Am J Cancer Res 2020;10(7):2160-2173
www.ajcr.us /ISSN:2156-6976/ajcr0116772

Original Article

Variants in *SNAI1*, *AMDHD1* and *CUBN* in vitamin D pathway genes are associated with breast cancer risk: a large-scale analysis of 14 GWASs in the DRIVE study

Haijiao Wang[1,2,3*], Lingling Zhao[2,3,4*], Hongliang Liu[2,3], Sheng Luo[5], Tomi Akinyemiju[3], Shelley Hwang[6], Qingyi Wei[2,3,7]

[1]Department of Gynecology Oncology, The First Hospital of Jilin University, Changchun 130021, Jilin, China; [2]Duke Cancer Institute, Duke University Medical Center, Durham 27710, NC, USA; [3]Department of Population Health Sciences, Duke University School of Medicine, Durham 27710, NC, USA; [4]Cancer Center, The First Hospital of Jilin University, Changchun 130021, Jilin, China; [5]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham 27710, NC, USA; [6]Department of Surgery, Duke University School of Medicine, Durham 27710, NC, USA; [7]Department of Medicine, Duke University School of Medicine, Durham 27710, NC, USA. [*]Equal contributors.

Received June 22, 2020; Accepted June 30, 2020; Epub July 1, 2020; Published July 15, 2020

Abstract: Vitamin D has a potential anticarcinogenic role, possibly through regulation of cell proliferation and differentiation, stimulation of apoptosis, immune modulation and regulation of estrogen receptor levels. Because breast cancer (BC) risk varies among individuals exposed to similar risk factors, we hypothesize that genetic variants in the vitamin D pathway genes are associated with BC risk. To test this hypothesis, we performed a larger meta-analysis using 14 published GWAS datasets in the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) Study. We assessed associations between 2,994 (237 genotyped in the DRIVE study and 2,757 imputed from the 1000 Genomes Project) single nucleotide polymorphisms (SNPs) in 33 vitamin D pathway genes and BC risk. In unconditional logistic regression analysis, we found 11 noteworthy SNPs to be associated with BC risk after multiple comparison correction by the Bayesian false-discovery probability method (<0.80). In stepwise logistic regression analysis, with adjustment for age, principal components and previously published SNPs in the same study populations, we identified three independent SNPs (*SNAI1* rs1047920 C>T, *AMDHD1* rs11826 C>T and *CUBN* rs3914238 C>T) to be associated with BC risk (P = 0.0014, 0.0020 and 0.0022, respectively). Additional expression quantitative trait loci analysis revealed that the rs73276407 A allele, in a high LD with the rs1047920 T allele, was associated with decreased *SNAI1* mRNA expression levels, while the rs11826 T allele was significantly associated with elevated *AMDHD1* mRNA expression levels. Once replicated by other investigators and additional mechanistic studies, these genetic variants may serve as new biomarkers for susceptibility to BC.

Keywords: Breast cancer susceptibility, single nucleotide polymorphism, Vitamin D, expression quantitative trait loci analysis

Introduction

Breast cancer (BC) is the most common malignancy and the second leading cause of cancer deaths in women in the United States [1]. In 2019, there were about 268,600 cases estimated to be diagnosed with BC in the United States, accounting for 15.2% of all new cancer cases, while an estimated 41,760 individuals died from BC, accounting for 6.9% of all cancer deaths [2]. Although the mortality of BC has declined due to early detection and advanced treatments, the incidence rate is still rising each year (https://seer.cancer.gov/statfacts/html/breast.html). Therefore, it is urgent to search for additional risk factors to identify individuals who are at high risk of BC for early screening and prevention that will reduce the incidence and mortality of BC.

There are several risk factors for developing female BC, including aging, unhealthy lifestyle, estrogen status, germ-line mutations and family history [3]. Genetic variation, such as single

nucleotide polymorphisms (SNPs) as reported by genome-wide association studies (GWASs) is believed to contribute to BC risk. However, the risk contribution of individual SNPs is considered small [4] as identified in the hypothesis-free GWASs, which have always focused on the most important SNPs with a rigorous *P* value after the most stringent multiple test correction. Furthermore, few of the GWAS-identified SNPs are found to be functionally annotated. In the post-GWAS era, hypothesis-driven and combined analyses of all published GWASs are performed to identify cancer-risk associated functional SNPs in a combination of pathway analysis, meta-analysis, and functional analysis. With such a hypothesis-driven approach, investigators can focus on SNPs with potential biological functions by using available genotyping data from previously published GWAS datasets with the hope to be able to identify truly cancer-risk associated functional variants.

Vitamin D is a fat-soluble steroid hormone obtained from both dietary sources and sun exposure to ultraviolet B radiation, and once present in the body, it regulates the expression of genes in many types of tissue [5-7]. In addition, vitamin D may have a potential anticarcinogenic role, by regulating cell proliferation and differentiation, apoptosis, immune modulation and estrogen receptor levels [8, 9]. Therefore, the vitamin D pathway plays a role in regulating cell growth and immune function, relevant to tumor progression. For example, studies have demonstrated that the vitamin D pathway has an effect on T cell function, monocyte/macrophage differentiation and cytokine production [10-12]. Other studies have found that vitamin D may affect the pathogenesis, prognosis and survival of BC at the cellular level [13, 14]. Many epidemiological studies have also attempted to determine associations of vitamin D levels with risk and mortality of various types of cancer [15-17]. In a Brazilian study of postmenopausal BC patients, low vitamin D levels were found to be a risk factor for individuals negative for estrogen receptor with a higher rate of cell proliferation and positive axillary lymph node [18]. However, few studies have comprehensively investigated the effect of genetic variation in vitamin D pathway genes on BC risk.

Considering the importance of the vitamin D pathway in cancer development, we hypothesize that genetic variants in vitamin D pathway genes are associated with BC risk, and we tested this hypothesis in a larger meta-analysis of 53,107 BC case-control study subjects with genotyping datasets from 14 previously published GWAS datasets in the DRIVE study.

Materials and methods

Study subjects

The subjects in this case-control meta-analysis were from 14 out of 17 previously published BC GWASs from the DRIVE study (phs001265.v1.p1), which is different from previously used by others named the DRIVE-Genome-Wide Association meta-analysis (phs001263.v1.p1); and the details of the specific differences between the two studies have been previously described [19]. The DRIVE study we used was one of five projects funded by the NCI's Genetic Associations and Mechanisms in Oncology (GAME-ON) in 2010. We removed three studies out of the 17 GWASs: one was "Women of African Ancestry Breast Cancer Study (WAABCS)", because it was on African ancestry study with a relatively small study population, and the other two were "The Sister Study (SISTER)" and "the Two Sister Study (2 SISTER)", because they had different research designs from others and used cases' sisters as controls. As a result, all the subjects of European ancestry in 14 GWAS studies were included in the final analysis, including 28,758 BC cases and 24,349 controls, whose characteristics are presented in Supplementary Table 1.

These 14 GWASs included 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). Illumina Infinium OncoArray-500k BeadChip genotyping platforms were used for these GWAS datasets, and information on both sex and age at interview was obtained for all the subjects. For the cases, there were three other variables

including age at diagnosis, estrogen receptor status, and tumor histology type; for age variables, we used the age at diagnosis for the cases and the age at interview for the controls. A written informed consent was obtained from subjects by each of the original studies that were approved by the Institutional Review Boards of the Participating institutions.

Identification of vitamin D pathway genes and their SNP extraction

We searched for candidate genes in the vitamin D pathway according to the online datasets "PathCards" (http://pathcards.genecards.org/) and the databases of KEGG, BIOCARTA, REACTOME, Canonical pathways and Gene Ontology (GO) in the "Molecular Signatures Database v7.0 (MsigDB)" (http://software.broadinstitute.org/gsea/msigdb/search.jsp) used by the keyword "vitamin D", and previously published studies (GWASs and vitamin D levels) [20]. In total, we identified 33 genes as candidate genes after excluding some duplicate genes (Supplementary Table 2).

We performed quality control before imputation with the following stringent criteria: (1) the minor allelic frequency (MAF) ≥1%, (2) genotyping success rate ≥95%, (3) missing rate ≤10% and (4) Hardy-Weinberg equilibrium (HWE) $P \geq 1 \times 10^{-6}$. SNPs located in the 33 candidate genes and their ± 500 kb flanking regions were extracted for imputation with the reference panel from the 1000 Genomes Project data (phase 3) by using IMPUTE2 software [21]. After quality control, SNPs within 2-kb up- and down-stream of genes in the vitamin D pathway were extracted for further analysis, and SNPs for the final meta-analysis with the following criteria: (1) Imputed SNPs with an information score ≥0.80; (2) MAF ≥5%; (3) a genotyping call rate ≥95%; and (4) HWE $P \geq 10^{-6}$.

Statistical analysis

Principal components (PCs) analysis was performed for each study separately and their combined dataset by using Genome-wide Complex Trait Analysis (GCTA) [22]. Through univariate logistic regression analysis, the associations between the top 20 PCs and BC risk were evaluated. In further SNP association analysis, significant PCs together with age as covariates were adjusted for in the final models. For each SNP, we calculated odds ratios (ORs) and 95% confidence intervals (CIs) with adjustment for covariates (age and PCs with significant associations) by unconditional logistic regression. We also adjusted for the four SNPs we previously published from the DRIVE Study to identify additional independent SNPs. A meta-analysis was further performed using the results of an additive model of the 14 studies with the inverse variance method. If the Cochran's Q test *P*-value >0.1 and $I^2 \leq 50\%$, a fixed-effects model was used; otherwise a random-effects model was employed. The results were first corrected for multiple comparisons by false discovery rate (FDR). Because many SNPs under investigation were in high LD as a result of imputations, the Bayesian false-discovery probability (BFDP) approach was also used in substitution for the false discovery rate, with a cut-off value of 0.8 was used to reduce the probability of false-positive findings for multiple comparisons, as recommended [23]. We applied a prior probability of 0.1 to detect an upper bound of 3.0 for an association with variant genotypes or minor alleles of the SNPs. The number of risk genotypes (NRGs) derived from the independent SNPs as a genetic score was subsequently used to assess the joint effects of these SNPs. Additionally, Haploview v4.2 [24] was employed to construct a Manhattan plot, and LocusZoom [25] was applied to produce regional association plots for significant SNPs. Other statistical analyses, if not specifically mentioned otherwise, were performed with SAS software Version 9.4 (SAS Institute, Cary, NC), R (version 3.5.0) and PLINK (version 1.90).

Functional analysis

Potential functions of independent SNPs were predicted by RegulomeDB (http://www.regulomedb.org/) and HaploReg [26] (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php). In addition, the expression quantitative trait locus (eQTL) analysis was performed to assess the associations between genotypes of independent SNPs and mRNA expression levels of their corresponding genes by using the genotyping and expression data from the lymphoblastoid cell lines of 373 European descendants available in the 1000 Genomes Project [27]. We also used data the Genotype-Tissue Expression (GTEx) Project to

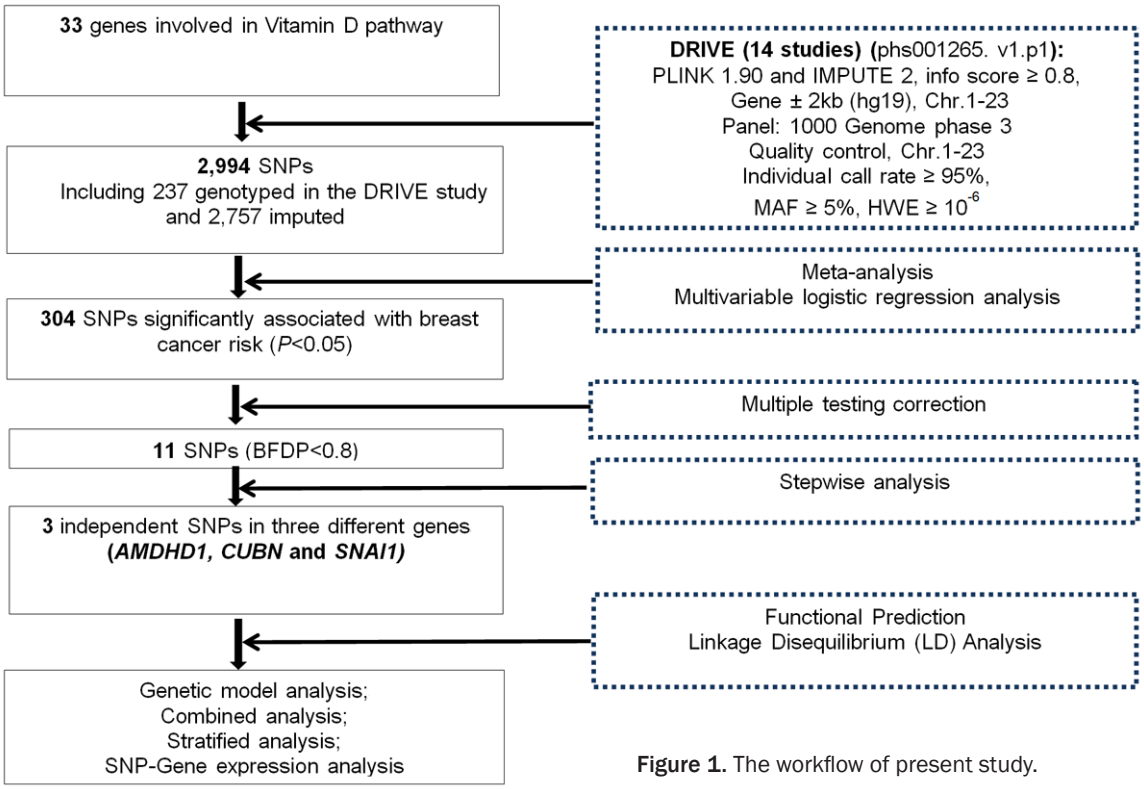


Figure 1. The workflow of present study.

assess the correlations between identified SNPs and their corresponding mRNA expression levels in whole blood and breast tissues (https://gtexportal.org/home/) [28].

