IJC
International Journal of Cancer

Genetic variants of genes in the NER pathway associated with risk of breast cancer: A large-scale analysis of 14 published GWAS datasets in the DRIVE study

Jie Ge[1,2,3], Hongliang Liu[2,3], Danwen Qian[2,3], Xiaomeng Wang[2,3], Patricia G. Moorman[2,4], Sheng Luo[5], Shelley Hwang[2,6] and Qingyi Wei[2,3,7]

[1]Department of Epidemiology and Statistics, Qiqihar Medical University, Qiqihar, Heilongjiang, China
[2]Duke Cancer Institute, Duke University Medical Center, Durham, NC
[3]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC
[4]Department of Community and Family Medicine, Duke University Medical Center, Durham, NC
[5]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC
[6]Department of Surgery, Duke University School of Medicine, Durham, NC
[7]Department of Medicine, Duke University School of Medicine, Durham, NC

A recent hypothesis-free pathway-level analysis of genome-wide association study (GWAS) datasets suggested that the overall genetic variation measured by single nucleotide polymorphisms (SNPs) in the nucleotide excision repair (NER) pathway genes was associated with breast cancer (BC) risk, but no detailed SNP information was provided. To substantiate this finding, we performed a larger meta-analysis of 14 previously published GWAS datasets in the Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) study with 53,107 subjects of European descent. Using a hypothesis-driven approach, we selected 138 candidate genes from the NER pathway using the "Molecular Signatures Database (MsigDB)" and "PathCards". All SNPs were imputed using IMPUTE2 with the 1000 Genomes Project Phase 3. Logistic regression was used to estimate BC risk, and pooled ORs for each SNP were obtained from the meta-analysis using the false discovery rate for multiple test correction. RegulomeDB, HaploReg, SNPinfo and expression quantitative trait loci (eQTL) analysis were used to assess the SNP functionality. We identified four independent SNPs associated with BC risk, *BIVM-ERCC5* rs1323697_C (OR = 1.06, 95% CI = 1.03–1.10), *GTF2H4* rs1264308_T (OR = 0.93, 95% CI = 0.89–0.97), *COPS2* rs141308737_C deletion (OR = 1.06, 95% CI = 1.03–1.09) and *ELL* rs1469412_C (OR = 0.93, 95% CI = 0.90–0.96). Their combined genetic score was also associated with BC risk (OR = 1.12, 95% CI = 1.08–1.16, $p_{trend} < 0.0001$). The eQTL analysis revealed that *BIVM-ERCC5* rs1323697 C and *ELL* rs1469412 C alleles were correlated with increased mRNA expression levels of their genes in 373 lymphoblastoid cell lines ($p = 0.022$ and 2.67×10^{-22}, respectively). These SNPs might have roles in the BC etiology, likely through modulating their corresponding gene expression.

Key words: breast cancer susceptibility, single nucleotide polymorphism, DNA repair, expression quantitative trait loci analysis

Abbreviations: BC: breast cancer; BREOGAN: Breast Oncology Galicia Network; CGPS: Copenhagen General Population Study; CI: confidence interval; CPSII: Cancer Prevention Study-II Nutrition Cohort; DRIVE: Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; EPIC: European Prospective Investigation Into Cancer and Nutrition; eQTL: expression quantitative trait loci; FDR: false discovery rate; GWAS: genome-wide association study; LD: linkage disequilibrium; MAF: minor allele frequency; MCCS: Melbourne Collaborative Cohort Study; MEC: multiethnic cohort; NBHS: Nashville Breast Health Study; NER: nucleotide excision repair; NHS: Nurses' Health Study; NHS2: Nurses' Health Study 2; NUGs: number of unfavorable genotypes; OR: odds ratio; PBCS: NCI Polish Breast Cancer Study; PCs: principal components; PLCO: The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; SEARCH: Study of Epidemiology and Risk factors in Cancer Heredity; SMC: Swedish Mammography Cohort; SNP: single nucleotide polymorphism; WHI: Women's Health Initiative

Additional Supporting Information may be found in the online version of this article.

Conflict of interest: The authors state no conflict of interest.

Grant sponsor: Qiqihar Medical University Support Grant; **Grant number:** QY2016B-13; **Grant sponsor:** Duke Cancer Institute; **Grant number:** P30 CA014236; **Grant sponsor:** National Institutes of Health; **Grant numbers:** U19 CA148065, X01 HG007491; **Grant sponsor:** Cancer Research UK; **Grant number:** C1287/A16563

DOI: 10.1002/ijc.32371

History: Received 10 Dec 2018; Accepted 27 Mar 2019; Online 26 Apr 2019

Correspondence to: Qingyi Wei, MD, PhD, Duke Cancer Institute, Duke University Medical Center and Department of Population Health Sciences, Duke University School of Medicine, 905 S. LaSalle Street, Durham, NC 27710, USA, Tel.: +1-919-660-0562, E-mail: qingyi.wei@duke.edu

What's new?

Although breast carcinogenesis is still not fully understood, a variety of risk factors have been identified, some of them with mechanisms likely involved in DNA damage and repair. This study identified four novel independent SNPs in the nucleotide excision repair (NER) pathway genes (*BIVM-ERCC5* rs1323697_C, *GTF2H4* rs1264308_T, *COPS2* rs141308737_C deletion and *ELL* rs1469412_C) to be associated with breast cancer risk. *BIVM-ERCC5* rs1323697_C and *ELL* rs1469412_C alleles were correlated with increased mRNA expression. These findings suggest that variants in the NER pathway genes play an important role in the development of breast cancer, possibly by influencing gene expression.

Introduction

Breast cancer (BC) is the most frequently diagnosed cancer and the leading cause of cancer deaths among women worldwide, with an estimated 1.7 million cases and 521,900 deaths in 2012, accounting for 25% of all cancer cases and 15% of all cancer deaths among women.[1] Despite the declining mortality rate due to early screenings and advanced medical therapies, the incidence rate of BC has remained steady over the past two decades in the US (https://seer.cancer.gov/statfacts/html/breast.html). Therefore, it is necessary to identify additional genetic factors that can be used for defining susceptible individuals at risk for BC.

