular sciences 2013;14(10):21087-21113.
3. Snell C, Krypuy M, Wong EM, *et al.* BRCA1 promoter methylation in peripheral blood DNA of mutation negative familial breast cancer patients with a BRCA1 tumour phenotype. Breast Cancer Research 2008;10(1):R12.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

4. Flanagan JM, Munoz-Alegre M, Henderson S, *et al.* Gene-body hypermethylation of ATM in peripheral blood DNA of bilateral breast cancer patients. Human molecular genetics 2009;18(7):1332-1342.

5. McCullough LE, Chen J, Cho YH, *et al.* DNA methylation modifies the association between obesity and survival after breast cancer diagnosis. Breast cancer research and treatment 2016;156(1):183-194.

6. Wong EM, Southey MC, Fox SB, *et al.* Constitutional methylation of the BRCA1 promoter is specifically associated with BRCA1 mutation-associated pathology in early-onset breast cancer. Cancer Prevention Research 2011;4(1):23-33.

7. Hansmann T, Pliushch G, Leubner M, *et al.* Constitutive promoter methylation of BRCA1 and RAD51C in patients with familial ovarian cancer and early-onset sporadic breast cancer. Human molecular genetics 2012;21(21):4669-4679.

8. Severi G, Southey MC, English DR, *et al.* Epigenome-wide methylation in DNA from peripheral blood as a marker of risk for breast cancer. Breast cancer research and treatment 2014;148(3):665-673.

9. van Veldhoven K, Polidoro S, Baglietto L, *et al.* Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clinical epigenetics 2015;7(1):67.

10. Xu Z, Bolick SC, DeRoo LA, *et al.* Epigenome-wide association study of breast cancer using prospectively collected sister study samples. Journal of the National Cancer Institute 2013;105(10):694-700.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

11. Dugué P-A, English DR, MacInnis RJ, *et al.* The repeatability of DNA methylation measures may also affect the power of epigenome-wide association studies. International journal of epidemiology 2015;44(4):1460-1461.

12. Van Baak TE, Coarfa C, Dugué P-A, *et al.* Epigenetic supersimilarity of monozygotic twin pairs. Genome biology 2018;19(1):2.

13. Joo JE, Dowty JG, Milne RL, *et al.* Heritable DNA methylation marks associated with susceptibility to breast cancer. Nature communications 2018;9(1):867.

14. Gaunt TR, Shihab HA, Hemani G, *et al.* Systematic identification of genetic influences on methylation across the human life course. Genome biology 2016;17(1):61.

15. McRae AF, Marioni RE, Shah S, *et al.* Identification of 55,000 replicated DNA methylation QTL. Scientific Reports 2018;8(1):17605.

16. Shi J, Marconett CN, Duan J, *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nature communications 2014;5:3365.

17. Richardson TG, Zheng J, Smith GD, *et al.* Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. The American Journal of Human Genetics 2017;101(4):590-602.

18. Richardson TG, Haycock PC, Zheng J, *et al.* Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. Human molecular genetics 2018.

19. Aryee MJ, Jaffe AE, Corrada-Bravo H, *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics 2014;30(10):1363-1369.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

20. Wu L, Shi W, Long J, *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nature genetics 2018;50(7):968.

21. Barbeira AN, Dickinson SP, Bonazzola R, *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nature communications 2018;9(1):1825.

22. Michailidou K, Lindström S, Dennis J, *et al.* Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551(7678):92.

23. Yang J, Ferreira T, Morris AP, *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature genetics 2012;44(4):369.

24. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 2010;38(16):e164-e164.

25. Breeze CE, Paul DS, van Dongen J, *et al.* eFORGE: a tool for identifying cell type-specific signal in epigenomic data. Cell reports 2016;17(8):2137-2150.

26. Kundaje A, Meuleman W, Ernst J, *et al.* Integrative analysis of 111 reference human epigenomes. Nature 2015;518(7539):317-330.

27. Consortium G. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 2015;348(6235):648-660.

28. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic epidemiology 2013;37(7):658-665.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

29. Harris RA, Nagy-Szakal D, Kellermayer R. Human metastable epiallele candidates link to common disorders. Epigenetics 2013;8(2):157-163.

