Supplementary Information

Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3, and 15q26.1

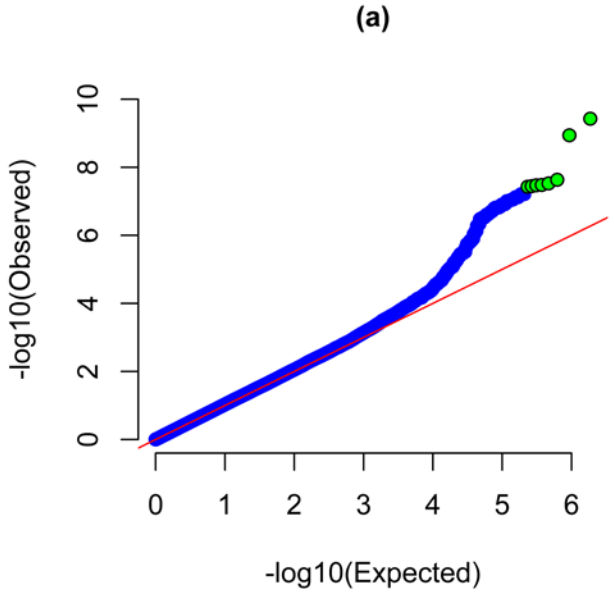
Qiuyin Cai, Ben Zhang, Hyuna Sung, Siew-Kee Low, Sun-Seog Kweon, Wei Lu, Jiajun Shi, Jirong Long, Wanqing Wen, Ji-Yeob Choi, Dong-Young Noh, Chen-Yang Shen, Keitaro Matsuo, Soo-Hwang Teo, Mi Kyung Kim, Ui Soon Khoo, Motoki Iwasaki, Mikael Hartman, Atsushi Takahashi, Kyota Ashikawa, Koichi Matsuda, Min-Ho Shin, Min Ho Park, Ying Zheng, Yong-Bing Xiang, Bu-Tian Ji, Sue K. Park, Pei-Ei Wu, Chia-Ni Hsiung, Hidemi Ito, Yoshio Kasuga, Peter Kang, Shivaani Mariapun, Sei Hyun Ahn, Han Sung Kang, Kelvin Y. K. Chan, Ellen P. S. Man, Hiroji Iwata, Shoichiro Tsugane, Hui Miao, Jiemin Liao, Yusuke Nakamura, Michiaki Kubo, The DRIVE GAME-ON Consortium, Ryan J. Delahanty, Yanfeng Zhang, Bingshan Li, Chun Li, Yu-Tang Gao, Xiao-Ou Shu, Daehee Kang, Wei Zheng

Corresponding author contact information:

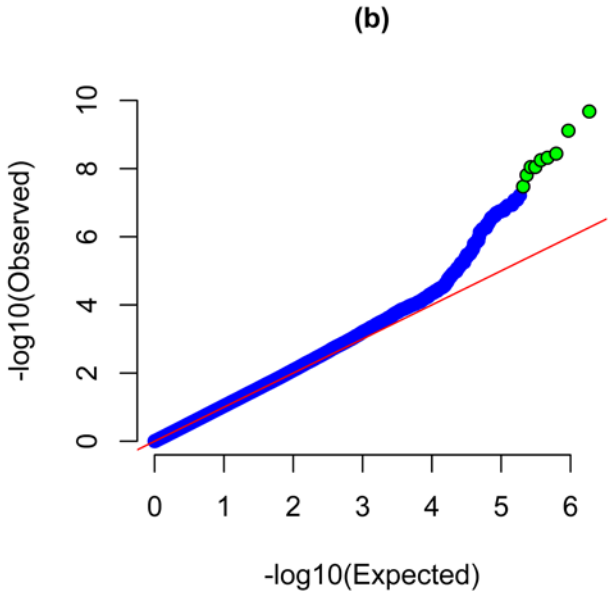
Qiuyin Cai, M.D., Ph.D.
Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center
Vanderbilt University School of Medicine
1161 21st Avenue South
Nashville, TN 37232
Phone: (615) 936-1351
Fax: (615)936-8291
E-mail: qiuyin.cai@vanderbilt.edu

Supplementary Information
Table of Contents

A. SBCGS-1

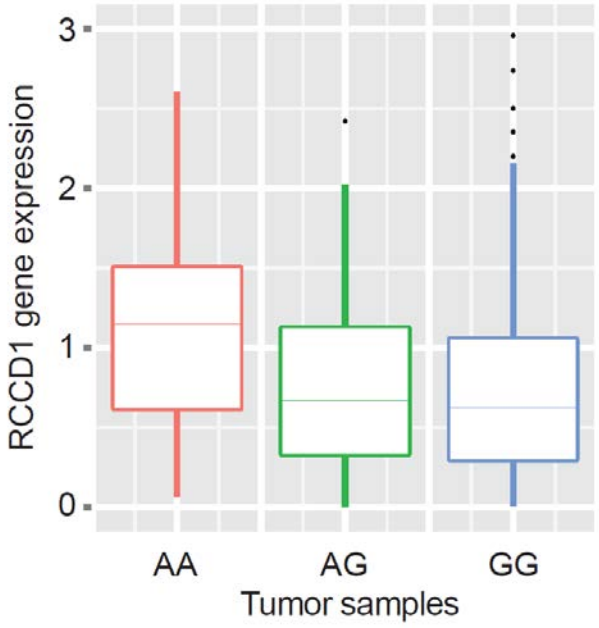


B. SeBCS1

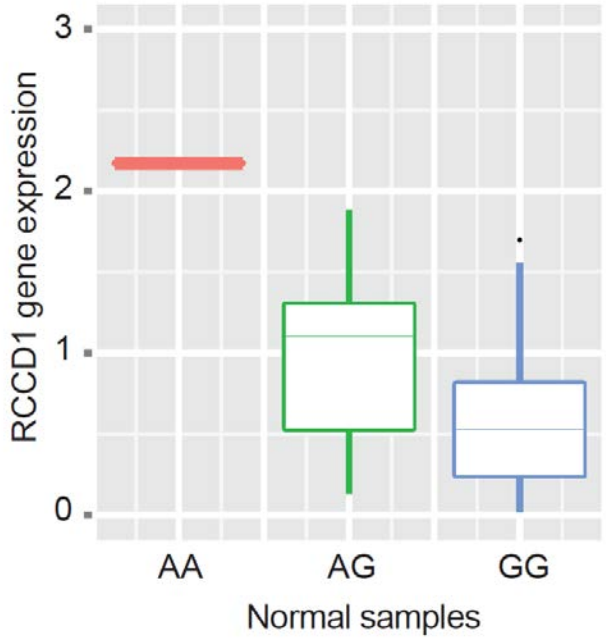


Supplementary Figure 1: Quantile-quantile plot of *P* values in –log10 scale among Stage 1 samples. Panel A, Chinese GWAS (SBCGS-1); Panel B, Korean GWAS (SeBCS1).

A: Tumor Tissue (P = 3.6 x 10^{-4})

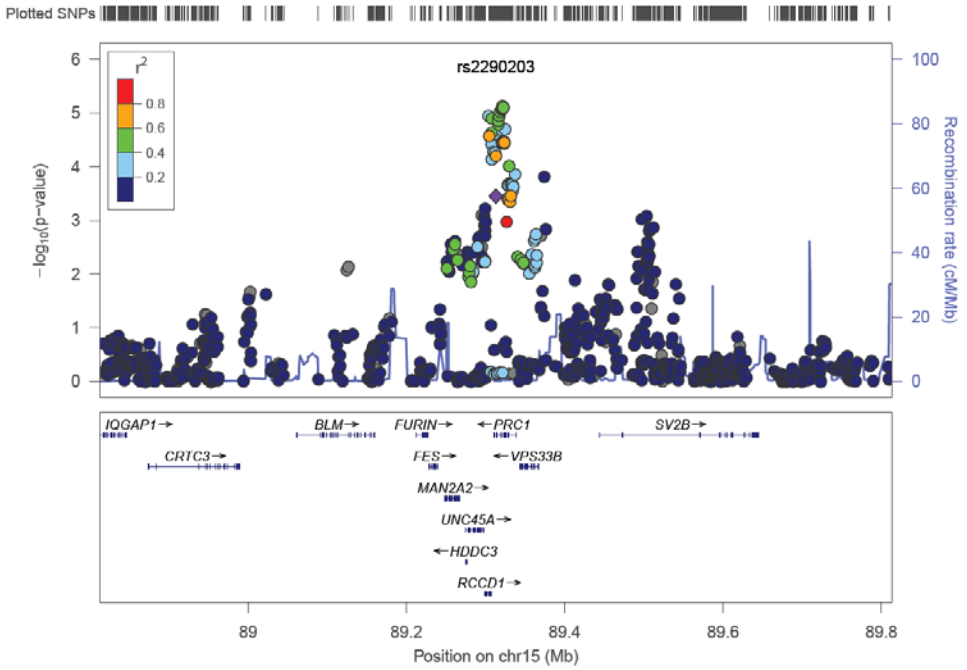


B: Adjacent Normal Tissue (P = 0.007)

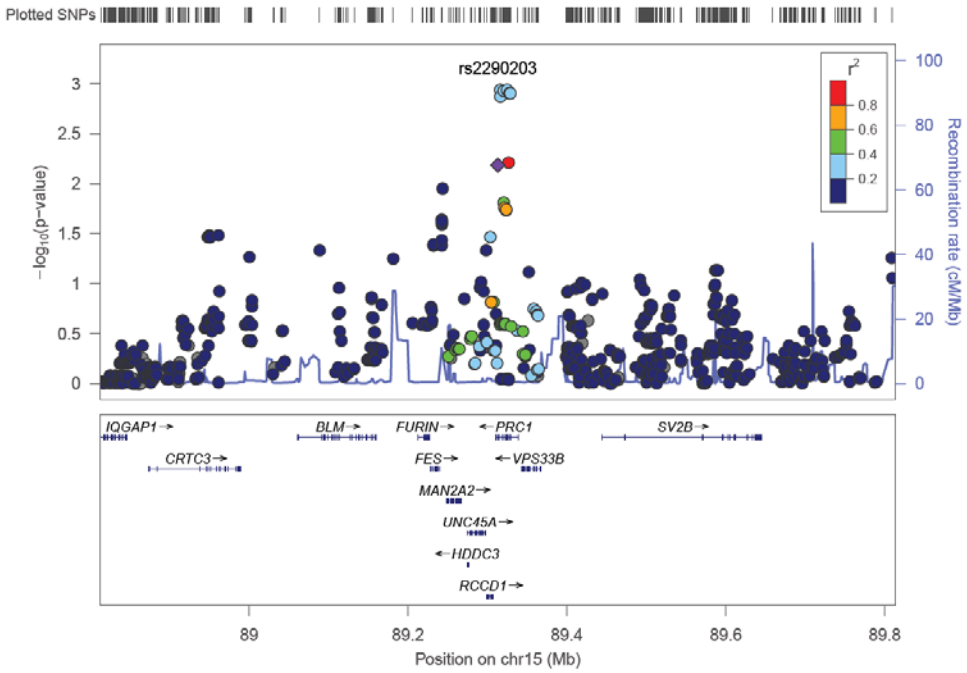


Supplementary Figure 2: Association of rs2290203 (15q26.1) with *RCCD1* gene expression. Data from 458 tumor tissues and 66 adjacent normal tissue samples included in TCGA. Panel A: tumor tissues; Panel B: adjacent normal tissue.