Although the mechanisms of breast carcinogenesis are still not fully understood, a variety of risk factors has already been identified.[2–5] Some studies have shown that mammalian cells can convert estrogen into related compounds that not only generate free radicals capable of damaging DNA but also bind to DNA, causing the loss of a nucleotide base, a process known as depurination. The resulting mutations can convert a normal cell into a cancerous one.[6–8]

Another putative risk factor is smoking. Although there are no consistent results about the association between smoking and BC risk, there are carcinogens in tobacco smoke, such as polycyclic aromatic hydrocarbons (PAH), aromatic amines, and nitrosoamines, and these carcinogens might cause DNA damage and adduct formation in mammary epithelial cells.[9,10]

In addition, many epidemiologic studies reported a positive association between BC risk and alcohol consumption. Animal models of BC, although not entirely consistent, do provide the support for an enhancing action of ethanol on mammary carcinogenesis.[11] Overall, evidence from human studies, animal studies and cell culture experiments support some biologically plausible mechanisms, such as an increase in circulating estrogens and androgens, enhancement of mammary gland susceptibility to carcinogenesis, increased mammary carcinogen-induced DNA damage, and a greater potential for invasiveness of BC cells.[11,12] These mechanisms are all likely involved in DNA damage leading to the initiation of mutations and carcinogenesis, and thus the DNA repair system plays a critical role in protecting against mutations, maintaining genomic integrity and preventing carcinogenesis of the breasts.[13]

One of the DNA repair pathways is nucleotide excision repair (NER), a highly versatile and sophisticated DNA damage removal mechanism that counteracts the deleterious effects of a multitude of DNA lesions, including major types of damage induced by environmental mutagens and carcinogens. The most relevant lesions to be repaired by NER are cyclobutane pyrimidine dimers (CPDs) and 6-4 photoproducts (6-4PPs) produced by the shortwave UV component of sunlight. In addition, numerous bulky chemical adducts are eliminated by this repair process as well.[14,15] Given the importance of NER in the repair of UV-induced DNA damage, it seems that the NER pathway may not be relevant to BC risk, because there is no evidence that UV light may cause BC; however, it is likely that tobacco smoke may cause DNA damage in breast tissues.

A recent large study with pathway-level analysis using hierarchical modeling across five cancers, including 11 Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) genome-wide association study (GWAS) datasets of 33,832 BC study subjects of European descent, did not find any specific risk-associated SNPs in genes involved in NER, but the limited study power did allow the investigators to find an association with the overall genetic variation of the NER pathway.[16] Other prior studies also investigated associations between SNPs in DNA repair pathway genes and BC risk, but these studies had relatively small sample sizes without a focus on the NER pathway, although they have some notable findings, such as *XRCC3* and *ERCC4*.[17–24]

Therefore, we hypothesize that genetic variants in the NER pathway genes are associated with BC risk. To assess the role of functional SNPs of the NER pathway genes in the BC etiology, we performed a much larger meta-analysis of 14 previously published DRIVE GWAS datasets with 53,107 study subjects of European descent. In contrast to the previously published studies, the present analysis had a much larger sample size to focus on functional SNPs in the NER pathway genes. Hence, using a hypothesis-driven pathway-based approach with a much increased study power, we expected to identify some susceptibility loci in the NER pathway genes that have biologically relevant functions and thus play a role in the BC etiology.

Populations and Methods

Study populations

This meta-analysis included a subdataset of SNPs in the NER pathway genes from each of 14 previously published BC GWASs for a total of 28,758 BC cases and 24,349 controls of European ancestry from the DRIVE study (phs001265.v1.p1), which is different from the DRIVE-Genome-Wide Association

meta-analysis (phs001263.v1.p1) previously used by others[16] (Supporting Information Table S1). The DRIVE study (phs001265.v1.p1), which included 17 GWASs, was one of the five projects funded in 2010 as part of the NCI's Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative. For this meta-analysis, we excluded three studies including the "Women of African Ancestry Breast Cancer Study (WAABCS)", which is a study of African ancestry, and "The Sister Study (SISTER)" and "The Two Sister Study (2 SISTER)", which had a different study design that used cases' sisters as the controls. These 14 GWAS studies consist of Breast Oncology Galicia Network (BREOGAN); Copenhagen General Population Study (CGPS); Cancer Prevention Study-II Nutrition Cohort (CPSII); European Prospective Investigation Into Cancer and Nutrition (EPIC); Melbourne Collaborative Cohort Study (MCCS); Multiethnic Cohort (MEC); Nashville Breast Health Study (NBHS); Nurses' Health Study (NHS); Nurses' Health Study 2 (NHS2); NCI Polish Breast Cancer Study (PBCS); The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO); Study of Epidemiology and Risk factors in Cancer Heredity (SEARCH); Swedish Mammography Cohort (SMC); and Women's Health Initiative (WHI). The details of case and control recruitment and their characteristics are summarized in Supporting Information Table S2. For all of the GWAS datasets, Illumina Infinium OncoArray-500k BeadChip genotyping platforms were used, and only two main variables (sex and age at interview) were available to us. For the cases, other three variables (age at diagnosis, estrogen receptor status and histology type) were available. Each of the 14 studies was reviewed and approved by the corresponding Institutional Review Board and thus exempted by Duke Institutional Review Board.

Gene and SNP selection

Candidate genes in the NER pathway were selected according to the online datasets "Molecular Signatures Database v6.1 (MsigDB)" (http://software.broadinstitute.org/gsea/msigdb/search.jsp) and "PathCards" (http://pathcards.genecards.org/) using the keywords "nucleotide excision repair". In total, we selected 138 candidate genes from eight NER-related pathways after excluding duplicate genes, pseudo genes and withdrawn genes (LOC652672 and LOC652857) in the National Center for Biotechnology Information (NCBI). The detailed genes selection results are listed in Supporting Information Table S3.

To avoid poor quality markers to be included in the imputation, we performed stringent quality control before imputation by including the following criteria: the minor allelic frequency (MAF) ≥1%, genotyping rate ≥95%, missing rate ≤90% and Hardy–Weinberg equilibrium (HWE) $\geq 1 \times 10^{-6}$. All SNPs were flipped to forward strand and aligned with the reference genome data. Ambiguous SNPs with A-T or G-C alleles that are hard to determine the strand orientation by allele frequency were removed. According to the multipopulation reference panels from the 1000 Genomes Project Phase 3, SNPs within the aforementioned 138 candidate genes and their ±500 kb flanking regions were also extracted, and imputation for each study was performed using IMPUTE2 software.[25] Imputed SNPs within 2-kb upstream and downstream of each gene's region were extracted for further analysis. After imputation, SNPs that met the following quality control criteria were included in further analysis: imputation SNPs with information score ≥0.80 in IMPUTE2; a minor allele frequency (MAF) ≥5%; and a p value for the Hardy–Weinberg Equilibrium test $\geq 10^{-6}$. Due to differences between the 14 studies, 8,433–9,016 common SNPs remained in each study for further analysis. The final analysis included 7,345 SNPs that were common to all 14 studies.

Functional analysis

To investigate the functions of candidate SNPs, we searched for functional annotation of the SNPs in three online functional prediction websites: RegulomeDB (http://regulomedb.org/), HaploReg (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php) and SNPinfo (https://snpinfo.niehs.nih.gov/snpinfo/snpfunc.html). In addition, we performed the expression quantitative trait loci (eQTL) analysis by using data from multiple sources: lymphoblastoid cell-line data of 373 subjects from the European Variation in Health and Disease Study (GEUVADIS) and the 1000 Genomes Project (phase I integrated release 3, March 2012).[26] Furthermore, we used Genotype-Tissue Expression project (GTEx) results to obtain the corresponding mRNA expression in whole blood and breast tissues (https://gtexportal.org/home/).[27]