30. Silver MJ, Kessler NJ, Hennig BJ, *et al.* Independent genomewide screens identify the tumor suppressor VTRNA2-1 as a human epiallele responsive to periconceptional environment. Genome biology 2015;16(1):118.

31. Kessler NJ, Waterland RA, Prentice AM, *et al.* Establishment of environmentally sensitive DNA methylation states in the very early human embryo. Science advances 2018;4(7):eaat2624.

32. Barros S, Mencia N, Rodríguez L, *et al.* The redox state of cytochrome c modulates resistance to methotrexate in human MCF7 breast cancer cells. PloS one 2013;8(5):e63276.

33. Lee J-H, Zhao X-M, Yoon I, *et al.* Integrative analysis of mutational and transcriptional profiles reveals driver mutations of metastatic breast cancers. Cell discovery 2016;2:16025.

34. Mohanty SK, Lai JP, Gordon OK, *et al.* BRCA- mutated Invasive Breast Carcinomas: Immunohistochemical Analysis of Insulin- like Growth Factor II mRNA- binding Protein (IMP3), Cytokeratin 8/18, and Cytokeratin 14. The breast journal 2015;21(6):596-603.

35. Stecher C, Battin C, Leitner J, *et al.* PD-1 Blockade Promotes Emerging Checkpoint Inhibitors in Enhancing T Cell Responses to Allogeneic Dendritic Cells. Frontiers in Immunology 2017;8.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

36. Matrone MA, Whipple RA, Thompson K, *et al.* Metastatic breast tumors express increased tau, which promotes microtentacle formation and the reattachment of detached breast tumor cells. Oncogene 2010;29(22):3217.

37. Pan Y, Pan Y, Cheng Y, *et al.* Knockdown of LncRNA MAPT-AS1 inhibites proliferation and migration and sensitizes cancer cells to paclitaxel by regulating MAPT expression in ER-negative breast cancers. Cell & bioscience 2018;8(1):7.

38. Evans DGR, van Veen EM, Byers HJ, *et al.* A dominantly inherited 5′ UTR variant causing methylation-associated silencing of BRCA1 as a cause of breast and ovarian cancer. The American Journal of Human Genetics 2018;103(2):213-220.

39. Network CGA. Comprehensive molecular portraits of human breast tumours. Nature 2012;490(7418):61-70.

40. Smith AK, Kilaru V, Kocak M, *et al.* Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics 2014;15:145.

41. Shi J, Marconett CN, Duan J, *et al.* Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun 2014;5:3365.

42. Stueve TR, Li WQ, Shi J, *et al.* Epigenome-wide analysis of DNA methylation in lung tissue shows concordance with blood studies and identifies tobacco smoke-inducible enhancers. Hum Mol Genet 2017;26(15):3014-3027.

43. Hannon E, Weedon M, Bray N, *et al.* Pleiotropic Effects of Trait-Associated Genetic Variation on DNA Methylation: Utility for Refining GWAS Loci. Am J Hum Genet 2017;100(6):954-959.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Tables