A: Tumor Tissue



B: Adjacent Normal Tissue



Supplementary Figure 3. **Regional plots of cis-eQTL association results for rs2290203 (15q26.1) with the *RCCD1* gene in breast cancer tissue included in TCGA.** Panel A: tumor tissue; Panel B: adjacent normal tissue. The LDs among SNPs located within ±500 kb of rs2290203 were calculated based on European-ancestry population data from the 1000 Genomes Project.

Supplementary Table 1. Selected characteristics of studies participating in the Asia Breast Cancer Consortium

Study	Cases	Controls	Population	Study design[a]	Age (years)[b]	Postmenopausal (%)[c]	ER+ (%)[d]
Stage 1	**5,113**	**4,337**					
SBCGS-1	2,867	2,285	Chinese	Population-based	51/50	43/42	65
SeBCS1	2,246	2,052	Korean	Hospital-based	48/51	36/56	63
Stage 2 - SBCGS-2	**3,472**	**3,595**	Chinese	Population-based	53/53	50/53	64
Stage 3	**14,195**	**16,249**					
Taiwan	1,066	1,065	Chinese	Hospital-based	52/48	52/40	66
Hong Kong	491	642	Chinese	Hospital-based	46/46	50/42	71
HCES-Br	3,387	3,186	Korean	Population-based	50/57	45/81	64
KOHBRA/KoGES	1,397	3,209	Korean	Hospital-based	40/50	23/NA	63
SeBCS2	777	1,104	Korean	Hospital-based	48/48	36/37	63
Korea-NCC	505	504	Korean	Hospital-based	49/49	50/45	65
BBJ1	2,642	2,099	Japanese	Hospital-based	57/56	79/72	63
BBJ2	2,885	3,395	Japanese	Hospital-based	60/44	76/41	79
Nagoya	644	644	Japanese	Hospital-based	51/51	49/49	73
Nagano	401	401	Japanese	Hospital-based	54/54	55/65	75
Total	**22,780**	**24,181**					

Abbreviations: SBCGS, Shanghai Breast Cancer Genetics Study (includes participants from the Shanghai Breast Cancer Study, Shanghai Breast Cancer Survival Study, Shanghai Endometrial Cancer Study, and Shanghai Women's Health Study); SeBCS, Seoul Breast Cancer Study; HCES-Br, Hwasun Cancer Epidemiology Study-Breast; KOHBRA, Korean Hereditary Breast Cancer; KoGES, Korea Genome Epidemiology Study; BBJ, The Biobank Japan Project; NA, Not available; ER, Estrogen receptor.

[a] Case-control study design was used.

[b] Mean age of cases/controls with available data.

[c] Proportion of postmenopausal status of cases/controls with available data.

[d] Among cases with ER data.

Supplementary Table 2. Results for the 50 SNPs analyzed in all three stages

SNP	Chr.	Position[a]	Alleles[b]	RAF[c]	Stage 1		Stage 2		Stage 3		Combined
					OR (95% CI)[d]	*P*[e]	OR (95% CI)[d]	*P*[e]	OR (95% CI)[d]	*P*[e]	*P*[e]
rs10889256	1	62042943	**G**/A	0.59	1.09 (1.03-1.16)	0.004	1.05 (0.98-1.12)	0.186	1.04 (1.00-1.08)	0.031	3.24×10^{-4}
rs17131635	1	84608287	**C**/T	0.45	1.09 (1.02-1.17)	0.007	1.07 (1.00-1.14)	0.043	0.99 (0.93-1.04)	0.606	0.029
rs12047887	1	110945530	**A**/G	0.14	1.12 (1.03-1.22)	0.007	1.15 (1.05-1.27)	0.003	1.02 (0.95-1.10)	0.612	7.57×10^{-4}
rs4951011	1	202032954	**G**/A	0.28	1.09 (1.02-1.17)	0.007	1.10 (1.02-1.18)	0.011	1.08 (1.05-1.12)	1.0×10^{-5}	8.82×10^{-9}
rs11693188	2	122269580	**A**/C	0.78	1.10 (1.02-1.18)	0.006	1.06 (0.98-1.15)	0.169	1.04 (1.00-1.08)	0.085	0.002
rs6430421	2	133934020	**A**/G	0.37	1.08 (1.02-1.15)	0.011	1.09 (1.02-1.17)	0.011	1.03 (0.97-1.08)	0.363	8.35×10^{-4}
rs9827291	3	31329880	**G**/A	0.79	1.15 (1.05-1.25)	7.8×10^{-4}	1.11 (1.03-1.21)	0.009	1.02 (0.97-1.08)	0.471	0.001
rs12635118	3	35506554	**G**/T	0.55	1.08 (1.01-1.14)	0.012	1.11 (1.04-1.19)	0.002	1.01 (0.95-1.06)	0.822	0.002
rs4234645	3	51902638	**C**/T	0.15	1.09 (1.01-1.19)	0.031	1.13 (1.03-1.25)	0.010	1.04 (0.97-1.12)	0.262	0.001
rs12233375	3	78566459	**C**/A	0.72	1.08 (1.01-1.16)	0.025	1.09 (1.01-1.18)	0.026	0.99 (0.94-1.05)	0.804	0.026
rs9828276	3	144723903	**T**/G	0.54	1.06 (1.00-1.12)	0.064	1.05 (0.98-1.12)	0.131	1.05 (1.01-1.09)	0.014	6.25×10^{-4}
rs10010358	4	5456706	**A**/C	0.16	1.09 (1.01-1.18)	0.033	1.17 (1.07-1.27)	6.2×10^{-4}	1.03 (0.98-1.08)	0.328	0.001
rs2350804	4	66758533	**C**/T	0.15	1.12 (1.04-1.21)	0.004	1.12 (1.03-1.23)	0.011	0.94 (0.87-1.01)	0.069	0.076
rs12512640	4	121831277	**A**/G	0.55	1.09 (1.02-1.16)	0.012	1.08 (1.01-1.16)	0.019	1.03 (0.97-1.08)	0.370	0.001
rs11929741	4	177144874	**T**/C	0.21	1.09 (1.02-1.17	0.017	1.12 (1.04-1.21)	0.005	0.97 (0.91-1.03)	0.350	0.022
rs6555218	5	3750286	**T**/C	0.48	1.06 (1.00-1.13)	0.051	1.10 (1.03-1.17)	0.005	1.02 (0.96-1.07)	0.577	0.003
rs7731095	5	3770907	**T**/C	0.31	1.11 (1.02-1.21)	0.012	1.10 (1.02-1.18)	0.009	1.03 (0.99-1.09)	0.163	5.56×10^{-4}
rs2937532	5	36437871	**A**/G	0.57	1.08 (1.02-1.15)	0.007	1.08 (1.01-1.15)	0.023	1.00 (0.94-1.05)	0.858	0.009
rs13179185	5	89650778	**A**/T	0.86	1.12 (1.03-1.21)	0.010	1.13 (1.03-1.24)	0.008	1.01 (0.93-1.10)	0.793	0.002
rs10474352	5	90767981	**C**/T	0.48	1.09 (1.03-1.17)	0.006	1.12 (1.05-1.20)	7.1×10^{-4}	1.08 (1.04-1.12)	1.9×10^{-5}	1.67×10^{-9}
rs3909301	5	125416129	**C**/A	0.20	1.10 (1.02-1.18)	0.015	1.13 (1.04-1.22)	0.003	0.98 (0.91-1.04)	0.478	0.010
rs31519	5	135314336	**A**/G	0.65	1.10 (1.03-1.16)	0.005	1.09 (1.02-1.17)	0.014	0.99 (0.93-1.04)	0.657	0.007
rs698287	5	142285615	**A**/G	0.73	1.08 (1.01-1.15)	0.028	1.10 (1.02-1.19)	0.013	1.02 (0.96-1.08)	0.513	0.003
rs9444163	6	85130450	**A**/C	0.81	1.10 (1.02-1.19)	0.008	1.13 (1.04-1.23)	0.004	1.06 (1.02-1.11)	0.008	1.11×10^{-5}
rs41336	7	28671303	**T**/C	0.48	1.09 (1.03-1.16)	0.005	1.10 (1.03-1.17)	0.006	1.02 (0.97-1.07)	0.401	5.18×10^{-4}
rs16874907	7	136445424	**A**/G	0.67	1.08 (1.02-1.15)	0.013	1.09 (1.01-1.17)	0.023	1.01 (0.96-1.07)	0.639	0.003
rs3779853	8	2003460	**G**/T	0.68	1.10 (1.03-1.17)	0.007	1.11 (1.03-1.19)	0.004	0.98 (0.93-1.04)	0.504	0.005
rs277778	8	94421978	**G**/A	0.76	1.13 (1.06-1.21)	4.4×10^{-4}	1.09 (1.01-1.19)	0.034	1.01 (0.96-1.07)	0.585	0.001
rs2025192	9	7969082	**T**/C	0.35	1.09 (1.02-1.16)	0.007	1.07 (1.00-1.15)	0.045	1.02 (0.97-1.08)	0.398	0.003