Statistical analysis

For each study and the combined dataset, principal components (PCs) were calculated using the Genome-wide Complex Trait Analysis (GCTA) on the LD-pruned subset of the whole-genome-typed dataset.[28] The top 20 PCs were assessed for their associations with BC risk using univariate logistic regression analysis. Those PCs with significant associations in each study were included as covariates in further analyses of associations between SNPs and BC risk. For each SNP, we estimated odds ratios (ORs) and 95% confidence intervals (CIs) by unconditional logistic regression of case/control groups with adjustment for age and PCs. We performed the meta-analysis by using the inverse variance method to combine the results of the 14 studies. We defined heterogeneity as a Cochran's Q test $p \leq 0.10$ or $I^2 > 50.0\%$. We used fixed-effects models, if no heterogeneity existed among the 14 studies, and random-effects models were used, when heterogeneity existed. To assess the robustness of the results, we performed a sensitivity analysis by omitting each study one by one.[29] The false discovery rate (FDR) with a critical cut-off value of 0.05 using the linear step-up method of Benjamini and Hochberg was mainly used to correct for multiple comparisons to reduce the probability of false-positive findings.[30] To observe the combined effect of significant SNPs, we used the number of

unfavorable genotypes (NUGs) of the significant SNPs as a genetic score to assess classification performance of the model. According to the frequency of each group and the effect values, we also dichotomized all the individuals into a low-risk group (0–2 NUGs) and a high-risk group (3–4 NUGs). In the eQTL analysis, we calculated the correlations between SNPs and specific mRNA expression levels by using a general linear regression model. Statistical analyses were performed using PLINK (version 1.9), SAS (version 9.3; SAS Institute, Cary, NC) and R (version 3.0.2). The Manhattan plots and linkage disequilibrium (LD) plots were generated by Haploview v4.2, and regional association plots were constructed by LocusZoom (http://locuszoom.sph.umich.edu/locuszoom/).

Results

General characteristics of the study populations

The overall analysis included 28,758 BC cases and 24,349 controls from 14 studies (Supporting Information Table S2). All subjects were women. The median age of controls was 60 years. The age distribution was statistically different between cases and controls ($p < 0.0001$), with the control group being younger than the case group (≤60 years: 52.47% *vs.* 49.73%). The proportion of estrogen receptor (ER) positive patients was 84%, and the proportion of patients with invasive tumors was 92% after deleting the missing data. The significant PCs among the first 20 in each study (Supporting Information Table S4) were included in the analyses of associations between SNPs and BC risk. Therefore, age and PCs were adjusted for as possible confounders in the multivariate logistic regression analysis.

Association analysis of single locus and BC risk

The workflow of the present study is shown in Figure 1. In total, we used 7,345 SNPs that passed QC in the analysis, including 442 genotyped SNPs and 6,903 imputed SNPs. Multivariate logistic regression and meta-analysis results showed that there were 666 SNPs significantly associated with BC risk ($p < 0.05$), of which 101 SNPs remained significant after FDR correction <0.05. The associations between SNPs of genes involved in the NER pathway and BC risk in the DRIVE study are shown in Figure 2. These 101 top SNPs were mapped to *BIVM-ERCC5*, *GTF2H4, COPS2*, *ELL* and *COPS4* (Supporting Information Table S5). *COPS4* rs75870305 was deleted due to mapping or clustering errors on the NCBI website. Seven tagSNPs remained for additional analysis after removal of SNPs in high pairwise LD (Supporting Information Fig. S1). To search for functional SNPs, we used stringent criteria with

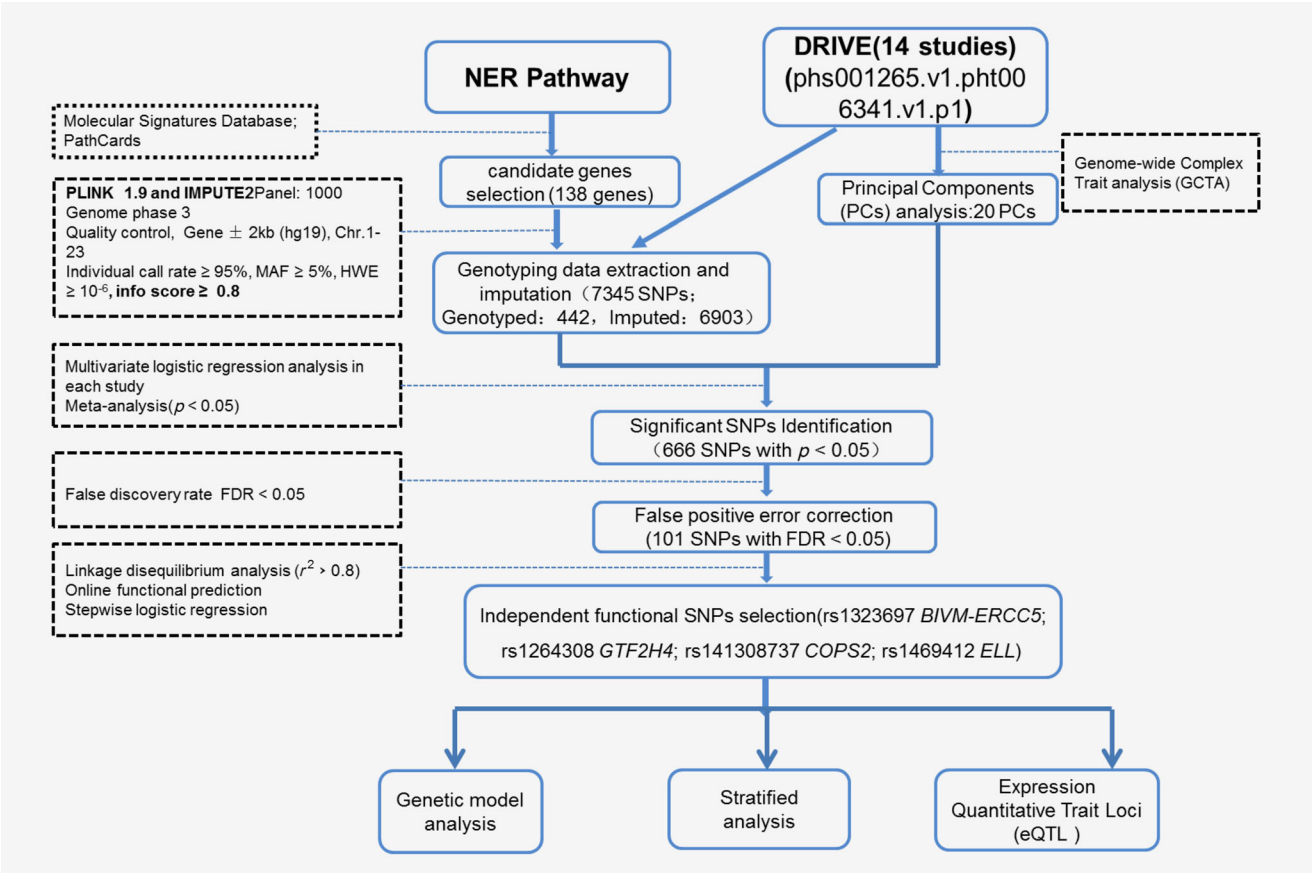

Figure 1. The workflow of present study. [Color figure can be viewed at wileyonlinelibrary.com]