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

CpG	Chr	Position	Closest gene	Classification	Z score*	OR (95% CI)*	*P* value*	$R_{FHS}^{2\dagger}$	$R_{WHI}^{2\dagger}$	Closest breast cancer risk SNP	Distance to the risk SNP (Kb)	*P* value* adjusted for the risk SNP
cg04794690	1	17,768,059	*RCC2*; *ARHGEF10L*	Intergenic	5.04	1.36 (1.20-1.53)	4.74×10^{-7}	0.01	0.04	rs2992756	-1,039	2.61×10^{-8}
cg22221025	1	110,186,044	*GSTM4*	Upstream	5.36	1.15 (1.09-1.21)	8.53×10^{-8}	0.05	0.03	rs11552449	-4,262	3.04×10^{-7}
cg26668989	1	110,186,163	*GSTM4*	Upstream	6.37	1.20 (1.13-1.27)	1.92×10^{-10}	0.04	0.04	rs11552449	-4,262	1.23×10^{-9}
cg04411307	2	69,391,395	*ANTXR1*	Intronic	5.10	1.05 (1.03-1.07)	3.38×10^{-7}	0.31	0.22	rs6725517	44,262	3.19×10^{-7}
cg16190888	2	69,428,235	*ANTXR1*	Intronic	-5.20	0.87 (0.82-0.91)	1.99×10^{-7}	0.04	0.03	rs6725517	44,299	2.06×10^{-7}
cg03277049	3	156,534,076	*LINC00886*	ncRNA_intronic	-5.01	0.92 (0.90-0.95)	5.52×10^{-7}	0.10	0.10	rs58058861	-15,751	5.01×10^{-7}
cg11359771	5	131,558,794	*P4HA2*	5'UTR	5.00	1.21 (1.13-1.31)	5.59×10^{-7}	0.01	0.02	rs6596100	-848	1.70×10^{-7}
cg16647868	5	131,706,066	*SLC22A5*	Intronic	4.95	1.05 (1.03-1.08)	7.36×10^{-7}	0.20	0.27	rs6596100	-701	2.02×10^{-7}
cg19040266	5	131,723,239	*SLC22A5*	Intronic	-5.05	0.93 (0.91-0.96)	4.31×10^{-7}	0.13	0.10	rs6596100	-684	1.32×10^{-7}
cg11920449	6	36,645,608	*CDKN1A*	TSS1500	-5.04	0.97 (0.96-0.98)	4.56×10^{-7}	0.69	0.67	rs9257408	7,719	7.78×10^{-7}
cg03714916	6	36,645,886	*CDKN1A*	TSS1500	-5.17	0.94 (0.92-0.96)	2.31×10^{-7}	0.17	0.16	rs9257408	7,720	3.95×10^{-7}
cg03171419	8	37,700,802	*GPR124*	3'UTR	-5.26	0.87 (0.82-0.92)	1.46×10^{-7}	0.04	0.04	rs13365225	842	7.86×10^{-8}
cg07540652	8	81,805,956	*ZNF704*; *PAG1*	Intergenic	5.05	1.19 (1.11-1.27)	4.48×10^{-7}	0.03	0.02	rs2943559	5,388	2.71×10^{-7}
cg25626611	12	115,102,065	*TBX5-AS1*; *TBX3*	Intergenic	6.01	1.12 (1.08-1.16)	1.91×10^{-9}	0.09	0.05	rs1292011	-734	1.09×10^{-7}
cg07211768	12	115,102,290	*TBX5-AS1*; *TBX3*	Intergenic	5.98	1.08 (1.06-1.11)	2.25×10^{-9}	0.16	0.14	rs1292011	-734	1.52×10^{-7}
cg25938347	15	75,639,163	*NEIL1*	TSS200	5.51	1.29 (1.18-1.42)	3.59×10^{-8}	0.01	0.02	rs151090251	8,231	1.84×10^{-7}
cg25839482	15	75,931,953	*IMP3*	3'UTR	-5.57	0.94 (0.93-0.96)	2.59×10^{-8}	0.21	0.25	rs151090251	8,524	1.57×10^{-7}

* MetaXcan was used to estimate association Z scores, ORs, 95% CIs and *P* values. All statistical tests were two-sided. OR, odds ratio per standard deviation increase in genetically predicted DNA methylation level (continuous variable); CI, confidence interval; Chr, chromosome; ncRNA, non-coding RNA; UTR, untranslated region; TSS, transcription start site; FHS, The Framingham Heart Study; WHI, The Women's Health Initiative; SNP, single nucleotide polymorphisms.

† Correlation between predicted and measured DNA methylation levels.

http://jnci.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Table 2. Eleven DNA methylation marks associated with breast cancer risk identified in genomic regions within 500Kb of known breast cancer risk variants but representing independent association signals

