Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer

A full list of authors and affiliations appears at the end of the article.

Abstract

Correspondence should be addressed to D.F.E. (dfe20@medschl.cam.ac.uk).

A full list of members and affiliations appears in the Supplementary Note

Competing Financial Interests

The authors confirm that they have no competing financial interests

URLs

BCAC http://ccge.medschl.cam.ac.uk/consortia/bcac/index.html
COGS http://www.cogseu.org/
ENCODEx http://www.genome.gov/encode/, genome.ucsc.edu/ENCODE/
icOGS http://www.nature.com/icogs/, http://ccge.medschl.cam.ac.uk/research/consortia/icogs/
IMPUTE http://mathgen.stats.ox.ac.uk/impute/impute.html
MACH http://www.sph.umich.edu/csg/abecasis/MACH/
SHAPEIT https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html

Author Contributions

Genome wide association studies (GWAS) and large scale replication studies have identified common variants in 79 loci associated with breast cancer, explaining ~14% of the familial risk of the disease. To identify new susceptibility loci, we performed a meta-analysis of 11 GWAS comprising of 15,748 breast cancer cases and 18,084 controls, and 46,785 cases and 42,892 controls from 41 studies genotyped on a 200K custom array (iCOGS). Analyses were restricted to women of European ancestry. Genotypes for more than 11M SNPs were generated by imputation using the 1000 Genomes Project reference panel. We identified 15 novel loci associated with breast cancer at $P<5\times10^{-8}$. Combining association analysis with ChIP-Seq data in mammary cell lines and ChIA-PET chromatin interaction data in ENCODE, we identified likely target genes in two regions: SETBP1 on 18q12.3 and RNF115 and PDZK1 on 1q21.1. One association appears to be driven by an amino-acid substitution in EXO1.

Breast cancer is the most common cancer in women worldwide\(^1\). The disease aggregates in families, and has an important inherited component. This inherited component is driven by a combination of rare variants, notably in BRCA1, BRCA2, PALB2, ATM and CHEK2 conferring a moderate or high lifetime risk of the disease, together with common variants at more than 70 loci, identified through GWAS and large scale replication studies\(^2–20\). Taken together, these loci explain approximately one-third of the excess familial risk of breast cancer.

The majority of susceptibility SNPs has been identified through the Breast Cancer Association Consortium (BCAC), a collaboration involving more than 50 case-control studies. We recently reported the results of a large-scale genotyping experiment within BCAC, which utilised a custom array (iCOGS) designed to study variants of interest for breast, ovarian and prostate cancers. iCOGS comprised more than 200,000 variants, of which 29,807 had been selected from combined analysis of nine breast cancer GWAS involving 10,052 breast cancer cases and 12,575 controls of European ancestry. In total, 45,290 breast cancer cases and 41,880 controls of European ancestry from 41 studies were genotyped with iCOGS, leading to the discovery of 41 novel susceptibility loci\(^16\). A parallel analysis identified four loci specific to oestrogen receptor (ER)-negative disease\(^17\). However, additional susceptibility loci may have been missed because they were not selected from the original GWAS, or not included on the array.

Genotype imputation is a powerful approach to infer missing genotypes using the genetic correlations defined in a densely genotyped reference panel, thus providing the opportunity to identify novel susceptibility variants even if not directly genotyped\(^21\). In this analysis we aimed to identify additional breast cancer susceptibility loci by utilising data from all 200k variants on the iCOGS array, and used imputation to estimate genotypes for more than 11M SNPs. We applied the same approach to data from 11 GWAS. After quality control (QC) exclusions, the dataset comprised 15,748 breast cancer cases and 18,084 controls from GWAS, and 46,785 cases and 42,892 controls from 41 studies genotyped with iCOGS (see Online Methods and Supplementary Tables 1a–1e). All subjects were women of European ancestry.
We imputed genotypes using the 1000 Genomes Project March 2012 release as the reference dataset (see Online Methods) The main analyses were based on ~11.6M SNPs that were imputed with imputation $r^2 > 0.3$ and had MAF$>0.005$ in at least one of the datasets. \(^\text{22}\)

Of common SNPs (MAF$>0.05$), 88% were imputed from the iCOGS array with $r^2>0.5$; this compared to 99% of variants for the largest GWAS (UK2), which was genotyped using a 670k SNP array (Figure 1a and 1b, Supplementary Table 2). Thirty-seven per cent of common SNPs were imputed on the iCOGS with $r^2>0.9$, compared with 85% for UK2. Thus, despite being designed as a follow-up of GWAS for different diseases rather than a genome-wide array, the majority of common variants could be imputed using the iCOGS, but the overall imputation quality was poorer than from a standard GWAS array. Imputation quality decreased with decreasing allele frequency (Figure 1c and 1d, Supplementary Table 2).