Results

Single locus analysis

The results of the top 20 PCs are shown in Supplementary Table 3. The workflow of the present analysis is showed in **Figure 1**. In total, 2,994 SNPs, including 237 genotyped SNPs and 2,757 imputed SNPs, were included in the meta-analysis of the 14 studies, and the distribution of information score for these SNPs is showed in Supplementary Figure 3, of which 304 SNPs were significantly associated with BC risk at $P<0.05$ in an additive model, and 11 SNPs remained significant after multiple comparison corrections with BFDP <0.8 (**Figure 2**). To further identify independent SNPs associated with BC risk, stepwise logistic regression analyses were performed to assess the effects of 11 significant SNPs with adjustment for age, significant PCs and four previously published risk-associated SNPs in the DRIVE study. As a result, three independent SNPs (i.e., *SNAI1* rs1047920 C>T, *AMDHD1* rs11826 C>T and *CUBN* rs3914238 C>T) remained statistically significantly associated with BC risk (P = 0.0014, 0.0020 and 0.0022, respectively) and used for further analysis (**Table 1**).

The locations of these three SNPs and their associations with BC risk are presented in Supplementary Table 4. There was no heterogeneity found among the 14 GWASs for the effects of these three independent SNPs. The forest plots of three independent SNPs by the meta-analysis are presented in Supplementary Figure 1. The results showed that two SNPs were associated with a significantly decreased risk of BC (*SNAI1* rs1047920 C>T: OR = 0.92, 95% CI = 0.88-0.97, $P = 6.14 \times 10^{-4}$; and *AMDHD1* rs11826 C>T: OR = 0.95, 95% CI = 0.92-0.98, $P = 2.77 \times 10^{-4}$), while the other SNP was associated with a significantly increased BC risk (*CUBN* rs3914238 C>T: OR = 1.05, 95% CI = 1.02-1.07, $P = 3.16 \times 10^{-4}$). Regional association plots of the three independent SNPs in the 100 kb up- and down-stream regions are summarized in Supplementary Figure 2. As shown in **Table 2**, the effects of the *SNAI1* rs1047920 T, *AMDHD1* rs11826 T and *CUBN* rs3914238 T alleles on BC risk were sta-

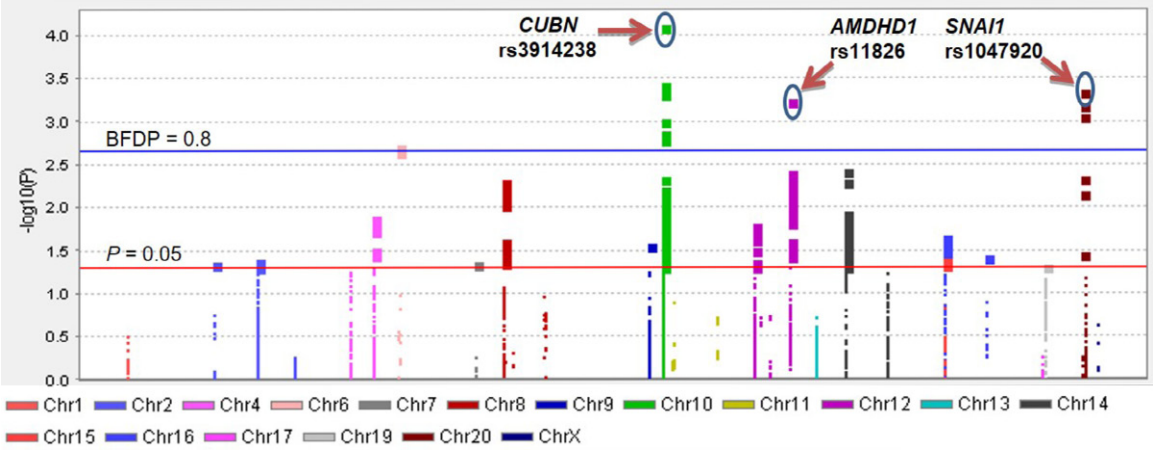


Figure 2. Manhattan Plot of the 2,994 SNPs of Vitamin D pathway Genes in the DRIVE study. The x-axis represents each chromosome. The y-axis represents the association *P* values (-log10 transformed) with breast cancer risk. The red horizontal line indicates *P* value equal to 0.05 and the blue horizontal line represents BFDP value equal to 0.8. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; BFDP, Bayesian false-discovery probability; *CUBN*, cubilin; DRIVE, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; *SNAI1*, snail family transcriptional repressor 1.

Table 1. Three genetic variants as independent BC risk predictors obtained from stepwise logistic regression analysis in the DRIVE study

SNP[1]	Location	MAF	Category[2]	Frequency	OR (95% CI)[1]	*P*[1]
SNAI1 rs1047920_T	20q13.13	0.07	CC/CT/TT	45730/6957/276	0.93 (0.88-0.97)	0.0014
AMDHD1 rs11826_T	12q23.1	0.24	CC/CT/TT	30973/18993/2997	0.96 (0.93-0.98)	0.0020
CUBN rs3914238_T	10p13	0.38	CC/CT/TT	20752/24693/7518	1.04 (1.01-1.07)	0.0022

[1]Stepwise logistic regression analysis included age, PC1, PC3, PC4, PC5, PC6, PC8, PC10, PC11, PC14, PC16, four previously published risk-associated SNPs (rs1323697, rs1264308, rs141308737 and rs1469412) in the same study by Jie Ge (PMID: 31026346) and 11 SNPs (rs3914238, rs1047920, rs7913144, rs7898138, rs11424438, rs2271464, rs77090490, rs4141977, rs11826, rs1800629 and rs1800628). [2]The most left-hand side "category" was used as the reference. Note: there were 20 PCs in the combined datasets as listed in Supplementary Table 3, of which ten remained significant and were adjusted in the final stepwise logistic regression analysis. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; BC, breast cancer; CI, confidence interval; *CUBN*, cubilin; DRIVE, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; SNP, single nucleotide polymorphism; MAF, minor allele frequency; OR, odds ratio; *SNAI1*, snail family transcriptional repressor 1.

tistically significant (trend test in univariate analysis: *P* = 0.001, 0.002 and 0.002, respectively; trend test in multivariate analysis: *P* = 0.001, 0.002 and 0.003, respectively).

Combined genotype analyses of the three independent SNPs

To assess the joint effect of the three independent SNPs on BC risk, we further combined risk genotypes of *SNAI1* rs1047920 CC, *AMDHD1* rs11826 CC and *CUBN* rs3914238 CT+TT into a genetic score as the NRG, which categorized all the individuals into four groups with 0 to 3 risk genotypes. The trend test indicated that the increased NRG was significantly associated with BC risk (*P*<0.0001, **Table 2**). According to the effect values and the frequency of each group, we further dichotomized all the individuals into two groups: low-risk (0-2 NRG) and high-risk (3 NRG). We found that the high-risk group had a higher BC risk (in multivariate analysis: OR = 1.07, 95% CI = 1.03-1.11, *P* = 0.0005, **Table 2**).

Stratified analysis of combined risk genotypes on BC risk

To further evaluate the interaction between genotypes and age, we stratified the analysis by age. As shown in **Table 3**, we found that the risk was more evident in the subgroup of age >60 (in univariate analysis: OR = 1.10, 95% CI = 1.04-1.16, P = 0.0005; in multivariate analysis: OR = 1.10, 95% CI = 1.04-1.15, *P* = 0.0007) compared with that for the subgroup of age

Table 2. Associations between three independent SNPs in the vitamin D pathway genes and BC risk in the DRIVE study

Genotype	$N_{Control}/N_{Case}$	Univariate analysis		Multivariate analysis[1]	
		OR (95% CI)	*P*	OR (95% CI)	*P*
SNAI1 rs1047920 C>T					
CC	20787/24943	1.00		1.00	
CT	3324/3633	0.91 (0.87-0.96)	0.0003	0.91 (0.87-0.96)	0.0003
TT	124/152	1.02 (0.81-1.30)	0.861	1.02 (0.80-1.30)	0.869
Trend test			0.001		0.001
CT+TT	3448/3785	0.92 (0.87-0.96)	0.0004	0.91 (0.87-0.96)	0.0004
AMDHD1 rs11826 C>T					
CC	14033/16940	1.00		1.00	
CT	8761/10232	0.97 (0.93-1.00)	0.074	0.97 (0.93-1.00)	0.069
TT	1441/1556	0.89 (0.83-0.96)	0.004	0.89 (0.82-0.96)	0.002
Trend test			0.002		0.002
CT+TT	10202/11788	0.96 (0.93-0.99)	0.013	0.96 (0.92-0.99)	0.011
CUBN rs3914238 C>T					
CC	9676/11076	1.00		1.00	
CT	11183/13510	1.06 (1.02-1.10)	0.004	1.05 (1.02-1.09)	0.006
TT	3376/4142	1.07 (1.02-1.13)	0.010	1.07 (1.01-1.13)	0.014
Trend test			0.002		0.003
CT+TT	14559/17652	1.06 (1.02-1.10)	0.001	1.06 (1.02-1.10)	0.002
Number of combined risk genotypes[2]					
0	596/628	1.00		1.00	
1	5128/5640	1.04 (0.93-1.18)	0.475	1.05 (0.93-1.18)	0.451
2	11284/13489	1.14 (1.01-1.27)	0.031	1.14 (1.01-1.28)	0.028
3	7232/8976	1.18 (1.05-1.32)	0.006	1.18 (1.05-1.33)	0.005
Trend test			<0.0001		<0.0001
0-2	17008/19757	1.00		1.00	
3	7232/8976	1.07 (1.03-1.11)	0.0005	1.07 (1.03-1.11)	0.0005

Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; BC: breast cancer; CI: confidence interval; *CUBN*, cubilin; DRIVE: Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; OR: odds ratio; *SNAI1*, snail family transcriptional repressor 1; SNP: single nucleotide polymorphism. [1]Adjusted for age, PC1, PC3, PC4, PC5, PC6, PC8, PC10, PC11, PC14 and PC16. [2]Risk genotypes were rs1047920 CC, rs11826 CC and rs3914238 CT+TT.

≥60. Further subgroup analysis (ER status and histological type) also revealed that the risk was more evident in the subgroup aged >60 than in the subgroup of age ≥60 among individuals with ER^+ and invasive tumor. Both ER status and histological type had any effect on BC risk associated with the genotypes, nor heterogeneity was found between these strata (all *P*>0.05) (**Table 3**).