rs7047037	9	113769674	**C**/T	0.27	1.10 (1.03-1.18)	0.006	1.09 (1.02-1.18)	0.017	0.99 (0.93-1.05)	0.689	0.013
rs7912035	10	3697315	**C**/T	0.49	1.09 (1.03-1.16)	0.002	1.07 (1.00-1.15)	0.037	0.98 (0.93-1.04)	0.577	0.018
rs12779590	10	63839924	**T**/C	0.67	1.08 (1.01-1.14)	0.012	1.09 (1.02-1.17)	0.012	1.03 (0.98-1.08)	0.187	7.56×10^{-4}
rs11033442	11	36103284	**A**/G	0.33	1.11 (1.04-1.18)	8.6×10^{-4}	1.12 (1.05-1.21)	0.001	1.00 (0.95-1.04)	0.858	0.002
rs1021311	12	29005155	**C**/A	0.64	1.15 (1.06-1.25)	0.001	1.10 (1.02-1.18)	0.008	1.03 (1.00-1.07)	0.054	1.06×10^{-4}
rs7295078	12	81066712	**T**/C	0.25	1.08 (1.01-1.16)	0.019	1.15 (1.07-1.25)	2.4×10^{-4}	0.98 (0.92-1.04)	0.538	0.005
rs9670198	13	30242453	**A**/T	0.17	1.11 (1.03-1.20)	0.004	1.14 (1.04-1.25)	0.005	1.02 (0.95-1.10)	0.507	8.91×10^{-4}
rs2406690	13	47387659	**G**/A	0.38	1.09 (1.02-1.16)	0.007	1.10 (1.03-1.17)	0.006	0.97 (0.91-1.02)	0.210	0.031
rs8010158	14	38066347	**T**/G	0.60	1.09 (1.02-1.17)	0.007	1.09 (1.02-1.16)	0.017	1.01 (0.96-1.07)	0.714	0.005
rs17123027	14	86682035	**T**/C	0.10	1.16 (1.05-1.28)	0.001	1.15 (1.03-1.29)	0.011	0.96 (0.89-1.05)	0.396	0.019
rs272797	15	51418580	**G**/A	0.71	1.08 (1.01-1.15)	0.016	1.13 (1.05-1.21)	0.001	1.02 (0.96-1.08)	0.501	5.06×10^{-4}
rs2970376	15	58024217	**C**/T	0.44	1.10 (1.03-1.19)	0.009	1.14 (1.06-1.22)	1.5×10^{-4}	1.03 (0.99-1.07)	0.146	8.66×10^{-5}
rs2290203	15	89313071	**G**/A	0.50	1.08 (1.02-1.14)	0.012	1.19 (1.10-1.30)	5.0×10^{-5}	1.06 (1.03-1.10)	2.4×10^{-4}	4.25×10^{-8}
rs11074880	16	27731110	**G**/T	0.08	1.13 (1.02-1.26)	0.006	1.11 (0.99-1.24)	0.073	1.08 (1.02-1.14)	0.012	1.54×10^{-4}
rs1122322	16	72333289	**G**/A	0.58	1.06 (1.00-1.13)	0.029	1.06 (0.99-1.14)	0.070	1.04 (1.00-1.07)	0.027	7.21×10^{-4}
rs7359598	17	38150996	**C**/T	0.43	1.09 (1.03-1.16)	0.003	1.11 (1.03-1.18)	0.003	1.03 (0.98-1.09)	0.292	9.73×10^{-5}
rs1078523	17	38166286	**A**/G	0.43	1.12 (1.04-1.19)	0.002	1.06 (0.99-1.13)	0.082	1.04 (1.01-1.08)	0.020	9.516×10^{-5}
rs11082321	18	19030649	**A**/G	0.21	1.09 (1.01-1.17)	0.020	1.10 (1.02-1.20)	0.017	1.08 (1.04-1.12)	2.0×10^{-4}	6.77×10^{-7}
rs1874381	18	32376408	**G**/A	0.51	1.08 (1.02-1.15)	0.009	1.09 (1.02-1.17)	0.009	0.98 (0.93-1.03)	0.479	0.018
rs16978325	18	41117753	**G**/A	0.94	1.21 (1.06-1.37)	0.005	1.21 (1.05-1.39)	0.010	0.99 (0.89-1.11)	0.897	0.005
rs5756386	22	20681183	**A**/G	0.67	1.08 (1.01-1.15)	0.016	1.07 (1.00-1.15)	0.055	1.05 (1.01-1.08)	0.014	1.30×10^{-4}

Abbreviations: Chr., Chromosome; RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] Chromosome position (bp) based on NCBI Human Genome Build 36.

[b] Risk/reference alleles; risk allele shown in bold.

[c] Risk allele frequency among all controls in Stages 1 to 3.

[d] Per-allele OR (95% CI) was adjusted for age and principal components in each study; summary OR (95% CI) was obtained from fixed-effect meta-analysis in each stage.

[e] Derived from a weighted z-statistic–based meta-analysis.

Supplementary Table 3. Association of breast cancer risk with newly identified risk variants by study population

SNP (Alleles[a])	Population	Cases	Controls	RAF[b]	Per-allele association		$P_{heterogeneity}$[e]
					OR (95% CI)[c]	P[d]	
rs4951011 (**G**/A)	Chinese	7,401	6,901	0.296	1.09 (1.04-1.15)	5.39×10^{-4}	0.756
	Japanese	6,568	6,515	0.300	1.07 (1.02-1.13)	0.010	
	Korean	8,240	9,951	0.261	1.10 (1.05-1.15)	1.01×10^{-4}	
rs10474352 (**C**/T)	Chinese	7,358	6,836	0.521	1.12 (1.06-1.17)	6.63×10^{-6}	0.082
	Japanese	6,557	6,512	0.439	1.11 (1.06-1.17)	1.80×10^{-5}	
	Korean	6,863	6,801	0.484	1.04 (0.99-1.09)	0.110	
rs2290203 (**G**/A)	Chinese	7,408	6,920	0.495	1.10 (1.05-1.16)	2.75×10^{-4}	0.182
	Japanese	6,569	6,514	0.525	1.04 (0.99-1.09)	0.120	
	Korean	8,237	9,957	0.496	1.09 (1.05-1.14)	2.43×10^{-5}	

Abbreviations: RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] Risk/reference alleles; risk allele shown in bold.

[b] Risk allele frequency among all controls in Stages 1 to 3.

[c] Per-allele OR (95% CI) was adjusted for age and principal components in each study; summary OR (95% CI) was obtained from fixed-effect meta-analysis in each stage.

[d] Derived from a weighted z-statistic–based meta-analysis.

[e] *P* for heterogeneity between different populations was calculated using a Cochran's *Q* test.

Supplementary Table 4. Association of breast cancer risk with newly identified risk variants by estrogen receptor status

SNP (Alleles[a])	ER status	Cases	Controls	RAF[b]	Per-allele association		$P_{heterogeneity}$[e]
					OR (95% CI)[c]	P[d]	
rs4951011 (**G**/A)	Positive	12,097	23,367	0.300	1.09 (1.05-1.12)	2.74×10^{-6}	0.969
	Negative	6,215	23,367	0.299	1.09 (1.04-1.13)	2.68×10^{-4}	
rs10474352 (**C**/T)	Positive	11,560	20,149	0.511	1.11 (1.07-1.15)	8.81×10^{-10}	0.085
	Negative	5,907	20,149	0.503	1.06 (1.01-1.10)	0.009	
rs2290203 (**G**/A)	Positive	12,101	23,391	0.524	1.09 (1.06-1.13)	1.36×10^{-7}	0.253
	Negative	6,213	23,391	0.514	1.06 (1.02-1.10)	0.008	

Abbreviations: RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] Risk/reference alleles; risk allele shown in bold.

[b] Risk allele frequency among all cases in Stages 1 to 3.

[c] Per-allele OR (95% CI) was adjusted for age and principal components in each study; summary OR (95% CI) was obtained from fixed-effect meta-analysis in each stage.

[d] Derived from a weighted z-statistic–based meta-analysis.

[e] *P* for heterogeneity between ER-positive and ER-negative breast cancer was calculated using a Cochran's *Q* test.

Supplementary Table 5. Association of breast cancer risk with newly identified risk variants among women of European ancestry (the DRIVE GAME-ON Consortium)[a]

SNP	Alleles[b]	RAF[c]	Per-allele association		*P* for Heterogeneity[e]
			OR (95% CI)[d]	*P*	
rs4951011	**G**/A	0.164	1.05 (1.01-1.11)	0.030	0.260
rs10474352	**C**/T	0.858	1.08 (1.02-1.14)	0.004	0.727
rs2290203	**G**/A	0.806	1.06 (1.01-1.10)	0.010	0.399

Abbreviations: RAF, risk allele frequency; OR, odds ratio; CI, confidence interval.

[a] Based on a meta-analysis of 12 breast cancer GWAS, including a total of 16,003 cases and 41,335 controls.

[b] Risk/reference alleles among East Asians; risk allele shown in bold.

[c] Risk allele frequency among controls.

[d] Per-allele OR (95% CI), adjusted for study and principal components.

[e] *P* for heterogeneity between Europeans and East Asians was calculated using a Cochran's *Q* test.

Supplementary Table 6. Gene expression levels in breast tumor tissue and paired adjacent normal tissue from 87 breast cancer cases included in TCGA

Gene	Chromosome Location	Log2 fold changes (tumor vs normal)	*P*[a]
ZC3H11A	*1q32.1*	0.14	0.0049
ARRDC3	*5q14.3*	-1.24	1.88×10^{-18}
PRC1	*15q26.1*	2.33	4.62×10^{-30}

[a] *P*-values were false discovery rate–adjusted using the Benjamini and Hochberg method.

