SCIENTIFIC REPORTS
natureresearch

OPEN

Identification of two novel breast cancer loci through large-scale genome-wide association study in the Japanese population

Siew-Kee Low[1,2*], Yoon Ming Chin[1], Hidemi Ito[4,6], Keitaro Matsuo[5,6], Chizu Tanikawa[7], Koichi Matsuda[9], Hiroko Saito[10], Mika Sakurai-Yageta[11], Naoki Nakaya[11], Atsushi Shimizu[12], Satoshi S. Nishizuka[12], Taiki Yamaji[13], Norie Sawada[13], Motoki Iwasaki[13], Shoichiro Tsugane[14], Toshiro Takezaki[15], Sadao Suzuki[16], Mariko Naito[17,18], Kenji Wakai[17], Yoichiro Kamatani[2], Yukihide Momozawa[3], Yoshinori Murakami[8], Johji Inazawa[19,21], Yusuke Nakamura[1], Michiaki Kubo[3], Toyomasa Katagiri[22] & Yoshio Miki[10,20]

Genome-wide association studies (GWAS) have successfully identified about 70 genomic loci associated with breast cancer. Owing to the complexity of linkage disequilibrium and environmental exposures in different populations, it is essential to perform regional GWAS for better risk prediction. This study aimed to investigate the genetic architecture and to assess common genetic risk model of breast cancer with 6,669 breast cancer patients and 21,930 female controls in the Japanese population. This GWAS identified 11 genomic loci that surpass genome-wide significance threshold of $P < 5.0 \times 10^{-8}$ with nine previously reported loci and two novel loci that include rs9862599 on 3q13.11 (*ALCAM*) and rs75286142 on 21q22.12 (*CLIC6-RUNX1*). Validation study was carried out with 981 breast cancer cases and 1,394 controls from the Aichi Cancer Center. Pathway analyses of GWAS signals identified association of dopamine receptor medicated signaling and protein amino acid deacetylation with breast cancer. Weighted genetic risk score showed that individuals who were categorized in the highest risk group are approximately 3.7 times more likely to develop breast cancer compared to individuals in the lowest risk group. This well-powered GWAS is a representative study to identify SNPs that are associated with breast cancer in the Japanese population.

[1]Cancer Precision Medicine Center, Japanese Foundation for Cancer Research, Tokyo, Japan. [2]Laboratory for Statistical Analysis, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [3]Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [4]Division of Cancer Information and Control, Aichi Cancer Center Research Institute, Nagoya, Japan. [5]Division of Cancer Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan. [6]Department of Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan. [7]Laboratory of Genome Technology, Human Genome Center, The University of Tokyo, Tokyo, Japan. [8]Division of Molecular Pathology, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [9]Graduate school of Frontier Sciences, The University of Tokyo, Tokyo, Japan. [10]Department of Genetic Diagnosis, The Cancer Institute of JFCR, Tokyo, Japan. [11]Tohoku Medical Megabank Organization, Tohoku University, Sendai, Japan. [12]Iwate Tohoku Medical Megabank Organization, Iwate Medical University, Iwate, Japan. [13]Division of Epidemiology, National Cancer Center, Tokyo, Japan. [14]Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. [15]Department of International Island and Community Medicine, Kagoshima University Graduate School of Medical and Dental Sciences, Kagoshima, Japan. [16]Department of Public Health, Nagoya City University Graduate School of Medical Sciences, Nagoya, Japan. [17]Department of Preventive Medicine, Nagoya University Graduate School of Medicine, Nagoya, Japan. [18]Department of Oral Epidemiology, Graduate School of Biomedical and Health Sciences, Hiroshima University, Hiroshima, Japan. [19]Department of Molecular Cytogenetics, Tokyo Medical & Dental University, Tokyo, Japan. [20]Department of Molecular Genetics, Medical Research Institute, Tokyo Medical & Dental University, Tokyo, Japan. [21]Bioresource Research Center, Tokyo Medical & Dental University, Tokyo, Japan. [22]Division of Genome Medicine, Institute for Genome Research, Tokushima University, Tokushima, Japan. *email: siewkee.low@jfcr.or.jp

Breast cancer is the most common malignancy among women worldwide. Based on the report of Cancer Statistics in Japan 2018[1], it is estimated that the incidence of breast cancer will rise to 86,500 in the year 2018, which comprise approximately 20% of all female cancers. Breast cancer is also the fifth leading cause of cancer death among women in Japan, with an estimated death of 14,285 in the year of 2017. Despite better 5-year survival rates for breast cancer compared to other malignancies, the age-adjusted incidence and mortality rate in Japan has increased steadily since the 1970s. Hence, predictive genetic markers and early detection screening methods to identify individuals at risk of breast cancer are crucial to reduce breast-cancer associated death.

Breast cancer is a complex polygenic disease with diverse risk factors that include lifestyle and genetic mutations. Common mutations linked to breast cancer include highly-penetrant *BRCA1* and *BRCA2* genes, moderate effect size genes (*CHEK2*, *PALB2*, *PTEN* and *ATM*) as well as common variants conferring small effect sizes[2–7]. A total of 28 genome-wide association studies (GWAS) showing association with breast cancer risk have been published[8]. These studies successfully identified common variants in 70 genetic loci from diverse worldwide populations: Europe (70 loci), East Asians (8 loci), Africans (3 loci), Latinos (2 loci) and Ashkenazi Jews (1 loci)[9–35]. Risk loci and risk variants differ across different populations due to several possible reasons. These include insufficient statistical power in individual studies, complexity in linkage disequilibrium, and differences in allele frequencies as well as environmental exposure. Notably, studies have indicated the importance of carrying out regional GWAS to identify specific genetic risk factors that are associated with complex disease, which could facilitate better risk assessment in the regional clinical settings[36]. Our group has published two breast cancer GWAS: The first reported association of chromosome 10q26 (*FGFR2*) and 16q12 (*TOX-LOC643714*) to breast cancer while the second showed association of 3q25.1 (*SIAH2*) with hormonal-positive breast cancer in the Japanese population[37,38]. The size of our current study is three times that of our previous two studies[37,38]. To the best of our knowledge, it is the largest and most well-powered GWAS to date that aims to investigate the genetic architecture as well as to assess the common genetic risk model of breast cancer in the Japanese population.