LD analysis and functional prediction of the three independent SNPs

Functional prediction by RegulomeDB showed that *SNAI1* rs1047920 C>T, *AMDHD1* rs11826 C>T and *CUBN* rs3914238 C>T had RegulomeDB scores of 4, 6 and no data, respectively. We further searched for SNPs in high-LD (r^2≥0.7) with these three independent SNPs and made functional prediction by using HaploReg. We found that *SNAI1* rs1047920 C>T located in the 3'-UTR may change the motifs of BDP1 and LF-A1, which is also located in DNase I hypersensitive sites and that *AMDHD1* rs11826 C>T is located in the 3'-UTR with the selected eQTL for 5 hits, whereas *CUBN* rs3914238 C>T may change the motifs of Hand1 and PU.1 (Supplementary Table 5). We further assessed potential functions of these three independent SNPs using data from the ENCODE Project. The results revealed that *SNAI1* rs1047920 C>T is

Table 3. Stratified analysis for associations between the combined risk genotypes and BC in the DRIVE study

Characteristics	NRG 0-2		NRG 3		Univariate analysis		Multivariate analysis[1]		P_{inter} [2]
	Control	Case	Control	Case	OR (95% CI)	*P*	OR (95% CI)	*P*	
Age									
≤60	8866	8541	3852	3872	1.04 (0.99-1.10)	0.120	1.04 (0.99-1.10)	0.156	0.163
>60	8142	11216	3380	5104	1.10 (1.04-1.16)	0.0005	1.10 (1.04-1.15)	0.0007	
ER^+ vs. control									
≤60	8866	5270	3852	2322	1.01 (0.95-1.08)	0.656	1.02 (0.96-1.09)	0.470	0.135
>60	8142	7925	3380	3570	1.09 (1.03-1.15)	0.005	1.09 (1.03-1.15)	0.004	
ER^- vs. control									
≤60	8866	1283	3852	634	1.14 (1.03-1.26)	0.014	1.09 (0.98-1.21)	0.118	0.655
>60	8142	1238	3380	551	1.07 (0.96-1.20)	0.207	1.05 (0.94-1.17)	0.356	
Invasiveness vs. control									
≤60	8866	7620	3852	3454	1.04 (0.99-1.10)	0.132	1.04 (0.99-1.10)	0.140	0.213
>60	8142	10212	3380	4634	1.09 (1.04-1.15)	0.001	1.09 (1.04-1.15)	0.001	
In-situ vs. control									
≤60	8866	747	3852	343	1.06 (0.93-1.21)	0.416	1.04 (0.91-1.19)	0.555	0.223
>60	8142	807	3380	393	1.17 (1.03-1.33)	0.014	1.16 (1.02-1.32)	0.020	

[1]Adjusted for PC1, PC3, PC4, PC5, PC6, PC8, PC10, PC11, PC14 and PC16. [2]P_{inter}: *P* value for interaction analysis between age and NRG. Abbreviations: BC: breast cancer; CI: confidence interval; DRIVE: Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; ER: estrogen receptor; NRG: number of risk genotypes; OR: odds ratio.

located at H3K27AC acetylation enriched region (Supplementary Figure 4).

Correlation analysis

We further performed correlation analysis between genotypes of the three independent SNPs and their corresponding mRNA expression levels in the publically available RNA-seq data of lymphoblastoid cell lines generated from 373 European descendants in the 1000 Genomes Project. We found that the rs10479-20 C>T SNP was not significantly associated with levels of mRNA expression of *SNAI1* in all additive, dominant and recessive genetic models (*P* = 0.093, *P* = 0.102, and *P* = 0.491, respectively; Supplementary Figure 5A-C) but that another SNP (rs73276407 G>A) with a high LD (r^2 = 0.71) of rs1047920 C>T was significantly associated with decreased levels of mRNA expression of *SNAI1* in additive and dominant models (*P* = 0.017 and *P* = 0.021, respectively; **Figure 3A** and **3B**), but not for the recessive model (*P* = 0.303, **Figure 3C**). We also found that rs11826 C>T was not significantly associated with levels of mRNA expression of *AMDHD1* in all additive, dominant and recessive genetic models (*P* = 0.894, *P* = 0.815, and *P* = 0.861, respectively; Supplementary Figure 5D-F), while rs3914238 C>T was also not significantly associated with levels of mRNA expression of *CUBN* in all additive, dominant and recessive genetic models (*P* = 0.889, *P* = 0.864, and *P* = 0.608, respectively; Supplementary Figure 5G-I).

To gain more evidence on the correlations between the independent SNPs and mRNA expression levels, we further evaluated eQTL using data from the GTEx Project (http://www.gtexportal.org/home). The results showed that rs11826 C>T had a positive correlation with *AMDHD1* mRNA expression levels in breast tissue ($P = 6.00 \times 10^{-4}$, **Figure 3D**) and the whole blood ($P = 1.30 \times 10^{-7}$, **Figure 3E**). However, there were no significant associations between rs1047920 genotypes and *SNAI1* mRNA expression levels in the whole blood (*P* = 0.670; Supplementary Figure 5J) or breast tissue (*P* = 0.900; Supplementary Figure 5K) from GTEx, nor there were any significant associations between rs3914238 genotypes and *CUBN* mRNA expression levels in the whole blood (*P* = 0.350; Supplementary Figure 5L) or breast tissue (*P* = 0.170; Supplementary Figure 5M) from GTEx.

Discussion

To investigate whether genetic variants in the vitamin D pathway genes contribute to BC risk, we assessed associations between 2,994

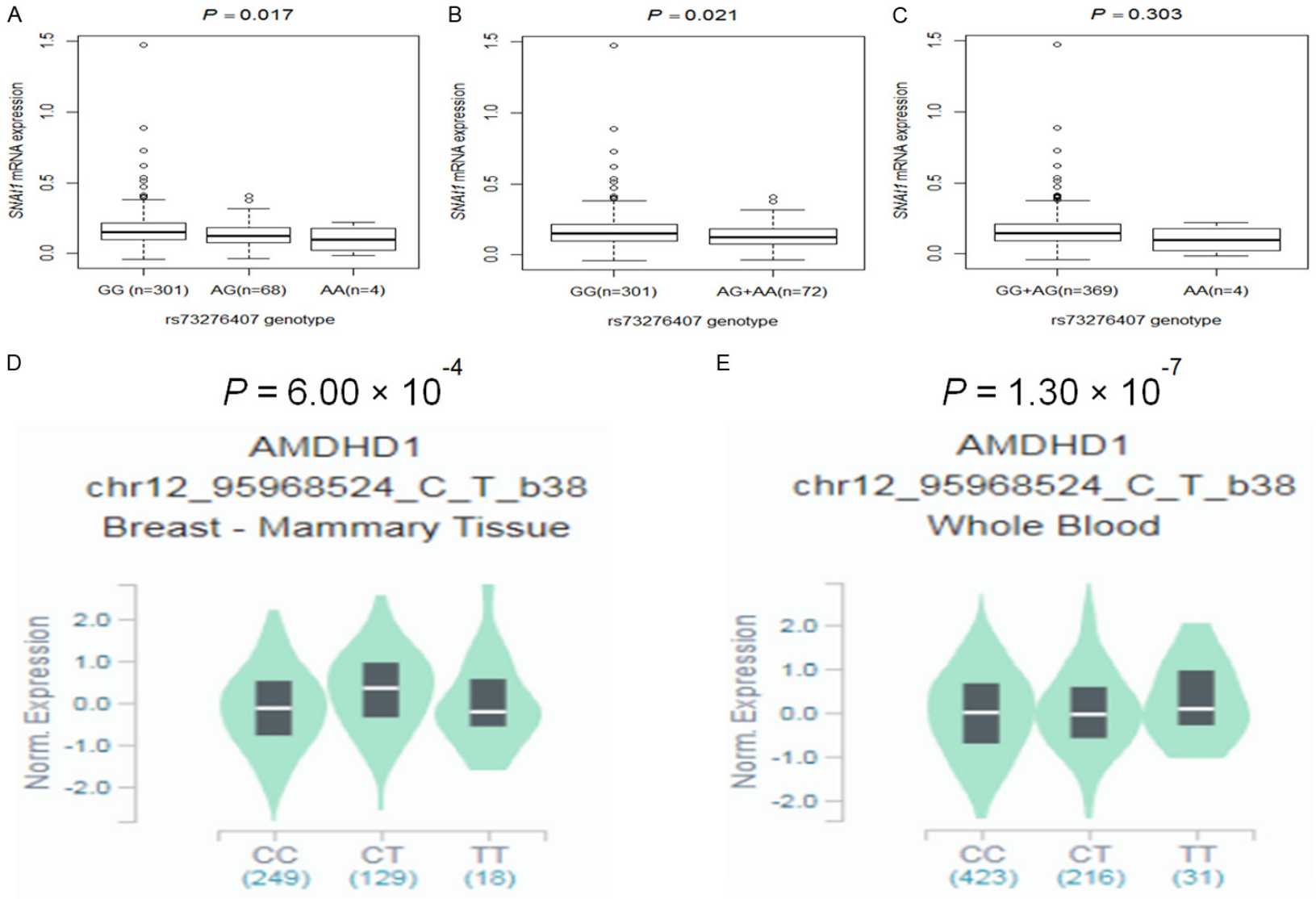


Figure 3. The expression quantitative trait loci (eQTLs) analysis for *SNAI1* rs73276407 (high LD with *SNAI1* rs1047920, $r^2 = 0.71$) and *AMDHD1* rs11826. Correlation between *SNAI1* mRNA expression and rs73276407 genotype in 373 Europeans from the 1000 Genomes Project in the (A) additive model, (B) dominant model and (C) recessive model; Correlation between *AMDHD1* mRNA expression and rs11826 genotype in the GTEx Project (D) Breast tissue, (E) Whole Blood. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; GTEx, Genotype-Tissue Expression; LD, Linkage disequilibrium; *SNAI1*, snail family transcriptional repressor 1.

SNPs in 33 vitamin D pathway genes and BC risk by using a large-scale meta-analysis of 14 GWASs in the DRIVE study. We identified three potential susceptibility variants (i.e., *SNAI1* rs1047920 at 20q13.13, *AMDHD1* rs11826 at 12q23.1, and *CUBN* rs3914238 at 10p13) that were independently or jointly associated with BC risk. Individuals with a higher NRG of these three genetic variants had a higher BC risk. Further eQTL analysis showed that the *SNAI1* rs73276407 A allele, in a high LD with the rs1047920 T allele, was associated with decreased *SNAI1* mRNA expression levels in lymphoblastoid cell lines, while the *AMDHD1* rs11826 T allele was found to be significantly associated with elevated *AMDHD1* mRNA expression levels in the whole blood cells and breast tissues. These findings suggest that SNPs in vitamin D pathway genes may play important biological roles in the development of BC, possibly by influencing their gene expression.

Vitamin D may play a controlling role in normal breast cell growth and has the ability to stop cancer cell growth in the breast. The bioactive form of vitamin D is 1,25-dihydroxyvitamin D that is largely considered a chemoprevention agent for its anti-cancer effect [29]. Several experimental studies have shown that 1,25-dihydroxyvitamin D can induce cell differentiation and apoptosis as well as inhibit angiogenesis and cell proliferation in normal and malignant breast cells [30, 31]. In the present study, we showed that genetic variants in some genes of the vitamin D pathway were associated with BC risk; however, their exact biological mechanisms need to be further investigated. Furthermore, the inconsistence observed for the correlations between genotypes of the risk-associated SNPs and mRNA expression levels of their corresponding genes in different human tissues (i.e., blood cells and breast tissues) remain to be resolved in more rigorously designed experimental and mechanistic studies.

SNAI1, also known as *SLUGH2*, *SNA*, *SNAH*, *SNAIL* or *SNAIL1*, encodes SNAIL that is a transcription factor and regulates the epithelial to mesenchymal transition by activating N-cadherin and repressing E-cadherin during embryonic development and cell migration [32, 33]. Few studies have investigated the roles of *SNAI1* in BC susceptibility. One study found that in the immune cells from the peripheral blood of BC patients, the expression of *SNAI1* was significantly higher in stage I patients than that in other higher stages [34], suggesting the involvement of this gene in early carcinogenesis. Another study showed that breast tumor tissue had lower levels of *SNAI1* protein product than normal breast tissue [35], further suggesting the involvement of this gene in breast carcinogenesis. One GWAS study has indicated that *MYT1* shares the same location of 20q13.13 as *SNAI1* [36], but the previously reported rs6062356 SNP in *MYT1* was not in LD with *SNAI1* rs1047920 (Supplementary Figure 6A), nor *MYT1* has any effect on the function of vitamin D. However, we found in the present study that the *SNAI1* rs73276407 A allele, in a high LD with the rs1047920 T allele, might down-regulate the expression of *SNAI1*, suggesting that *SNAI1* may play an oncogenic role in BC risk. Nevertheless, how *SNAI1* rs73276407 C>T influences BC risk needs additional experimental investigation.

AMDHD1 is also known as *HMFT1272*, and there were few studies focusing on the roles of *AMDHD1* in oncogenesis, especially in BC. In one study on quantitative analysis of gene expression, *AMDHD1* was found to be overexpressed in adrenal adenoma, compared with adrenal cancer [37]. Data from one GWAS of esophageal squamous cell carcinoma showed a statistically significant association between SNPs in *AMDHD1* (other than the ones identified in the present study) and cancer risk [38]. Although the GWAS study indicated that *LINC02452* shared the same location at 12q23.1 as *AMDHD1* [36], *AMDHD1* rs11826 has no LD with the previously published *LINC02452* rs77034926 (Supplementary Figure 6B) and lack of functional analysis. In the present study, we found that the rs11826 T allele might up-regulate the expression of *AMDHD1*, suggesting that *AMDHD1* may play a protective role in BC risk, but we did not have additional experimental data to explain how *AMDHD1* rs11826 C>T influenced BC risk.