CpG	Chr	Position	Closest gene	Classification	Z score*	OR (95% CI)*	*P* value*	$R_{FHS}^{2\dagger}$	$R_{WHI}^{2\dagger}$	Breast cancer risk SNPs	Distance to the risk SNPs (Kb)	*P* value* adjusted for the risk SNPs
cg18789177	2	217,729,408	*TNP1*; *LOC105373876*	Intergenic	5.73	1.28 (1.17-1.39)	1.01×10^{-8}	0.02	0.01	rs4442975; rs34005590	191; 233	7.44×10^{-8}
cg16971831	5	56,110,935	*MAP3K1*	5'UTR	-11.90	0.51 (0.46-0.57)	1.23×10^{-32}	0.01	0.06	rs16886397; rs16886113; rs2229882; rs7726354; rs16886034; rs16886181; rs12655019; rs889312	23; -115; 57; 145; -127; -81; 84; -79	7.42×10^{-8}
cg20580673	14	91,735,665	*GPR68*; *CCDC88C*	Intergenic	-6.16	0.76 (0.69-0.83)	7.26×10^{-10}	0.02	0.03	rs941764	105	8.03×10^{-8}
cg00787180	14	91,751,731	*CCDC88C*	Intronic	-5.57	0.88 (0.84-0.92)	2.54×10^{-8}	0.05	0.06	rs941764	89	6.72×10^{-8}
cg09032423	16	4,015,231	*ADCY9*	3'UTR	-5.01	0.82 (0.77-0.89)	5.56×10^{-7}	0.02	0.01	rs11076805	91	1.02×10^{-7}
cg12776287	18	24,125,939	*KCTD1*	Intronic	-5.36	0.95 (0.93-0.97)	8.12×10^{-8}	0.27	0.21	rs1436904; rs527616	444; 211	1.45×10^{-7}
cg19738924	18	24,126,072	*KCTD1*	Intronic	-5.60	0.93 (0.91-0.96)	2.20×10^{-8}	0.14	0.14	rs1436904; rs527616	444; 211	2.91×10^{-7}
cg15073853	19	18,549,131	*ISYNA1*	TSS200	9.28	1.09 (1.07-1.11)	1.77×10^{-20}	0.26	0.21	rs4808801	22	9.24×10^{-8}
cg21962901	19	18,549,134	*ISYNA1*	TSS200	9.37	1.11 (1.09-1.13)	7.27×10^{-21}	0.19	0.15	rs4808801	22	9.24×10^{-8}
cg11102782	19	18,549,136	*ISYNA1*	TSS200	8.80	1.08 (1.06-1.10)	1.38×10^{-18}	0.32	0.26	rs4808801	22	9.24×10^{-8}
cg09232727	22	29,140,725	*HSCB*	Intronic	-6.23	0.76 (0.70-0.83)	4.60×10^{-10}	0.02	0.03	rs17879961; rs132390	-19; 480	2.12×10^{-7}

* MetaXcan was used to estimate association Z scores, ORs, 95% CIs and *P* values. All statistical tests were two-sided. OR, odds ratio per standard deviation increase in genetically predicted DNA methylation level (continuous variable); CI, confidence interval; Chr, chromosome; UTR, untranslated region; TSS, transcription start site; FHS, The Framingham Heart Study; WHI, The Women's Health Initiative; SNPs, single nucleotide polymorphisms.

† Correlation between predicted and measured DNA methylation levels.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Table 3. Selected * consistent directions of associations across DNA methylation, gene expression and breast cancer risk