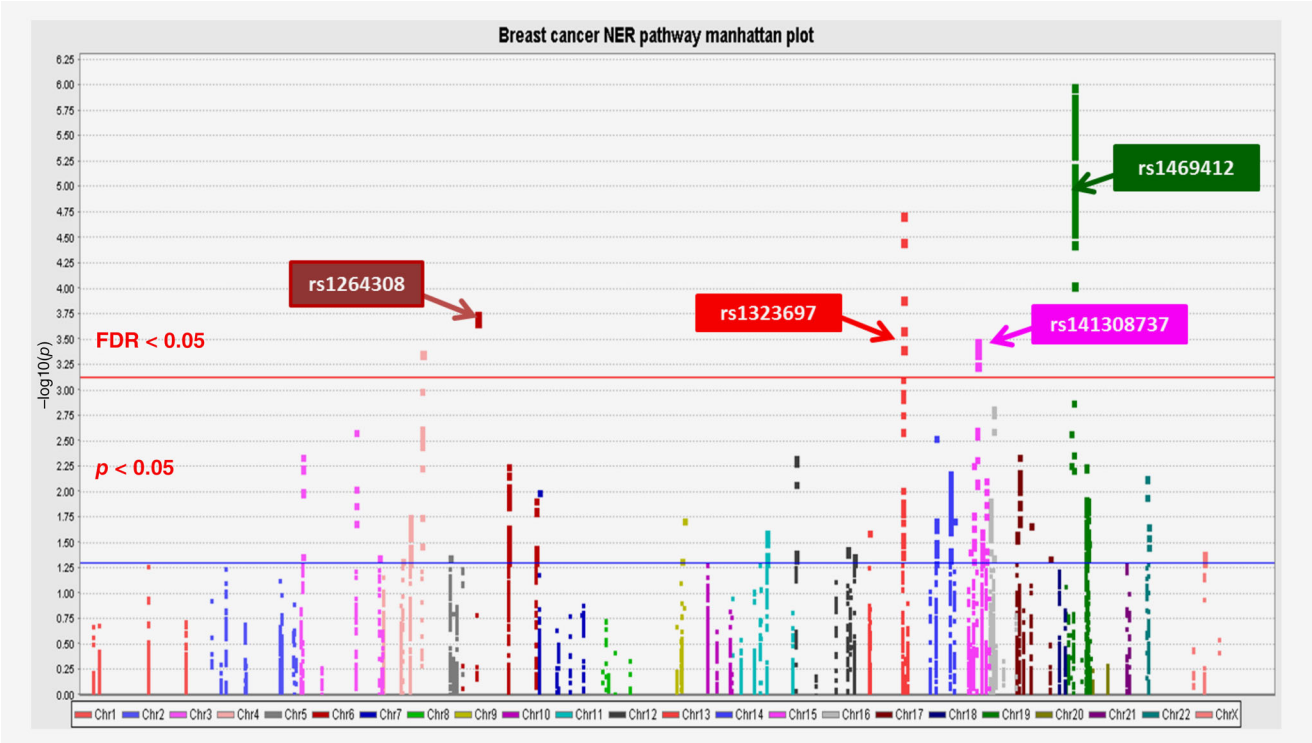

Figure 2. Manhattan Plot of the 7,345 SNPs of NER pathway Genes in the DRIVE. The *x*-axis represents each chromosome. The y-axis represents the association *p* values with breast cancer risk. Horizontal blue line means nominal *p* values of 0.05 and red line means FDR threshold 0.05. [Color figure can be viewed at wileyonlinelibrary.com]

RegulomeDB scores ≤3 and with eQTL evidence in breast tissues or blood cells (Supporting Information Tables S6 and S7). As a result, *BIVM-ERCC5* rs12870541 and *ELL* rs34664859 were left out, because of no functional annotation. Finally, the stepwise analysis kept four independent, potentially functional SNPs for further analysis (Table 1).

Supporting Information Figure S2 presents the forest plots of the meta-analysis of the four independent SNPs. The results showed that SNP rs1323697 G>C and rs141308737 C>deletion were associated with a significantly increased risk of BC (OR = 1.06, 95% CI = 1.03–1.10, $p = 2.66 \times 10^{-4}$; OR = 1.06, 95% CI = 1.03–1.09, $p = 3.60 \times 10^{-4}$, respectively), while two other SNPs were associated with a significantly decreased BC risk (rs1264308 C>T: OR = 0.93, 95% CI = 0.89–0.97, $p = 2.21 \times 10^{-4}$; and rs1469412 T>C: OR = 0.93, 95% CI = 0.90–0.96, $p = 3.09 \times 10^{-6}$). There was no heterogeneity observed for the effect estimates of these four SNPs from the 14 GWASs. Dropping any one of the studies in the DRIVE study, did not change the pooled ORs and their 95% CIs (Supporting Information Table S8). The association results

Table 1. Predictors of risk obtained from stepwise logistic regression analysis of selected variables in the DRIVE study

Variables	Category[1]	Frequency[2]	OR (95% CI)	*p*[3]
Age (in years)		52,994	1.003 (1.001–1.005)	0.0003
BIVM-ERCC5 rs1323697	GG/GC/CC	35,461/15,712/1,821	1.064 (1.031–1.097)	0.0001
GTF2H4 rs1264308	CC/CT/TT	41,872/10,401/721	0.928 (0.893–0.964)	0.0001
COPS2 rs141308737[4]	CC/C–/––	32,971/17,645/2,378	1.062 (1.031–1.094)	‹0.0001
ELL rs1469412	TT/TC/CC	33,899/16,933/2,162	0.934 (0.906–0.962)	‹0.0001

There were 20 PCs in the combined datasets as listed in Supporting Information Table S4, of which seven remained significant and were adjusted in the final stepwise logistic regression analysis.
[1]The most left-hand side "category" was used as the reference.
[2]In total, 52,994 subjects were included in the final stepwise analysis with deletion of the missing data.
[3]Stepwise logistic regression analysis included age, PC1, PC3, PC4, PC5, PC6, PC10, PC16 and 5 SNPs (rs1323697, rs1264308, rs141308737, rs1469412 and rs4808136).
[4]This SNP is a base deletion.
Abbreviations: CI, confidence interval; DRIVE, Discovery, Biology and Risk of Inherited Variants in Breast Cancer; OR, odds ratio.

from different genetic models for each SNP, including additive and dominant models, showed that all of the SNPs were significantly associated with BC risk in all of the genetic models (Table 2). Although the SNP rs4808801 (in the chromosome region 19p13.11 where *ELL* is located) has been previously reported by a GWAS,[31] the *ELL* rs1469412 that we identified in the present study has a moderate LD with rs4808801 (r^2 = 0.471).

Association analysis of the combined score of SNPs and BC risk

The effect size (beta) values of each SNPs are very similar, from 0.056 to 0.074, so we did not consider the weight of each SNP in our analysis of the combined risk genotypes. Using a dominant model, we combined risk genotypes of rs1323697 GC + CC, rs141308737 C–/––, rs1264308 CC and rs1469412 TT into a genetic score as the number of unfavorable genotypes (NUGs). The trend test indicated a significant association between an increased NUGs and an increased risk of BC (p < 0.0001, Table 3). Stratified analyses were performed to assess subgroups defined by age, ER status and invasiveness. We found that the risk associated with NUGs was more evident in the younger group (OR = 1.14, 95% CI = 1.08–1.20, p < 0.0001, Supporting Information Table S9), but no heterogeneity or interaction were observed between these strata (p = 0.227 and 0.274, respectively, Supporting Information - Table S9). Subgroups analysis (ER status and histological type) also showed similar results by age among patients with ER^+ and invasive tumors (Supporting Information Table S9). Additionally, we found no significant differences between ER^+ and ER^- patients (p = 0.990) or between invasive and *in situ* carcinomas (p = 0.945).

