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 15,748 breast cancer cases and 18,084 controls together with 46,785 cases and 42,892 controls from 41 studies genotyped on a 211,155-marker custom array (iCOGS). Analyses were restricted to women of European ancestry. We generated genotypes for more than 11 million SNPs by imputation using the 1000 Genomes Project reference panel, and we identified 15 new loci associated with breast cancer at $P < 5 \times 10^{-8}$. Combining association analysis with ChIP-seq chromatin binding data in mammary cell lines and ChIA-PET chromatin interaction data from ENCODE, we identified likely target genes in two regions: SETBP1 at 18q12.3 and RNF115 and PDZK1 at 1q21.1. One association appears to be driven by an amino acid substitution encoded in EXO1.

Breast cancer is the most common cancer in women worldwide. 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. Taken together, these loci explain approximately one-third of the excess familial risk of breast cancer.

The majority of susceptibility SNPs have 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 used 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 9 breast cancer GWAS involving 10,052 breast cancer cases and 12,575 controls of European ancestry. In total, 45,290 breast cancer cases and 12,575 controls of European ancestry from 41 studies were genotyped with iCOGS, leading to the discovery of 41 new susceptibility loci. A parallel analysis identified four loci specific to estrogen receptor (ER)-negative disease. However, additional susceptibility loci may have been missed because they were not selected from the original GWAS or were 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 new susceptibility variants even if they are not directly genotyped. In this analysis, we aimed to identify additional breast cancer susceptibility loci by using data from all ~200,000 variants on the iCOGS array, and we used imputation to estimate genotypes for more than 11 million SNPs. We applied the same approach to data from 11 GWAS. After quality control exclusions, the data set comprised 15,748 breast cancer cases and 18,084 controls from GWAS together with 46,785 cases and 42,892 controls from 41 studies genotyped with iCOGS (Online Methods and Supplementary Table 1). All subjects were women of European ancestry.

We imputed genotypes using the 1000 Genomes Project March 2012 release as the reference data set (Online Methods). The main analyses were based on ~11.6 million SNPs that were imputed with imputation $r^2 > 0.3$ and had minor allele frequency (MAF) $> 0.005$ in at least one of the data sets.

Of common SNPs (MAF > 0.05), 88% were imputed from the iCOGS array with $r^2 > 0.5$, compared with 99% of variants for the largest GWAS (UK2), which was genotyped using arrays with 594,375 SNPs (Fig. 1a,b and Supplementary Table 2). Of common SNPs, 37% were imputed on the iCOGS platform with $r^2 > 0.9$, in comparison to 85% for UK2. Thus, despite the iCOGS array 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 this array, but the overall imputation quality was poorer than from a standard GWAS array. Imputation quality decreased with decreasing allele frequency (Fig. 1c,d and Supplementary Table 2).

We calculated log odds ratio (OR) estimates and standard errors for each data set using logistic regression, adjusting for principal components when this adjustment was found to substantially reduce the inflation factor. We then combined the results from each data set for variants with MAF > 0.5% using a fixed-effects meta-analysis. We identified more than 7,000 variants with combined $P < 5 \times 10^{-8}$ for association, the large majority of which were 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 8 showed evidence...
of association at $P < 5 \times 10^{-8}$ for overall, ER-positive or ER-negative disease risk (Supplementary Table 3). For four of the eight variants (rs1550623 at 2q31, rs11571833 at 13q13.1, rs12422552 at 12p13.1 and rs11242674 at 6p25.3), we observed slightly weaker evidence of association. One reported variant, rs7726159, did not show association reaching $P < 5 \times 10^{-8}$ in this data set (ref. 25). One other variant in AKAP9, rs6964587, reported previously by another group, did not show association reaching $P < 5 \times 10^{-8}$, but another variant correlated with it at $P = 3.67 \times 10^{-8}$ for chr7:91681597:D; $r^2$ between the two markers = 0.98). The two remaining variants (rs2380205 at 10p15 and rs1045485 at CASP8) were reported in earlier analyses but did not have associations even reaching $P < 0.0001$, suggesting that these loci might have been false positive reports. An alternative variant at CASP8, rs1830298 ($r^2 = 0.06, D' = 1$ with rs1045485 in the 1000 Genomes Project CEU population (Utah residents of Northern and Western European ancestry)), did show association reaching $P < 5 \times 10^{-8}$ in this data set (ref. 3).