CpG	Chr	Position	Gene	Classification	CpG Vs. breast cancer risk		CpG Vs. Gex		Gex Vs. breast cancer risk		Gex Vs. breast cancer risk adjusted§ for DNA methylation	
					Dir	*P* value†	Dir	*P* value‡	Dir	*P* value†	Dir	*P* value†
cg26668989	1	110,186,163	*GSTM4*	Upstream	Positive	1.92×10^{-10}	Negative	5.37×10^{-5}	Negative	2.04×10^{-5}	Negative	0.52
cg08614201	1	145,715,134	*CD160*	5'UTR	Positive	5.00×10^{-7}	Negative	3.26×10^{-51}	Negative	9.35×10^{-4}	Negative	0.93
cg20311333	1	155,197,753	*GBAP1*	TSS1500	Positive	2.54×10^{-11}	Negative	4.55×10^{-12}	Negative	2.22×10^{-9}	Positive	0.86
cg02834765	1	155,214,859	*GBA*	TSS1500	Negative	1.16×10^{-7}	Negative	2.44×10^{-5}	Negative	5.77×10^{-8}	Negative	0.62
cg16030869	4	38,867,304	*FAM114A1*	Upstream	Positive	1.72×10^{-8}	Positive	6.47×10^{-5}	Positive	1.72×10^{-8}	Positive	0.22
cg17942617	5	81,327,376	*ATG10*	Intronic	Negative	6.81×10^{-13}	Positive	9.66×10^{-11}	Negative	6.84×10^{-11}	Negative	0.75
cg16647868	5	131,706,066	*SLC22A5*	Intronic	Positive	7.36×10^{-7}	Negative	9.84×10^{-9}	Negative	1.81×10^{-6}	Negative	0.90
cg12078157	6	13,612,218	*SIRT5*	3'UTR	Negative	1.44×10^{-7}	Negative	1.62×10^{-5}	Positive	1.74×10^{-4}	Positive	0.51
cg05216056	6	28,887,836	*TRIM27*	Exonic	Negative	3.71×10^{-7}	Negative	1.45×10^{-29}	Positive	7.34×10^{-4}	Negative	0.99
cg14701867	10	64,193,068	*ZNF365*	Intronic	Negative	5.92×10^{-10}	Negative	1.15×10^{-6}	Positive	0.02	Positive	0.26
cg23754390	11	835,074	*CD151*	Intronic	Positive	2.51×10^{-7}	Positive	9.14×10^{-22}	Positive	2.03×10^{-3}	Negative	0.81
cg04111478	11	1,991,677	*MRPL23*	Downstream	Positive	2.20×10^{-8}	Positive	0.008	Positive	3.63×10^{-6}	Positive	0.57
cg15531562	11	65,601,754	*SNX32*	Intronic	Positive	1.41×10^{-13}	Positive	0.002	Positive	6.15×10^{-8}	Positive	0.74
cg06065225	11	65,640,137	*EFEMP2*	Intronic	Negative	2.37×10^{-9}	Negative	0.009	Positive	1.93×10^{-11}	Positive	0.42
cg23526087	14	68,973,466	*RAD51B*	Intronic	Negative	4.21×10^{-21}	Negative	5.62×10^{-13}	Positive	4.55×10^{-4}	Positive	1.00
cg25839482	15	75,931,953	*IMP3*	3'UTR	Negative	2.59×10^{-8}	Negative	1.27×10^{-4}	Positive	6.18×10^{-6}	Negative	0.78
cg18878992	17	43,974,344	*MAPT*	TSS1500	Negative	7.07×10^{-10}	Negative	0.003	Positive	8.78×10^{-6}	Positive	0.09
cg21757127	19	18,525,886	*LRRC25*	Downstream	Negative	1.33×10^{-15}	Negative	1.71×10^{-5}	Positive	4.25×10^{-16}	Positive	2.21×10^{-5}
cg09516349	19	18,529,339	*SSBP4*	TSS1500	Positive	4.95×10^{-25}	Positive	9.70×10^{-4}	Positive	2.55×10^{-23}	Positive	0.49
cg22161383	19	18,545,441	*ISYNA1*	3'UTR	Positive	3.85×10^{-27}	Negative	0.003	Negative	9.62×10^{-10}	Negative	0.01
cg14066757	19	44,285,568	*KCNN4*	TSS200	Negative	3.55×10^{-16}	Negative	0.01	Positive	6.12×10^{-15}	Positive	0.23

* Selected from 38 consistent directions of associations across DNA methylation, gene expression and breast cancer risk. Complete list is available in **Supplementary Table 8**. Chr, chromosome; UTR, untranslated region; TSS, transcription start site; Dir, direction of association and/or correlation; Gex, gene expression.

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

‡ *P* values were calculated using Spearman correlation test. All statistical tests were two-sided.

§ Adjusted for all predicting variants included in prediction models of corresponding CpGs.

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Figure Legends

Figure 1. Study design flow chart. FHS = The Framingham Heart Study; WHI = The Women's Health Initiative; BCAC = The Breast Cancer Association Consortium; CpGs = CpG sites; ER = Estrogen receptor; GWAS = Genome-wide association study; meQTL = DNA methylation quantitative trait loci; SNP = single nucleotide polymorphism; GTEx = Genotype-Tissue Expression.

Figure 2. Performances of DNA methylation prediction models in the prediction dataset and the validation dataset. A total of 81,361 models had a prediction performance in FHS (R_{FHS}^2) ≥0.01. This figure shows the performance of these models in the prediction dataset, FHS, and in the validation dataset, WHI. The *X*-axis represents the R_{FHS}^2 (squared correlation coefficients of predicted and measured DNA methylation levels). We then apply these models into the genetic data in WHI to predict the DNA methylation levels of these 81,361 CpGs. The *Y*-axis represents the performance of these models in the WHI (R_{WHI}^2, squared correlation coefficients of predicted and measured DNA methylation levels). The black line represents the identity line ($Y = X$). FHS = The Framingham Heart Study; WHI = The Women's Health Initiative.