Table 2. Associations genotypes of the four independent SNPs and risk of BC in the DRIVE study

	Univariate analysis			Multivariate analysis[1]		
Genotype	$n_{Control}/n_{Case}$	**OR (95% CI)**	p	$n_{Control}/n_{Case}$	**OR (95% CI)**	p
BIVM-ERCC5 rs1323697 G>C[2]						
GG	16,495/19,029	Reference		16,432/19029	Reference	
GC	7,079/8,666	1.06 (1.02–1.10)	0.0020	7,049/8,666	1.05 (1.01–1.09)	0.0071
CC	766/1,057	1.20 (1.09–1.32)	0.0002	764/1,057	1.17 (1.07–1.29)	0.0010
Trend test			<0.0001			<0.0001
GC+CC	7,845/9,723	1.07 (1.04–1.11)	0.0001	7,813/9,723	1.07 (1.03–1.11)	0.0007
GTF2H4 rs1264308 C>T[3]						
CC	19,106/22,861	Reference		19,028/22,861	Reference	
CT	4,880/5,539	0.95 (0.91–0.99)	0.0161	4,863/5,539	0.94 (0.90–0.98)	0.0032
TT	363/358	0.82 (0.71–0.95)	0.0098	363/358	0.80 (0.69–0.93)	0.0029
Trend test			0.0011			<0.0001
CT + TT	5,243/5,897	0.94 (0.90–0.98)	0.0037	5,226/5,897	0.93 (0.89–0.97)	0.0005
CC[4]	19,106/22,861	1.06 (1.02–1.11)	0.0037	19,028/22,861	1.08 (1.03–1.13)	0.0005
COPS2 rs141308737 C>_[5]						
CC	15,327/17,705	Reference		15,278/17,705	Reference	
C–	7,994/9,693	1.05 (1.01–1.09)	0.0096	7,957/9,693	1.05 (1.02–1.09)	0.0064
––	1,028/1,360	1.15 (1.05–1.25)	0.0016	1,019/1,360	1.15 (1.06–1.26)	0.0008
Trend test			0.0002			<0.0001
C–/––	9,022/11,053	1.06 (1.02–1.10)	0.0011	8,976/11,053	1.06 (1.03–1.10)	0.0006
ELL rs1469412 T>C[2,3]						
TT	15,316/18,646	Reference		15,263/18,646	Reference	
TC	8,003/8,973	0.92 (0.89–0.96)	<0.0001	7,965/8,973	0.92 (0.89–0.95)	<0.0001
CC	1,030/1,136	0.91 (0.83–0.99)	0.0259	1,026/1,136	0.91 (0.83–0.99)	0.0328
Trend test			<0.0001			<0.0001
TC + CC	9,033/10,109	0.92 (0.89–0.95)	<0.0001	8,991/10,109	0.92 (0.89–0.95)	<0.0001
TT[4]	15,316/18,646	1.09 (1.05–1.13)	<0.0001	15,263/18,646	1.09 (1.05–1.13)	<0.0001

[1]Adjusted for age, PC1, PC3, PC4, PC5, PC6, PC10 and PC16.
[2]rs1323697 has 9 controls and 6 cases missing; rs1469412 has 3 cases missing.
[3]Risk genotypes were rs1323697 GC + CC, rs1264308 CC, rs141308737 C–/–– and rs1469412 TT.
[4]For consistent with the risk SNPs, we transfer the protected SNPs into risk ones.
[5]This SNP is a base deletion.
Abbreviations: BC, breast cancer; CI, confidence interval; DRIVE, Discovery, Biology and Risk of Inherited Variants in Breast Cancer; NUGs, number of unfavorable genotypes; OR, odd ratio; SNP, single nucleotide polymorphism.

Table 3. Combined risk genotypes of the four validated SNPs and risk of BC in the DRIVE study

Genotype	Univariate analysis			Multivariate analysis[1]		
	$n_{Control}/n_{Case}$	OR (95% CI)	*p*	$n_{Control}/n_{Case}$	OR (95% CI)	*p*
NUG[2]						
0	854/877	Reference		852/877	Reference	
1	5,268/5,769	1.07 (0.96–1.18)	0.2139	5,252/5,769	1.07 (0.97–1.19)	0.1805
2	10,067/11,776	1.14 (1.03–1.26)	0.0091	10,024/11,776	1.15 (1.04–1.27)	0.0057
3	6,734/8,365	1.21 (1.10–1.34)	0.0002	6,705/8,365	1.22 (1.10–1.35)	‹0.0001
4	1,417/1,962	1.35 (1.20–1.52)	‹0.0001	1,412/1,962	1.36 (1.21–1.53)	‹0.0001
Trend			‹0.0001			‹0.0001
0–2	16,189/18,422	Reference		16,128/18,422	Reference	
3–4	8,151/10,327	1.11 (1.07–1.15)	‹0.0001	8,117/10,327	1.12 (1.08–1.16)	‹0.0001

[1]Multivariate logistic regression analyses were adjusted for age and PCs.
[2]Risk genotypes were rs1323697 GC + CC, rs1264308 CC, rs141308737 C–/–– and rs1469412 TT.
Abbreviations: BC, breast cancer; CI, confidence interval; DRIVE, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; NUG, number of unfavorable genotypes; OR, odd ratio; SNPs, single nucleotide polymorphism.

In silico functional validation

The *in silico* eQTL analysis among 373 European descendants with both SNP genotype and mRNA expression data showed that *BIVM-ERCC5* rs1323697 C allele demonstrated a significant association with increased mRNA expression levels of *BIVM* in both additive ($p = 0.022$) and dominant models ($p = 0.025$; Figs. 3*a* and 3*b*). The *ELL* rs1469412 C allele also demonstrated a significant association with increased mRNA expression levels of *ELL* in all genetic models (Figs. 3*f*–3*h*, $p = 2.67E{-}22$, 1.14E−17 and 3.01E−11, respectively). However, no significant associations between the other two SNPs and corresponding mRNA expression levels were found (Figs. 3*d* and 3*e*). In addition, *GTF2H4* rs114596632, the same SNP with rs1264308, has been reported significantly associated with a decreased mRNA expression levels in 270 lymphoblastoid cell lines from HapMap.[32]

To further examine the correlation between the significant SNPs and mRNA expression levels, we searched GTEx as well and found that *BIVM-ERCC5* rs1323697, *GTF2H4* rs1264308 and *ELL* rs1469412 were correlated with their specific mRNA expression levels in the whole blood cells ($p = 0.003$, 0.032