Supplementary Table 7. Functional annotation of SNPs correlated with newly-identified risk variants (r^2 >0.8) using data from ENCODE

Locus	SNP[a]	Chr.	Position[b]	Promoter histone marks[c]	Enhancer histone marks[d]	DNAse[e]	Proteins bound[f]	Motifs changed[g]	Gene	Annotation
1q32.1	rs3737971	1	202030961	6 cell types	K562, NHEK, HMEC		BAF155	5 altered motifs	443bp 5' of ZC3H11A	
1q32.1	**rs4951011**	1	**202032954**	**6 cell types**	**HMEC, NHEK, K562**	**MCF-7,HCPEpiC,HMVEC-dBl-Neo**		**HNF1**	**ZBED6**	**intronic**
1q32.1	rs7552670	1	202034951		K562	GM12878,NHEK		DMRT1,DMRT7,Sox	ZBED6	missense
1q32.1	rs7540041	1	202036155	GM12878	K562, NHEK		GATA1	Pax-6	ZBED6	missense
1q32.1	rs10494844	1	202036514		K562, NHEK, GM12878			Hmbox1	ZC3H11A	intronic
1q32.1	rs6662925	1	202039412			NHEK		Hic1,Sin3Ak-20,Zbtb3	ZC3H11A	intronic
1q32.1	rs7532505	1	202041416					PPAR	ZC3H11A	intronic
1q32.1	rs6685918	1	202043932					HNF4	ZC3H11A	intronic
1q32.1	rs35762392	1	202043993					8 altered motifs	ZC3H11A	intronic
1q32.1	rs7546400	1	202044422					Foxa,Pou5f1	ZC3H11A	intronic
1q32.1	rs7542453	1	202044800					4 altered motifs	ZC3H11A	intronic
1q32.1	rs61827262	1	202045997						ZC3H11A	intronic
1q32.1	rs61827263	1	202046751						ZC3H11A	intronic
1q32.1	rs61827264	1	202050124					4 altered motifs	ZC3H11A	intronic
1q32.1	rs59366140	1	202051207					Nkx2	ZC3H11A	intronic
1q32.1	rs4951259	1	202051792					5 altered motifs	ZC3H11A	intronic
1q32.1	rs3753590	1	202054113					Hand1,TCF4	ZC3H11A	intronic
1q32.1	rs12403727	1	202054976					16 altered motifs	ZC3H11A	intronic
1q32.1	rs12403777	1	202055379					5 altered motifs	ZC3H11A	intronic
1q32.1	rs67642275	1	202056219					Foxp1,Pou2f2,RREB-1	ZC3H11A	intronic
1q32.1	rs12403366	1	202056864					9 altered motifs	ZC3H11A	intronic
1q32.1	rs12402315	1	202056900						ZC3H11A	intronic
1q32.1	rs12403365	1	202059205			12 cell types		Irx	ZC3H11A	intronic
1q32.1	rs61827266	1	202061376					SZF1-1,Spz1,ZBTB33	ZC3H11A	intronic
1q32.1	rs6673230	1	202062320					4 altered motifs	ZC3H11A	intronic
1q32.1	rs61559537	1	202062549					TAL1,Zfp691,p300	ZC3H11A	intronic
1q32.1	rs12068113	1	202063380						ZC3H11A	intronic
1q32.1	rs111509603	1	202066763		HSMM	18 cell types		CEBPB,CEBPD	ZC3H11A	intronic
1q32.1	rs6689679	1	202073076					Irx	ZC3H11A	intronic

1q32.1	rs10494847	1	202086063					5 altered motifs	ZC3H11A	intronic
5q14.3	rs11948075	5	90716335	5 cell types	4 cell types	5 cell types	CFOS,MAFF	4 altered motifs	1.4kb 5' of ARRDC3	intronic
5q14.3	rs7705465	5	90716705		K562	104 cell types	13 bound proteins	Hltf	1.8kb 5' of ARRDC3	intronic
5q14.3	rs746637	5	90717393					GR	2.5kb 5' of ARRDC3	intronic
5q14.3	rs889213	5	90717907			27 cell types	CTCF	4 altered motifs	3kb 5' of ARRDC3	intronic
5q14.3	rs2367353	5	90722324		K562			9 altered motifs	7.4kb 5' of ARRDC3	intronic
5q14.3	rs11950530	5	90724496					CDP,CEBPA	9.6kb 5' of ARRDC3	intronic
5q14.3	rs10058025	5	90724808					HNF1,Irf,Pou2f2	9.9kb 5' of ARRDC3	intronic
5q14.3	rs10073657	5	90725505					6 altered motifs	11kb 5' of ARRDC3	intronic
5q14.3	rs6869717	5	90730553					Dobox4	16kb 5' of ARRDC3	intronic
5q14.3	rs6860948	5	90732053					5 altered motifs	17kb 5' of ARRDC3	intronic
5q14.3	rs11959525	5	90739596					STAT	25kb 5' of ARRDC3	intronic
5q14.3	rs1477349	5	90743529					Bbx,Pou2f2,Pou3f2	29kb 5' of ARRDC3	intronic
5q14.3	rs1477348	5	90743579					YY1,p53	29kb 5' of ARRDC3	intronic
5q14.3	rs11955120	5	90747775					7 altered motifs	33kb 5' of ARRDC3	intronic
5q14.3	rs12654787	5	90749015		NHLF	19 cell types		MZF1::1-4	34kb 5' of ARRDC3	intronic
5q14.3	rs10063683	5	90750351			WI-38		Foxc1,PRDM1,Pbx-1	35kb 5' of ARRDC3	intronic
5q14.3	rs10045280	5	90753431					AhR::Arnt,Arnt	38kb 5' of ARRDC3	
5q14.3	rs6867219	5	90753764			SK-N-MC		4 altered motifs	39kb 5' of ARRDC3	
5q14.3	rs74914393	5	90753765			SK-N-MC		4 altered motifs	39kb 5' of ARRDC3	
5q14.3	rs12110065	5	90754230					4 altered motifs	39kb 5' of ARRDC3	
5q14.3	rs7708782	5	90754273					Hand1	39kb 5' of ARRDC3	
5q14.3	rs6452953	5	90758790					DMRT2,INSM1,Pou2f2	44kb 5' of ARRDC3	
5q14.3	rs6862125	5	90760354					11 altered motifs	45kb 5' of ARRDC3	
5q14.3	rs2004483	5	90761961					GATA,Maf,Pitx2	47kb 5' of ARRDC3	
5q14.3	rs35945986	5	90762205					9 altered motifs	47kb 5' of ARRDC3	

5q14.3	rs2839728	5	90763168			Pou2f2	48kb 5' of ARRDC3
5q14.3	rs60513756	5	90763448			17 altered motifs	49kb 5' of ARRDC3
5q14.3	rs12651840	5	90765348		H7-hESC	AP-1,Sox	50kb 5' of ARRDC3
5q14.3	rs2052553	5	90765662		Osteobl,GM12864	Rhox11	51kb 5' of ARRDC3
5q14.3	rs2059207	5	90767687				53kb 5' of ARRDC3
5q14.3	rs2059208	5	90767821			5 altered motifs	53kb 5' of ARRDC3
5q14.3	**rs10474352**	5	**90767981**		**Huh-7,HMVEC-dLy-Ad**	**AP-1,VDR**	**53kb 5' of ARRDC3**
5q14.3	rs10474353	5	90768041		HMVEC-dLy-Ad	Pou5f1,SIX5,Znf143	53kb 5' of ARRDC3
5q14.3	rs6867091	5	90770089			AP-3,DMRT1	55kb 5' of ARRDC3
5q14.3	rs6452954	5	90771448			HEY1	57kb 5' of ARRDC3
5q14.3	rs12655973	5	90772293			TCF4	57kb 5' of ARRDC3
5q14.3	rs9293568	5	90773601			CHOP::CEBPalpha,Evi-1	59kb 5' of ARRDC3
5q14.3	rs10041402	5	90774704	HSMM, NHLF	7 cell types	Gfi1,PLZF	60kb 5' of ARRDC3
5q14.3	rs10054820	5	90776162			STAT	61kb 5' of ARRDC3
5q14.3	rs9686361	5	90776173			ZID	61kb 5' of ARRDC3
5q14.3	rs2367526	5	90776281			4 altered motifs	61kb 5' of ARRDC3
5q14.3	rs2887007	5	90776287			4 altered motifs	61kb 5' of ARRDC3
5q14.3	rs202184387	5	90777222			41 altered motifs	62kb 5' of ARRDC3
5q14.3	rs200206153	5	90777225			37 altered motifs	62kb 5' of ARRDC3
5q14.3	rs10073935	5	90777944			4 altered motifs	63kb 5' of ARRDC3
5q14.3	rs10059163	5	90778200			COMP1	63kb 5' of ARRDC3
5q14.3	rs1035477	5	90779081			PLAG1	64kb 5' of ARRDC3
5q14.3	rs1964292	5	90779364			Hand1	64kb 5' of ARRDC3
5q14.3	rs1895448	5	90779610			TCF11::MafG	65kb 5' of ARRDC3
5q14.3	rs1895449	5	90779862			7 altered motifs	65kb 5' of ARRDC3
5q14.3	rs4382199	5	90779984			Pou5f1	65kb 5' of ARRDC3
5q14.3	rs985434	5	90780265			HNF4,RREB-1,TCF4	65kb 5' of ARRDC3

5q14.3	rs7720132	5	90780861			HPAF		4 altered motifs	66kb 5' of ARRDC3
5q14.3	rs1035476	5	90781732					AP-1,Hdx,SRF	67kb 5' of ARRDC3
5q14.3	rs6865495	5	90782180					13 altered motifs	67kb 5' of ARRDC3
5q14.3	rs6866191	5	90782608					GATA	68kb 5' of ARRDC3
5q14.3	rs140115903	5	90786306					10 altered motifs	71kb 5' of ARRDC3
5q14.3	rs12656350	5	90787099				BATF		72kb 5' of ARRDC3
5q14.3	rs10052492	5	90795686			LNCaP		10 altered motifs	81kb 5' of ARRDC3
5q14.3	rs933688	5	90798504			LNCaP		LBP-1,Pou2f2	84kb 5' of ARRDC3
5q14.3	rs12188467	5	90801388	Huvec, K562, HSMM		PANC-1	GR,JUNB,JUND	Cdx,Pbx-1,p300	86kb 5' of ARRDC3
5q14.3	rs332544	5	90814604					BAF155,Znf143	100kb 5' of ARRDC3
5q14.3	rs186717	5	90816223					9 altered motifs	101kb 5' of ARRDC3
5q14.3	rs332529	5	90825226			Melano		10 altered motifs	110kb 5' of ARRDC3
15q26.1	rs1105292	15	89303359	HSMM		Medullo			RCCD1	intronic
15q26.1	rs1105291	15	89303387	HSMM		Medullo		ERalpha-a,VDR	RCCD1	intronic
15q26.1	rs8038661	15	89303853	HSMM				Zbtb3	RCCD1	intronic
15q26.1	rs3743451	15	89305185			5 cell types		Hic1	RCCD1	intronic
15q26.1	rs7166130	15	89306720					7 altered motifs	RCCD1	3'-UTR
15q26.1	rs79548680	15	89306783					7 altered motifs	RCCD1	3'-UTR
15q26.1	rs5814452	15	89307411				BCL3	Hsf,Pax-3,SREBP	57bp 3' of RCCD1	
15q26.1	rs8028409	15	89307426				BCL3	Hand1,Hsf	72bp 3' of RCCD1	
15q26.1	rs8037137	15	89307641			K562,CMK,Jurkat		Pou3f3	287bp 3' of RCCD1	
15q26.1	rs112770790	15	89308499					TATA,TEF-1	1.1kb 3' of RCCD1	
15q26.1	rs56665089	15	89309583				POL2,POL24H8	4 altered motifs	690bp 3' of PRC1	
15q26.1	rs77554484	15	89310219						54bp 3' of PRC1	
15q26.1	**rs2290203**	15	**89313071**					**Gfi1b,Myc,PLZF**	**PRC1**	**intronic**
15q26.1	rs2290202	15	89313271					4 altered motifs	PRC1	intronic
15q26.1	rs2890156	15	89314161					4 altered motifs	PRC1	intronic
15q26.1	rs4517768	15	89315039					HDAC2,Irf,Mef2	PRC1	intronic