To assess evidence for additional susceptibility loci, we removed all SNPs within 500 kb of susceptibility variants identified previously in women of European ancestry, leaving 314 variants from 27 regions associated with breast cancer at $P < 5 \times 10^{-8}$ (Supplementary Figs. 1 and 2). We observed the strongest associations in a 610-kb interval (Build 37 coordinates 28,314,612–28,928,858) on chromosome 22 (smallest $P = 8.2 \times 10^{-22}$, for rs62237573). This interval lies approximately 100 kb centromeric to CHEK2, and further analysis showed that the associated SNPs were correlated with the CHEK2 founder variant c.1100delC (strongest correlation $r^2 = 0.39$ for SNP rs62237573). CHEK2 c.1100delC is known to be associated with breast cancer from candidate gene analysis but has not previously generated an association in GWAS. We performed an analysis adjusting for CHEK2 c.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 c.1100delC, the statistical significance of the rs140914118 association was markedly diminished ($P = 3.1 \times 10^{-9}$ to $P = 0.78$; Supplementary Fig. 3a,b), suggesting that this signal is driven by CHEK2 c.1100delC.

Variants in four regions (DNAJC1, 5p12, PTNLH and MKL1) lay within 2 Mb of a previously published breast cancer–associated SNP. In each case, these associations became weaker (no longer having $P < 5 \times 10^{-8}$) after adjustment for the previously associated SNP(s) in the region (data not shown). For four other regions, the significantly associated variants were identified in just one GWAS and failed imputation ($r^2 < 0.3$) in the remaining data sets, 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 data set without pre-phasing (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, whereas the other two were selected on the basis of 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 ~4,000 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 $r^2$ estimates of correlation between the observed and imputed genotypes were close to the $r^2$ values estimated in imputation. Furthermore, the estimated effect sizes in the subset of individuals 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 OR estimates among the studies genotyped using iCOGS (Supplementary Fig. 4 and Supplementary Table 6). 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 OR estimates for heterozygotes and homozygotes for the risk-associated allele were similar (Supplementary Table 6).

The estimated OR values 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...
Table 1 Results for the 15 newly discovered regions with combined $P \leq 5 \times 10^{-8}$

| Chromosome | Region | Best variant | log 10 odds ratio | SNP ID | Genotype | Alleles | Position | log 10 effect allele frequency | Combined GWAS OR (95% CI) | GWAS P | iCOGS P | $r^2$ | Annotation-Dependent Depletion (CADD) score | Combined GWAS OR (95% CI) | iCOGS P | $r^2$ |
|------------|--------|--------------|------------------|--------|---------|---------|---------|-----------------------------|---------------------------|--------|---------|------|--------------------------------------------|---------------------------|--------|------|
| chr1 | 1q21.2 | rs12048493 | 1.21 | 0.34 | A/T | 0.03 | 149,927,034 | 1.04 | 1.26 | 0.46 | 0.72 | 1.05 | 0.94 | 2.34 | 2.41 | 2.41 |
| chr17 | 2q14 | rs514 | 0.69 | 0.62 | G/C | 0.55 | 123,657,890 | 0.17 | 0.97 | 0.46 | 0.7 | 1.05 | 0.94 | 2.34 | 2.41 | 2.41 |
| chr8 | 1p11 | rs12051792 | 0.77 | 1.0 | G/A | 0.09 | 15,678,901 | 0.15 | 0.7 | 1.05 | 0.94 | 2.34 | 2.41 | 2.41 |
| chr3 | 22q11 | rs6507583 | 0.79 | 0.79 | G/A | 0.09 | 6,789,012 | 0.15 | 0.7 | 1.05 | 0.94 | 2.34 | 2.41 | 2.41 |

Results are shown for the most strongly associated variant in the region.