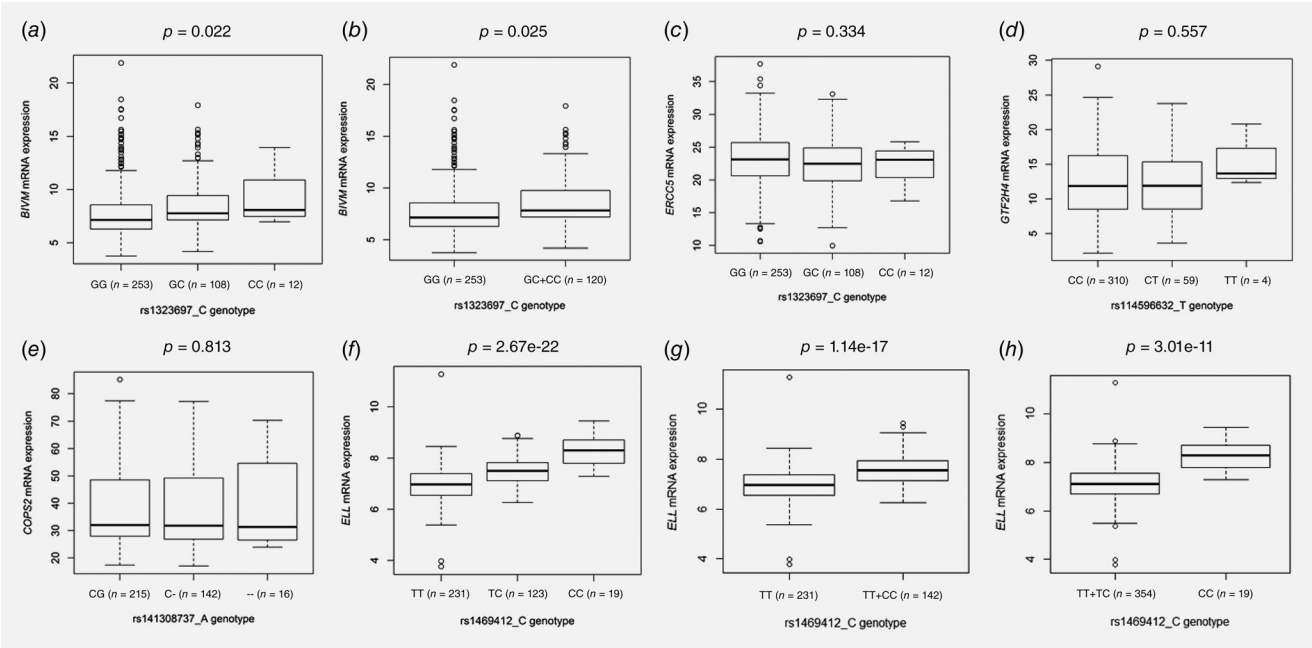

Figure 3. Correlations between identified putative functional SNPs and corresponding gene's mRNA expression in the 1000 Genome Project. rs1323697 ((*a*) additive model, $p = 0.022$; (*b*) dominant model, $p = 0.025$; (*c*) additive model, $p = 0.334$), rs114596632, which was merge into rs1264308 ((*d*) additive model, $p = 0.557$), rs141308737 ((*e*) additive model, $p = 0.813$), rs1469412 ((*f*) additive model, $p = 2.67e{-}22$; (*g*) dominant model, $p = 1.14e{-}17$; (*h*) recessive model, $p = 3.01e{-}11$).

and <0.0001, respectively), but *COPS2* rs141308737 was unrelated to its gene expression levels. In addition, *COPS2* rs141308737 (p = 0.026) and *BIVM-ERCC5* rs1323697 (p = 0.001) had a positive correlation with their gene specific mRNA expression levels in breast tissues (Supporting Information Table S10).

Discussion

To determine whether genetic variants in the NER pathway genes contribute to BC susceptibility, we performed association analyses between 7,345 SNPs in 138 NER genes and BC risk with a large sample size of 28,758 cases and 24,349 controls of European descent. As a result, we identified four novel susceptibility variants, *BIVM-ERCC5* rs1323697 at 13q33.1, *GTF2H4* rs1264308 at 6p21.33, *COPS2* rs141308737 at 15q21.2 and *ELL* rs1469412 at 19p13.11. In addition, the eQTL analysis results revealed that *BIVM-ERCC5* rs1323697 C allele was associated with increased mRNA expression levels, as was the *ELL* rs1469412 C allele. These results indicate that these two SNPs may influence mRNA expression levels and thus the functions of the genes, a possible mechanism underlying the observed associations. These findings suggest that variants in the NER pathway genes play an important role in the development of BC possibly by influencing mRNA expression.

The NER pathway is a mechanism that recognizes and repairs bulky DNA damage caused by chemical compounds, environmental carcinogens and exposure to UV-light. The repair of damaged DNA involves at least 30 polypeptides within two different subpathways of NER known as transcription-coupled repair (TC-NER) and global genome repair (GG-NER).[33] The TCR and GGR processes are different in terms of damage recognition: RNA polymerase II (RNAP II) is needed in TC-NER, while XPC-hHR23B complexes together with XPE complex are needed in GG-NER. In general, genes of GG-NER have been associated with cancer predisposition.[15] However, the present study indicated some genes of TC-NER also might be involved in BC susceptibility, such as *ELL*, *COPS2* and *GTF2H4*, but their exact mechanisms involved in the BC etiology need to be further studied.

BIVM-ERCC5 rs1323697 is located on 13q33.1, which has not been reported by any of the GWASs included in the present analysis. Based on the NCBI website (https://www.ncbi.nlm.nih.gov/gene/100533467), this locus represents naturally occurring read-through transcription between the neighboring basic, immunoglobulin-like variable motif containing (*BIVM*) and excision repair cross-complementing rodent repair deficiency, complementation group 5 (*ERCC5*) genes on chromosome 13. The read-through transcript encodes a fusion protein that shares sequence identity with the products of each individual gene (Supporting Information Fig. S3). Because the present study mainly indicated that rs1323697 was correlated with *BIVM* gene expression levels, the discussion will focus on the function of *BIVM* only. Previous studies have shown that *BIVM* possesses virtually no sequence similar to any currently described protein, making the prediction of a function challenging.[34] It is highly likely that BIVM is essential for some aspect of basic cellular functioning and is expressed in a near-ubiquitous manner.[34] The presence of a CpG island at the 5′-end of *BIVM* and its wide tissue distribution suggest that it may function as a housekeeping gene.[34,35] While others think it is likely that the immunoglobulin-like motif in BIVM may have functions similar to an immunoglobulin, but this remains to be experimentally confirmed.[36] Furthermore, we found that SNP rs1323697 is located at the LUN-1 motif, as shown by the position weight matrix (PWM) based Sequence Logo (Supporting Information Fig. S4 and Table S7).