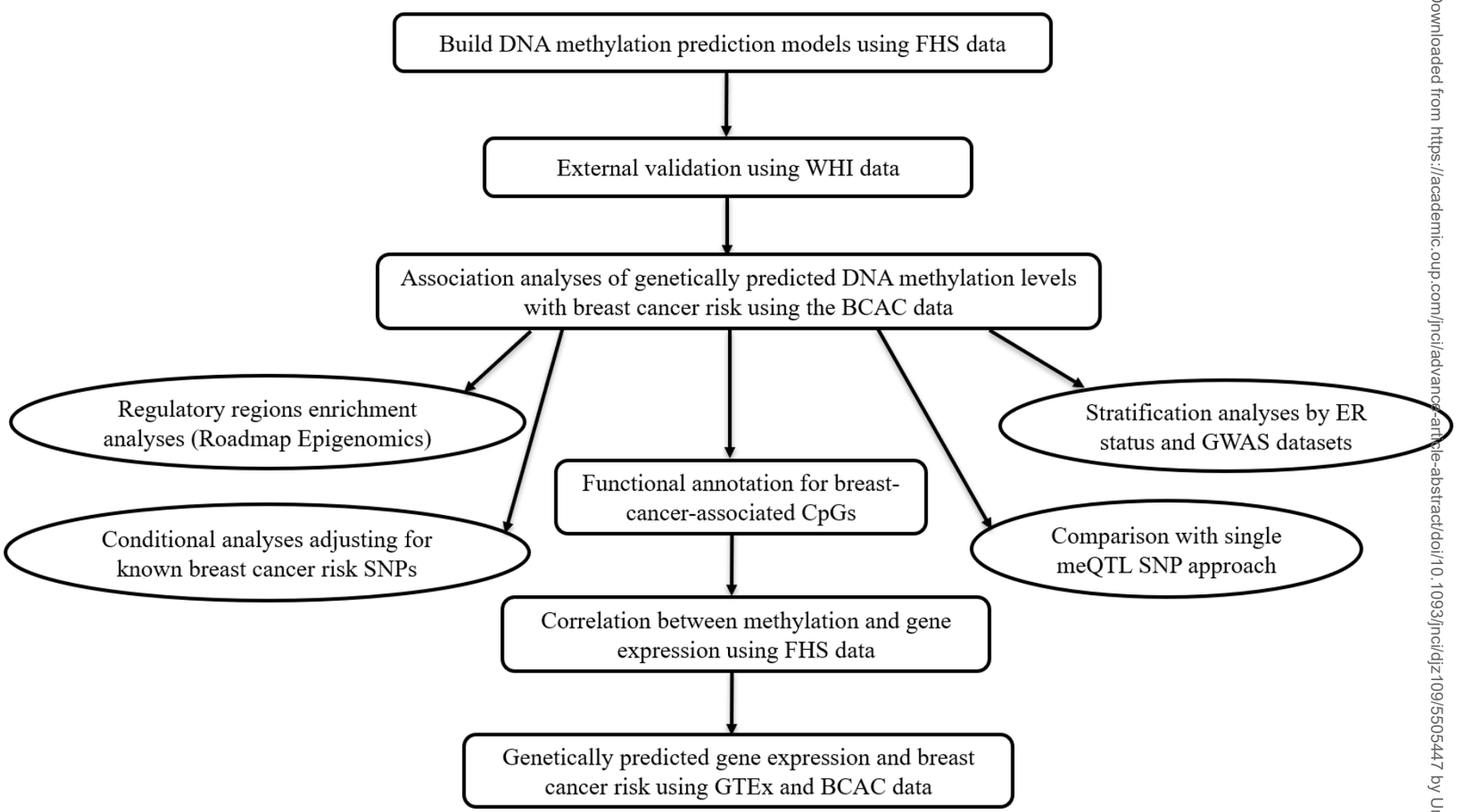
Figure 3. The performance of DNA methylation prediction using the prediction model approach and using the single meQTL SNP approach. For a total of 24,845 CpGs, prediction models were built in the present study and meQTLs were identified in a previous study. This figure presents the prediction performances of models and meQTLs for these CpGs. The *X*-axis represents the performance (R^2) of DNA methylation prediction using the single meQTL SNP

Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

approach, i.e. squared correlation coefficients of predicted and measured DNA methylation levels in the meQTL data. The *Y*-axis represents performance (R^2) of DNA methylation prediction using the prediction model approach, i.e. squared correlation coefficients of predicted and measured DNA methylation levels in the FHS data. The black line represents the identity line ($Y = X$). meQTL = DNA methylation quantitative trait loci; SNP = single nucleotide polymorphism.