Log odds ratio estimates and standard errors were calculated for each dataset using logistic regression, adjusting for principal components where it was found to reduce substantially the inflation factor. We then combined the results from each dataset for variants with MAF $>0.5\%$ using a fixed effects meta-analysis. \(^\text{23}\) More than 7,000 variants with a combined $P<5\times10^{-8}$ for association were identified, the large majority of which was in regions previously shown to be associated with breast cancer susceptibility. Of the 79 previously published breast cancer susceptibility loci identified in women of European ancestry, all but eight show evidence of association at $P<5\times10^{-8}$ for overall, ER-positive or ER-negative disease risk (Supplementary Tables 3a, 3b and 3c). For four of the eight variants, \(rs1550623\) on 2q31, \(rs11571833\) on 13q13.1, \(rs12422552\) on 12p13.1 and \(rs11242674\) on 6p25.3, slightly weaker evidence of association was observed. One reported variant, \(rs7726159\) did not reach $P<5\times10^{-8}$ in this ($P=0.0017$) or the previous analysis – it was identified through fine-mapping of the \(TERT\) region on 5p15.33. \(^\text{18}\) One other variant in \(AKAP9\), \(rs6964587\) reported previously did not reach $P<5\times10^{-8}$ but an alternative correlated with it did ($P=3.67\times10^{-8}$ for chr7:91681597:D; $r^2$ between the two markers = 0.98). The two remaining variants \((rs2380205\) on 10p15 and \(rs1045485\) at \(CASP8\)) were reported in earlier analysis, \(^\text{9,24}\) but did not even reach $P<0.0001$, suggesting that they may have been false positive reports. An alternative variant at \(CASP8\), \(rs1830298\) ($r^2=0.06$, $D'=1$ with $rs1045485$ in 1000G CEU) did reach $P<5\times10^{-8}$ in this dataset. \(^\text{25}\)

To assess evidence for additional susceptibility loci, we removed all SNPs within 500kb of susceptibility variants identified previously in women of European ancestry, \(^\text{2-14,16-19}\), leaving 314 variants from 27 regions associated with breast cancer at $P<5\times10^{-8}$ (Supplementary Figures 1 and 2). The strongest associations were observed in a 610kb (\(b37\) 28,314,612- 28,928,858) interval on chromosome 22 (smallest $P=8.2\times10^{-22}$, for \(rs62237573\)). This interval lies approximately 100kb centromeric to \(CHEK2\), and further analysis revealed that the associated SNPs were correlated with the \(CHEK2\) founder variant 1100delC (strongest correlation $r^2=0.39$ for SNP \(rs62235635\)). \(CHEK2\) 1100delC is known to be associated with breast cancer through candidate gene analysis, but has not previously generated an association in GWAS. \(^\text{26,27}\) We performed an analysis adjusting for \(CHEK2\) 1100delC using data on ~40,000 samples that had been genotyped for this variant. The strongest associated variant in this subset was \(rs140914118\); after adjustment for 1100delC

\(Nat Genet.\) Author manuscript; available in PMC 2015 October 01.
the statistical significance diminished markedly \( (P=3.1 \times 10^{-9} \text{ to } P=0.78; \text{ Supplementary Figures 3a and 3b}) \), suggesting that this signal is driven by \textit{CHEK2} 1100delC.

Variants in four regions (\textit{DNAJC1}, 5p12, \textit{PTHLH} and \textit{MKL1}) lay within 2Mb of a previously published susceptibility-associated SNP. In each case, these associations became weaker (no longer \( P<5 \times 10^{-8} \)) after adjustment for the previously associated SNP(s) in the region (data not shown). For four other regions, the significant variants were identified in just one GWAS, and failed imputation \( (r^2<0.3) \) in the remaining datasets, including iCOGS; we did not consider these variants further.

To confirm the results for the remaining 18 regions, we performed re-imputation in the iCOGS dataset without phasing (See Online Methods). Fifteen loci remained associated with breast cancer at \( P<5 \times 10^{-8} \) (Table 1 and Supplementary Table 4). For three of the loci, the most significant SNP, or a highly correlated SNP, had been directly genotyped on iCOGS (Supplementary Table 5); one, rs11205277, had been included on the array because it is associated with adult height\(^{28}\), while the other two were selected based on evidence from the combined breast cancer GWAS but failed to reach genome-wide significance in the earlier analyses. We attempted to genotype the 12 remaining variants on a subset of ~4K samples to confirm the quality of the imputation (10 variants could be directly genotyped, for one region an alternative correlated variant was selected (Supplementary Table 5). For the 11 variants that could be assessed, the \( r^2 \) between the observed and imputed genotypes were close to the \( r^2 \) estimated in the imputation. Furthermore, the estimated effect sizes in the subset of individuals that we genotyped were similar to those obtained from the imputed genotypes (Supplementary Table 5). These results indicate that the analyses based on imputed genotype data were reliable.

There was little or no evidence of heterogeneity in the per-allele odds ratios (ORs) among studies genotyped using iCOGS (Supplementary Table 6 and Supplementary Figure 4). There was little evidence for departure from a log-additive model for any locus, except for a borderline departure for rs6796502 \( (P=0.049) \) for which the ORs for heterozygotes and homozygotes for the risk associated allele were similar (Supplementary Table 6).