CUBN encodes cubilin that is a large protein with three types of domain, an N-terminal stretch, 27 CUB domains and eight epidermal growth-factor-like repeats [39]. One study showed that mutations of *CUBN* caused glycosylation, endoplasmic reticulum retention, and abrogation of surface expression [40]. A meta-exome-wide association study of 33,985 Europeans found that a novel *CUBN* mutation was

associated with albuminuria levels, with a 3.5-fold increase in the effect in type 2 diabetes compared with non-diabetic individuals [41], and *CUBN* variants have also been associated with albuminuria [42, 43]. A genome-wide colorectal cancer association study has identified a promising cancer risk-associated intronic *CUBN* rs10904849 SNP [44]. However, there is no report on an association between SNPs in *CUBN* and BC risk. One BC GWAS study indicated that *FRMD4A* shared the same location 10p13 as *CUBN* [36], but the previously reported *FRMD4A* rs10906522 is not in LD with *CUBN* rs3914238 (Supplementary Figure 6C). In the present study, we did not find any evidence for an association between *CUBN* rs3914238 and its mRNA expression levels. Therefore, further experimental studies are needed to explain biological mechanisms underlying the observed associations between *CUBN* rs3914238 and BC risk.

The present study has several limitations that should be mentioned. First, some known risk factors, such as unhealthy lifestyle, physical activities, and incomplete information on estrogen status prevented us from complete adjustment in the analysis. Second, the biological mechanisms by which these three independent SNPs may influence BC risk remain unclear. Third, the risk-predicting model was built in non-Hispanic white, which may not be generalizable to the general population. Finally, our study was based on a signaling pathway and did not adjust the previously published BC risk associated SNPs, but only adjust previously published SNPs from the DRIVE study.

In summary, we have comprehensively analyzed SNPs in the vitamin D pathway genes for their associations with BC risk using genotyping data from 14 previously published GWASs among 53,107 subjects of European descendants, and we identified three independent BC susceptibility loci in the vitamin D pathway genes (i.e., *SNAI1* rs1047920, *AMDHD1* rs11826, and *CUBN* rs3914238). Once replicated by other investigators, these genetic variants may serve as new biomarkers for predicting BC risk.

Acknowledgements

DRIVE (dbGaP Study Accession: phs001265.v1.p1): The genotyping and phenotype data obtained from the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) breast-cancer case control samples was supported by U19 CA148065 and X01 HG007491 and by Cancer Research UK (C1287/A16563). Genotyping was conducted by the Centre for Inherited Disease Research (CIDR), University of Cambridge, Centre for Cancer Genetic Epidemiology, and the National Cancer Institute. The following studies provided germline DNA for 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), Multi-ethnic 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). GTEx: The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: GTEx Analysis V7 (dbGaP Accession phs000424.v7.p2). Qingyi Wei was supported by the P30 Cancer Center Support Grant from the Duke Cancer Institute (Grant ID: NIH CA014236). Hai-jiao Wang and Ling-ling Zhao were supported by funds from The First Hospital of Jilin University in China.

Disclosure of conflict of interest

None.

Abbreviations

AMDHD1, amidohydrolase domain containing 1; BC, breast cancer; BFDP, Bayesian false-discovery probability; BREOGAN, Breast Oncology Galicia Network; CGPS, Copenhagen General Population Study; CI, confidence interval; CPSII, Cancer Prevention Study-II Nutrition Cohort; CUBN, cubilin; DRIVE, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer;

EPIC, European Prospective Investigation Into Cancer and Nutrition; eQTL, expression quantitative trait loci; 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; NHS, Nurses' Health Study; NHS2, Nurses' Health Study 2; NRGs, number of risk 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; SNAI1, snail family transcriptional repressor 1; SNP, single nucleotide polymorphism; WHI, Women's Health Initiative.

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

References

[1] DeSantis CE, Ma J, Gaudet MM, Newman LA, Miller KD, Goding Sauer A, Jemal A and Siegel RL. Breast cancer statistics, 2019. CA Cancer J Clin 2019; 69: 438-451.

[2] Siegel RL, Miller KD and Jemal A. Cancer statistics, 2019. CA Cancer J Clin 2019; 69: 7-34.

[3] Sun YS, Zhao Z, Yang ZN, Xu F, Lu HJ, Zhu ZY, Shi W, Jiang J, Yao PP and Zhu HP. Risk factors and preventions of breast cancer. Int J Biol Sci 2017; 13: 1387-1397.

[4] Shiovitz S and Korde LA. Genetics of breast cancer: a topic in evolution. Ann Oncol 2015; 26: 1291-1299.

[5] Feldman D, Krishnan AV, Swami S, Giovannucci E and Feldman BJ. The role of vitamin D in reducing cancer risk and progression. Nat Rev Cancer 2014; 14: 342-357.

[6] Dawson-Hughes B, Heaney RP, Holick MF, Lips P, Meunier PJ and Vieth R. Estimates of optimal vitamin D status. Osteoporos Int 2005; 16: 713-716.

[7] Shao T, Klein P and Grossbard ML. Vitamin D and breast cancer. Oncologist 2012; 17: 36-45.

[8] Deeb KK, Trump DL and Johnson CS. Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. Nat Rev Cancer 2007; 7: 684-700.

[9] Krishnan AV, Swami S and Feldman D. Vitamin D and breast cancer: inhibition of estrogen synthesis and signaling. J Steroid Biochem Mol Biol 2010; 121: 343-348.

[10] Baeke F, Takiishi T, Korf H, Gysemans C and Mathieu C. Vitamin D: modulator of the immune system. Curr Opin Pharmacol 2010; 10: 482-496.

[11] Shirvani SS, Nouri M, Sakhinia E, Babaloo Z, Mohammadzaeh A, Alipour S, Jadideslam G and Khabbazi A. The molecular and clinical evidence of vitamin D signaling as a modulator of the immune system: role in Behcet's disease. Immunol Lett 2019; 210: 10-19.

[12] Provvedini DM, Tsoukas CD, Deftos LJ and Manolagas SC. 1,25-dihydroxyvitamin D3 receptors in human leukocytes. Science 1983; 221: 1181-1183.

[13] Goodwin PJ, Ennis M, Pritchard KI, Koo J and Hood N. Prognostic effects of 25-hydroxyvitamin D levels in early breast cancer. J Clin Oncol 2009; 27: 3757-3763.

[14] Neuhouser ML, Sorensen B, Hollis BW, Ambs A, Ulrich CM, McTiernan A, Bernstein L, Wayne S, Gilliland F, Baumgartner K, Baumgartner R and Ballard-Barbash R. Vitamin D insufficiency in a multiethnic cohort of breast cancer survivors. Am J Clin Nutr 2008; 88: 133-139.

[15] Ahn J, Peters U, Albanes D, Purdue MP, Abnet CC, Chatterjee N, Horst RL, Hollis BW, Huang WY, Shikany JM and Hayes RB. Serum vitamin D concentration and prostate cancer risk: a nested case-control study. J Natl Cancer Inst 2008; 100: 796-804.

[16] Cui Y and Rohan TE. Vitamin D, calcium, and breast cancer risk: a review. Cancer Epidemiol Biomarkers Prev 2006; 15: 1427-1437.

[17] Weinstein SJ, Yu K, Horst RL, Ashby J, Virtamo J and Albanes D. Serum 25-hydroxyvitamin D and risks of colon and rectal cancer in Finnish men. Am J Epidemiol 2011; 173: 499-508.

[18] de Sousa Almeida-Filho B, De Luca Vespoli H, Pessoa EC, Machado M, Nahas-Neto J and Nahas EAP. Vitamin D deficiency is associated with poor breast cancer prognostic features in postmenopausal women. J Steroid Biochem Mol Biol 2017; 174: 284-289.

[19] Ge J, Liu H, Qian D, Wang X, Moorman PG, Luo S, Hwang S and Wei Q. 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. Int J Cancer 2019; 145: 1270-1279.

[20] Jiang X, O'Reilly PF, Aschard H, Hsu YH, Richards JB, Dupuis J, Ingelsson E, Karasik D, Pilz S, Berry D, Kestenbaum B, Zheng J, Luan J, Sofianopoulou E, Streeten EA, Albanes D, Lutsey PL, Yao L, Tang W, Econs MJ, Wallaschofski H, Volzke H, Zhou A, Power C, McCarthy MI, Mi-

chos ED, Boerwinkle E, Weinstein SJ, Freedman ND, Huang WY, Van Schoor NM, van der Velde N, Groot L, Enneman A, Cupples LA, Booth SL, Vasan RS, Liu CT, Zhou Y, Ripatti S, Ohlsson C, Vandenput L, Lorentzon M, Eriksson JG, Shea MK, Houston DK, Kritchevsky SB, Liu Y, Lohman KK, Ferrucci L, Peacock M, Gieger C, Beekman M, Slagboom E, Deelen J, Heemst DV, Kleber ME, Marz W, de Boer IH, Wood AC, Rotter JI, Rich SS, Robinson-Cohen C, den Heijer M, Jarvelin MR, Cavadino A, Joshi PK, Wilson JF, Hayward C, Lind L, Michaelsson K, Trompet S, Zillikens MC, Uitterlinden AG, Rivadeneira F, Broer L, Zgaga L, Campbell H, Theodoratou E, Farrington SM, Timofeeva M, Dunlop MG, Valdes AM, Tikkanen E, Lehtimaki T, Lyytikainen LP, Kahonen M, Raitakari OT, Mikkila V, Ikram MA, Sattar N, Jukema JW, Wareham NJ, Langenberg C, Forouhi NG, Gundersen TE, Khaw KT, Butterworth AS, Danesh J, Spector T, Wang TJ, Hypponen E, Kraft P and Kiel DP. Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels. Nat Commun 2018; 9: 260.

[21] Howie B, Marchini J and Stephens M. Genotype imputation with thousands of genomes. G3 (Bethesda) 2011; 1: 457-470.

[22] Yang J, Lee SH, Goddard ME and Visscher PM. Genome-wide complex trait analysis (GCTA): methods, data analyses, and interpretations. Methods Mol Biol 2013; 1019: 215-236.

[23] Wakefield J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. Am J Hum Genet 2007; 81: 208-227.

[24] Barrett JC, Fry B, Maller J and Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005; 21: 263-265.

[25] Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, Boehnke M, Abecasis GR and Willer CJ. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics 2010; 26: 2336-2337.

[26] Ward LD and Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012; 40: D930-934.

[27] Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlof J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Hasler R, Syvanen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigo R, Gut IG, Estivill X and Dermitzakis ET. Transcriptome and genome sequencing uncovers functional variation in humans. Nature 2013; 501: 506-511.

[28] Gibson G. Human genetics. GTEx detects genetic effects. Science 2015; 348: 640-641.

[29] Wulaningsih W, Sagoo HK, Hamza M, Melvin J, Holmberg L, Garmo H, Malmstrom H, Lambe M, Hammar N, Walldius G, Jungner I and Van Hemelrijck M. Serum calcium and the risk of breast cancer: findings from the Swedish AMORIS study and a meta-analysis of prospective studies. Int J Mol Sci 2016; 17: 1487.

[30] Mantell DJ, Owens PE, Bundred NJ, Mawer EB and Canfield AE. 1 alpha,25-dihydroxyvitamin D(3) inhibits angiogenesis in vitro and in vivo. Circ Res 2000; 87: 214-220.

[31] Colston KW, Berger U and Coombes RC. Possible role for vitamin D in controlling breast cancer cell proliferation. Lancet 1989; 1: 188-191.

[32] Wang X, Liu R, Zhu W, Chu H, Yu H, Wei P, Wu X, Zhu H, Gao H, Liang J, Li G and Yang W. UDP-glucose accelerates SNAI1 mRNA decay and impairs lung cancer metastasis. Nature 2019; 571: 127-131.

[33] Barrallo-Gimeno A and Nieto MA. The Snail genes as inducers of cell movement and survival: implications in development and cancer. Development 2005; 132: 3151-3161.

[34] Naik A, Al-Zeheimi N, Bakheit CS, Al Riyami M, Al Jarrah A, Al Moundhri MS, Al Habsi Z, Basheer M and Adham SA. Neuropilin-1 associated molecules in the blood distinguish poor prognosis breast cancer: a cross-sectional study. Sci Rep 2017; 7: 3301.

[35] Chen S, Itoh T, Wu K, Zhou D and Yang C. Transcriptional regulation of aromatase expression in human breast tissue. J Steroid Biochem Mol Biol 2002; 83: 93-99.