Results

Evaluation of two GWAS with samples obtained from Phase I-II and Phase III Biobank Japan.

Two sets of GWAS were performed in this study with samples obtained from Phase I-II and Phase III Biobank Japan that consist of 6,669 breast cases and 21,930 female controls.

For sample quality control, identity-by-state analysis was carried out to assess close relatedness in the sample population, no samples were removed as all the samples were independent from each other (Data not shown). Subsequently, principal component analysis (PCA) was performed to assess population substructure of the sample populations. PCA revealed that case and control subjects that participate in this study were clustered into two major clusters, the mainland (Hondo) cluster and the Ryuukyu (southern island of Japan) cluster[39] (Supplementary Fig. 1). Association analyses were carried out by incorporating principal components as covariates to avoid the bias effects from population substructure. The quantile-quantile (Q-Q) plot and the genomic inflation factor (λ_{GC}) of the test statistic for the GWAS of Phase I-II and Phase III were 1.202 and 1.019, respectively (Supplementary Fig. 2). As λ_{GC} value increase correspondingly with sample size, the λ_{GC} value adjusting to a sample size of 1000 was evaluated for GWAS of Phase I-II[40]. The adjusted λ_{1000} value was 1.025, indicating a low possibility of false-positive association by population stratification. The two GWAS studies were subsequently combined by meta-analysis after whole-genome imputation and quality control. A total of 4,946,503 SNPs was evaluated to identify common genetic variations that are associated with breast cancer susceptibility. The Manhattan plot of the whole-genome meta-analysis was plotted by using $-\log_{10}(P\text{-value})$ against chromosome location (Supplementary Fig. 3).

Identification of two novel loci associated with- breast cancer patients.

This study identified a total of 11 genomic loci that surpassed the genome-wide significance level with *P*-value threshold of 5.0×10^{-8} to be associated with breast cancer. These loci included 2q33.1 (*TRAK2-ALS2CR11*), 3q13.11 (*ALCAM*), 3q25.1 (*SIAH2*), 5p12 (*FGF10-MRPS30*), 5q11.2 (*MAP3K1*), 6q25.1 (*CCDC170-ESR1*), 10q26.13 (*FGFR2*), 12p11.22 (*PTHLH*), 12q24.21 (*UBA52P7*), 16q12.1 (2 independent SNPs on *TOX3-CASC16*) and 21q22.12 (*CLIC6-RUNX1*) shown in Table 1. We compared them with previously reported breast cancer susceptibility loci that are identified from GWAS of multiple ethnicities, European, East Asian, African and Ashkenazi Jews, and found that nine loci identified from this study were previously reported while 3q13.11 (*ALCAM*) and 21q22.12 (*CLIC6-RUNX1*) are novel associated loci that surpassed genome-wide significance threshold (Table 1 and Fig. 1). In addition, a total of 40 previously reported SNPs from 37 genomic loci remained to be suggestively associated with the range of *P*-value from 4.88×10^{-2} to 6.01×10^{-6} in this GWAS study (Supplementary Table 2).

The association of the 12 genome-wide significant SNPs comprises breast cancer patients of various cancer subtypes. To evaluate the effect of cancer subtype towards SNP association, a subset analysis was performed for ER+, PR+ and HER2+ respectively (Supplementary Table 3). In general, breast cancer subtype associations for all SNPs were weaker compared to the cumulative cohort (Supplementary Table 3). Validity of the associations were confirmed through permutation, with the permutated *P*-values of combined cohorts for all SNPs in different subtypes showing less than 5% false positives (Supplementary Table 3). However, the effect size shows a similar trend across different cancer subtypes. The results suggest that association of the 12 genome-wide significant SNPs show a consistent trend across different cancer subtypes and the lack of genome-wide significance is due to insufficient statistical power.

In our replication study, none of the 12 genome-wide significant SNPs was statistically significant after adjusting for multiple testing at $P < 0.002$ (0.05/22 independent tests). Despite this, the 12 SNPs showed similar effect size trends with Phase I-II and III cohorts. In addition, the inclusion of the replication cohort improved association *P*-values in the meta-analysis combining GWAS and replication studies (Supplementary Table 4). For example, the association of rs9862599 and rs75286142 on *ALCAM* and *CLIC6-RUNX1* were not replicated after

CHR	MARKER	POS	Risk Allele	Reference Allele	RAF_ Case	RAF_ Control	pMeta	OR	95% CI (lower)	95% CI (upper)	Gene/Nearby gene
2	rs2540431	202367621	T	G	0.330	0.313	2.38E-08	1.131	1.083	1.181	*TRAK2-ALS2CR11*
*3	rs9862599	104864331	G	T	0.092	0.078	8.11E-09	1.234	1.149	1.325	*ALCAM*
3	rs1838337	150479816	G	T	0.657	0.627	1.33E-08	1.128	1.082	1.176	*SIAH2*
5	rs7701466	44663137	T	C	0.518	0.492	2.22E-10	1.138	1.094	1.184	*FGF10-MRPS30*
5	rs79160707	56052938	T	C	0.113	0.100	4.98E-08	1.200	1.124	1.282	*MAP3K1*
6	rs6900157	151954127	C	T	0.305	0.269	5.01E-10	1.147	1.099	1.198	*CCDC170-ESR1*
10	rs2912778	123338654	G	A	0.560	0.507	1.43E-19	1.201	1.155	1.250	*FGFR2*
12	rs805583	28152993	G	A	0.790	0.760	2.86E-10	1.170	1.114	1.228	*PTHLH*
12	rs10744856	115835385	G	A	0.664	0.626	1.04E-08	1.130	1.083	1.178	*UBA52P7*
16	rs3803662	52586341	A	G	0.575	0.523	1.36E-16	1.183	1.136	1.231	*TOX3-CASC16*
16	rs4784227	52599188	T	C	0.279	0.235	2.18E-24	1.271	1.214	1.331	*TOX3-CASC16*
*21	rs75286142	36098645	G	C	0.198	0.178	1.28E-08	1.158	1.101	1.218	*CLIC6-RUNX1*

Table 1. SNPs that surpassed genome-wide significance threshold after meta-analysis of GWAS I + II. RAF: risk allele frequency; P-meta: Meta-analysis of GWAS I + II; OR referred to reference allele: odds ratio; CI: confidence interval; *Novel loci from this GWAS.