The estimated ORs for invasive versus in-situ disease were similar for all the loci \( (P>0.05) \) (Supplementary Table 7). For four of the loci, rs12405132, rs12048493, rs4593472 and rs6507583 the association was stronger for ER positive disease (case only \( P<0.05 \)) (Supplementary Table 8). Seven of the loci were associated with ER-negative disease \( (P<0.05) \) but none had a stronger association for ER-negative than ER-positive disease. Two of the loci showed significant trends in the OR by age at diagnosis: for rs13162653, the OR was higher at younger ages \( (P=0.007) \), while for rs6507583, the OR was higher at older ages \( (P=0.006) \) (Supplementary Table 9). One of the variants, chr17:29230520:D in \textit{ATAD5} is correlated with a variant that has also been shown to be associated with serous ovarian cancer in a meta-analysis\(^{29}\) \( (r^2=0.93 \text{ between chr17:29230520:D and chr17:29181220:I}) \).

To approach the task of identifying the likely causal variants and genes underlying these associations, we first defined the set of all SNPs correlated with each of the 15 lead SNPs and that could not be ruled out as potentially causal (based on a likelihood ratio 100:1\(^{30}\)).
resulting in a subset of 522 variants (Supplementary Table 10). One of the variants, rs72755295, lies in an intron of EXO1, encoding a protein involved in mismatch repair. It is strongly correlated with only one other variant, rs4149909, coding for an amino-acid substitution in EXO1 (p.Asn279Ser; CADD score 33), suggesting that this variant is likely to be functionally related to breast cancer risk. None of the remaining SNPs lay within gene coding sequences, consistent with previous observations that most common cancer susceptibility variants are regulatory. For each of the remaining 520 variants, we then looked for enhancer elements in mammary cell lines, based on ENCODE ChIP-Seq data32,33. To identify potential gene targets, we combined this information with ENCODE ChIA-PET chromatin interaction data. We identified two regions in which the associated variants overlapped with putative enhancer sequences and for which consistent promoter interactions were predicted (Table 1). For rs12405132 at 1q21.1, we identified four potential interacting genes, RNF115, POLR3C, PDZK1 and PIAS3 (Figure 2). Of these, the strongest evidence was for RNF115 and PDZK1: three of the 64 potentially causal variants lay in interacting enhancer regions. RNF115 (also known as BCA2) is an E3 ubiquitin ligase RING finger protein that is overexpressed in ER-positive breast cancers34. PDZK1 is a scaffold protein that connects plasma membrane proteins and regulatory components, regulating their surface expression in epithelial cells apical domains, and has been proposed to act as an oncogene in breast cancer35.

SNPs correlated with rs6507583 at 18q12.3 lay in regions interacting with the promoter of SETBP1 (Supplementary Figure 5). The encoded protein has been shown to bind the SET nuclear oncogene which is involved in DNA replication.

We utilised data from TCGA to assess associations between the 15 novel susceptibility variants and expression of neighbouring genes in breast tumors and normal breast tissue. One SNP, rs7707921, was strongly associated with RPS23 expression in all tissues (Supplementary Table 11, Supplementary Figure 6). However, stronger associations with expression were observed with more telomeric SNPs that were less strongly associated with disease risk (top eQTL SNP rs3739: P=10−23, P-risk=5.28×10−7), suggesting that this association may be coincidental. SNP, rs7707921 was also more weakly associated with expression of ATP6AP1L (P=5.6×10−5 in tumours, P=0.066 in normal tissue).

Based on the estimated ORs in the iCOGS stage (all but one of which were in the range 1.05–1.10), the 15 novel loci identified here would explain a further ~2% of the 2-fold familial risk of breast cancer. Taken together with previously identified loci, more than 90 independent common susceptibility loci for breast cancer have been identified, explaining ~16% of the familial risk. We estimate assuming a log-additive model that, based on genotypes for variants at these loci, approximately 5% of women in the general population have a >2 fold increased risk of breast cancer and 0.7% of women have a >3 fold increased risk. In the current analyses, more than 50% of variants with MAF>0.005 in subjects of European ancestry were well imputable (r^2>0.5) These results suggest that, while there may be further susceptibility variants with comparable associated effects that were not well imputed, the identification of many additional loci will require larger association studies. In the meantime, inclusion of these additional loci in polygenic risk scores will improve our
ability to discriminate between high and low risk individuals, potentially improving breast cancer screening and prevention.

**Online Methods**

Details of the subjects, genotyping and QC measures for the GWAS and iCOGS data are described elsewhere. All participating studies were approved by their appropriate ethics review board and all subjects provided informed consent. Analyses were restricted to women of European ancestry. All imputations were performed using the 1000 Genomes Project March 2012 release as the reference panel. Of the 11 GWAS, 8 (C-BCAC) plus a subset of the BPC3 GWAS (CGEMS) were used in the combined GWAS analysis that nominated 29,807 SNPs for the array. The BPC3 and TNBCC GWAS nominated additional SNPs with evidence for association with ER-negative or triple-negative (ER-, PR- and HER2- negative) breast cancer. The EBCG GWAS was not used to nominate SNPs for the iCOGS array.