[36] Michailidou K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A, Bolla MK, Wang Q, Tyrer J, Dicks E, Lee A, Wang Z, Allen J, Keeman R, Eilber U, French JD, Qing Chen X, Fachal L, McCue K, McCart Reed AE, Ghoussaini M, Carroll JS, Jiang X, Finucane H, Adams M, Adank MA, Ahsan H, Aittomaki K, Anton-Culver H, Antonenkova NN, Arndt V, Aronson KJ, Arun B, Auer PL, Bacot F, Barrdahl M, Baynes C, Beckmann MW, Behrens S, Benitez J, Bermisheva M, Bernstein L, Blomqvist C, Bogdanova NV, Bojesen SE, Bonanni B, Borresen-Dale AL, Brand JS, Brauch H, Brennan P, Brenner H,

Brinton L, Broberg P, Brock IW, Broeks A, Brooks-Wilson A, Brucker SY, Bruning T, Burwinkel B, Butterbach K, Cai Q, Cai H, Caldes T, Canzian F, Carracedo A, Carter BD, Castelao JE, Chan TL, David Cheng TY, Seng Chia K, Choi JY, Christiansen H, Clarke CL, Collee M, Conroy DM, Cordina-Duverger E, Cornelissen S, Cox DG, Cox A, Cross SS, Cunningham JM, Czene K, Daly MB, Devilee P, Doheny KF, Dork T, Dos-Santos-Silva I, Dumont M, Durcan L, Dwek M, Eccles DM, Ekici AB, Eliassen AH, Ellberg C, Elvira M, Engel C, Eriksson M, Fasching PA, Figueroa J, Flesch-Janys D, Fletcher O, Flyger H, Fritschi L, Gaborieau V, Gabrielson M, Gago-Dominguez M, Gao YT, Gapstur SM, Garcia-Saenz JA, Gaudet MM, Georgoulias V, Giles GG, Glendon G, Goldberg MS, Goldgar DE, Gonzalez-Neira A, Grenaker Alnaes GI, Grip M, Gronwald J, Grundy A, Guenel P, Haeberle L, Hahnen E, Haiman CA, Hakansson N, Hamann U, Hamel N, Hankinson S, Harrington P, Hart SN, Hartikainen JM, Hartman M, Hein A, Heyworth J, Hicks B, Hillemanns P, Ho DN, Hollestelle A, Hooning MJ, Hoover RN, Hopper JL, Hou MF, Hsiung CN, Huang G, Humphreys K, Ishiguro J, Ito H, Iwasaki M, Iwata H, Jakubowska A, Janni W, John EM, Johnson N, Jones K, Jones M, Jukkola-Vuorinen A, Kaaks R, Kabisch M, Kaczmarek K, Kang D, Kasuga Y, Kerin MJ, Khan S, Khusnutdinova E, Kiiski JI, Kim SW, Knight JA, Kosma VM, Kristensen VN, Kruger U, Kwong A, Lambrechts D, Le Marchand L, Lee E, Lee MH, Lee JW, Neng Lee C, Lejbkowicz F, Li J, Lilyquist J, Lindblom A, Lissowska J, Lo WY, Loibl S, Long J, Lophatananon A, Lubinski J, Luccarini C, Lux MP, Ma ESK, MacInnis RJ, Maishman T, Makalic E, Malone KE, Kostovska IM, Mannermaa A, Manoukian S, Manson JE, Margolin S, Mariapun S, Martinez ME, Matsuo K, Mavroudis D, McKay J, McLean C, Meijers-Heijboer H, Meindl A, Menendez P, Menon U, Meyer J, Miao H, Miller N, Taib NAM, Muir K, Mulligan AM, Mulot C, Neuhausen SL, Nevanlinna H, Neven P, Nielsen SF, Noh DY, Nordestgaard BG, Norman A, Olopade OI, Olson JE, Olsson H, Olswold C, Orr N, Pankratz VS, Park SK, Park-Simon TW, Lloyd R, Perez JIA, Peterlongo P, Peto J, Phillips KA, Pinchev M, Plaseska-Karanfilska D, Prentice R, Presneau N, Prokofyeva D, Pugh E, Pylkas K, Rack B, Radice P, Rahman N, Rennert G, Rennert HS, Rhenius V, Romero A, Romm J, Ruddy KJ, Rudiger T, Rudolph A, Ruebner M, Rutgers EJT, Saloustros E, Sandler DP, Sangrajrang S, Sawyer EJ, Schmidt DF, Schmutzler RK, Schneeweiss A, Schoemaker MJ, Schumacher F, Schurmann P, Scott RJ, Scott C, Seal S, Seynaeve C, Shah M, Sharma P, Shen CY, Sheng G, Sherman ME, Shrubsole MJ, Shu XO, Smeets A, Sohn C, Southey MC, Spinelli JJ, Stegmaier C, Stewart-Brown S, Stone J, Stram DO, Surowy H, Swerdlow A, Tamimi R, Taylor JA, Tengstrom M, Teo SH, Beth Terry M, Tessier DC, Thanasitthichai S, Thone K, Tollenaar R, Tomlinson I, Tong L, Torres D, Truong T, Tseng CC, Tsugane S, Ulmer HU, Ursin G, Untch M, Vachon C, van Asperen CJ, Van Den Berg D, van den Ouweland AMW, van der Kolk L, van der Luijt RB, Vincent D, Vollenweider J, Waisfisz Q, Wang-Gohrke S, Weinberg CR, Wendt C, Whittemore AS, Wildiers H, Willett W, Winqvist R, Wolk A, Wu AH, Xia L, Yamaji T, Yang XR, Har Yip C, Yoo KY, Yu JC, Zheng W, Zheng Y, Zhu B, Ziogas A, Ziv E, Lakhani SR, Antoniou AC, Droit A, Andrulis IL, Amos CI, Couch FJ, Pharoah PDP, Chang-Claude J, Hall P, Hunter DJ, Milne RL, Garcia-Closas M, Schmidt MK, Chanock SJ, Dunning AM, Edwards SL, Bader GD, Chenevix-Trench G, Simard J, Kraft P and Easton DF. Association analysis identifies 65 new breast cancer risk loci. Nature 2017; 551: 92-94.

[37] Assie G, Guillaud-Bataille M, Ragazzon B, Bertagna X, Bertherat J and Clauser E. The pathophysiology, diagnosis and prognosis of adrenocortical tumors revisited by transcriptome analyses. Trends Endocrinol Metab 2010; 21: 325-334.

[38] Hyland PL, Zhang H, Yang Q, Yang HH, Hu N, Lin SW, Su H, Wang L, Wang C, Ding T, Fan JH, Qiao YL, Sung H, Wheeler W, Giffen C, Burdett L, Wang Z, Lee MP, Chanock SJ, Dawsey SM, Freedman ND, Abnet CC, Goldstein AM, Yu K and Taylor PR. Pathway, in silico and tissue-specific expression quantitative analyses of oesophageal squamous cell carcinoma genome-wide association studies data. Int J Epidemiol 2016; 45: 206-220.

[39] Moestrup SK, Kozyraki R, Kristiansen M, Kaysen JH, Rasmussen HH, Brault D, Pontillon F, Goda FO, Christensen EI, Hammond TG and Verroust PJ. The intrinsic factor-vitamin B12 receptor and target of teratogenic antibodies is a megalin-binding peripheral membrane protein with homology to developmental proteins. J Biol Chem 1998; 273: 5235-5242.

[40] Udagawa T, Harita Y, Miura K, Mitsui J, Ode KL, Morishita S, Urae S, Kanda S, Kajiho Y, Tsurumi H, Ueda HR, Tsuji S, Saito A and Oka A. Amnionless-mediated glycosylation is crucial for cell surface targeting of cubilin in renal and intestinal cells. Sci Rep 2018; 8: 2351.

[41] Ahluwalia TS, Schulz CA, Waage J, Skaaby T, Sandholm N, van Zuydam N, Charmet R, Bork-Jensen J, Almgren P, Thuesen BH, Bedin M, Brandslund I, Christensen CK, Linneberg A, Ahlqvist E, Groop PH, Hadjadj S, Tregouet DA, Jorgensen ME, Grarup N, Pedersen O, Simons M, Groop L, Orho-Melander M, McCarthy MI,

Melander O, Rossing P, Kilpelainen TO and Hansen T. A novel rare CUBN variant and three additional genes identified in Europeans with and without diabetes: results from an exome-wide association study of albuminuria. Diabetologia 2019; 62: 292-305.

[42] Bedin M, Boyer O, Servais A, Li Y, Villoing-Gaude L, Tete MJ, Cambier A, Hogan J, Baudouin V, Krid S, Bensman A, Lammens F, Louillet F, Ranchin B, Vigneau C, Bouteau I, Isnard-Bagnis C, Mache CJ, Schafer T, Pape L, Godel M, Huber TB, Benz M, Klaus G, Hansen M, Latta K, Gribouval O, Moriniere V, Tournant C, Grohmann M, Kuhn E, Wagner T, Bole-Feysot C, Jabot-Hanin F, Nitschke P, Ahluwalia TS, Kottgen A, Andersen CBF, Bergmann C, Antignac C and Simons M. Human C-terminal CUBN variants associate with chronic proteinuria and normal renal function. J Clin Invest 2020; 130: 335-344.

[43] Boger CA, Chen MH, Tin A, Olden M, Kottgen A, de Boer IH, Fuchsberger C, O'Seaghdha CM, Pattaro C, Teumer A, Liu CT, Glazer NL, Li M, O'Connell JR, Tanaka T, Peralta CA, Kutalik Z, Luan J, Zhao JH, Hwang SJ, Akylbekova E, Kramer H, van der Harst P, Smith AV, Lohman K, de Andrade M, Hayward C, Kollerits B, Tonjes A, Aspelund T, Ingelsson E, Eiriksdottir G, Launer LJ, Harris TB, Shuldiner AR, Mitchell BD, Arking DE, Franceschini N, Boerwinkle E, Egan J, Hernandez D, Reilly M, Townsend RR, Lumley T, Siscovick DS, Psaty BM, Kestenbaum B, Haritunians T, Bergmann S, Vollenweider P, Waeber G, Mooser V, Waterworth D, Johnson AD, Florez JC, Meigs JB, Lu X, Turner ST, Atkinson EJ, Leak TS, Aasarod K, Skorpen F, Syvanen AC, Illig T, Baumert J, Koenig W, Kramer BK, Devuyst O, Mychaleckyj JC, Minelli C, Bakker SJ, Kedenko L, Paulweber B, Coassin S, Endlich K, Kroemer HK, Biffar R, Stracke S, Volzke H, Stumvoll M, Magi R, Campbell H, Vitart V, Hastie ND, Gudnason V, Kardia SL, Liu Y, Polasek O, Curhan G, Kronenberg F, Prokopenko I, Rudan I, Arnlov J, Hallan S, Navis G, Parsa A, Ferrucci L, Coresh J, Shlipak MG, Bull SB, Paterson NJ, Wichmann HE, Wareham NJ, Loos RJ, Rotter JI, Pramstaller PP, Cupples LA, Beckmann JS, Yang Q, Heid IM, Rettig R, Dreisbach AW, Bochud M, Fox CS and Kao WH. CUBN is a gene locus for albuminuria. J Am Soc Nephrol 2011; 22: 555-570.

[44] Al-Tassan NA, Whiffin N, Hosking FJ, Palles C, Farrington SM, Dobbins SE, Harris R, Gorman M, Tenesa A, Meyer BF, Wakil SM, Kinnersley B, Campbell H, Martin L, Smith CG, Idziaszczyk S, Barclay E, Maughan TS, Kaplan R, Kerr R, Kerr D, Buchanan DD, Win AK, Hopper J, Jenkins M, Lindor NM, Newcomb PA, Gallinger S, Conti D, Schumacher F, Casey G, Dunlop MG, Tomlinson IP, Cheadle JP and Houlston RS. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci Rep 2015; 5: 10442.

Supplementary Table 1. Summary of 14 breast cancer GWASs from the DRIVE Study*

Study	Full name	Sample size (cases/controls)	Countries
BREOGAN	Breast Oncology Galicia Network	1,387/755	Spain
CGPS	Copenhagen General Population Study	1,415/725	Denmark
CPSII	Cancer Prevention Study-II Nutrition Cohort	3,067/3,036	USA
EPIC	European Prospective Investigation Into Cancer and Nutrition	3,863/3,662	France, Germany, Greece, Italy, Netherlands, Spain, UK
MCCS	Melbourne Collaborative Cohort Study	872/821	Australia
MEC	Multiethnic Cohort	682/732	USA
NBHS	Nashville Breast Health Study	889/797	USA
NHS	Nurses' Health Study	1,598/1,806	USA
NHS2	Nurses' Health Study 2	1,615/1,910	USA
PBCS	NCI Polish Breast Cancer Study	1,934/2,052	Poland
PLCO	The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial	870/860	USA
SEARCH	Study of Epidemiology and Risk factors in Cancer Heredity	4,062/1,841	UK
SMC	Swedish Mammography Cohort	1,522/716	Sweden
WHI	Women's Health Initiative	4,982/4,636	USA
Total		28,758/24,349	

Abbreviations: DRIVE: Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; GWAS: Genome-wide association studies; *Other 3 studies (2 SISTER, SISTER and WAABCS) were excluded, because they did not have either the cases or controls of European Descent.