15q26.1	rs59278520	15	89315711						PRC1	intronic
15q26.1	rs11852999	15	89315764					GATA,MZF1::1-4,STAT	PRC1	intronic
15q26.1	rs59902579	15	89316493		HSMM		CTCF		PRC1	intronic
15q26.1	rs4932381	15	89316744						PRC1	intronic
15q26.1	rs17636091	15	89318483					RFX5	PRC1	intronic
15q26.1	rs8035844	15	89320646					8 altered motifs	PRC1	intronic
15q26.1	rs8036430	15	89320664					Rad21,Zfx,Znf143	PRC1	intronic
15q26.1	rs12595025	15	89321135					4 altered motifs	PRC1	intronic
15q26.1	rs12594925	15	89321291						PRC1	intronic
15q26.1	rs12595052	15	89321312					4 altered motifs	PRC1	intronic
15q26.1	rs60640465	15	89323026			LNCaP		CCNT2,Spz1,YY1	PRC1	intronic
15q26.1	rs8032722	15	89323074			LNCaP		Egr-1,ZNF219	PRC1	intronic
15q26.1	rs8026714	15	89323257			Th1		4 altered motifs	PRC1	intronic
15q26.1	rs6496742	15	89324040			6 cell types	MAFK,NFE2	Nkx2	PRC1	intronic
15q26.1	rs1867226	15	89324717					4 altered motifs	PRC1	intronic
15q26.1	rs2301826	15	89326201		GM12878	GM06990,Monocytes-CD14+_RO01746	7 bound proteins	YY1	PRC1	synonymous
15q26.1	rs59524191	15	89326889		GM12878			4 altered motifs	PRC1	intronic
15q26.1	rs7180016	15	89327279						PRC1	intronic
15q26.1	rs2285937	15	89328156					Ets	PRC1	intronic
15q26.1	rs8031684	15	89328455					4 altered motifs	PRC1	intronic
15q26.1	rs2301825	15	89329074						PRC1	intronic
15q26.1	rs150333865	15	89329879					6 altered motifs	PRC1	intronic
15q26.1	rs8036115	15	89329965					6 altered motifs	PRC1	intronic
15q26.1	rs8042518	15	89330126					MIZF	PRC1	intronic
15q26.1	rs60290423	15	89330456			HMEC		Pou5f1,Sox	PRC1	intronic
15q26.1	rs11853073	15	89331162	GM12878	NHEK	4 cell types		5 altered motifs	PRC1	intronic
15q26.1	rs8026781	15	89331679		4 cell types	PanIslets			PRC1	intronic
15q26.1	rs8028856	15	89331824		7 cell types	HA-sp		Hmx,PLZF	PRC1	intronic
15q26.1	rs3803563	15	89332356		6 cell types			PLZF,Pou2f2	PRC1	intronic
15q26.1	rs17692742	15	89332802		6 cell types	13 cell types		6 altered motifs	PRC1	intronic
15q26.1	rs12594752	15	89332999		6 cell types	GM12892,Hepatocytes		Pou1f1	PRC1	intronic
15q26.1	rs7179428	15	89333569	K562, GM12878	Huvec	GM12878,CD20+,HL-60		AP-2,Pax-5	PRC1	intronic
15q26.1	rs6496745	15	89333634	K562, GM12878	Huvec			NF-Y,Pax-4,Sox	PRC1	intronic

15q26.1	rs6496746	15	89333873	K562, GM12878	Huvec, NHLF			Cdx2,Mef2	PRC1	intronic
15q26.1	rs8025699	15	89334820		GM12878, K562		CEBPB	4 altered motifs	PRC1	intronic
15q26.1	rs112553294	15	89335298		K562, GM12878			4 altered motifs	PRC1	intronic
15q26.1	rs72630488	15	89335377		K562, GM12878			10 altered motifs	PRC1	intronic
15q26.1	rs76119208	15	89336333	HepG2, GM12878	5 cell types	6 cell types			PRC1	intronic
15q26.1	rs73502775	15	89337229	5 cell types	4 cell types	14 cell types	CTCF	HNF4,Irf	PRC1	intronic
15q26.1	rs72630489	15	89338410	9 cell types		25 cell types	TAF1	12 altered motifs	PRC1	intronic
15q26.1	rs78029804	15	89339924	7 cell types	Huvec, HMEC	Hepatocytes		NF-Y,Sox	60bp 5' of PRC1	
15q26.1	rs72759321	15	89345421					6 altered motifs	VPS33B	intronic
15q26.1	rs3826033	15	89349605			Th1		Myb	VPS33B	intronic
15q26.1	rs79253100	15	89352880			Melano			VPS33B	intronic

Abbreviations: ENCODE, the Encyclopedia of DNA Elements; Chr., chromosome.

[a] Index SNPs are shown in bold.

[b] Chromosome position (bp) is based on NCBI Human Genome Build 36.

[c] Evidence of local H3K4Me1 and H3K27Ac modification (cell lines/types: if >3, only the number is included).

[d] Evidence of local H3K4Me3 modification (cell lines/types: if >3, only the number is included).

[e] Evidence of chromatin hypersensitivity to DNase (cell lines/types: if >3, only the number is included).

[f] ChIP-seq experiments indicate alteration in binding of transcription factor (if >3, only the number is included).

[g] Evidence of alteration in regulatory motif (if >3, only the number is included).

Supplementary Table 8. Association of breast cancer risk with 16 genetic variants in the Malaysia/Singapore study

SNP	Chr.	Position[a]	Alleles[b]	Cases			Controls			Per-allele association	
				N	RAF	HWE *P*	N	RAF	HWE *P*	OR (95% CI)	*P*
rs10889256	1	62042943	**G**/A	2020	0.53	5.8×10^{-10}	1958	0.534	8.9×10^{-16}	0.98 (0.90-1.06)	0.554
rs4951011	1	202032954	**G**/A	2018	0.30	0.953	1957	0.320	0.857	0.90 (0.82-0.99)	0.030
rs11693188	2	122269580	**A**/C	2019	0.80	0.990	1958	0.802	0.504	0.97 (0.87-1.08)	0.551
rs9828276	3	144723903	**T**/G	2014	0.60	0.965	1954	0.591	0.341	1.05 (0.96-1.15)	0.273
rs10010358	4	5456706	**A**/C	2021	0.17	0.539	1955	0.168	0.857	1.05 (0.94-1.18)	0.406
rs10474352	5	90767981	**C**/T	2019	0.55	0.916	1958	0.538	0.372	1.03 (0.94-1.13)	0.491
rs9444163	6	85130450	**A**/C	2020	0.79	0.210	1958	0.784	0.242	1.02 (0.91-1.13)	0.775
rs1021311	12	29005155	**C**/A	2018	0.76	0.325	1955	0.742	0.019	1.10 (0.99-1.21)	0.072
rs2970376	15	58024217	**C**/T	2019	0.45	0.956	1958	0.458	0.350	0.97 (0.89-1.06)	0.527
rs2290203	15	89313071	**G**/A	2015	0.52	0.881	1955	0.517	0.436	0.99 (0.91-1.08)	0.868
rs1108351	15	89363941	**C**/A	2019	0.51	0.384	1955	0.492	0.734	1.06 (0.97-1.15)	0.220
rs11074880	16	27731110	**G**/T	2021	0.13	0.483	1958	0.131	0.426	0.99 (0.87-1.13)	0.885
rs1122322	16	72333289	**G**/A	2018	0.68	0.916	1958	0.684	0.064	1.00 (0.91-1.10)	0.984
rs1078523	17	38166286	**A**/G	2020	0.55	0.307	1958	0.530	0.305	1.08 (0.99-1.18)	0.076
rs11082321	18	19030649	**A**/G	2016	0.16	0.415	1954	0.169	0.805	0.91 (0.81-1.03)	0.125
rs5756386	22	20681183	**A**/G	2017	0.65	0.917	1957	0.660	0.833	0.97 (0.89-1.07)	0.580

Abbreviations: Chr., Chromosome; RAF, risk allele frequency; HWE, Hardy–Weinberg equilibrium, OR, odds ratio; CI, confidence interval.

[a] Chromosome position (bp) based on NCBI Human Genome Build 36.

[b] Risk/reference alleles, with risk allele shown in bold.

Supplementary Note

1. Description of Study Participants

Shanghai Breast Cancer Study (SBCS): The SBCS is a population-based, case-control study conducted in urban Shanghai, China's largest commercial center[1, 2]. For the SBCS-I, subjects were recruited between 1996 and 1998. Through a rapid case-ascertainment system and the population-based Shanghai Cancer Registry, 1,602 eligible breast cancer cases diagnosed during the study period were identified, of which 1,459 cases (91.1%) completed in-person interviews. Cancer diagnoses for all cases were reviewed and confirmed by two senior pathologists. Controls were randomly selected from the general population using the Shanghai Resident Registry, a population registry containing demographic information for all residents of urban Shanghai. The inclusion criteria for controls were identical to those for cases, with the exception of a breast cancer diagnosis. Of the 1,724 eligible controls, 1,556 (90.3%) completed in-person interviews. A structured questionnaire was used to elicit detailed information on demographic factors and known/suspected risk factors for breast cancer. All participants were measured for their current weight, height, and circumference of the waist and hips. All interviews were tape-recorded and reviewed by the field supervisor and quality-control staff to monitor the quality of interview data. Blood samples (10 ml from each woman) were obtained from 1,193 (82%) cases and 1,310 (84%) controls who completed the in-person interview. A sample of exfoliated buccal cells was obtained using cotton swabs from virtually all study participants who did not provide a blood sample. Because DNA yield from buccal cell samples collected in cotton swabs is low, the current study is limited to those who provided a blood sample.