For eight GWAS (C-BCAC), genotypes were imputed in a two-stage procedure, using SHAPEIT to derive phased genotypes and IMPUTEv2 to perform the imputation on the phased data. We performed the imputation using 5Mb non-overlapping intervals for the whole genome. OR estimates and standard errors where obtained using logistic regression with SNPTEST. For two of the studies we adjusted for the 3 leading principal components as it was found to reduce materially the inflation factor; for the rest of the studies no such adjustment was necessary. For the remaining three GWAS (BPC3, TNBCC and EBCG), imputation was performed using MACH and Minimac. Genomic control adjustment was applied to each GWAS as previously described. The iCOGS data were also imputed in a two-stage procedure using SHAPEIT and IMPUTEv2, again using 5Mb non-overlapping intervals. We split the ~90K samples into 10 subsets, where possible keeping subjects from the same study in the same subset. We obtained OR estimates and standard errors using logistic regression adjusting for study and 9 principal components.

For the regions showing evidence of association we repeated the imputation in iCOGS, using IMPUTEv2 but without pre-phasing in SHAPEIT to improve imputation accuracy. We also increased the number of MCMC iterations from 30 to 90, and increased the buffer region from 250kb to 500kb.

**Meta-analysis**

OR estimates and standard errors were combined in a fixed effects inverse variance meta-analysis using METAL. For the GWAS, results were included in the analysis for all SNPs with MAF>0.01 and imputation r²>0.3, except for the TN GWAS where the criteria were r²>0.9 and MAF>0.05. For iCOGS, we included all SNPs with r²>=0.3 and MAF>0.005.

**Confirmatory genotyping**

The best variant in each region after the re-imputation and meta-analysis was genotyped in 4123 samples from SEARCH, using Taqman according to the manufacturer’s instructions. The squared correlations between the observed genotypes and the genotypes estimated by imputation are shown in Supplementary Table 5. For all the imputed SNPs the squared
correlations was greater than 0.7, the call-rates were $\geq 0.98$ and there was no evidence of departure of genotype frequencies from those expected under HWE ($p>0.1$).

**eQTL analyses**

Germline genotype, mRNA expression, and somatic copy number data for samples taken from breast tumours and tumour-adjacent normal tissue were obtained from The Cancer Genome Atlas$^{38}$. The copy number and genotype data were measured using the Affymetrix Genome-Wide Human SNP 6.0 platform. For the mRNA expression data, we used the expression profiles obtained using the Agilent G4502A-07-3 microarray. The genotype data were subjected to the following quality control filters. SNPs were excluded in case of low frequency (MAF < 1%), low call rate (< 95%), or departure from Hardy-Weinberg equilibrium at $P < 1 \times 10^{-13}$. Individuals were excluded based on low call rate (< 95%), or high heterozygosity (false discovery rate < 1%). Furthermore, individuals were also excluded in case of non-European ancestry, or male gender. Quality control and intersection with the other genomic data types resulted in 380 tumour samples and 56 normal samples.

The genotype data were imputed as described above. eQTL analysis was performed using linear regression with SNPTEST, regressing the mRNA expression of selected candidate genes on the imputed genotype. For each gene, we performed the eQTL analysis against every microarray probe that uniquely maps to that gene. We adjusted the analyses for somatic copy number of the gene, and for SNPs that intersect the probe sequence, provided that their MAF exceeds 1% in individuals of European ancestry in the 1,000 Genomes data.

**Enhancer analyses**

Maps of enhancer regions with predicted target genes were obtained from Hnisz et al.$^{33}$, and Corradin et al.$^{32}$. Enhancers active in the mammary cell types MCF7, HMEC and HCC1954 were intersected with candidate causal variants using Galaxy. ENCODE ChIA-PET chromatin interaction data from MCF7 cells (mediated by RNAPolII and ERα) were downloaded using the UCSC Table browser. Galaxy was used to identify ChIA-PET interactions between an implicated mammary cell enhancer (containing a strongly associated variant) and a predicted gene promoter (defined as regions 3 kb upstream and 1 kb downstream of the transcription start site).

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**Authors**

Kyriaki Michailidou$^1$, Jonathan Beesley$^2$, Sara Lindstrom$^3$, Sander Canisius$^4$, Joe Dennis$^1$, Michael Lush$^1$, Mel J Maranian$^5$, Manjeet K Bolla$^1$, Qin Wang$^1$, Mitul Shah$^5$, Barbara J Perkins$^5$, Kamila Czene$^6$, Mikael Eriksson$^6$, Hafed Darabi$^6$, Judith S Brand$^6$, Stig E Bojesen$^{7,8,9}$, Børge G Nordestgaard$^{7,8,9}$, Henrik Flyger$^{10}$, Sune F Nielsen$^{7,8}$, Nazneen Rahman$^{11}$, Clare Turnbull$^{11}$, BOCS$^{12}$, Olivia Fletcher$^{13}$, Julian Peto$^{14}$, Lorna Gibson$^{14}$, Isabel dos-Santos-Silva$^{14}$, Jenny Chang-Claude$^{15}$, Dieter Flesch-Janys$^{16,17}$, Anja Rudolph$^{15}$, Ursula Eilber$^{15}$, Sabine Behrens$^{15}$, Heli
Affiliations

1Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
2Department of Genetics, QIMR (Queensland Institute for Medical Research) Berghofer Medical Research Institute, Brisbane, Queensland, Australia
3Program in Genetic Epidemiology and Statistical Genetics, Harvard School of Public Health, Boston, MA, USA
4Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands
5Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK
6Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden
7Copenhagen General Population Study, Herlev Hospital, Copenhagen University Hospital, Herlev, Denmark
8Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, Herlev, Denmark
9Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
10Department of Breast Surgery, Herlev Hospital, Copenhagen University Hospital, Herlev, Denmark
11Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK
12Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, UK
13Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK
14Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
15Department of Cancer Epidemiology/Clinical Cancer Registry, University Clinic Hamburg-Eppendorf, Hamburg, Germany
16Institute for Medical Biometrics and Epidemiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany
17Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
18Department of Clinical Genetics, Helsinki University Central Hospital, Helsinki, Finland
19Department of Oncology, University of Helsinki and Helsinki University Central Hospital, Helsinki, Finland
20Center for Cancer Epidemiology and Prevention, University of Chicago, Chicago, Illinois
21Department of Health Studies, University of Chicago, Chicago, Illinois
22Department of Medicine, University of Chicago, Chicago, Illinois
23Department of Human Genetics, University of Chicago, Chicago, Illinois
24Comprehensive Cancer Center, University of Chicago, Chicago, Illinois
25Department of Health Research and Policy - Epidemiology, Stanford University School of Medicine, Stanford, CA.
USA 27Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA 28Department of Epidemiology, Cancer Prevention Institute of California, Fremont, CA, USA 29Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington 30Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 31Department of Environmental Health Sciences, Columbia University Mailman School of Public Health, New York, New York 32Norway Cancer Registry, Norway 33Centre for Molecular, Environmental, Genetic, and Analytic Epidemiology, Melbourne School of Population Health, Melbourne, Australia 34Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA 35Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA 36Epidemiology Program, University of Hawaii Cancer Center, Honolulu, Hawaii, USA 37Department of Radiation Oncology and Molecular Radiation Sciences, Johns Hopkins Medicine, Baltimore, MD, USA 38Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA 39Vesalius Research Center, VIB, Leuven, Belgium 40Laboratory for Translational Genetics, Department of Oncology, University of Leuven, Leuven, Belgium 41Multidisciplinary Breast Center, University Hospitals Leuven, Leuven, Belgium 42Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands 43Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA 44Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA 45Department of Clinical Genetics, VU University Medical Center, section Oncogenetics, Amsterdam, The Netherlands 46Division of Biomedical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands 47Human Genetics Division, Genome Institute of Singapore, Singapore 48Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea 49Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Korea 50Cancer Research Institute, Seoul National University, Seoul, Korea 51Seoul National University College of Medicine, Seoul, Korea 52Department of Preventive Medicine, Kyushu University Faculty of Medical Sciences, Fukuoka, Japan 53Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Aichi, Japan 54Department of Breast Oncology, Aichi Cancer Center Hospital, Nagoya, Japan 55Epidemiology Center for Disease Control and Prevention, Mie University Hospital, Tsu, Mie, Japan 56Inserm (National Institute of Health and Medical Research), CESP (Center for Research in Epidemiology and Population Health), U1018, Environmental Epidemiology of Cancer, 94807 Villejuif, France 57University Paris-Sud, UMRS 1018, 94807 Villejuif, France 58Université Paris Sorbonne Cité, UMR-S775 Inserm, Paris, France 59Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, Germany 60Molecular Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany 61National Center for Tumor Diseases, University of Heidelberg, Heidelberg, Germany 62Human Genotyping-CEGEN Unit, Human Cancer Genetics Program, Spanish National Cancer Research Centre (CNIO), Madrid, Spain 63Centro de Investigación en Red de Enfermedades Raras
(CIBERER), Valencia, Spain 64Servicio de Oncología Médica, Hospital Universitario La Paz, Madrid, Spain 65Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, Oviedo, Spain 66Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA 67Shanghai Center for Disease Control and Prevention, Changning, Shanghai, China 68Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China 69Sheffield Cancer Research Centre, University of Sheffield, Sheffield, UK 70Department of Oncology, University of Sheffield, Sheffield, UK 71Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK 72Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada 73Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada 74Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada 75Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada 76Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada 77Laboratory Medicine Program, University Health Network, Toronto, ON, Canada 78Research Oncology, Division of Cancer Studies, King’s College London, Guy’s Hospital, London, UK 79Wellcome Trust Centre for Human Genetics, University of Oxford, UK 80Oxford Biomedical Research Centre, University of Oxford, UK 81Surgery, School of Medicine, National University of Ireland, Galway 82Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden 83Department of Oncology - Pathology, Karolinska Institutet, Stockholm, Sweden 84Cancer Research Initiatives Foundation, Sime Darby Medical Centre, Subang Jaya, Malaysia 85Breast Cancer Research Unit, University Malaya Cancer Research Institute, University Malaya Medical Centre (UMMC), Kuala Lumpur, Malaysia 86Department of Medical Oncology, Erasmus MC Cancer Institute, 3008 AE Rotterdam, The Netherlands 87Department of Clinical Genetics, Erasmus University Medical Center, 3000 CA Rotterdam, The Netherlands 88International Epidemiology Institute, Rockville, MD, USA 89Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA 90Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC, Australia 91Department of Pathology, The University of Melbourne, Melbourne, Australia 92Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 93Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan 94School of Public Health, China Medical University, Taichung, Taiwan 95Cancer Center, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan 96Department of Surgery, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung, Taiwan 97Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Radiumhospitalet, Oslo, Norway 98Institute of Clinical Medicine, University of Oslo (UiO), Oslo, Norway 99Department of Clinical Molecular Biology (EpiGen), University of Oslo (UiO), Oslo, Norway 100Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, VIC, Australia 101Anatomical Pathology, The Alfred Hospital, Melbourne, VIC, Australia
Genomic Epidemiology Group, German Cancer Research Center (Deutsches Krebsforschungszentrum; DKFZ), Heidelberg, Germany