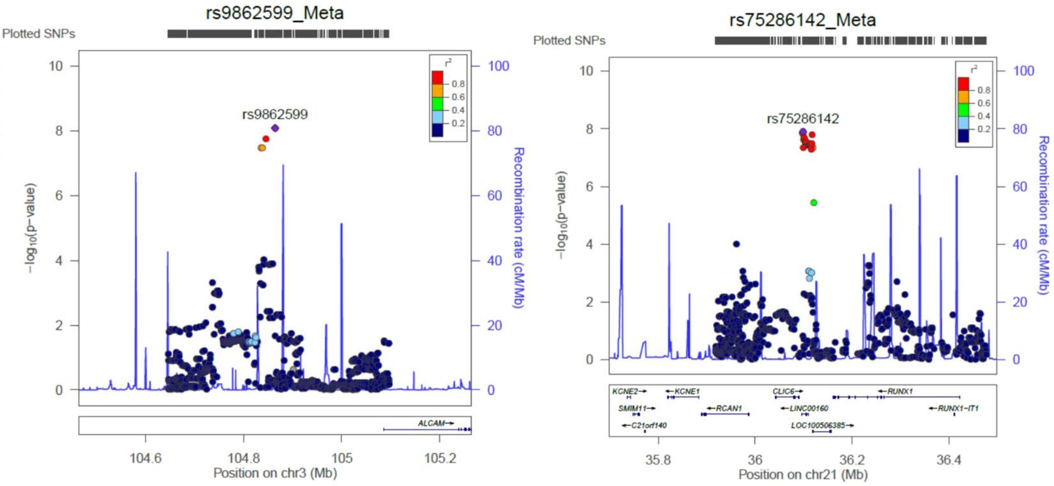

Figure 1. Regional plot for two novel associated loci that include 3q13.11 (ALCAM) and 21q22.12 (CLIC6-RUNX1).

considering multiple testing. However, the combined *P*-values of rs9862599 ($P = 1.14 \times 10^{-8}$; OR = 1.216; 95% CI = 1.136–1.301) and rs75286142 ($P = 2.42 \times 10^{-9}$; OR = 1.157; 95% CI = 1.103–1.214) imply an established link between these SNPs and breast cancer susceptibility (Table 2 and Fig. 1).

Pathway analysis identified two significant pathways associated with breast cancer. Pathway analysis that incorporated whole-genome imputed SNPs to MAGENTA software have identified two significant pathways to be associated with breast cancer. The first associated pathway is dopamine receptor medicated signaling pathway (GSEA *P*-value = 8.10×10^{-5}, FDR = 2.90×10^{-3}) from Panther database that encompasses with genes mostly from the *CLIC*, *EPB41* and *FRMD* families (Supplementary Table 5). The second pathway is protein amino acid deacetylation from GOTERM database (GSEA *P*-value = 1.00×10^{-4}, FDR = 3.72×10^{-2}), which mostly consist of genes from the *SIRT* family (Supplementary Table 5).

Association of rs2540431-linked SNPs with lower expression of *CASP8* and *ALS2CR12* in breast mammary tissue. In order to assess the effects of SNPs with the expression of the nearby genes within the locus, eQTL analyses were performed for the 12 SNPs that surpassed the genome-wide significant threshold by using GTEx Portal, focusing with breast mammary tissue. The analysis also included all LD-linked (LD > = 0.8) SNPs as well. eQTL was detected for rs2540431-linked SNPs rs2714486 and rs2540334 (Supplementary Fig. 4A,B). No eQTL was detected for rs2540431 itself. The risk allele rs2714486-A is correlated with lower expression of *CASP8* ($P = 4.2 \times 10^{-12}$; Normalized effect size = −0.44) and higher expression of *ALS2CR12*($P = 5.2 \times 10^{-7}$; Normalized effect size = 0.35) in breast mammary tissue (Supplementary Fig. 4A). The same trend was observed for risk allele rs2540334-T with lower expression of *CASP8* ($P = 6.9 \times 10^{-12}$; Normalized effect size = −0.43) and higher expression of *ALS2CR12*($P = 7.8 \times 10^{-7}$; Normalized effect size = 0.34) (Supplementary Fig. 4B)

CHR	MARKER	POS	Gene	Stage	Risk_ Allele	Reference_ Allele	RAF_ Case	RAF_ Ctrl	*P*-value	OR	L95	U95	P_Hetero
3	rs9862599	104864331	*ALCAM*	GWAS1	G	T	0.089	0.076	2.95E-06	1.215	1.119	1.319	
3				GWAS2	G	T	0.101	0.081	3.86E-04	1.295	1.123	1.495	
3				Meta-1	G	T	—	—	8.11E-09	1.234	1.325	1.149	
3				Replication	G	T	0.086	0.081	5.58E-01	1.065	0.864	1.313	
3				Meta-2	G	T	—	—	1.44E-08	1.216	1.136	1.301	0.318
21	rs75286142	36098645	*CLIC-RUNX1*	GWAS1	G	C	0.200	0.176	2.36E-07	1.164	1.100	1.232	
21				GWAS2	G	C	0.199	0.18	2.49E-02	1.135	1.017	1.267	
21				Meta-1	G	C	—	—	1.28E-08	1.158	1.218	1.101	
21				Replication	G	C	0.192	0.172	7.36E-02	1.148	0.987	1.336	
21				Meta-2	G	C	—	—	2.42E-09	1.157	1.103	1.214	0.918

Table 2. Validation study of the two novel loci. *P*-value for meta-analysis: Inverse-variance meta-analysis method; *P*-value for heterogeneity: Cochran's Q-test; OR is referred to reference allele: odds ratio; L95: lower 95% confidence interval; U95: upper 95% confidence interval.