Supplementary Table 2. List of 33 selected genes in the vitamin D pathway

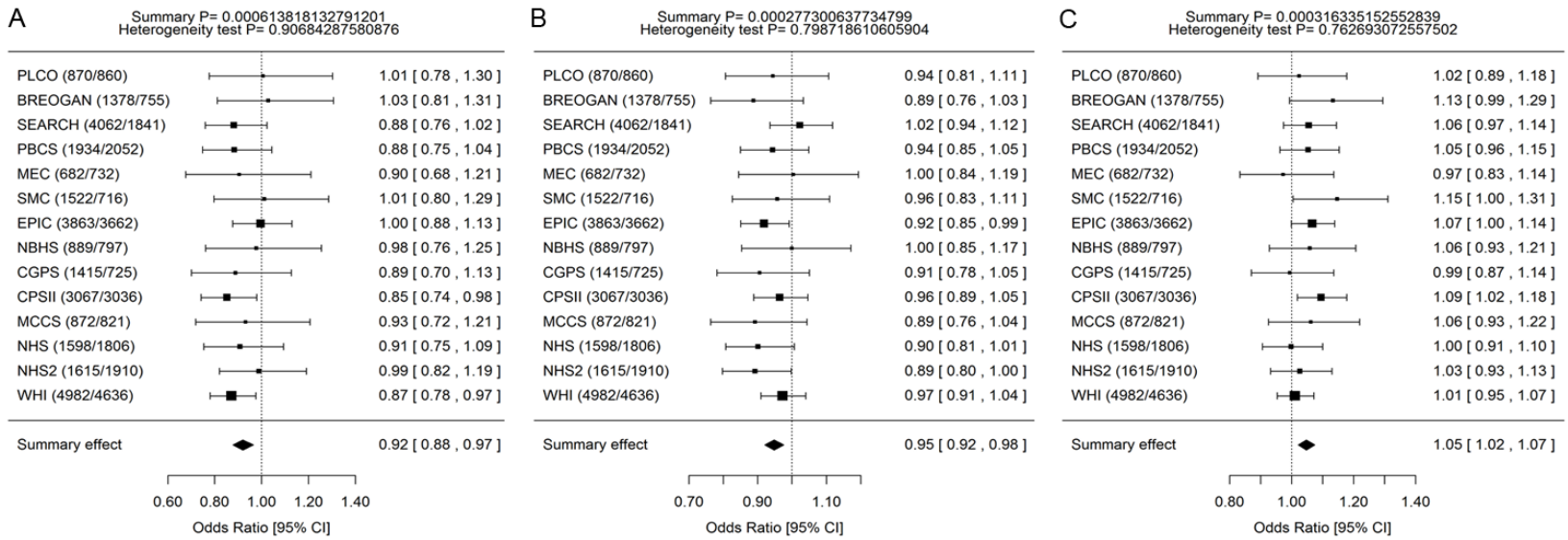
Dataset/Journal	Name of pathway/article	Name of genes	Number of genes
REACTOME	REACTOME_VITAMIN_D_CALCIFEROL_METABOLISM	*CUBN, CYP24A1, CYP27B1, CYP2R1, GC, LGMN, LRP2, PIAS4, SUMO2, UBE2I, VDR*	11
GO	GO_NEGATIVE_REGULATION_OF_VITAMIN_D_BIOSYNTHETIC_PROCESS	*CYP27B1, GFI1, NFKB1, SNAI1, SNAI2*	5
GO	GO_REGULATION_OF_VITAMIN_D_BIOSYNTHETIC_PROCESS	*CYP27B1, GFI1, IFNG, IL1B, NFKB1, SNAI1, SNAI2, TNF, VDR*	9
GO	GO_VITAMIN_D_BINDING	*CALB1, GC, IRX5, KL, S100G, VDR*	6
GO	GO_VITAMIN_D_BIOSYNTHETIC_PROCESS	*CYP27A1, CYP27B1, CYP2R1, CYP3A4, GFI1, IFNG, IL1B, NFKB1, SNAI1, SNAI2, TNF, VDR*	12
GO	GO_VITAMIN_D_METABOLIC_PROCESS	*CUBN, CYP11A1, CYP1A1, CYP24A1, CYP27A1, CYP27B1, CYP2R1, CYP3A4, FGFR1, GC, LGMN, LRP2, PIAS4, VDR*	14
BIOCARTA	NO DATA	-	0
KEGG	NO DATA	-	0
Canonical pathways	NO DATA	-	0
Pathcards	VITAMIN D METABOLISM	*CYP24A1, CYP27A1, CYP27B1, DHCR7, GC, PTH, RXRA, RXRB, VDR, CYP2R1*	10
Nature communication	Genome-wide association study in 79,366 European-ancestry individuals informs the genetic architecture of 25-hydroxyvitamin D levels	*AMDHD1, CYP2R1, CYP24A1, DHCR7, GC, SEC23A*	6
Total	*AMDHD1, CALB1, CUBN, CYP24A1, CYP27B1, CYP2R1, CYP11A1, CYP1A1, CYP27A1, CYP3A4, DHCR7, FGFR1, GC, GFI1, IFNG, IL1B, IRX5, KL, LGMN, LRP2, NFKB1, PTH, PIAS4, RXRA, RXRB, S100G, SEC23A, SNAI1, SNAI2, SUMO2, TNF, UBE2I, VDR* (after removing the duplicated 40 genes)		33

Keyword: Vitamin D. Organism: Homo sapiens. Website: http://software.broadinstitute.org/gsea/msigdb/search.jsp; https://pathcards.genecards.org/.

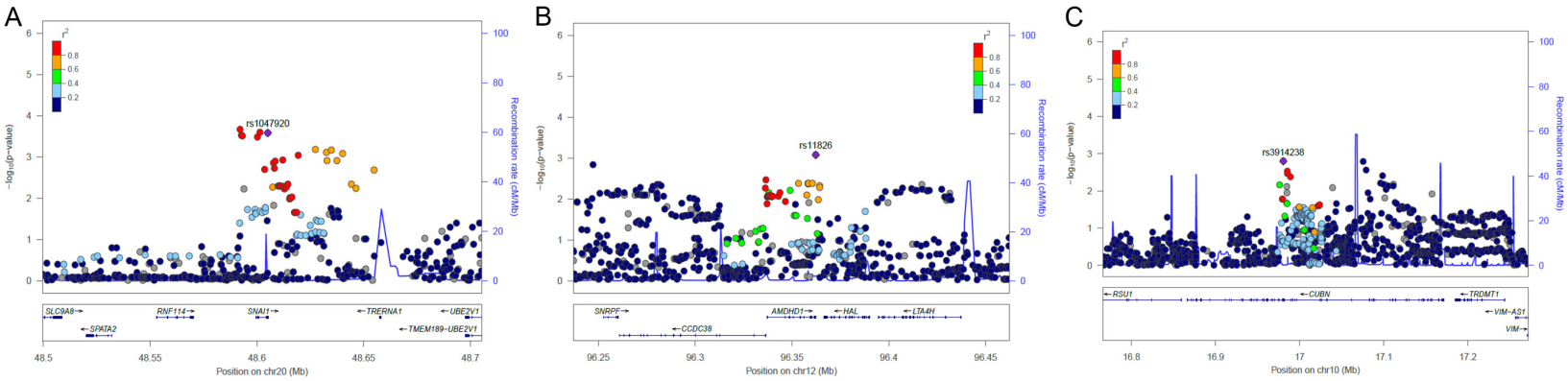
Supplementary Table 3. *P* values for the associations between the first 20 principal components and breast cancer risk across the 14 studies obtained from logistic regression analysis

PC	PLCO	BREOGAN	SEARCH	PBCS	MEC	SMC	EPIC	NBHS	CGPS	CPSII	MCCS	NHS	NHS2	WHI	Total
PC1	0.1037	0.5805	0.2867	0.0454	0.3991	0.0380	0.7795	0.0021	0.0006	0.4725	0.8840	0.0450	0.1618	0.0046	0.0073
PC2	0.0568	<0.0001	0.0487	0.0054	0.5331	0.5216	0.7077	0.2561	0.0018	0.9709	0.3143	0.0034	0.0000	0.0553	0.4113
PC3	0.8118	0.0000	0.0131	0.0002	0.2660	0.3160	0.1530	0.9698	0.5786	0.0527	0.3018	0.7830	0.5546	0.7260	0.0052
PC4	0.0446	0.0003	0.0189	0.1641	0.7901	0.0093	0.2301	0.4936	0.0087	0.8038	0.1680	0.1080	0.8805	0.6848	<0.0001
PC5	0.4040	<0.0001	0.2412	0.0005	0.6816	0.0016	0.0833	0.9152	0.9242	0.1130	0.7407	0.8735	0.9646	0.5264	0.0001
PC6	0.9355	0.8375	0.3452	0.0300	0.4022	0.0514	0.3276	0.2717	0.2902	0.0489	0.4852	0.4453	0.4242	0.4947	0.0348
PC7	0.0064	0.0172	0.3320	0.9963	0.7884	0.0301	0.2413	0.1694	0.0114	0.2128	0.4673	0.2337	0.4330	0.7594	0.8664
PC8	0.6881	<0.0001	0.6223	0.6397	0.3538	0.3700	0.2941	0.2188	0.0098	0.9488	0.7606	0.8417	0.3357	0.5643	0.0097
PC9	0.9842	0.0028	0.8906	0.4724	0.3383	0.0001	0.4834	0.3539	0.1410	0.8688	0.1291	0.0627	0.7938	0.2913	0.6651
PC10	0.0024	0.0001	0.0047	0.5835	0.2688	0.0092	0.2498	0.2566	0.3851	0.1504	0.2482	0.0742	0.1640	0.6210	<0.0001
PC11	0.2690	0.3033	0.8024	0.3455	0.3232	0.0201	0.7979	0.2314	0.1016	0.0609	0.4782	0.4643	0.1699	0.4135	0.0001
PC12	0.0193	0.0070	0.5973	0.4480	0.3409	0.0171	0.0557	0.0079	0.0360	0.3422	0.2218	0.1781	0.7040	0.2996	0.2278
PC13	0.5667	0.7368	0.4179	0.1628	0.6539	0.4115	0.5987	0.6322	0.2025	0.4899	0.3618	0.3892	0.6420	0.2667	0.6936
PC14	0.5210	0.0773	0.6156	0.9417	0.9678	0.7219	0.0637	0.6152	0.0029	0.7316	0.3297	0.1117	0.4411	0.0416	0.0026
PC15	0.3947	0.0430	0.0008	0.7612	0.9218	0.0039	0.1798	0.1498	0.6950	0.4773	0.2890	0.7490	0.6949	0.8926	0.2451
PC16	0.3687	0.1415	0.1605	0.6169	0.7086	0.7873	0.2597	0.9809	0.5170	0.1966	0.5652	0.3401	0.6612	0.1433	0.0449
PC17	0.0914	0.0388	0.0120	0.0489	0.1571	0.3439	0.1780	0.0407	0.5787	0.4500	0.2360	0.7130	0.8080	0.5594	0.4563
PC18	0.1300	0.3079	0.0953	0.2101	0.6809	0.7605	0.3856	0.5882	0.6820	0.4223	0.5085	0.1953	0.2607	0.3737	0.1143
PC19	0.4476	0.2744	<0.0001	0.2677	0.7638	0.4543	0.3552	0.0640	0.9202	0.7738	0.4906	0.9082	0.7693	0.2010	0.1980
PC20	0.1147	<0.0001	0.0030	0.8132	0.1847	0.7416	0.2996	0.3378	0.0001	0.9941	0.3284	0.1682	0.0205	0.9548	0.3403