Bureau of Epidemiologic Research, Academy of Athens, Athens, Greece; Hellenic Health Foundation, Athens, Greece

Department of Epidemiology, Julius Center for Health Sciences and Primary Care, University Medical Center, Utrecht, The Netherlands

Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, UK

Department of Community Medicine, Faculty of Health Sciences, University of Tromsø - The Arctic University of Norway, Tromsø, Norway

Department of Surgical and Perioperative Sciences, Umea University, Sweden

School of Clinical Medicine, Cambridge Institute of Public Health, University of Cambridge, UK

Molecular and Nutritional Epidemiology Unit, Cancer Research and Prevention Institute, ISPO, Florence, Italy

Section for Epidemiology, Aarhus University, Aarhus, Denmark

Inserm Centre for research in Epidemiology and Population Health (CESP), U1018, Nutrition, Hormones and Women’s Health team, Villejuif, France

Univ Paris Sud, Villejuif, France

Department of Epidemiology, Consejería de Sanidad y Política Social, CIBER de Epidemiología y Salud Publica, Murcia, Spain

Division of Gynaecology and Obstetrics, Technische Universität München, Munich, Germany

Center for Hereditary Breast and Ovarian Cancer, University Hospital Cologne, Cologne, Germany

Center for Integrated Oncology (CIO), University Hospital Cologne, Cologne, Germany

Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine, University of Cologne, Cologne, Germany

Institute of Human Genetics, University Hospital Heidelberg, Heidelberg, Germany

Division of Health Sciences, Warwick Medical school, Warwick University, Coventry, UK

Institute of Population Health, University of Manchester, Manchester, UK

Ministry of Public Health, Nonthaburi, Thailand

Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore

Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore

Division of General Surgery, National University Health System

University Breast Center Franconia, Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany

University of California at Los Angeles, David Geffen School of Medicine, Department of Medicine, Division of Hematology and Oncology, Los Angeles, CA, USA