Calculation of weighted genetic risk score (wGRS) with SNPs significantly associated with breast cancer risk. wGRS was calculated to evaluate the cumulative effect of the significantly associated SNPs with breast cancer risk. The 12 SNPs that surpassed $P < 5.0 \times 10^{-8}$ were selected and their corresponding weight were calculated to be incorporated into the regression model; these SNPs include rs2540431 (0.08853), rs9862599 (0.18249), rs1838337 (0.11120), rs7701466 (0.10993), rs79160707 (0.18772), rs6900157 (0.17682), rs2912778 (0.19709), rs805583 (0.16442), rs10744856 (0.12644), rs3803662 (0.12069), rs4784227 (0.18744) and rs75286142 (0.15262). The risk score groups were divided into 5 categories, odds ratio of each category increased concordantly to the level of risk score. Individuals (4% of cases and 2% of controls) in group 5, who carry the most risk alleles, have approximately 3.7 times higher risk to develop when comparing group 1 as reference with AUC of 0.593 (Supplementary Table 6 and Supplementary Fig. 5). After genotyping the selected 12 SNPs in the replication set from Aichi Cancer Centre, the same wGRS model was used, and individuals categorized in group 5 was 4.7 times higher risk comparing with group 1 with the AUC of 0.595 (Supplementary Table 6 and Supplementary Fig. 5).

Discussion

This large-scale GWAS, which utilizes a total of 7,650 breast cancer cases and 23,324 female controls, validated nine previously reported loci and suggested two novel loci to be associated with breast cancer in the Japanese population.

Among the previously reported loci, *FGFR2* on chr10q26.13 and *TOX3-CASC16* on chr16q12.1 are the two most significant associated breast cancer susceptible loci across different populations, followed by *MAP3K1* on chr5q11.2 and *CCDC170-ESR1* on chr6q25.1. The locus of *FGFR2* carries a significant disease burden, contributing approximately 16% of all breast cancers[41]. *FGFR2* encodes for fibroblast growth factor receptor type 2, a receptor tyrosine kinase that play a role in growth and differentiation of cells in various tissues. Recent systems biology approach identified SPDEF, ERα, FOXA1, GATA3 and PTTG1 as master regulators of FGFR2 signaling and demonstrated that ERα occupancy responds to FGFR2 signaling[42]. Another follow-up study indicated that risk alleles of SNPs on *FGFR2* augment silencer activity after map to transcriptional silencer elements and the presence of risk variants results in reduced FGFR2 expression and increased estrogen responsiveness[43].

Another susceptibility locus identified in our study is located on chr16q12.1, close to *TOX3* and *CASC16* genes. TOX3 was reported to bind with BRCA1 promoter and negatively regulates BRCA1 expression[44]. Ectopic expression of *TOX3* is associated with tumor progression in breast cancer mouse model[44]. In addition, hypomethylation of the promoter upregulate *TOX3* luminal subtype breast cancer[45]. Taken together, both genetic and epigenetic factors play a role in TOX3 overexpression in breast cancer.

Mitogen-Activated Protein Kinase Kinase Kinase 1 (MAP3K1), is a serine/threonine kinase that involved in the mitogen-activated protein kinase (MAPK) pathway that involves Ras, Raf, Mek, and Erk. MAPK cascade is known to be an important pathway for cancer cell survival, dissemination, and resistance to drug therapy[46]. Lastly, *ESR1* encode for ER-alpha known to acts as a transcriptional regulator by interacting with estrogen and other coactivator proteins. Interestingly, previous study has reported that neoplastic *ESR1–CCDC170* fusions is related to a more aggressive subset of ER + breast cancer[47].

The first novel SNP, rs9862599, identified to be suggestively associated with breast cancer from this GWAS is located on chromosome 3q13.11 near to the 5′ end of *ALCAM* gene. *ALCAM* encode for Activated Leukocyte Cell Adhesion Molecule, which is a glycoprotein that binds to T-cell differentiation antigen CD6 as well as plays a role in the process of cell adhesion and migration[48]. Decreased expression of ALCAM protein is reported as an indicator of poor prognosis in breast cancer[49]. The second novel SNP, rs75286142, is located on *CLIC6-RUNX1*. CLIC6 is one of the family members of chloride intracellular channels and CLIC6 expression profile was shown to be altered in breast cancer[50]. Runx1, a transcription factor, regulates various physiological processes that include cell proliferation, survival, differentiation and cell cycle progression. Importantly, RUNX1 somatic mutations were found in ER+, luminal subtype of breast cancer and indicate a tumor suppressor role for RUNX1[51,52]. Further *in-silico* or functional analysis should be carried out to further investigate the effect of the identified SNP to *ALCAM* and *RUNX1* gene as well as the crosstalk in between germline variations and somatic mutations in these breast cancer-associated genes.

Pathway analysis identified two pathways, dopamine receptor mediated signaling pathway and protein amino acid deacetylation, to be associated with breast cancer in this GWAS study. Dopamine receptor mediated signaling pathway consists of Chloride intracellular ionic channels family (CLIC1-6), Proteins of the 4.1 family (EPB41), Serine/threonine-protein phosphatase family (PPP1C) and FERM Domain Containing (FRDM) family. Among these gene sets, there are substantial reports about the involvement of the CLIC protein in tumorigenic process[53]. For instance, the expression of CLIC4 transcript is regulated by p53 and tumor necrosis factor α as well as related to Myc-induced apoptosis[54,55]. Additionally, CLIC1 protein levels were detected to increase in multiple cancers[53]. The second associated pathway, protein amino acid deacetylation, consist of SIRT and HDAC families. Sirtuins (SIRT1-7) play a significant role in cancer by regulating cancer-associated metabolism, modifying tumor microenvironment and affecting the response to genomic instability[56]. In breast cancer, besides SIRT6 that shows to have increased expression and act as oncogene, SIRT1, SIRT2, SIRT3 and SIRT4 exhibits to have reduced expression and act as tumor suppressor genes[56]. Although SNPs in these gene sets showed only moderate association with breast cancer, gene-set enrichment *P*-value indicated their cumulative association might be of significant for further investigations.

As an apical caspase, CASP8 functions to initiate a caspase cascade upon receipt of apoptosis signaling from a death receptor–ligand interaction[57]. In addition to its central role in apoptosis, CASP8 also plays a number of non-apoptotic roles in cells, namely promoting activation NFκB signaling, regulating autophagy and altering endosomal trafficking, and enhancing cellular adhesion and migration[57]. The role of CASP8 varies and is highly dependent on cellular context[57]. Based on the *in silico* GTEx eQTL analysis, the risk allele of rs2714486-A and rs2540334-T is correlated with lower expression of CASP8. GWAS data show that both risk alleles are more frequent in breast cancer patients compared to healthy controls. All factors considered, our data suggests that CASP8 plays an apoptotic role in breast cancer, suppressing tumor malignancy. The *ALS2CR12* gene product is a structural component of the sperm flagellum[58]. With no previous reports linking it to breast cancer, this makes *ALSCR12* a weak candidate for breast cancer susceptibility.