Using a protocol similar to the SBCS-I, the SBCS-II recruited 1,989 incident breast cancer cases and 1,989 community controls between 2002 and 2005 with a response rate of 83.7% and 70.4%, respectively. Similar to subject recruitment in the SBCS-I, the majority of newly recruited cases (n=1,932, 97.1%) and controls (n=1,857, 93.4%) provided a blood sample or an exfoliated buccal cell sample to the study. The mouthwash method used in the study was modified from that reported initially by Lum and Le Marchand [3] and provided, on average, approximately 34 μg of DNA per sample. With the exception of age, eligibility criteria for study participation were identical for SBCS-I and SBCS-II. The age range was expanded from 25 to 65 years in SBCS-I, to 25 to 70 years in SBCS-II.

Shanghai Breast Cancer Survival Study (SBCSS) and Shanghai Endometrial Cancer Study (SECS): The SBCSS also used the population-based Shanghai Cancer Registry to identify newly diagnosed breast cancer cases for the study[2]. A total of 6,303 cases were diagnosed between April 1, 2002, and December 31, 2006, and were approached for the study approximately six months after cancer diagnosis; 5,046 were recruited (response rate: 80.1%). In-person interviews were conducted to collect information on known breast cancer risk factors and anthropometrics by using a protocol and questionnaire similar to those used in the SBCS. Buccal cell samples were collected from 96% of study participants using the modified mouthwash method described above. Because of a time overlap in the participant recruitment period for the SBCS-II and the SBCSS, 1,469 breast cancer patients participated in both studies. The remaining 3,466 SBCSS cases were included in the current study.

Controls for this group of cases were draw from the SECS[2], a population-based case-control study conducted between 1997 and 2003, a time period that overlaps the recruitment period of the SBCS and SBCSS. With the exception of a few questions related specifically to breast or endometrial cancer risk, the questionnaires used in the SECS and the SBCS were virtually identical. Using a protocol similar to the one used in the SBCS, eligible cases were identified through the population-based Shanghai Cancer Registry and controls were randomly selected from the general population of Shanghai using the Shanghai Resident Registry and were age frequency-matched to cases. Women with a history of cancer or hysterectomy were not eligible. In-person interviews were conducted by trained interviewers to collect detailed information on demographic factors as well as known and suspected risk factors. Of the study participants who completed an in-person interview, 1,039 controls provided a blood sample or buccal cell sample using the mouthwash method, and these women were included in the current study.

Shanghai Women's Health Study (SWHS): The SWHS is a population-based prospective cohort study of approximately 75,000 adult women who were recruited between 1997 and 2000[4]. At the baseline recruitment, a roster of all women aged 40 to 70 years was obtained from the resident registry offices in the study communities. Of the 81,170 eligible women, 75,221 participated in the study, with a participation rate of 92.7%. Detailed exposure data were collected during the baseline survey through an in-person interview. Body weight/height and circumferences of the waist and hips were measured. Among those who completed the survey, 56,831 (75.8%) donated a blood sample, and 65,754 (87.7%) donated a urine sample. An exfoliated buccal cell sample was collected from an additional 8,934 (49.3%) of the 18,111 subjects who did not provide a blood sample at baseline. Therefore, we have genomic DNA from about 88% of cohort members.

The cohort has been followed by a combination of record linkage and active follow-ups. Every two years, an interviewer visits the last known address of each living cohort member and records details of the interim health history, including cancer and several other chronic diseases that occurred since the last in-person contact. In addition to the interim health history, the survey questionnaire also includes a module to obtain information related to selected lifestyle factors, including a FFQ. Data routinely collected by the cancer registry and death certificates are also used to assure a timely and complete ascertainment of new cancer cases and deceased subjects in the study cohort. All possible matches are checked manually and verified through home visits. For cohort members who are diagnosed with cancer, information on date and hospital of diagnosis is collected. Copies of medical charts from the diagnostic hospital are obtained to verify the diagnosis and collect detailed information on the pathology characteristics of the tumor. In addition, pathology slides and tumor tissue blocks are being collected to verify cancer diagnosis and for future studies of biomarkers. The first follow-up survey was conducted from 2000 to 2002. Approximately 99.8% of cohort members (or their next of kin, if subjects were deceased) were interviewed. The response rates were 98.7% for the second follow-up survey (2002-04), 96.7% for the third follow-up survey (2004-07), and approximately 93% for the fourth follow-up (2007-2010). For non-respondents, cancer diagnosis and vital status can still be identified through the linkage of data from cancer and vital statistics registries, and thus

ascertainment for cancer outcomes and total mortality is virtually complete in this cohort. Breast cancer patients identified in the SWHS and non-cases were included in the current study.

Taiwan Study[5, 6]**:** This case-control study is part of an on-going, cooperative study aimed at understanding the causes of breast cancer in Taiwan, which is characterized by low incidence, early tumor onset, hormone dependency, and novel genomic alterations. The study included 1,001 female breast cancer patients and 1,013 healthy female controls. All breast cancer patients had pathologically confirmed incident primary breast cancer and were diagnosed and treated at the Tri-Service General Hospital or the Changhua Christian Hospital between March 2002 and August 2005. The participation rate was over 90%. Patients with inadequate blood specimens were excluded from the study. Women included in the study were similar to those excluded in the distribution of major breast cancer risk factors. Because these are two of the major breast cancer clinics in northern and central Taiwan, patients recruited for the study accounted for a significant proportion (~40%) of all breast cancer cases diagnosed during the study period in these regions. Controls were randomly selected from women attending the health examination clinics of the same hospitals during the same period. These women underwent a one-day comprehensive health examination (including regular breast screening using X-ray mammography and ultrasonic examination), and those showing any evidence of breast cancer, suspicious precancerous lesions of the breast, or other cancers were excluded from the control group. Almost all women (>95%) initially identified as potential controls participated in the study, and the controls accounted for ~20% of all women attending the clinics. No significant differences in socioeconomic status were found between those included and those excluded from the study. Informed consent was obtained from all study participants before collection of epidemiologic data through in-person interviews. At the completion of each interview, blood was taken for DNA isolation and genotyping. Two experienced research nurses were assigned to administer a structured questionnaire to cases and controls. The information collected included age, family history of breast cancer, age at menarche and/or menopause, history of full-term pregnancy, menopausal status, and body mass index.

Hong Kong Study[7]: This is a hospital-based study consisting of women with incident breast cancer, recruited during the period of June 2003 to March 2009 from patients attending follow-up surgical and oncology outpatient clinics at three major public hospitals on Hong Kong Island (Queen Mary Hospital), and Kowloon (Queen Elizabeth Hospital and Kwong Wah Hospital) and the period 2009-2011 from oncology outpatient clinic at Pamela Youde Nethersole Hospital. All participants completed face-to-face interviews. Control participants matched for age on 10-year intervals were recruited from outpatients attending the general gynecological clinic at Queen Mary Hospital and from the Well-Women Clinic at Kwong Wah Hospital, who had no personal history of cancer. They were also questioned about any family history of breast and/or ovarian cancer. About 70% of cases and controls interviewed agreed to participate in this project. Blood samples were obtained from 517 cases and 651 controls, which were subsequently used for DNA extraction by proteinase K digestion followed by conventional phenol-chloroform-ethanol extraction. The protocol was approved by the Institutional Review Boards of the University of Hong Kong Hospital Authority. Patient consent was obtained for study participation and blood collection.

Seoul Breast Cancer Study (SeBCS)[8, 9]: The SeBCS is a hospital-based case-control study conducted in two teaching hospitals in Seoul. Included in this project were 2,342 incident breast patients histopathologically diagnosed with primary breast cancer and consecutively recruited between 2001 and 2007. In-person interviews were conducted to collect information on known breast cancer risk factors and anthropometrics by using a protocol and questionnaire. Medical charts were reviewed to verify clinical information. Eligible controls were derived from a large urban cohort that is participating in the Korea Genome Epidemiology Study (KoGES), which is an ongoing cohort study that has sought to understand the causes and risk factors of disease in Korea. Controls were 2,052 women selected between May 2006 and December 2007. They were frequency-matched to cases on the case's age at diagnosis in five-year intervals. Trained interviewers using a structured questionnaire determined the demographic characteristics of the controls, their family histories with regard to breast cancer in first-degree relatives, reproductive and menstrual factors, and life-style habits, using a protocol similar to the SeBCS. Women with a history of cancer were not included.

Hwasun Cancer Epidemiology Study-Breast (HCES-Br)[10-12]: The Hwasun Cancer Epidemiology Study (HCES) is a hospital-based case-control study whose goal is to identify factors of the cancer development and clinical progression in a Korean population[10, 11]. Participants of the HCES-Br consisted of 3,387 female breast cancer cases, who were newly diagnosed between April 2004 and February 2013 at Chonnam National University Hwasun Hospital, a cancer specified hospital in Jeollanam-do province, South Korea. Patients with secondary or recurrent tumor were excluded. Control (n=3,186) were randomly selected from among women with no previous cancer diagnosis at enrollment in the Namwon Study and the Dong-gu study, ongoing community-based cohort studies in Korea[12]. Genomic DNA was extracted from their peripheral blood. Data for demographics and conventional factors of breast cancer were collected by structured questionnaire and review of medical records. All cases and control subjects provided the informed consent to participate in the study and Institutional Review Board of Chonnam National University Hwasun Hospital approved this study.

Korea Genome Epidemiology Study (KoGES)[13]: The KoGES is ongoing study since 2001 to investigate major genetic and environmental factors for common diseases in the Korean population. Of 10,038 subjects surveyed at baseline enrollment in 2001, 1,536 women with sufficient DNA concentrations were analyzed. Of 7,861 subjects recruited from 2005 to 2006, 1,673 women were analyzed. Thus, a total of 3,209 control subjects were selected and analyzed for Stage IV.

Korean Hereditary Breast Cancer (KOHBRA)[14]: The KOHBRA study is an ongoing cohort study since 2007 to examine high risk groups for hereditary breast cancer such as female breast cancer patients with a family history, ovarian cancer, or other coincidental cancers, male breast cancer patients, and family members of breast cancer patients with *BRCA*1/2 mutation4. We finally selected 1,397 female cancer patients without *BRCA*1, 2 mutation among KOHBRA subjects recruited in 2007-2009.