GTF2H4, known as a general transcription factor IIH subunit 4, encodes a subunit of transcription factor IIH (*TFIIH*), a helicase that is responsible for unwinding DNA structure, allowing repair of the damaged DNA, and it is involved in both NER process and transcription control interacting with variable factors important in carcinogenesis.[37] The TFIIH complex has both ATPase and helicase activities and opens DNA at sites of DNA distorting damage, and the TFIIH4 subunit may regulate the ATPase activity of the TFIIH subunit (XPB, a protein coded by *ERCC3*).[38] Previous studies have found that some *GTF2H4* SNPs were significantly associated with lung cancer risk and survival, multiple sclerosis risk and cervical cancer,[32,39–41] but there is no report on the associations between *GTF2H4* SNPs and BC risk to date. There is an interesting finding that BC and lung cancer risk was associated with the same SNP, *GTF2H4* rs1264308.[32] GWAS catalog results indicated that some of the adjacent genes shared the same location 6p21.33, including *ABCF1*, *PPP1P18* and *LOC105375013*, have a high LD with rs1264308 (Supporting Information Table S11). As an intron SNP, *GTF2H4* rs1264308 may have an effect on the disease by changing motif FOXJ3 (Supporting Information Fig. S4) or by mechanisms of interacting with other above-mentioned genes. However, none of the other genes has a known function in NER.

In the present study, *COPS2* rs141308737 has no functional clues from mRNA expression levels. However, a study showed that overexpression of *COPS2* was linked to chromosome instability.[42] Functional prediction software shows that rs141308737 is located at the ER motif (Supporting Information Fig. S4) and can bind to the CJUN protein (Supporting Information Table S7). It has been reported that endogenous c-Jun plays a key role in ErbB2-induced migration and invasion of mammary epithelial cells and mediates the expansion of a self-renewing population of mammary tumor stem cells *via* the production of CCL5 and SCF to enhance BC tumor invasiveness.[43]

As for *ELL* rs1469412, although some SNPs in this region have been reported by GWASs, it is necessary to include this SNP, because it was only in moderate-to-low LD with other reported SNPs (Supporting Information Table S11) and lack of functional analysis in the previously published study. *ELL* is known as an elongation factor for RNA polymerase II, which is an important gene in the TC-NER subpathway. One

study reported that ELL encoded an elongation factor that could increase the catalytic rate of RNA polymerase II transcription by suppressing transient pausing by polymerase at multiple sites along the DNA.[44] Another study showed that ELL was a key regulator of transcriptional elongation, suggesting that, as an E3 ubiquitin ligase for c-Myc and a potential tumor suppressor, ELL may function as a partner of steroid receptors, hypoxia-inducible factor 1-alpha (HIF-1α), E2F1 and the TFIIH complex, modulating their binding partner's activity.[45] The present study showed that the *ELL* rs1469412 C allele was associated with an increase in mRNA expression levels, exerting a protective effect on BC risk. However, further studies are needed to investigate biological mechanisms underlying the observed associations between *ELL* rs1469412 and BC risk.

It should also be mentioned that the present study has some limitations. First, due to the limited access to phenotypes of the published GWAS datasets with many PCs included in the analysis, we could not adjust for some known risk factors, such as smoking, menstrual, reproductive and lactational history,[46] and the findings need to be verified in other BC studies with more detailed information about the known risk factors. Second, we did not have access to the target tissues collected by the participating GWAS studies, and we only did *in silico* analysis using published data for the functional prediction of the identified SNPs. Therefore, the biological mechanisms by which the four identified SNPs may influence BC risk remain unclear. Third, the study populations were of non-Hispanic whites, and thus the findings may not generalizable to other ethnic groups, and thus additional studies in other ethnic groups are warranted.

In conclusion, this large-scale meta-analysis of 14 published GWASs among 53,107 subjects of European descent identified four novel BC susceptibility loci in the NER pathway genes (i.e., *BIVM-ERCC5* rs1323697, *GTF2H4* rs1264308, *COPS2* rs141308737 and *ELL* rs1469412) and also provided some evidence for their functional relevance. Further studies on the exact biological mechanisms and functional analysis of these SNPs in the BC etiology are needed.

Acknowledgements

DRIVE: OncoArray genotyping and phenotype data harmonization for the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE; dbGaP#: phs001265.v1.p1) breast-cancer case control samples was supported by X01 HG007491 and U19 CA148065 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Centre for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and the National Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls: the Two Sister Study (2SISTER), Breast Oncology Galicia Network (BREOGAN), Copenhagen General Population Study (CGPS), Cancer Prevention Study 2 (CPSII), The European Prospective Investigation into Cancer and Nutrition (EPIC), Melbourne Collaborative Cohort Study (MCCS), Multiethnic Cohort (MEC), Nashville Breast Health Study (NBHS), Nurses' Health Study (NHS), Nurses' Health Study 2 (NHS2), Polish Breast Cancer Study (PBCS), Prostate Lung Colorectal and Ovarian Cancer Screening Trial (PLCO), Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH), The Sister Study (SISTER), Swedish Mammographic Cohort (SMC), Women of African Ancestry Breast Cancer Study (WAABCS), Women's Health Initiative (WHI). Jie Ge was supported by the Qiqihar Medical University's Support Grant (QY2016B-13). Qingyi Wei was in part supported by the Duke Cancer Institute's P30 Cancer Center Support Grant (NIH CA014236).

References

1. Torre LA, Bray F, Siegel RL, et al. Global cancer statistics, 2012. *CA Cancer J Clin* 2015;65:87–108.
2. Maas P, Barrdahl M, Joshi AD, et al. Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol* 2016;2:1295–302.
3. Bjerkaas E, Parajuli R, Weiderpass E, et al. Smoking duration before first childbirth: an emerging risk factor for breast cancer? Results from 302,865 Norwegian women. *Cancer Causes Control* 2013;24:1347–56.
4. Sun YS, Zhao Z, Yang ZN, et al. Risk factors and preventions of breast cancer. *Int J Biol Sci* 2017; 13:1387–97.
5. Knight JA, Fan J, Malone KE, et al. Alcohol consumption and cigarette smoking in combination: a predictor of contralateral breast cancer risk in the WECARE study. *Int J Cancer* 2017;141: 916–24.
6. Miller K. Estrogen and DNA damage: the silent source of breast cancer? *J Natl Cancer Inst* 2003; 95:100–2.
7. Ahmed S, Thomas G, Ghoussaini M, et al. Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat Genet* 2009;41:585–90.
8. Yager JD, Davidson NE. Estrogen carcinogenesis in breast cancer. *N Engl J Med* 2006;354:270–82.
9. Kawai M, Malone KE, Tang MT, et al. Active smoking and the risk of estrogen receptor-positive and triple-negative breast cancer among women ages 20 to 44 years. *Cancer* 2014;120:1026–34.
10. Catsburg C, Kirsh VA, Soskolne CL, et al. Active cigarette smoking and the risk of breast cancer: a cohort study. *Cancer Epidemiol* 2014;38:376–81.
11. Singletary KW, Gapstur SM. Alcohol and breast cancer: review of epidemiologic and experimental evidence and potential mechanisms. *JAMA* 2001; 286:2143–51.
12. Ellingjord-Dale M, Vos L, Hjerkind KV, et al. Dos-Santos-Silva I, Ursin G. alcohol, physical activity, smoking, and breast cancer subtypes in a large, nested case-control study from the Norwegian breast cancer screening program. *Cancer Epidemiol Biomarkers Prev* 2017;26:1736–44.
13. Kennedy DO, Agrawal M, Shen J, et al. DNA repair capacity of lymphoblastoid cell lines from sisters discordant for breast cancer. *J Natl Cancer Inst* 2005;97:127–32.
14. de Laat WL, Jaspers NG, Hoeijmakers JH. Molecular mechanism of nucleotide excision repair. *Genes Dev* 1999;13:768–85.
15. Marteijn JA, Lans H, Vermeulen W, et al. Understanding nucleotide excision repair and its roles in cancer and ageing. *Nat Rev Mol Cell Biol* 2014;15: 465–81.
16. Scarbrough PM, Weber RP, Iversen ES, et al. A cross-cancer genetic association analysis of the DNA repair and DNA damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. *Cancer Epidemiol Biomarkers Prev* 2016; 25:193–200.
17. Tang LL, Chen FY, Wang H, et al. Haplotype analysis of eight genes of the monoubiquitinated FANCD2-DNA damage-repair pathway in breast cancer patients. *Cancer Epidemiol* 2013;37: 311–7.
18. Sapkota Y, Mackey JR, Lai R, et al. Assessing SNP-SNP interactions among DNA repair, modification and metabolism related pathway genes in breast cancer susceptibility. *PLoS One* 2014;8: e64896.
19. Silva SN, Tomar M, Paulo C, et al. Breast cancer risk and common single nucleotide polymorphisms in homologous recombination DNA repair pathway genes XRCC2, XRCC3, NBS1 and RAD51. *Cancer Epidemiol* 2010;34:85–92.
20. Sehl ME, Langer LR, Papp JC, et al. Associations between single nucleotide polymorphisms in double-stranded DNA repair pathway genes and familial breast cancer. *Clin Cancer Res* 2009;15: 2192–203.
21. Monsees GM, Kraft P, Chanock SJ, et al. Comprehensive screen of genetic variation in DNA repair pathway genes and postmenopausal breast cancer risk. *Breast Cancer Res Treat* 2011; 125:207–14.