The AUC-value from wGRS analysis by utilizing 12 SNPs is 0.593, which indicates the current prediction model required further improvement by identifying additional markers that are associated with breast cancer susceptibility. Nevertheless, this well-powered GWAS is a representative study for the Japanese population to identify common genetic variations that are associated with breast cancer.

Methods

Participants in this study.

Breast cancer case samples for the discovery study were recruited from the Biobank Japan (Phase I to III, http://biobankjp.org). Biobank Japan collaboratively collects and stores DNA and serum samples throughout Japan. For the discovery set, a total of 5,272 and 1,397 breast cancer patients were recruited from Phase I-II and Phase III Biobank Japan Project, respectively[59]. In the Phase I-II study, there were 2412 estrogen receptor (ER+), 2010 progesterone receptor (PR+) and 2059 human epidermal growth factor receptor (HER+) breast cancer patients. In the Phase III study, there were 998 ER+, 790 PR+ and 1143 HER+ breast cancer patients. As for control, genotyping information of 16,496 and 5,434 female individuals who do not have cancer history from population-based cohorts of the Tohoku Medical Megabank organization (ToMMo) (http://www.megabank.tohoku.ac.jp/english/), Iwate Tohoku Medical Megabank Organization (IMM), Japan Multi-Institutional Collaborative Cohort (J-MICC) Study and the Japan Public Health Center-based Prospective (JPHC) Study[60], and a hospital-based cohort of Biobank Japan were collected, respectively.

To validate the associations identified from GWAS, a total of 981 breast cancer cases and 1,394 females were collected from Aichi Cancer Centre. The detailed sample demographic and clinical parameters are summarized in Supplementary Table 1.

All participating studies obtained informed consent from all participants by following the protocols approved by their institutional ethical committees before enrollment. The ethical committees from the Institute of Medical Science, the University of Tokyo, RIKEN Center for Integrative Medical Sciences, Tohoku Medical Megabank organization, Iwate Medical University, National Cancer Center, and Aichi Cancer Center have approved this project.

GWAS, quality control and genotype imputation.

For genome-wide genotyping in the discovery study, all Phase I-II and Phase III subjects were genotyped by Illumina HumanOmniExpress v1.1. To perform sample quality control, samples with call rates $< 98\%$ were excluded from the study. We evaluated cryptic relatedness of our samples using identity-by-state. To assess population stratification, principal component analysis (PCA) by using EIGENSTRAT software (ver3.0) was carried out to compare the distribution of principal component scores between samples with four major reference population obtained from the HapMap Database that consist of Europeans (represented by Utah Residents (CEPH) with Northern and Western European Ancestry, CEU), Africans (represented by Yoruba in Ibadan, YRI) and East Asian (represented by Japanese from Tokyo, JPT, and Han Chinese in Beijing, CHB)[61]. The top two principal components that could distinguish the clusters were used to produce scatter plot for the evaluation of distribution (Supplementary Fig. 1). Based on the PCA, 3 outliers were excluded from Phase III GWAS. For SNP quality control, SNPs with call rate < 0.99, SNPs that deviated from Hardy-Weinberg equilibrium among control samples at the threshold of 1.0×10^{-6}, non-polymorphic SNPs, and SNPs from chromosome X, Y as well as mitochondrial SNPs were excluded from further study.

Statistical analysis for both case-control GWAS, Phase I-II and Phase III, were performed by using logistic regression analysis by incorporating associated principal components as covariates. Quantile-quantile plot (Q-Q plot) of each GWAS was constructed between observed *P*-value versus expected *P*-value to evaluate potential population substructure (Supplementary Fig. 2). The genomic inflation factor (λ_{GC}) values were calculated to evaluate the deviation of the GWAS distribution from the null distribution. Since inflation factor scales with sample size and considering our large sample size, λ_{1000} was also calculated. Calculation was done as previously

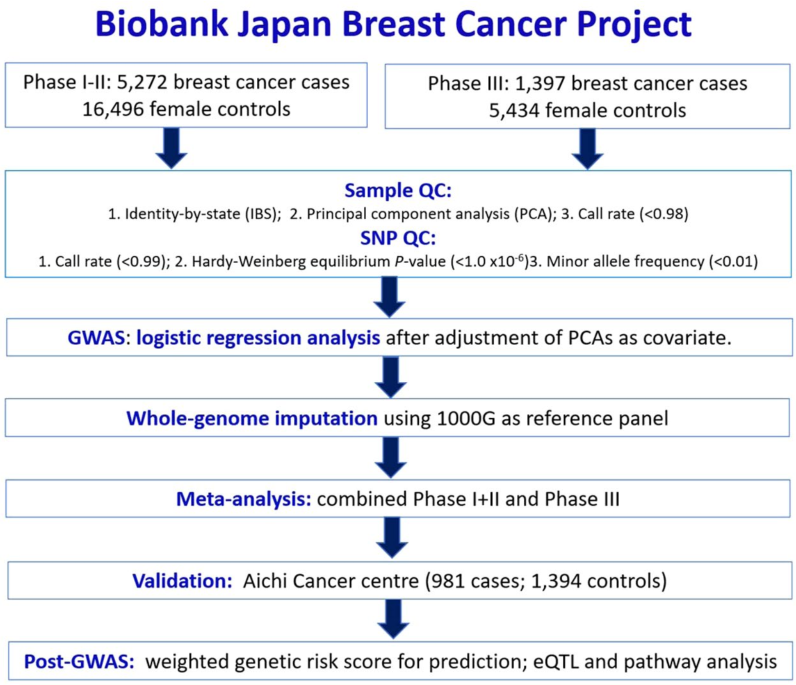
Biobank Japan Breast Cancer Project
Phase I-II: 5,272 breast cancer cases
16,496 female controls
Phase III: 1,397 breast cancer cases
5,434 female controls
Sample QC:
1. Identity-by-state (IBS); 2. Principal component analysis (PCA); 3. Call rate (<0.98)
SNP QC:
1. Call rate (<0.99); 2. Hardy-Weinberg equilibrium P-value (<1.0 x10-6)3. Minor allele frequency (<0.01)
GWAS: logistic regression analysis after adjustment of PCAs as covariate.
Whole-genome imputation using 1000G as reference panel
Meta-analysis: combined Phase I+II and Phase III
Validation: Aichi Cancer centre (981 cases; 1,394 controls)
Post-GWAS: weighted genetic risk score for prediction; eQTL and pathway analysis