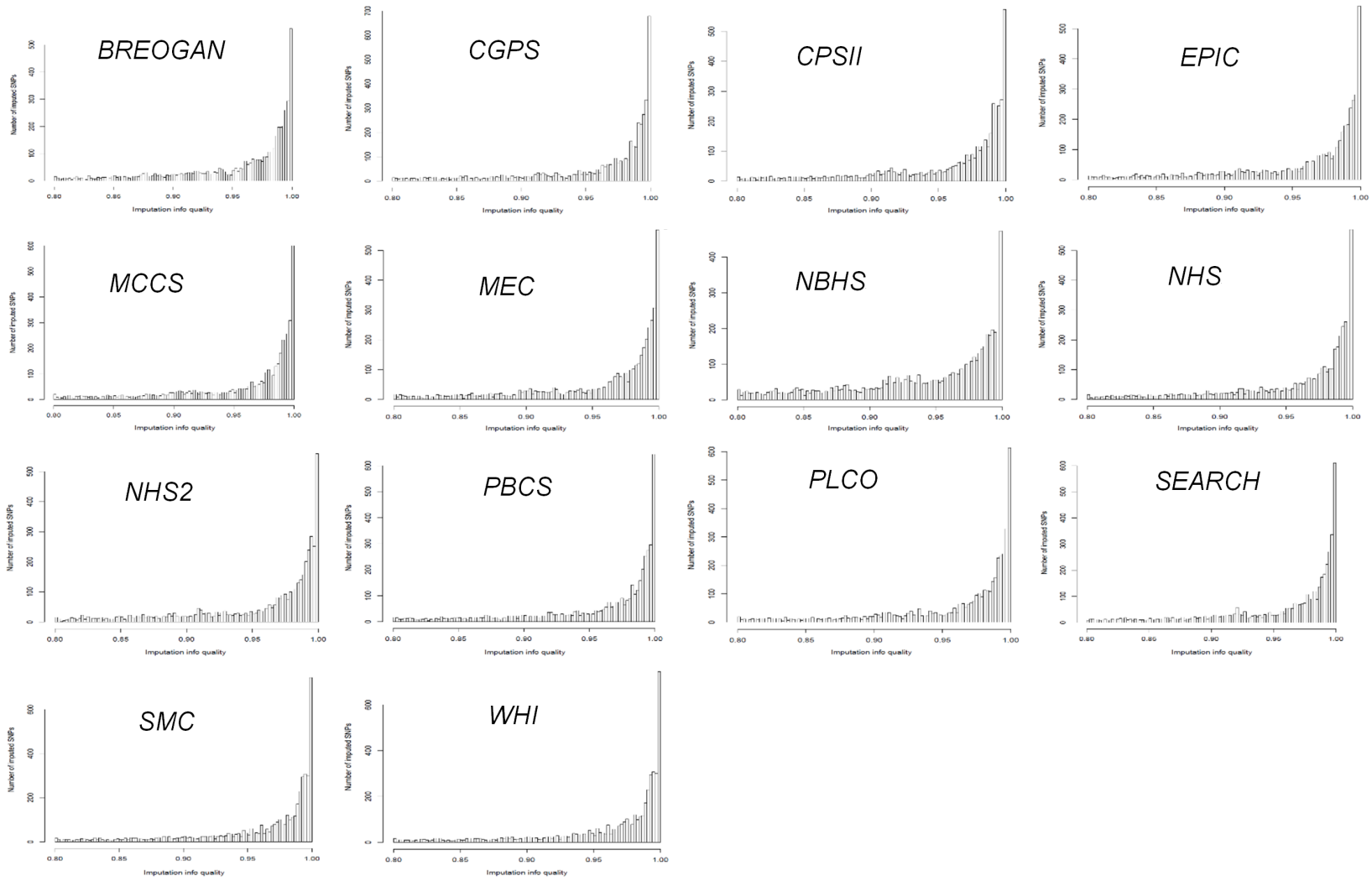
The DRIVE Studies: BREOGAN, Breast Oncology Galicia Network; CGPS, Copenhagen General Population Study; CPSII, Cancer Prevention Study-II Nutrition Cohort; EPIC, European Prospective Investigation Into Cancer and Nutrition; MCCS, Melbourne Collaborative Cohort Study; MEC, Multiethnic Cohort; NBHS, Nashville Breast Health Study; NHS, Nurses' Health Study; NHS2, Nurses' Health Study 2; PBCS, NCI Polish Breast Cancer Study; PLCO, The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; SEARCH, Study of Epidemiology and Risk factors in Cancer Heredity; SMC, Swedish Mammography Cohort; WHI, Women's Health Initiative. Abbreviations: PC: principal component.



Supplementary Figure 1. Forest plot for three independent SNPs. (A) *SNAI1* rs1047920; (B) *AMDHD1* rs11826; (C) *CUBN* rs3914238. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; *CUBN*, cubilin; *SNAI1*, snail family transcriptional repressor 1; SNP, single nucleotide polymorphism.



Supplementary Figure 2. Regional association plots contained 100 kb up and downstream of the gene regions in (A) *SNAI1*, (B) *AMDHD1* and (C) *CUBN*. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; *CUBN*, cubilin; *SNAI1*, snail family transcriptional repressor 1.



Supplementary Figure 3. Distribution plot for imputation info quality. Abbreviations: BREOGAN, Breast Oncology Galicia Network; CGPS, Copenhagen General Population Study; CPSII, Cancer Prevention Study-II Nutrition Cohort; EPIC, European Prospective Investigation Into Cancer and Nutrition; MCCS, Melbourne Collaborative Cohort Study; MEC, Multiethnic Cohort; NBHS, Nashville Breast Health Study; NHS, Nurses' Health Study; NHS2, Nurses' Health Study 2; PBCS, NCI Polish Breast Cancer Study; PLCO, The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial; SEARCH, Study of Epidemiology and Risk factors in Cancer Heredity; SMC, Swedish Mammography Cohort; WHI, Women's Health Initiative.

Supplementary Table 4. Associations between three independent SNPs in the vitamin D pathway genes and breast cancer risk in the DRIVE study with multiple testing corrections

SNP	Location	Gene	Type	Allele[a]	MAF	P_{het}[b]	I^2	P[c]	FDR	BFDP (Prior probability[d])		
										0.1	0.01	0.001
rs3914238	10p13	*CUBN*	Imputed	C/T	0.38	0.38	6.21	7.97E-05	0.234	4.16E-06	4.58E-05	4.62E-04
rs1047920	20q13.13	*SNAI1*	Imputed	C/T	0.07	0.94	0	4.61E-04	0.234	0.324	0.840	0.982
rs11826	12q23.1	*AMDHD1*	Imputed	C/T	0.24	0.60	0	5.79E-04	0.248	0.631	0.950	0.995

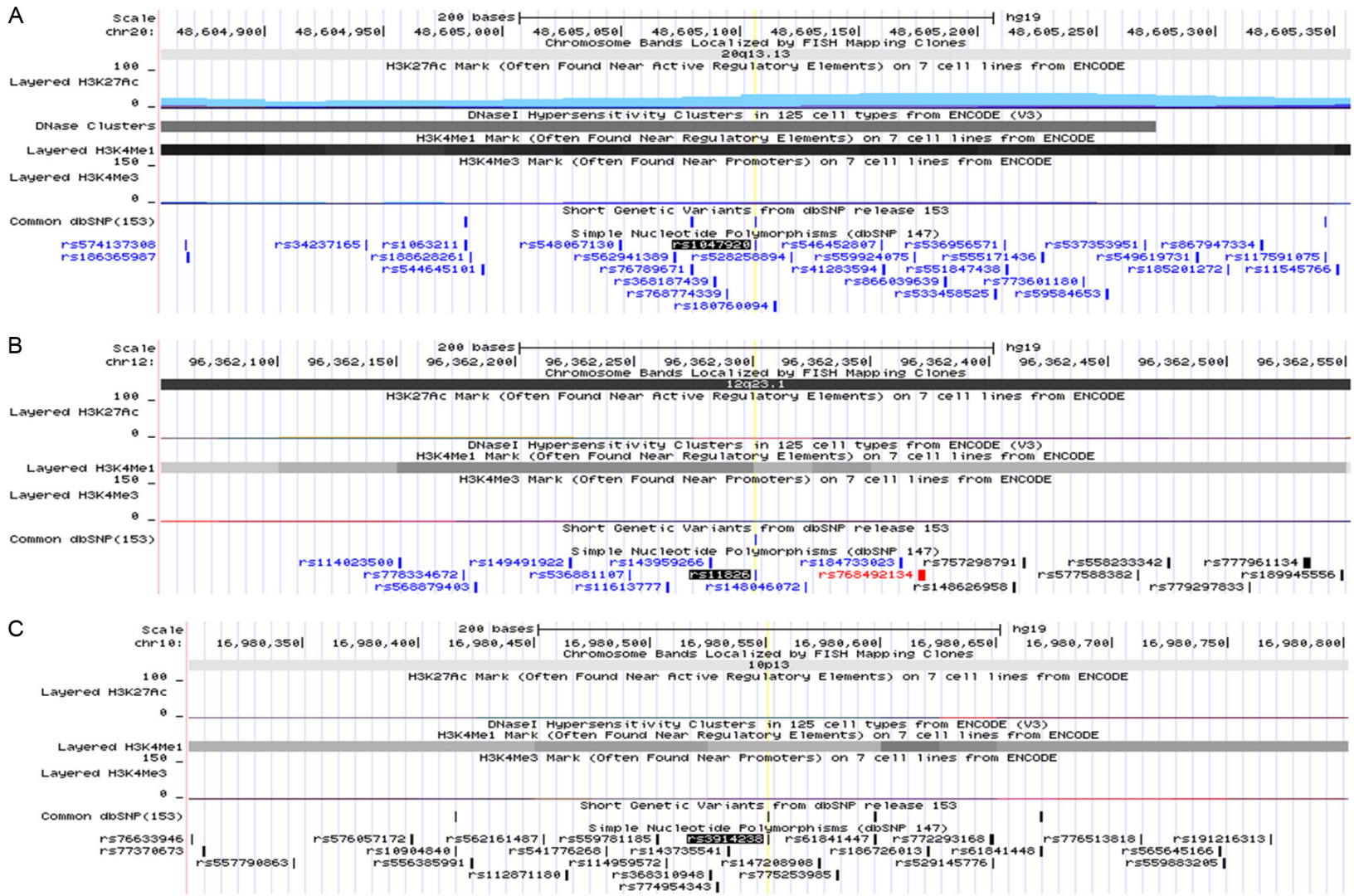
[a]Reference allele/effect (minor) allele; [b]*P* value for heterogeneity by Cochrane's Q test; [c]Meta-analysis in the fixed-effects model if Q test *P*>0.100 and I^2<50.0%; otherwise: random-effects model; [d]Calculated using study subjects to detect an upper bound of 3.0 and a prior probability of 0.1. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; BFDP, Bayesian false-discovery probability; *CUBN*, cubilin; FDR, false discovery rate; *SNAI1*, snail family transcriptional repressor 1; SNP, single nucleotide polymorphism.

Supplementary Table 5. Function prediction of three independent SNPs and other SNPs in high linkage disequilibrium (LD) ($r^2 \geq 0.7$) in the same genes