Korea NCC study: Newly diagnosed breast cancer patients (cases, n=505) were recruited from the Breast Cancer Clinic at the Asan Medical Center in Seoul, Korea between February 2006 and July 2010. Each case member had received a histologically confirmed

diagnosis of their first primary breast cancer and participated in the study before the treatment was started. Ineligibility criteria were a previous malignancy (at either the same site or a different site) and an age greater than 80 years. The hospital controls (n=505) were women free of any malignant neoplasms and free of any clinical, biochemical, or hematological manifestations of cardiovascular, hepatic, renal, or endocrinal disorders. All case and control subjects completed a questionnaire on lifestyle and dietary intake and provided blood samples. Informed consent was obtained from all subjects after a full explanation of the study, which had been previously approved by the institutional review board of the Korea National Cancer Center. Both case and control subjects were interviewed by one trained interviewer who was unaware of the subject's status. Using both a non-dietary questionnaire and a 95-item semi-quantitative food frequency questionnaire, information was collected on socio-demographic characteristics, anthropometric measures, individual medical history, family cancer history, and dietary factors detailing their usual food intake over the year prior to enrolment in the study. Socio-demographic characteristics included education level, occupation, cigarette smoking status, alcohol consumption, and physical activity. Pathological and laboratory data for each subject were collected, recorded, and entered into an epidemiological database. Medical charts and pathology reports were examined to ensure that control subjects had no known history of cancer. A peripheral venous blood sample (20 ml aliquot in an anticoagulant tube) was obtained from each enrolled subject. Laboratory assays of the blood samples were performed before the initiation of any treatment or therapy. Blood samples were wrapped in aluminum foil to protect against photo-oxidation, and transported to the laboratory without revealing the subject's case/control status prior to performing the antioxidant micronutrient assay. After separating plasma, samples were stored at -80°C until assayed.

The Biobank Japan Project (BBJ)[15, 16]: The DNA samples were recruited from the Biobank Japan Project (http://biobankjp.org). A total of 5,527 individuals were registered as breast cancer patients from the Biobank Japan, these patients' DNA were subsequently divided into 2,642 and 2,885 to be screen as discovery (BBJ1) and validation phase (BBJ2), respectively. For controls, a total of 5,494 (2,099 for BBJ1 and 3,395 for BBJ2) females consisting of healthy volunteers from Midosuji Rotary Club, Osaka, Japan, Health Science Research Resource Bank and individuals in the Biobank who do not have history of cancer for discovery and validation phase, respectively. BBJ study imputation: For Imputation analysis, 1000 Genomes Project Phase I Integrated Release Version 2 dataset that include individuals of JPT (Japanese in Tokyo), CHB (Chinese in Beijing) and CHD (Chinese in Denver) was utilized as reference panel to infer missing genotypes. Firstly, allele frequencies of the reference allele were confirmed to be comparable between the GWAS dataset and the reference panel with differences not more than 0.15. MACH1.0 (http://www.sph.umich.edu/csg/abecasis/MACH/index.html) was used to phase haplotypes, which relate the samples to the reference haplotypes; map crossover and estimate error rates using 20 iterations of the Markov chain. Subsequently, missing genotypes were imputed using Minimac (http://genome.sph.umich.edu/wiki/Minimac).

Nagoya Study (Hospital-based Epidemiologic Research Program at Aichi Cancer Center, HERPACC-II)[17]: This is a hospital-based, comprehensive epidemiologic research program at the Aichi Cancer Center (ACC), Japan. All first-visit outpatients 20-79 years of age at the ACC from December 2000 to November 2005 were asked to participate in the HERPACC-II. A total of 29,736 eligible patients were approached, and 28,766 participated in

the study, with a response rate of 96.7%. Subjects completed a self-administered questionnaire about their lifestyle and demographic characteristics and to provide blood samples. Dietary habits were investigated using a 47-item semi-quantitative food frequency questionnaire. ER status for cases was taken from medical records. ER status is routinely determined by pathologists by using commercially based immunohistochemistry tests at the ACC. Case status was confirmed by linkage of the HERPACC-II database and the hospital-based cancer registry database. 1,850 histologically-confirmed breast cancer cases were identified, and 644 were selected for the Asia Breast Cancer Consortium analysis based on availability of DNA samples. Of 14,260 non-cancer subjects in the HERPACC-II database, 644 subjects matched for age and menopausal status were randomly selected. The study protocol was approved by the Institutional Review Board at the ACC (Nagoya, Japan).

Nagano Breast Cancer Study[18]: This multicenter, hospital-based case-control study was conducted from May 2001 to September 2005 at four hospitals in Nagano Prefecture, Japan. The cases – a consecutive series of women ages 20-74 years with newly diagnosed, histologically confirmed invasive breast cancer – were admitted to the four hospitals during the survey period. Of the 412 eligible patients, 405 (98%) agreed to participate. Healthy controls were selected from medical checkup examinees in two of the hospitals who were confirmed as not having any cancer, with one control matched for each case by age (within three years) and residential area during the study period. Among potential control subjects, one declined to participate. Written informed consent was obtained from 405 matched pairs. Because two controls refused to provide blood samples, the analysis was restricted to 403 matched pairs. Participants completed a self-administered questionnaire, which included questions on demographic characteristics, anthropometric factors, smoking habits, family history of cancer, physical activity, medical history, and menstrual and reproductive history. Dietary habits were investigated using a 136-item semi-quantitative food-frequency questionnaire (FFQ), which was developed and validated in the Japanese population. The ER and PR status of the patient's breast cancer tissue was obtained from medical records. Hormone receptor positivity values were determined either as specified by the laboratory that performed the assay, in accordance with the laboratory's written interpretation thereof, or both. The study protocol was approved by the Institutional Review Board of the National Cancer Center (Tokyo, Japan).

2. Genotyping Methods – Quality Control and Exclusion Criteria

2.1. Stage 1 Genotyping

Chinese GWAS (SBCGS-1): We included one negative control and at least three positive QC samples from Coriell Cell Repositories (http://ccr.coriell.org/) in each of the 96-well plates for Affymetrix SNP Array 6.0 genotyping. A total of 273 positive QC samples were successfully genotyped with an average concordance of 99.9% with a median value of 100%. We excluded samples that (1) had a genotype call rate per sample of < 95%; (2) were genetically identical (i.e., PI_HAT >0.9) or duplicates; (3) had an assay-determined sex that did not concur with reported sex; (4) were first- or second-degree relatives (i.e., PI_HAT >0.25); (5) had an outlier value for ethnicity; or (6) had an outlier value for heterozygosity. We excluded SNPs that had (1) a genotype call rate of < 95%; (2) a MAF of < 0.01; (3) condordance with the QC sample

genotyping of < 95%; (4) a Hardy-Weinberg equilibrium (HWE) test *P* of < 1×10^{-5} in controls; or (5) a poor cluster plot in either cases or controls.

Korean GWAS (SeBCS1): A total of 30 QC samples were successfully genotyped with a concordance of 99.8%. Similar criteria were used to exclude samples in the SeBCS1[19]. We excluded SNPs that had (1) a genotype call rate of < 95%; (2) a MAF of < 1% among either cases or controls; (3) evidence of deviation from HWE at a *P*-value of < 10^{-6}; (4) a poor cluster plot in either cases or controls; (5) a missing rate that differed between cases and controls (*P* < 10^{-4}); or (6) multiple positioning or mitochondrial SNPs or both.

2.2. Stage 2 Genotyping

On each 96-well plate, two blinded duplicate samples and two HapMap samples were included for QC. We evaluated concordance between HapMap samples genotyped in our study and data from the 1000 Genomes Project (http://www.1000genomes.org/). Principal components analyses (PCA) were conducted based on 3,200 ancestry informative markers (AIMs) on the HumanExome Beadchip using EIGENSTRAT (http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm) to identify population outliers with 1000 Genomes Project data as the reference. We also estimated the pair-wise proportion of IBD to identify potential genetically identical or unexpected duplicate samples or close relatives. We excluded samples that had (1) a genotype call rate of < 98%, (2) concordance with the QC sample genotyping of < 99%, (3) an outlier value for heterozygosity, or (4) an outlier value for ethnicity, as well as (5) samples that appeared to be closely related, (6) duplicate samples with a consistency of < 99%, and (7) samples whose assay-determined sex did not concur with reported sex. We excluded SNPs that had (1) a MAF = 0, (2) a genotype call rate of < 98%, (3) concordance with the QC sample genotyping of < 98%, or (4) a HWE test *P* of <1 x 10^{-5}, as well as (5) SNPs that were redundant and (6) SNPs designated by the Illumina Infinium assay design group as “treat with caution”.

2.3. Stage 3 Genotyping

Genotyping assays for the 50 SNPs in Stage 3 were completed at the Vanderbilt Molecular Epidemiology Laboratory using the iPLEX Sequenom MassArray platform for 19,423 samples from the Taiwan, Hong Kong, HCES-Br, KOHBRA/KoGES, SeBCS2, Korea-NCC, Nagoya, and Nagano studies. QC samples were used in the Sequenom assay, including one negative control (water), two blinded duplicates, and two samples from the HapMap project in each 96-well plate.We excluded samples or SNPs that had a genotyping call rate of <95%. We further excluded SNPs that had (1) concordance with the QC samples of < 95%; (2) an unclear genotype call; or (3) a *P* value for HWE of <0.001. The mean genotype call rate of all samples was 98.1% (median: 98.5%). The mean concordance was 99.4% for the blinded duplicates and 99.3% for HapMap samples with a median value of 100% for all QC samples.

3. TCGA Data Resource

We downloaded RNA-Seq V2 data (level 3) for 1,006 breast cancer tumor tissue samples and 94 adjacent normal tissue samples from TCGA data portal (http://cancergenome.nih.gov/). DNA methylation data were also retrieved from TCGA level 3 data. Only those with DNA methylation data measured using the Illumina HumanMethylation450 BeadChip were included in the present study. The average of CpG methylation levels across probes in the promoter region (+/- 2 kb to transcription start site) was used to represent the methylation level of the corresponding gene. We also downloaded level 3 SNP data genotyped using the Affymetrix SNP Array 6.0. Genotype data from the 1Mb region flanking the three loci were extracted and then imputed using 1000 Genomes Project data with Minimac (http://genome.sph.umich.edu/wiki/Minimac). Only common SNPs (MAF > 0.05) with high imputation quality (RSQR > 0.3) were included in the present study. Copy number variation (CNV) data for each of the 87 genes in the 1Mb regions flanking the three loci were collected from cBioPortal (http://www.cbioportal.org/public-portal/) for tumor tissue samples. A total of 5 copy number categories were considered, including -2 (homozygous deletion), -1 (heterozygous deletion), 0 (diploid), 1 (low level gain), and 2 (high level amplification). In the current analysis, we used only samples from breast cancer patients with European ancestry. For the downstream eQTL analysis, we included only tumor samples with data available for RNA expression, DNA methylation, CNV, and SNP data. In total, 458 tumor tissue samples were included, of which 66 had matching adjacent normal tissue samples with both RNA expression and SNP data available. For the differential gene expression analysis, we included 87 breast cancer cases for which gene expression and SNP data were available for both the tumor tissue sample and the matched adjacent normal tissue sample.