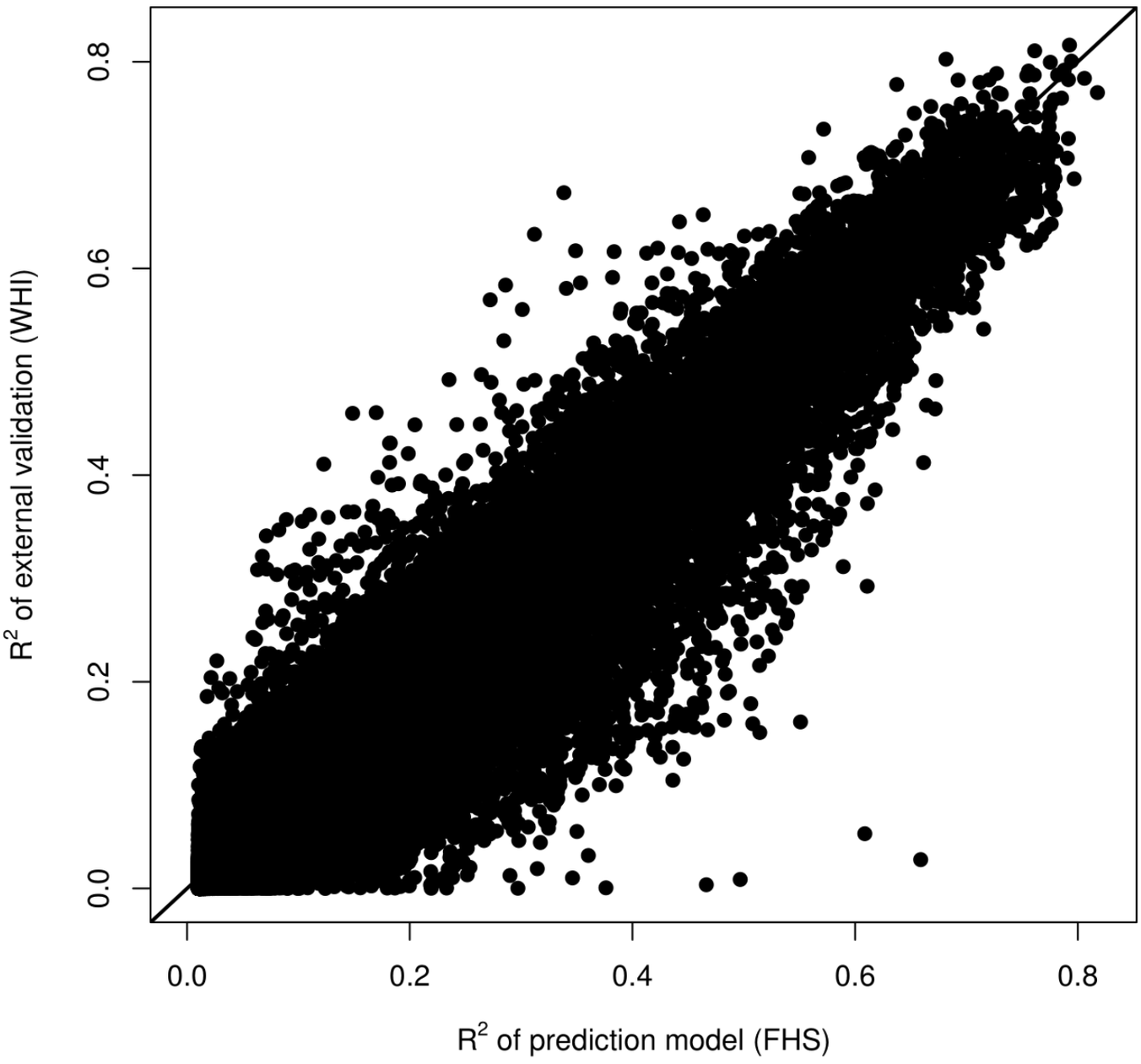
Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Figure 1