Figure 2. Schematic study workflow for the Biobank Japan Breast Cancer project.

described[62]. We used Haploview 4.2 to visualize all SNP association *P*-values in a Manhattan plot, expressed as $-\log_{10}$(*P*-value) against chromosome position (Supplementary Fig. 3)[63]. Whole genome imputation analysis was performed to infer missing genotypes that are not included in the genotyping SNP array. The 1000 G Phase I integrated release version 3 from East Asian descendant that include Japanese from Tokyo, Chinese from Beijing and Chinese from Southern China was used as a reference for this imputation analysis. In brief, SNPs with allele frequency differences that > 0.16 between GWAS and reference panel were excluded. Haplotypes of the samples were phased by using MaCH1.0 before imputation analysis was carried out referring to 1000 G reference panel with map crossover and error rates using 20 iterations of the Markov chain by using Minimac (2013.7.17) software[64,65]. Association study of the imputed genotype dosage was carried out by using mach2dat (ver1.2.4)[66]. Stringent imputation quality (R^2) threshold was applied by excluding SNPs with $R^2 < 0.9$ for further studies. To combine Phase I-II GWAS and Phase III GWAS, whole-genome meta-analysis was carried out by using Inverse-variance meta-analysis method. The schematic study workflow is summarized in Fig. 2. To address the effect of breast cancer subtypes on the association analyses, subset analysis for estrogen receptor (ER+), progesterone receptor (PR+) and human epidermal growth factor receptor 2 (HER2+) breast cancer subtypes was carried out. Minimac dosages for Phase I-II and Phase III GWAS were converted to hard genotypes. Subset association analysis of ER+, PR+ and HER2+ breast cancer was calculated in PLINK using logistic regression adjusting for covariates PC1, PC2 and age. Permutation was performed for 1000 iterations to confirm the validity of the associations, with permutation *P*-value < 0.05 considered a valid association.

Validation study. To select SNPs for validation study, logistic regression analysis was performed by including the effects of primary associated SNPs of a genomic locus in order to exclude SNPs that have the similar effects and to identify SNPs that are independently associated with breast cancer from the primary associated SNPs.

A total of 22 SNPs with $P_{\mathrm{meta}<} 1.0 \times 10^{-5}$ that are not published previously to be associated with breast cancer were selected for validation study by using an independent samples group of 981 breast cases and 1,394 controls from the Aichi Cancer Centre, Japan. Genotyping of the SNPs were performed by using Multiplex Invader Assay. Considering the multiple testing for validation study, Bonferroni correction threshold at $P < 5.00 \times 10^{-3}$ was applied.

To evaluate the combined effects of discovery Phase I-II, Phase III GWAS and validation study, meta-analysis was performed using weighted inverse-variance[62].

Pathway and eQTL analysis. Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA, ver2.4)[67] software was used to assess potential pathways that are associated with breast cancer by using GWAS and 1000 G imputed dataset (RSQ > 0.9). In brief, gene boundary between 110 kb upstream of the gene's most extreme transcript start site and 40 kb downstream to the gene's most extreme transcript end site was set. This boundary was suggested by a comprehensive study of putative functional regulatory element (cis-eQTLs) using expression data from human lymphoblastoid cell line. After assigning SNPs within the gene boundary, cumulative *P*-value of the SNPs within individual genes were calculated after correcting for confounders such as gene size, variant number and LD properties using step-wise multiple linear regression analysis. Gene-set enrichment analysis

P-value was calculated by referring to a total of 3,217 gene-sets from Gene ontology, KEGG, PANTHER, Biocarta and Reactome databases. False discovery rate ($FDR < 0.05$) was used to evaluate the significance of associations of the pathway with breast cancer.

To assess expression quantitative trait loci (eQTL), correlation between SNP genotypes and expression of the nearby genes in a genomic locus was evaluated from the GTEx portal V8 (http://www.gtexportal.org/home/) focusing specifically on breast-mammary tissue. The eQTL analysis was expanded to include all 12 genome-wide level linked variants (LD >= 0.8 ASN 1 K genomes) due to potential differences in the database SNP repository.

Weighted genetic risk score (wGRS). wGRS analysis was carried out to evaluate the cumulative effects of genetic variants associated with breast cancer risk. A total of 12 SNPs that are with $P < 5.0 \times 10^{-8}$ from the whole-genome meta-analysis of Phase I + II and Phase III GWAS were utilized to establish the model. These SNPs include rs2540431 on chromosome 2q33.1, rs9862599 on 3q13.11, rs1838337 on 3q25.1, rs7701466 on 5p12, rs79160707 on 5q11.2, rs6900157 on 6q25.1, rs2912778 on 10q26.13, rs805583 on 12p11.22, rs10744856 on 12q24.21, rs3803662 and rs4784227 on 16q12.1, rs75286142 on 21q22.12. The estimate (weight) of each associated SNP was evaluated by multivariate logistic regression analysis after incorporating 12 SNPs into the model. The cumulative risk scores were calculated by multiplying the weight (estimate) of the SNPs with the number of risk alleles (0/1/2) of the SNPs carried by each of the individual, subsequently the sum of the scores were taken across the number of SNPs in the model. The risk scores were then classified into four different categories that derived from mean and standard deviation (SD); group 1, < mean-1SD; group 2, mean-1SD to mean; group 3 mean to mean + 1 SD; group 4, mean + 1 SD to mean + 2 SD and group 5 > mean + 2 SD. Odds ratio and 95% confidence interval was evaluated by using group 1 as reference. Validation of this model was carried out by genotyping the 12 SNPs using the sample groups from Aichi Cancer Centre. Similar categorization was performed to evaluate the validity of this model. Lastly, receiving operating characteristic (ROC) curve was plotted to observe how well this model could be used as prediction model for breast cancer.