22. Liu C, Srihari S, Lal S, et al. Personalised pathway analysis reveals association between DNA repair pathway dysregulation and chromosomal instability in sporadic breast cancer. *Mol Oncol* 2016;10:179–93.
23. Han J, Haiman C, Niu T, et al. Genetic variation in DNA repair pathway genes and premenopausal breast cancer risk. *Breast Cancer Res Treat* 2009; 115:613–22.
24. Haiman CA, Hsu C, de Bakker PI, et al. Comprehensive association testing of common genetic variation in DNA repair pathway genes in relationship with breast cancer risk in multiple populations. *Hum Mol Genet* 2008;17:825–34.
25. Howie B, Marchini J, Stephens M. Genotype imputation with thousands of genomes. *G3* 2011; 1:457–70.
26. Liu H, Liu Z, Wang Y, et al. Functional variants in DCAF4 associated with lung cancer risk in European populations. *Carcinogenesis* 2017;38: 541–51.
27. Gibson G. Human genetics. GTEx detects genetic effects. *Science* 2015;348:640–1.
28. Yang J, Lee SH, Goddard ME, et al. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. *Methods Mol Biol* 2013;1019:215–36.
29. Teasdale N, Elhussein A, Butcher F, et al. Systematic review and meta-analysis of remotely delivered interventions using self-monitoring or tailored feedback to change dietary behavior. *Am J Clin Nutr* 2018;107:247–56.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc B Methodol* 1995; 57:289–300.
31. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 2013;45: 353–61.
32. Wang M, Liu H, Liu Z, et al. Genetic variant in DNA repair gene GTF2H4 is associated with lung cancer risk: a large-scale analysis of six published GWAS datasets in the TRICL consortium. *Carcinogenesis* 2016;37:888–96.
33. Nadkarni A, Burns JA, Gandolfi A, et al. Nucleotide excision repair and transcription-coupled DNA repair abrogate the impact of DNA damage on transcription. *J Biol Chem* 2016;291:848–61.
34. Yoder JA, Hawke NA, Eason DD, et al. BIVM, a novel gene widely distributed among deuterostomes, shares a core sequence with an unusual gene in *Giardia lamblia*. *Genomics* 2002;79:750–5.
35. Mitra PS, Ghosh S, Zang S, et al. Analysis of the toxicogenomic effects of exposure to persistent organic pollutants (POPs) in Slovakian girls: correlations between gene expression and disease risk. *Environ Int* 2012;39:188–99.
36. Ferraren DO, Liu C, Badner JA, et al. Linkage disequilibrium analysis in the LOC93081-KDELC1-BIVM region on 13q in bipolar disorder. *Am J Med Genet B Neuropsychiatr Genet* 2005;133B:12–7.
37. Gervais V, Lamour V, Jawhari A, et al. TFIIH contains a PH domain involved in DNA nucleotide excision repair. *Nat Struct Mol Biol* 2004;11: 616–22.
38. Oksenych V, Coin F. The long unwinding road: XPB and XPD helicases in damaged DNA opening. *Cell Cycle* 2010;9:90–6.
39. Buch SC, Diergaarde B, Nukui T, et al. Genetic variability in DNA repair and cell cycle control pathway genes and risk of smoking-related lung cancer. *Mol Carcinog* 2012;51(Suppl 1):E11–20.
40. Briggs FB, Goldstein BA, McCauley JL, et al. Variation within DNA repair pathway genes and risk of multiple sclerosis. *Am J Epidemiol* 2010;172: 217–24.
41. Wang SS, Gonzalez P, Yu K, et al. Common genetic variants and risk for HPV persistence and progression to cervical cancer. *PLoS One* 2010;5:e8667.
42. Wicker CA, Izumi T. Analysis of RNA expression of normal and cancer tissues reveals high correlation of COP9 gene expression with respiratory chain complex components. *BMC Genomics* 2016;17:983.
43. Jiao X, Katiyar S, Willmarth NE, et al. C-Jun induces mammary epithelial cellular invasion and breast cancer stem cell expansion. *J Biol Chem* 2010;285:8218–26.
44. Shilatifard A, Lane WS, Jackson KW, et al. An RNA polymerase II elongation factor encoded by the human ELL gene. *Science* 1996;271:1873–6.
45. Chen Y, Zhou C, Ji W, et al. ELL targets c-Myc for proteasomal degradation and suppresses tumour growth. *Nat Commun* 2016;7:11057.
46. Persson I. Estrogens in the causation of breast, endometrial and ovarian cancers—evidence and hypotheses from epidemiological findings. *J Steroid Biochem Mol Biol* 2000;74:357–64.

gibco

24 days of stem cells

Shape the future of stem cell innovation

October 1- November 1, 2019

Join us for 24 Days of Stem Cells; a premiere virtual event featuring the latest advances in stem cell research.

This year's format will feature a new hour of cutting edge content every week day starting October 1st. Attend the sessions that are most relevant to your work - at your convenience and at your pace.

During the 24-day long event, you can:

- Access leading scientific presentations from thought leaders around the world
- Watch live training demonstrations from our stem cell experts
- Download key stem cell tools and resources
- Complete weekly challenges to earn points towards certification and prizes

Register today at
www.24daysofstemcells.com

ThermoFisher SCIENTIFIC

WILEY