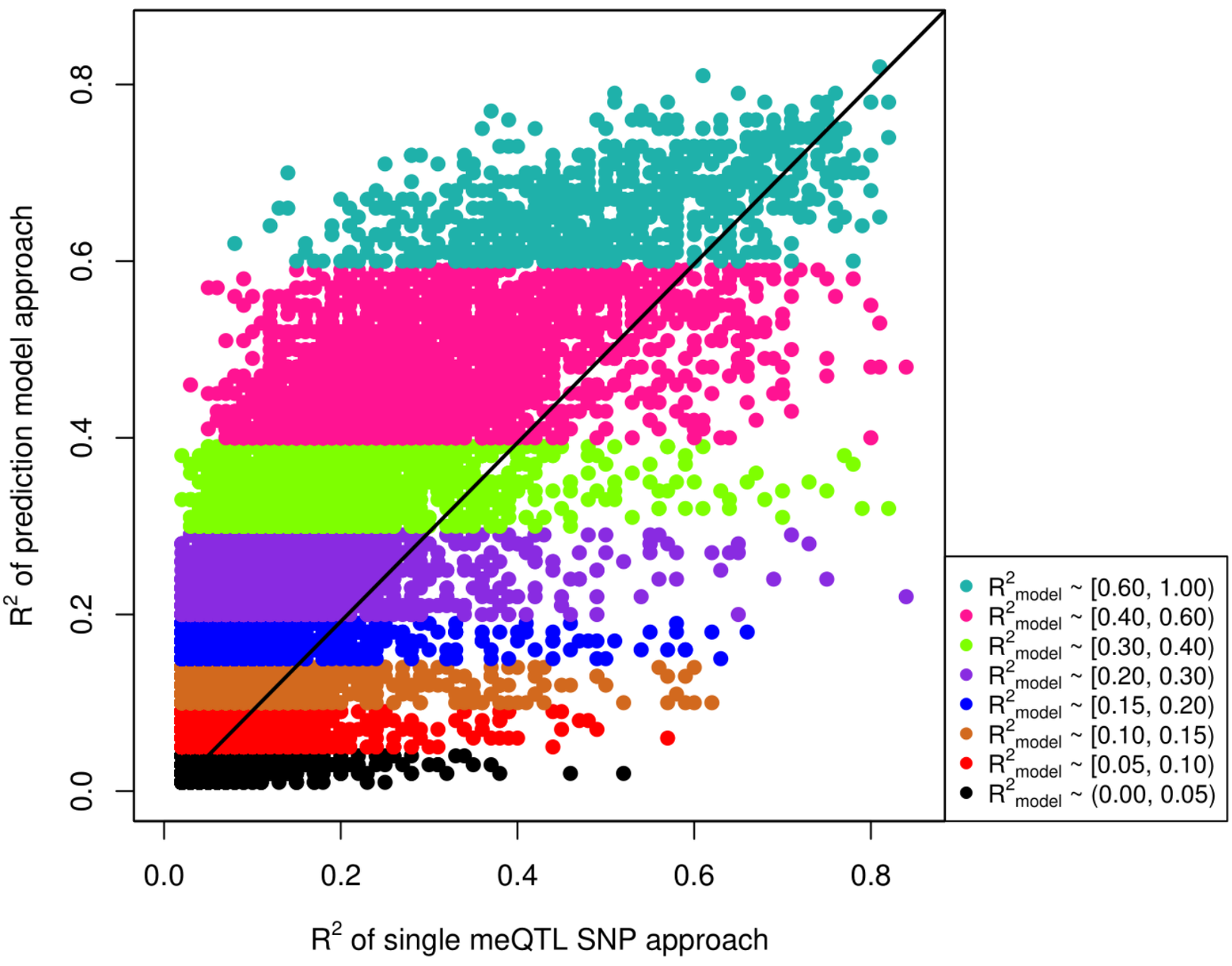
Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Figure 2



Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019

Figure 3



Downloaded from https://academic.oup.com/jnci/advance-article-abstract/doi/10.1093/jnci/djz109/5505447 by University of Cambridge user on 19 June 2019