4. eQTL Analysis

We used the TCGA breast cancer data described above to perform an eQTL analysis for normal and tumor tissue samples separately. We applied several steps, following Pickrell *et al.*'s approach, to reduce batch and other technical effects on the gene expression data[20]. First, the RNA-Seq by Expectation-Maximization value of each gene was log2 transformed, and genes with a median expression level of 0 across all tissue samples were removed. We then performed the principal component correction on gene expression data for tumor and adjacent normal tissue separately to remove potential batch effects. A linear regression of expression values on the first 5 principal components was constructed, and the residuals were used to replace the expression values of each gene among the tissue samples.

To make the data better conform to the linear model for the eQTL analysis, we further transformed the gene expression level to fit a quantile of N(0,1) distributions based on the rank of the expression values to their respective quantiles. To further adjust for the potential effects of methylation and CNV on the expression of each gene in tumor tissue, we constructed both full (equation (1a)) and residual (equation (1b)) linear regression models to detect eQTLs according to the approach used by Li *et al.* [21]. In the analysis of gene expression in adjacent normal tissue, this adjustment was not performed and the regular linear model was used with SNPs included (equation (2)).

$$T_i = Sc_i + M_i + G_i + \omega_i \quad (1a)$$

$$T_i = Sc_i + M_i + \varepsilon_i \quad (1b)$$
$$\varepsilon_i = G_i + \omega_i$$

$$T_i = G_i + \omega_i \quad (2)$$

We focused only on the SNPs and genes located within the 1Mb regions flanking the three newly identified risk loci to identify cis-eQTLs. A significance threshold *P* value of < 0.01 was used to determine candidate cis-eQTLs.

5. Differential Gene Expression Analysis

To identify differentially expressed genes located in the three identified risk loci, we analyzed data from a total of 87 pairs of breast tumor-normal tissue samples included in TCGA described above. We did not apply principal component analysis to remove batch or other technical effects for tumor and adjacent normal tissue in the differential expression analysis, since this approach can also remove the outcome from real differences between tumor and adjacent normal tissue. Instead, we utilized a specialized tool, surrogate variable analysis[22], which was developed to perform paired comparisons of gene expression between tumor and normal tissue, while removing potential batch effects and other artifacts. Specifically, we created a full model including the tumor-normal comparison of interest adjusted for the paired design, and a null model, which was only adjusted for the paired design. The two models were used to estimate the total number of latent factors and the values of the surrogate variables. After adjustment for the surrogate variables, we used the limma package available through Bioconductor (www.bioconductor.org) to determine differential expression of genes[23]. False discovery rate–adjusted (Benjamini and Hochberg method) *P*-values are presented.

6. The DRIVE GAME-ON Consortium

The DRIVE GAME-ON consortium (http://epi.grants.cancer.gov/gameon/) includes 16,003 breast cancer cases and 41,335 controls from 12 breast cancer GWAS: ABCFS, BBCS, DFBBCS, MARIE, SASBAC, HEBCS, GC-HBOC, UK2, BPC3, BCFR, and SardiNIA. The key investigators (institution and location) from the DRIVE GAME-ON consortium included (in alphabetical order): Muriel Adank (VU University Medical Center, Amsterdam, The Netherlands), Habibul Ahsan (University of Chicago, Chicago, IL, USA), Kristiina Aittomäki (University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland), Lars Beckman (Institute for Quality and Efficieny in Health Care, Cologne, Germany), Carl Blomquist (University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland), Federico Canzian (German Cancer Research Center, Heidelberg, Germany), Jenny Chang-Claude (Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany), Stephen J. Chanock (National Cancer Institute, Bethesda, MD, USA), Laura Crisponi (Istituto di Ricerca Genetica e Biomedica, Consiglio Nazionale delle Ricerche, Cagliari, Italy), Kamila Czene (Karolinska Institut, Stockholm, Sweden), Norbert Dahmen (University of Mainz, Mainz, Germany), Isabel dos Santos Silva (London School of Hygiene and Tropical Medicine, London,

U.K.), Douglas Easton (University of Cambridge, Cambridge, UK), Olivia Fletcher (The Institute of Cancer Research, London, U.K.), Lorna Gibson (London School of Hygiene and Tropical Medicine, London, U.K.), Christopher A. Haiman (University of Southern California, Los Angeles, CA, USA), Per Hall (Karolinska Institut, Stockholm, Sweden), HEBON (Hereditary Breast and Ovarian Cancer Research Group Netherlands), Rebecca Hein (Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany, and University of Cologne, Cologne, Germany), Brian E. Henderson (University of Southern California, Los Angeles, CA, USA), Albert Hofman (Erasmus Medical Center, Rotterdam, The Netherlands), John L. Hopper (Melbourne School of Population Health, University of Melbourne, Melbourne, Victoria, Australia), David J. Hunter (Harvard University School of Public Health, Boston, MA), Astrid Irwanto (Genome Institute of Singapore, Singapore), Rudolf Kaaks (German Cancer Research Center, Heidelberg, Germany), Muhammad G. Kibriya (University of Chicago, Chicago, IL, USA), Peter Lichtner (German Research Center for Environmental Health, Neuherberg, Germany), Jianjun Liu (Genome Institute of Singapore, Singapore), Enes Makalic (Melbourne School of Population Health, University of Melbourne, Melbourne, Victoria, Australia), Alfons Meindl (Technische Universität München, Munich, Germany), Hanne Meijers-Heijboer (VU University Medical Center, Amsterdam, The Netherlands), Bertram Müller-Myhsok (Max Planck Institute of Psychiatry, Munich, Germany), Taru A. Muranen (University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland), Heli Nevanlinna (Univesity of Helsinki and Helsinki University Central Hospital, Helsinki, Finland), Julian Peto (London School of Hygiene and Tropical Medicine, London, U.K.), Ross L. Prentice (Fred Hutchinson Cancer Research Center, Seattle, WA, USA), Nazneen Rahman (Institute of Cancer Research, Sutton, U.K.), Daniel F. Schmidt (Melbourne School of Population Health, University of Melbourne, Melbourne, Victoria, Australia), Rita K. Schmutzler (University of Cologne, Cologne, Germany), Melissa C. Southey (The University of Melbourne, Melbourne, Victoria, Australia), Clare Turnbull (Institute of Cancer Research, Sutton, U.K.), Andre G. Uitterlinden (Erasmus Medical Center, Rotterdam, The Netherlands), Rob B. van der Luijt (University Medical Center Utrecht, Utrecht, The Netherlands), Quinten Waisfisz (VU University Medical Center, Amsterdam, The Netherlands), Alice S. Whittemore (Stanford University, Stanford, CA, USA), and Wei Zheng (Vanderbilt University, Nashville, TN, USA).

Reference List

1. Gao,Y.T. *et al.* Association of menstrual and reproductive factors with breast cancer risk: results from the Shanghai Breast Cancer Study. *Int J Cancer* **87**, 295-300 (2000).
2. Zheng,W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324-328 (2009).
3. Lum,A. & Le Marchand,L. A simple mouthwash method for obtaining genomic DNA in molecular epidemiological studies. *Cancer Epidemiol. Biomarkers Prev.* **7**, 719-724 (1998).
4. Zheng,W. *et al.* The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am. J. Epidemiol.* **162**, 1123-1131 (2005).
5. Ding,S.L. *et al.* Genetic variants of BLM interact with RAD51 to increase breast cancer susceptibility. *Carcinogenesis* **30**, 43-49 (2009).
6. Hsu,H.M. *et al.* Breast cancer risk is associated with the genes encoding the DNA double-strand break repair Mre11/Rad50/Nbs1 complex. *Cancer Epidemiol Biomarkers Prev* **16**, 2024-2032 (2007).
7. Chan,K.Y. *et al.* Functional polymorphisms in the BRCA1 promoter influence transcription and are associated with decreased risk for breast cancer in Chinese women. *J Med Genet* **46**, 32-39 (2009).
8. Han,S. *et al.* CASP8 polymorphisms, estrogen and progesterone receptor status, and breast cancer risk. *Breast Cancer Res Treat* **110**, 387-393 (2008).
9. Cho,Y.S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* **41**, 527-534 (2009).
10. Song,H.R. *et al.* Sex-specific differences in the association between ABO genotype and gastric cancer risk in a Korean population. *Gastric. Cancer* **16**, 254-260 (2013).
11. Cui,L.H. *et al.* Methylenetetrahydrofolate reductase C677T polymorphism in patients with gastric and colorectal cancer in a Korean population. *BMC. Cancer* **10**, 236 (2010).
12. Kweon,S.S. *et al.* Cohort Profile: The Namwon Study and the Dong-gu Study. *Int. J Epidemiol.* **43**, 558-567 (2014).
13. Cho,Y.S. *et al.* A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat. Genet.* **41**, 527-534 (2009).
14. Han,S.A. *et al.* The Korean Hereditary Breast Cancer (KOHBRA) Study: Protocols and Interim Report. *Clin Oncol (R. Coll. Radiol.)* **23**, 434-441 (2011).
15. Elgazzar,S. *et al.* A genome-wide association study identifies a genetic variant in the SIAH2 locus associated with hormonal receptor-positive breast cancer in Japanese. *J Hum Genet* **57**, 766-771 (2012).
16. Low,S.K. *et al.* Genome-wide association study of breast cancer in the Japanese population. *PLoS. One.* **8**, e76463 (2013).
17. Hamajima,N. *et al.* Gene-environment Interactions and Polymorphism Studies of Cancer Risk in the Hospital-based Epidemiologic Research Program at Aichi Cancer Center II (HERPACC-II). *Asian Pac. J Cancer Prev* **2**, 99-107 (2001).

18. Itoh,H. *et al.* Serum organochlorines and breast cancer risk in Japanese women: a case-control study. *Cancer Causes Control* **20**, 567-580 (2009).
19. Kim,H.C. *et al.* A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. *Breast Cancer Res.* **14**, R56 (2012).
20. Pickrell,J.K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 (2010).
21. Li,Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-641 (2013).
22. Leek,J.T., Johnson,W.E., Parker,H.S., Jaffe,A.E., & Storey,J.D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics.* **28**, 882-883 (2012).
23. Smyth,G.K., Michaud,J., & Scott,H.S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics.* **21**, 2067-2075 (2005).