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

German Cancer Consortium (DKTK), Heidelberg, Germany

Saarland Cancer Registry, Saarbrücken, Germany

Division of Breast Cancer Research, Institute of Cancer Research, London, UK

Department of Medicine, McGill University, Montreal, QC, Canada

Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, Montreal, QC, Canada

Département de Santé environnementale et santé au travail, École de santé publique, Université de Montréal, Montreal, Quebec, Canada
Canada 135Centre Hospitalier Universitaire de Québec Research Center, Quebec City, Quebec, Canada 136Laval University, Quebec City, Quebec, Canada
137Laboratory of Cancer Genetics and Tumor Biology, Department of Clinical Chemistry and Biocenter Oulu, University of Oulu, NordLab Oulu/Oulu University Hospital, Oulu, Finland 138Department of Oncology, Oulu University Hospital, University of Oulu, Oulu, Finland 139Department of Surgery, Oulu University Hospital, University of Oulu, Oulu, Finland
140Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany 141University of Tübingen, Tübingen, Germany 142German Cancer Research Center (DKFZ), Heidelberg, Germany
143Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany 144Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany 145Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS (Istituto di Ricerca e Cura a Carattere Scientifico) Istituto Nazionale Tumori (INT), Milan, Italy 146Fondazione Istituto FIRC (Italian Foundation for Cancer Research) di Oncologia Molecolare (IFOM), Milan, Italy 147Unit of Medical Genetics, Department of Preventive and Predictive Medicine, Fondazione IRCCS (Istituto di Ricerca e Cura a Carattere Scientifico) Istituto Nazionale Tumori (INT), Milan, Italy 148Department of Experimental Oncology, Istituto Europeo di Oncologia, Milan, Italy 149Cogentech Cancer Genetic Test Laboratory, Milan, Italy 150Department of Radiation Oncology, Hannover Medical School, Hannover, Germany 151Gynaecology Research Unit, Hannover Medical School, Hannover, Germany 152School of Medicine, Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland 153Cancer Center of Eastern Finland, University of Eastern Finland, Kuopio, Finland 154Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland 155Cancer Center, Kuopio University Hospital, Kuopio, Finland 156Central Finland Hospital District, Jyväskylä Central Hospital, Jyväskylä, Finland 157Department of Human Genetics and Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands 158Department of Surgical Oncology, Leiden University Medical Center, Leiden, The Netherlands 159Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands 160Department of Clinical Genetics, Leiden University Medical Center Leiden, The Netherlands 161Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland 162National Cancer Institute, Bangkok, Thailand 163International Agency for Research on Cancer, Lyon, France 164Department of Molecular Virology, Immunology and Medical Genetics, The Ohio State University, Columbus, OH, USA 165Roswell Park Cancer Institute, Buffalo, NY, USA 166Molecular Diagnostics Laboratory, IRRP, National Centre for Scientific Research “Demokritos”, Aghia Paraskevi Attikis, Athens, Greece 167Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia 168Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA 169Department of Epidemiology, University of California Irvine, Irvine, CA, USA
Acknowledgments

The authors wish to thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. BCAC is funded by Cancer Research UK [C1287/A10118, C1287/A12014] and by the European Community’s Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS). Meetings of the BCAC have been funded by the European Union COST programme [BM0606]. Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710, C8197/A16565), the Canadian Institutes of Health Research for the “CHR Team in Familial Risks of Breast Cancer” program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec – grant # PSR-SIIRI-701. Combining the GWAS data was supported in part by The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148065-01 (DRIVE, part of the GAME-ON initiative). For a full description of funding and acknowledgments, see Supplementary Note.

References

Figure 1.
Histograms of the imputation $r^2$ a) Histogram of the imputation $r^2$ for the iCOGS for variants with MAF>0.05 b) Histogram of the imputation $r^2$ for the UK2 GWAS for variants with MAF>0.05 c) Histogram of the imputation $r^2$ for the iCOGS for variants with MAF<=0.05 d) Histogram of the imputation $r^2$ for the UK2 GWAS for variants with MAF<=0.05.
Figure 2.
The chromosome 1 locus tagged by rs12405132 a) The Manhattan Plot displays the strength of genetic association ($-\log_{10} P$) versus chromosomal position (Mb), where each dot presents a genotyped (solid black dot) or imputed (red circle) SNP (in the iCOGS stage). The purple horizontal line represents the threshold for genome-wide significance ($P=5\times10^{-8}$). Gene structures are depicted as well as the location of SNPs with MAF>0.01 which were neither imputed reliably nor genotyped. b) Mammary cell enhancer locations as defined in Corradin et al.\textsuperscript{32} and Hnisz et al.\textsuperscript{33}, are shown where elements overlapping the best associated SNPs are labelled with their predicted target genes. A subset of ChiA-PET interactions in MCF7 cells (mediated by either RNApolII or ERα) between enhancers and their target gene promoters are also shown.
## Table 1

Results for the 15 regions with combined P<5×10^-8. Results are shown for the strongest associated variant in the region.