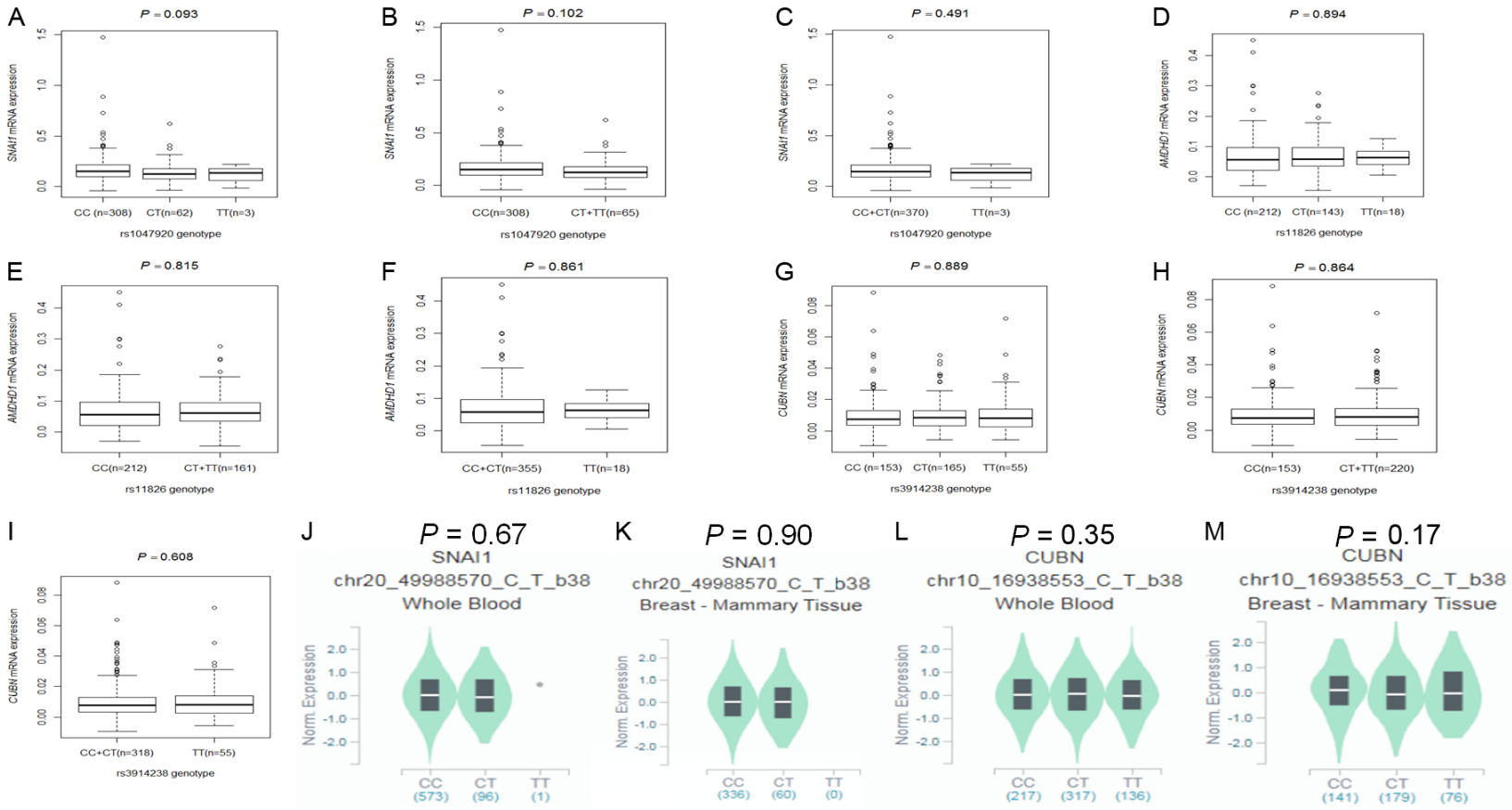
				Haploreg v4.1[b]								
SNPs	Chr	Gene	RegDB[a]	LD (r^2)	Promoter histone marks	Enhancer histone marks	DNAse	Proteins bound	Motifs changed	GRASP QTL hits	Selected eQTL hits	dbSNP func annot
rs2356587	10	*CUBN*	6	0.84					16 altered motifs			intronic
rs1801234	10	*CUBN*	No Data	0.84			IPSC					synonymous
rs3914238	10	*CUBN*	No Data	1					Hand1, PU.1			intronic
rs144169849	10	*CUBN*	3b	0.91	BLD	7 tissues			4 altered motifs			intronic
rs11424438	10	*CUBN*	-	0.94	FAT, SKIN	13 tissues			7 altered motifs			intronic
rs7898138	10	*CUBN*	5	0.96	FAT, SKIN	13 tissues	6 tissues		Foxj1, Foxk1, TEF			intronic
rs7913144	10	*CUBN*	4	0.96	4 tissues	12 tissues	8 tissues	CTCF, RAD21	CDP			intronic
rs2271464	10	*CUBN*	No Data	0.96		5 tissues	ESDR					intronic
rs2932908	10	*CUBN*	No Data	0.75			IPSC		Pax-5			intronic
rs1512701	10	*CUBN*	5	0.81		SPLN		CEBPB	5 altered motifs			intronic
rs79322116	12	*AMDHD1*	4	0.8	24 tissues	BRN	33 tissues	4 bound proteins	8 altered motifs		4 hits	5'-UTR
rs79175383	12	*AMDHD1*	4	0.81	24 tissues		34 tissues	4 bound proteins			4 hits	
rs7955759	12	*AMDHD1*	4	0.77	24 tissues	BRN	37 tissues	9 bound proteins	CHD2, ZBTB33	8 hits	6 hits	5'-UTR
rs1436121	12	*AMDHD1*	2b	0.8	24 tissues	BRN	49 tissues	10 bound proteins	5 altered motifs	9 hits	7 hits	synonymous
rs7956040	12	*AMDHD1*	3a	0.81	15 tissues	12 tissues	4 tissues				4 hits	intronic
rs7956351	12	*AMDHD1*	2b	0.81	15 tissues	12 tissues	BLD, BLD, LNG		8 altered motifs		4 hits	intronic
rs75497549	12	*AMDHD1*	4	0.82	15 tissues	12 tissues	SKIN		GATA, RXRA, p300		4 hits	intronic
rs76989777	12	*AMDHD1*	3a	0.81	LIV, ADRL	BLD, MUS, SPLN	BLD, ADRL, KID		Hbp1, Mtf1		4 hits	intronic
rs7966871	12	*AMDHD1*	No Data	0.81	LIV, ADRL	BLD, MUS, SPLN	IPSC, LIV		CHOP::CEBPalpha, Ets		3 hits	intronic
rs7966688	12	*AMDHD1*	6	0.81	LIV, ADRL	BLD, MUS, SPLN			5 altered motifs		4 hits	intronic
rs7953691	12	*AMDHD1*	5	0.81	LIV, ADRL				4 altered motifs		4 hits	intronic

rs12579607	12	*AMDHD1*	5	0.81	LIV, ADRL				Nkx2, Nkx3, TCF4	8 hits	7 hits	intronic
rs35929186	12	*AMDHD1*	5	0.81		ESDR, ESC, LIV			5 altered motifs		4 hits	intronic
rs12582506	12	*AMDHD1*	No Data	0.81		ESDR, ESC, LIV			Brachyury, RXR:LXR, Sox	8 hits	7 hits	intronic
rs35756705	12	*AMDHD1*	1f	0.83		5 tissues				2 hits	4 hits	intronic
rs4762643	12	*AMDHD1*	4	0.83		5 tissues	IPSC, LIV	10 bound proteins	CTCF	10 hits	8 hits	intronic
rs1982138	12	*AMDHD1*	No Data	0.84					4 altered motifs	9 hits	8 hits	synonymous
rs4762256	12	*AMDHD1*	1f	0.84		IPSC, LIV	LIV	FOXA1, FOXA2, P300	Duxl, FXR, Pbx-1	10 hits	8 hits	intronic
rs11826	12	*AMDHD1*	6	1							5 hits	3'-UTR
rs4141977	20	*SNAI1*	4	0.95	20 tissues	6 tissues	7 tissues	POL2	Klf7			intronic
rs77090490	20	*SNAI1*	2b	0.95	15 tissues	11 tissues	SKIN, MUS, MUS		CDP, Pou2f2, Pou3f3			intronic
rs6091079	20	*SNAI1*	5	0.92	ESDR, ESC, FAT	16 tissues	12 tissues		LF-A1, Pax-6			intronic
rs1047920	20	*SNAI1*	4	1			7 tissues	EGR1	BDP1, LF-A1			3'-UTR
rs73276407	20	*SNAI1*	No Data	0.71		17 tissues	ESDR, ESDR, ESC		6 altered motifs			
rs16995019	20	*SNAI1*	5	0.8		14 tissues			HEN1			
rs11481333	20	*SNAI1*	5	0.83		8 tissues	ESDR		8 altered motifs			
rs73276408	20	*SNAI1*	No Data	0.83		8 tissues						
rs60578213	20	*SNAI1*	3a	0.83		8 tissues	6 tissues	USF1	GATA, Pax-4, Zfp740			
rs8116864	20	*SNAI1*	4	0.83		4 tissues	11 tissues	MAX	BCL			
rs73910366	20	*SNAI1*	5	0.8		10 tissues	6 tissues		4 altered motifs			
rs16995020	20	*SNAI1*	5	0.8		10 tissues	6 tissues		Ehf, Mef2, p53			
rs6091082	20	*SNAI1*	5	0.8		6 tissues			Fox, TATA			
rs77971891	20	*SNAI1*	6	0.8			LNG		6 altered motifs			
rs6020182	20	*SNAI1*	No Data	0.8		BLD			5 altered motifs			
rs7263384	20	*SNAI1*	No Data	0.8					RFX5, TATA			
rs6095732	20	*SNAI1*	6	0.8					YY1			
rs6020184	20	*SNAI1*	5	0.8		ESDR, PLCNT, LNG			STAT, Znf143			
rs7274092	20	*SNAI1*	5	0.8		5 tissues	ESDR		Hsf			
rs3886627	20	*SNAI1*	5	0.8		ESDR, BLD, LIV						
rs3886626	20	*SNAI1*	5	0.8		ESDR, BLD, LIV			5 altered motifs			
rs6091083	20	*SNAI1*	5	0.8	BLD	4 tissues	4 tissues		Pax-5			
rs6095733	20	*SNAI1*	6	0.78		BLD			7 altered motifs			
rs6091084	20	*SNAI1*	4	0.79		5 tissues	5 tissues	EBF1	Brachyury, Eomes			
rs6095734	20	*SNAI1*	No Data	0.79		4 tissues			E2A, SP2, p300			
rs56851604	20	*SNAI1*	5	0.79		ESDR, BLD	7 tissues		E2A, Pax-5, p300			
rs112694468	20	*SNAI1*	6	0.79		BLD			9 altered motifs			
rs78487338	20	*SNAI1*	5	0.77		7 tissues	IPSC, SKIN		LBP-1, SRF			

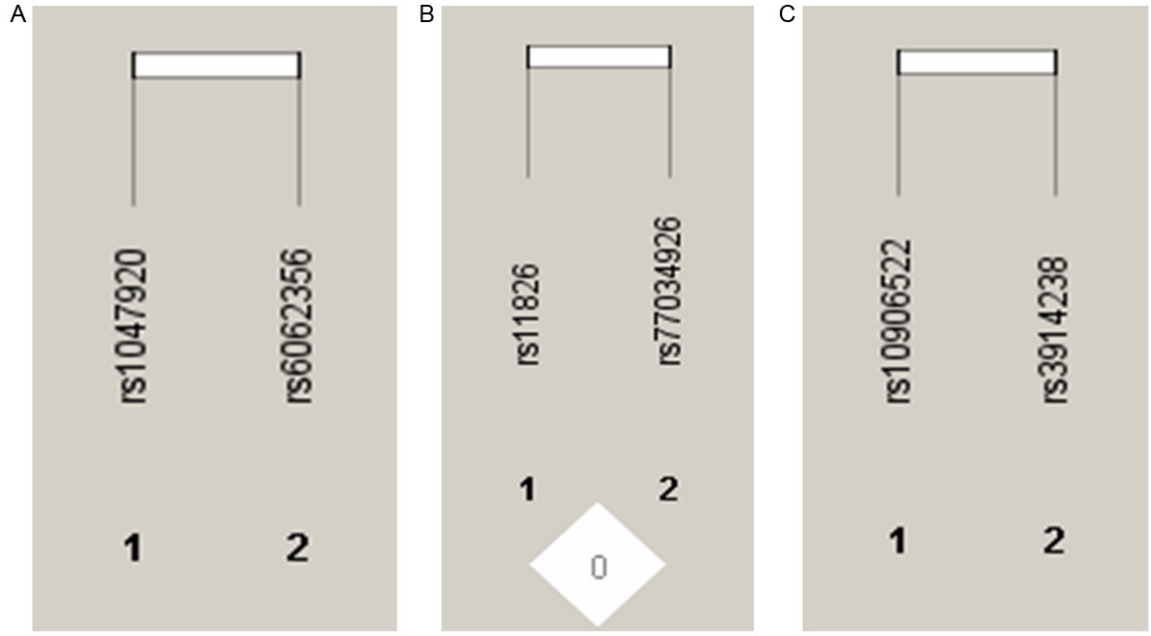
[a]RegulomeDB (http://www.regulomedb.org); [b]HaploReg v4.1 (http://archive.broadinstitute.org/mammals/haploreg/haploreg.php). Abbreviations: Chr, chromosome; dbSNP funcannot, dbSNP function annotation; SNP, single-nucleotide polymorphism.



Supplementary Figure 4. Functional prediction of SNPs in the ENCODE Project. (A) Location and functional prediction of SNPs rs1047920, (B) Location and functional prediction of SNPs rs11826. (C) Location and functional prediction of SNPs rs3914238. Abbreviations: SNP, single nucleotide polymorphism.



Supplementary Figure 5. Correlation between *SNAI1* mRNA expression and rs1047920 genotype in 373 Europeans from the 1000 Genomes Project in the (A) additive model, (B) dominant model and (C) recessive model; Correlation between *AMDHD1* mRNA expression and rs11826 genotype in 373 Europeans from the 1000 Genomes Project in the (D) additive model, (E) dominant model and (F) recessive model; Correlation between *CUBN* mRNA expression and rs3914238 genotype in 373 Europeans from the 1000 Genomes Project in the (G) additive model, (H) dominant model and (I) recessive model; Correlation between *SNAI1* mRNA expression and rs1047920 genotype in the GTEx Project (J) Whole Blood (K) Breast tissue; Correlation between *CUBN* mRNA expression and rs3914238 genotype in the GTEx Project (L) Whole Blood (M) Breast tissue. Abbreviations: *AMDHD1*, amidohydrolase domain containing 1; *CUBN*, cubilin; GTEx: Genotype-Tissue Expression; *SNAI1*, snail family transcriptional repressor 1.



Supplementary Figure 6. LD analysis between independent SNPs and the previously reported SNPs in the same location (PMID: 29059683). (A) rs1047920, (B) rs11826 and (C) rs3914238.