Data availability

The summary statistics for the GWAS will be publicly available from the National Bioscience Database Center (NBDC) Human Database (https://humandbs.biosciencedbc.jp/en/). The genotype data of case subjects in BBJ_Phase I-II and case and control subjects in BBJ_Phase III are available at the Japanese Genotype-phenotype Archive (JGA; http://trace.ddbj.nig.ac.jp/jga/index_e.html) with accession codes JGAS00000000114 for the study and JGAD00000000123 for the genotype data. The other datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 12 November 2018; Accepted: 26 October 2019;
Published online: 22 November 2019

References

1. Foundation for Promotion of Cancer Research. *Cancer statistics in Japan-2018*. 1-130 (FPCR c/o National Cancer Center, 2019).
2. Miki, Y. *et al*. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66–71, https://doi.org/10.1126/science.7545954 (1994).
3. Wooster, R. *et al*. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792, https://doi.org/10.1038/378789a0 (1995).
4. CHEK2 B Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am J Hum Genet* **74**, 1175–1182, https://doi.org/10.1086/421251 (2004).
5. Hofstatter, E. W. *et al*. PALB2 mutations in familial breast and pancreatic cancer. *Fam Cancer* **10**, 225–231, https://doi.org/10.1007/s10689-011-9426-1 (2011).
6. Liaw, D. *et al*. Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet* **16**, 64–67, https://doi.org/10.1038/ng0597-64 (1997).
7. Renwick, A. *et al*. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet* **38**, 873–875, https://doi.org/10.1038/ng1837 (2006).
8. Buniello, A. *et al*. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012, https://doi.org/10.1093/nar/gky1120 (2019).
9. Michailidou, K. *et al*. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94, https://doi.org/10.1038/nature24284 (2017).
10. Cai, Q. *et al*. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat Genet* **46**, 886–890, https://doi.org/10.1038/ng.3041 (2014).
11. Stacey, S. N. *et al*. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat Genet* **39**, 865–869, https://doi.org/10.1038/ng2064 (2007).
12. Haiman, C. A. *et al*. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* **43**, 1210–1214, https://doi.org/10.1038/ng.985 (2011).
13. Easton, D. F. *et al*. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093, https://doi.org/10.1038/nature05887 (2007).
14. Siddiq, A. *et al*. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum Mol Genet* **21**, 5373–5384, https://doi.org/10.1093/hmg/dds381 (2012).
15. Michailidou, K. *et al*. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**(353–361), 361e351–352, https://doi.org/10.1038/ng.2563 (2013).
16. Couch, F. J. *et al*. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. *Nat Commun* **7**, 11375, https://doi.org/10.1038/ncomms11375 (2016).
17. Thomas, G. *et al*. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet* **41**, 579–584, https://doi.org/10.1038/ng.353 (2009).
18. Garcia-Closas, M. *et al*. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45**(392–398), 398e391–392, https://doi.org/10.1038/ng.2561 (2013).
19. Fletcher, O. *et al*. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J Natl Cancer Inst* **103**, 425–435, https://doi.org/10.1093/jnci/djq563 (2011).