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
<th>r²</th>
<th>GWAS P</th>
<th>iCOGS OR (95% CI)</th>
<th>iCOGS P</th>
<th>Combined GWAS + iCOGS P</th>
<th>Genes within ±3kb</th>
<th>Enhancers in MCF7/HMEC</th>
<th>eQTLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1q21.1</td>
<td>C/T</td>
<td>0.36</td>
<td>0.96</td>
<td>0.96 (0.92–0.99)</td>
<td>0.00962</td>
<td>0.95 (0.93–0.97)</td>
<td>2.34×10^-17</td>
<td>7.92×10^-9</td>
<td>-</td>
</tr>
<tr>
<td>1q21.2</td>
<td>A/C</td>
<td>0.34</td>
<td>0.76</td>
<td>1.04 (0.99–1.09)</td>
<td>0.121</td>
<td>1.07 (1.05–1.10)</td>
<td>1.66×10^-9</td>
<td>1.10×10^-9</td>
<td>-</td>
</tr>
<tr>
<td>1q43</td>
<td>A/G</td>
<td>0.03</td>
<td>0.94</td>
<td>1.19 (1.03–1.35)</td>
<td>0.021</td>
<td>1.15 (1.09–1.22)</td>
<td>2.60×10^-9</td>
<td>1.82×10^-9</td>
<td>EXO1</td>
</tr>
<tr>
<td>2q11.3</td>
<td>G/A</td>
<td>0.09</td>
<td>0.91</td>
<td>0.92 (0.87–0.98)</td>
<td>0.00657</td>
<td>0.92 (0.89–0.95)</td>
<td>8.13×10^-9</td>
<td>1.84×10^-8</td>
<td>-</td>
</tr>
<tr>
<td>2q5.1</td>
<td>G/C</td>
<td>0.15</td>
<td>0.72</td>
<td>0.92 (0.88–0.95)</td>
<td>0.0524</td>
<td>0.95 (0.93–0.97)</td>
<td>5.18×10^-9</td>
<td>1.71×10^-9</td>
<td>-</td>
</tr>
<tr>
<td>3p21.3</td>
<td>A/T</td>
<td>0.03</td>
<td>0.88</td>
<td>0.94 (0.91–0.98)</td>
<td>0.00302</td>
<td>0.93 (0.91–0.95)</td>
<td>4.09×10^-9</td>
<td>5.00×10^-11</td>
<td>AGT10</td>
</tr>
<tr>
<td>3p13.3</td>
<td>C/T</td>
<td>0.04</td>
<td>0.81</td>
<td>1.06 (1.02–1.09)</td>
<td>0.00101</td>
<td>1.05 (1.03–1.08)</td>
<td>1.66×10^-6</td>
<td>6.38×10^-9</td>
<td>-</td>
</tr>
<tr>
<td>4p14</td>
<td>A/T</td>
<td>0.23</td>
<td>0.88</td>
<td>0.94 (0.90–0.98)</td>
<td>0.00302</td>
<td>0.93 (0.91–0.95)</td>
<td>4.09×10^-9</td>
<td>5.00×10^-11</td>
<td>-</td>
</tr>
<tr>
<td>4q21.1</td>
<td>C/T</td>
<td>0.03</td>
<td>0.92</td>
<td>1.05 (1–1.1)</td>
<td>0.0372</td>
<td>1.05 (1.03–1.08)</td>
<td>4.53×10^-9</td>
<td>4.84×10^-8</td>
<td>-</td>
</tr>
<tr>
<td>6q2.3</td>
<td>C/T</td>
<td>0.35</td>
<td>1.00</td>
<td>0.92 (0.88–0.96)</td>
<td>2.57×10^-4</td>
<td>0.95 (0.94–0.97)</td>
<td>3.97×10^-6</td>
<td>1.83×10^-9</td>
<td>KL40660</td>
</tr>
<tr>
<td>8p11.33</td>
<td>G/A</td>
<td>0.17</td>
<td>0.94</td>
<td>0.89 (0.85–0.93)</td>
<td>6.32×10^-6</td>
<td>0.95 (0.93–0.97)</td>
<td>0.00059</td>
<td>1.06×10^-8</td>
<td>-</td>
</tr>
<tr>
<td>8q2.3</td>
<td>G/A</td>
<td>0.36</td>
<td>0.97</td>
<td>1.07 (1.03–1.12)</td>
<td>0.00557</td>
<td>1.05 (1.03–1.07)</td>
<td>4.87×10^-6</td>
<td>1.72×10^-8</td>
<td>LINCO8536</td>
</tr>
<tr>
<td>11q22.12</td>
<td>T/C</td>
<td>0.06</td>
<td>0.73</td>
<td>0.94 (0.90–0.98)</td>
<td>0.0069</td>
<td>0.94 (0.92–0.96)</td>
<td>1.06×10^-6</td>
<td>4.48×10^-9</td>
<td>RN3</td>
</tr>
<tr>
<td>11q22.12</td>
<td>T/C</td>
<td>0.06</td>
<td>0.73</td>
<td>0.94 (0.90–0.98)</td>
<td>0.0069</td>
<td>0.94 (0.92–0.96)</td>
<td>1.06×10^-6</td>
<td>4.48×10^-9</td>
<td>RN3</td>
</tr>
<tr>
<td>17q12.12</td>
<td>G/T/G</td>
<td>0.20</td>
<td>0.77</td>
<td>0.94 (0.89–0.98)</td>
<td>0.009</td>
<td>0.93 (0.91–0.96)</td>
<td>1.11×10^-6</td>
<td>3.34×10^-8</td>
<td>ATAD5</td>
</tr>
<tr>
<td>17q25.3</td>
<td>A/G</td>
<td>0.50</td>
<td>0.93</td>
<td>0.94 (0.91–0.98)</td>
<td>0.00754</td>
<td>0.95 (0.93–0.97)</td>
<td>4.52×10^-7</td>
<td>1.40×10^-9</td>
<td>-</td>
</tr>
<tr>
<td>18q2.3</td>
<td>A/G</td>
<td>0.07</td>
<td>0.96</td>
<td>0.91 (0.85–0.98)</td>
<td>0.00803</td>
<td>0.91 (0.88–0.95)</td>
<td>1.23×10^-6</td>
<td>3.20×10^-8</td>
<td>SETBP1</td>
</tr>
</tbody>
</table>

1. Chromosome
2. Build 37 position
3. Reference/effect allele, based on the forward strand
4. Mean effect allele frequency over all controls
5. Imputation r² in the iCOGS samples (calculated by the average info score from IMPUTEv2)
6. Per allele odds ratio for the minor allele relative to the major allele
7. P value for the 1df trend test