20. Turnbull, C. *et al*. Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**, 504–507, https://doi.org/10.1038/ng.586 (2010).
21. Ahsan, H. *et al.* A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. *Cancer Epidemiol Biomarkers Prev* **23**, 658–669, https://doi.org/10.1158/1055-9965.EPI-13-0340 (2014).
22. Gaudet, M. M. *et al*. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. *PLoS Genet* **6**, e1001183, https://doi.org/10.1371/journal.pgen.1001183 (2010).
23. Antoniou, A. C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**, 885–892, https://doi.org/10.1038/ng.669 (2010).
24. Li, J. *et al.* A combined analysis of genome-wide association studies in breast cancer. *Breast Cancer Res Treat* **126**, 717–727, https://doi.org/10.1007/s10549-010-1172-9 (2011).
25. Purrington, K. S. *et al*. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. *Carcinogenesis* **35**, 1012–1019, https://doi.org/10.1093/carcin/bgt404 (2014).
26. Orr, N. *et al*. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. *Nat Genet* **44**, 1182–1184, https://doi.org/10.1038/ng.2417 (2012).
27. Han, M. R. *et al*. Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. *Hum Mol Genet* **25**, 3361–3371, https://doi.org/10.1093/hmg/ddw164 (2016).
28. 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, https://doi.org/10.1186/bcr3158 (2012).
29. Zheng, W. *et al*. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet* **41**, 324–328, https://doi.org/10.1038/ng.318 (2009).
30. Long, J. *et al*. Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. *PLoS Genet* **6**, e1001002, https://doi.org/10.1371/journal.pgen.1001002 (2010).
31. Long, J. *et al*. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. *PLoS Genet* **8**, e1002532, https://doi.org/10.1371/journal.pgen.1002532 (2012).
32. Cai, Q. *et al*. Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum Mol Genet* **20**, 4991–4999, https://doi.org/10.1093/hmg/ddr405 (2011).
33. Huo, D. *et al*. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. *Hum Mol Genet* **25**, 4835–4846, https://doi.org/10.1093/hmg/ddw305 (2016).
34. Fejerman, L. *et al*. Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. *Nat Commun* **5**, 5260, https://doi.org/10.1038/ncomms6260 (2014).
35. Gold, B. *et al*. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc Natl Acad Sci USA* **105**, 4340–4345, https://doi.org/10.1073/pnas.0800441105 (2008).
36. Low, S. K. *et al*. Identification of six new genetic loci associated with atrial fibrillation in the Japanese population. *Nat Genet* **49**, 953–958, https://doi.org/10.1038/ng.3842 (2017).
37. Low, S. K. *et al*. Genome-wide association study of breast cancer in the Japanese population. *PLoS One* **8**, e76463, https://doi.org/10.1371/journal.pone.0076463 (2013).
38. 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, https://doi.org/10.1038/jhg.2012.108 (2012).
39. Yamaguchi-Kabata, Y. *et al*. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet* **83**, 445–456, https://doi.org/10.1016/j.ajhg.2008.08.019 (2008).
40. Freedman, M. L. *et al*. Assessing the impact of population stratification on genetic association studies. *Nat Genet* **36**, 388–393, https://doi.org/10.1038/ng1333 (2004).
41. Hunter, D. J. *et al*. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat Genet* **39**, 870–874, https://doi.org/10.1038/ng2075 (2007).
42. Fletcher, M. N. *et al*. Master regulators of FGFR2 signalling and breast cancer risk. *Nat Commun* **4**, 2464, https://doi.org/10.1038/ncomms3464 (2013).
43. Campbell, T. M. *et al*. FGFR2 risk SNPs confer breast cancer risk by augmenting oestrogen responsiveness. *Carcinogenesis* **37**, 741–750, https://doi.org/10.1093/carcin/bgw065 (2016).
44. Shan, J. *et al*. TNRC9 downregulates BRCA1 expression and promotes breast cancer aggressiveness. *Cancer Res* **73**, 2840–2849, https://doi.org/10.1158/0008-5472.CAN-12-4313 (2013).
45. Han, Y. J., Zhang, J., Zheng, Y., Huo, D. & Olopade, O. I. Genetic and Epigenetic Regulation of TOX3 Expression in Breast Cancer. *PLoS One* **11**, e0165559, https://doi.org/10.1371/journal.pone.0165559 (2016).
46. Burotto, M., Chiou, V. L., Lee, J. M. & Kohn, E. C. The MAPK pathway across different malignancies: a new perspective. *Cancer* **120**, 3446–3456, https://doi.org/10.1002/cncr.28864 (2014).
47. Veeraraghavan, J. *et al*. Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nat Commun* **5**, 4577, https://doi.org/10.1038/ncomms5577 (2014).
48. Davies, S. & Jiang, W. G. ALCAM, activated leukocyte cell adhesion molecule, influences the aggressive nature of breast cancer cells, a potential connection to bone metastasis. *Anticancer Res* **30**, 1163–1168 (2010).
49. Witzel, I. *et al*. Detection of activated leukocyte cell adhesion molecule in the serum of breast cancer patients and implications for prognosis. *Oncology* **82**, 305–312, https://doi.org/10.1159/000337222 (2012).
50. Ko, J. H. *et al*. Expression profiling of ion channel genes predicts clinical outcome in breast cancer. *Mol Cancer* **12**, 106, https://doi.org/10.1186/1476-4598-12-106 (2013).
51. Banerji, S. *et al*. Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature* **486**, 405–409, https://doi.org/10.1038/nature11154 (2012).
52. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70, https://doi.org/10.1038/nature11412 (2012).
53. Peretti, M. *et al*. Chloride channels in cancer: Focus on chloride intracellular channel 1 and 4 (CLIC1 AND CLIC4) proteins in tumor development and as novel therapeutic targets. *Biochim Biophys Acta* **1848**, 2523–2531, https://doi.org/10.1016/j.bbamem.2014.12.012 (2015).
54. Fernandez-Salas, E. *et al*. mtCLIC/CLIC4, an organellular chloride channel protein, is increased by DNA damage and participates in the apoptotic response to p53. *Mol Cell Biol* **22**, 3610–3620, https://doi.org/10.1128/mcb.22.11.3610-3620.2002 (2002).
55. Shiio, Y. *et al*. Quantitative proteomic analysis of myc-induced apoptosis: a direct role for Myc induction of the mitochondrial chloride ion channel, mtCLIC/CLIC4. *J Biol Chem* **281**, 2750–2756, https://doi.org/10.1074/jbc.M509349200 (2006).
56. Chalkiadaki, A. & Guarente, L. The multifaceted functions of sirtuins in cancer. *Nat Rev Cancer* **15**, 608–624, https://doi.org/10.1038/nrc3985 (2015).
57. Stupack, D. G. Caspase-8 as a therapeutic target in cancer. *Cancer Lett* **332**, 133–140, https://doi.org/10.1016/j.canlet.2010.07.022 (2013).
58. Choi, E. & Cho, C. Expression of a sperm flagellum component encoded by the Als2cr12 gene. *Gene Expr Patterns* **11**, 327–333, https://doi.org/10.1016/j.gep.2011.03.003 (2011).

59. Nakamura, K. *et al.* Characteristics and prognosis of Japanese female breast cancer patients: The BioBank Japan project. *J Epidemiol* **27**, S58–S64, https://doi.org/10.1016/j.je.2016.12.009 (2017).
60. Tsugane, S. & Sawada, N. The JPHC study: design and some findings on the typical Japanese diet. *Jpn J Clin Oncol* **44**, 777–782, https://doi.org/10.1093/jjco/hyu096 (2014).
61. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909, https://doi.org/10.1038/ng1847 (2006).
62. de Bakker, P. I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122–128, https://doi.org/10.1093/hmg/ddn288 (2008).
63. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265, https://doi.org/10.1093/bioinformatics/bth457 (2005).
64. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959, https://doi.org/10.1038/ng.2354 (2012).
65. Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am J Hum Genet* **93**, 687–696, https://doi.org/10.1016/j.ajhg.2013.09.002 (2013).
66. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387–406, https://doi.org/10.1146/annurev.genom.9.081307.164242 (2009).
67. Segre, A. V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**, https://doi.org/10.1371/journal.pgen.1001058 (2010).

Acknowledgements

We express our heartfelt gratitude to all the patients who participate in this study. We would like to express our gratefulness to the staff from Biobank Japan, Tohoku Medical Megabank, Iwate Tohoku Medical Megabank, J-MICC study and the JPHC Study for their outstanding assistance. We extend our appreciation to members from the laboratory for statistical analysis for their support. This study was partially supported by the BioBank Japan project and the Tohoku Medical Megabank project, which is supported by the Ministry of Education, Culture, Sports, Sciences and Technology Japan and the Japan Agency for Medical Research and Development. The JPHC Study has been supported by the National Cancer Center Research and Development Fund since 2011 and was supported by a Grant-in-Aid for Cancer Research from the Ministry of Health, Labour and Welfare of Japan from 1989 to 2010. This study was also supported by Nanken-Kyoten, TMDU.

Author contributions

S.-K.L., M.K., T.K. and Y. Mi. conceived and designed the experiments. C.T., K. Matsud., S.-Y.M., N.N., A.S., S.S.N., T.Y., N.S., M.I., S.T., T.T., S.S., M.N. and K.W. initiated and administered tissue collection, storage, and sample data collection. Y.K. and Y. Mo. provided bioinformatic support. S.-K.L. performed genotype imputation and regression analyses. S.-K.L., H.I., K. Matsuo., H.S., Y. Mo., M.K., T.K. and Y. Mi. performed replication analyses. J.I., Y. Mu. Y.N., M.K., T.K. and Y. Mi. supervised this study. S.-K.L. wrote the draft manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-53654-9.

Correspondence and requests for materials should be addressed to S.-K.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.