We used data from The Cancer Genome Atlas (TCGA) to assess associations between the 15 newly discovered susceptibility variants and expression of neighboring genes in breast tumors and normal breast tissue. One SNP, rs7707921, was strongly associated with expression of ENO1 (p.Ass279Ser; with a Combined Annotation-Dependent Depletion (CADD) score of 33, indicating that it is likely to be deleterious), 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, on the basis of Encyclopedia of DNA Elements (ENCODE) chromatin immunoprecipitation and sequencing (ChIP-seq) data. To identify potential gene targets, we combined this information with ENCODE chromatin interaction analysis with paired-end tag (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 1q12.1, we identified four potential interacting genes: RNF115, POLR3B, PDZK1 and IAS3 (Fig. 2). Of these, the strongest evidence was for RNF115 and PDZK1: 3 of the 64 potentially causal variants lay in interacting enhancer regions. RNF115 (also known as BC2A) encodes an E3 ubiquitin ligase RING finger protein that is overexpressed in ER-positive breast cancers. PDZK1 encodes a scaffold protein that connects plasma membrane proteins and regulatory components, regulating their surface expression in epithelial cell apical domains, and has been proposed to act as an oncogene in breast cancer.

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

On the basis of the estimated OR values in the iCOGS stage (all but one of which were in the range 1.05–1.10) and assuming that all loci combine multiplicatively, the 15 new loci identified here would explain a further ~2% of the twofold risk of breast cancer in the first-degree relatives of women with the disease. Taking the newly identified loci together with previously identified ones, 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 on the basis of genotypes for variants at these loci approximately 5% of women in the general population have >2-fold increased risk of breast cancer and 0.7% of women have >3-fold increased risk. In the current analyses, more than 50% of variants with MAF >0.005 in subjects of European ancestry could be imputed well ($r^2 > 0.5$). These results suggest that, although there may be further susceptibility variants with comparably 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 the ability to discriminate between individuals at high and low risk, potentially improving breast cancer screening and prevention.

**METHODS**

Methods and any associated references are available in the online version of the paper.

**ACKNOWLEDGMENTS**

The authors 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 223175 (HEALTH-F2-2009-223175) (COGS). Meetings of the BCAC have been funded by the European Union COST programme (BM0606). Genotyping on the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10170, C18197/A16565), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer program and the Ministry of Economic Development, Innovation and Export Trade of Quebec, grant PSRII-701. Combination of the GWAS data was supported in part by the US National Institutes of Health (NIH) Cancer Post-Cancer GWAS initiative, grant U19 CA148065-01 (DRIVE, part of the GAME-ON initiative). For a full description of funding and acknowledgments, see the Supplementary Note.

**AUTHOR CONTRIBUTIONS**

K. Michailidou and D.F.E. performed the statistical analysis and drafted the manuscript. D.F.E. conceived and coordinated the synthesis of the iCOGS array and led the BCAC. PH. coordinated COGS. J. Benitez led the iCOGS genotyping working group. A.-N., G.P., M.R.A., N.A., D.H., J. Benitez, D.V., F.B., D.C.T., J.S., A.M.D., C.L., C. Baynes, S.A., C.S.H. and M.J.M. coordinated genotyping of the iCOGS array. M.G.-C., P.D.D.P. and M.K.S. led the BCAC pathology and survival working group. J.C.-C. led the BCAC risk factor working group. A.M.D. and G.C.-T. led the iCOGS quality control working group. J. Beesley, J.D. and M.J.L. provided bioinformatics support. M.K.B. and Q. Wang provided data management support for BCAC. S. Canisius provided analysis of the TCGA expression data. J.L.H., M.C.S., H.T. and C.A. coordinated ABCFS. M.K.S., A.B., S.V. and S. Cornellisson coordinated ABCS. K. Mait, A. Lophatananon, S.S.-B. and P.S. coordinated ACP. P.A.E., A. Hein, M.W.B. and L.H. coordinated BRC-C. T.P. and L.-S.-S., O.F. and I.G. coordinated BRCs. E.J.S., I.T., M.J.K. and N.M. coordinated BIGGS. P.K., D.J.H., S.L., S.M.G., M.M.G., W.R.D., A.H., J.M.S. and L.M., C.D.B., S.J.C., E.J.F. and R.N.H. coordinated BPCR. B.B., K.M., S. and C. Cohn coordinated BSUCH. N.R. and C. Turnbull coordinated BOCS. S.P., T.T., C. Mullot and M. Sanchez coordinated CECC. S.E.B., B.G.N. and S.P. coordinated CGPS. A.G.-N., J. Benitez, M.P.Z. and J.I.A.P. coordinated CNIO-BCS. H.A.-C. and S.L.N. coordinated CTS. H. Brenner, A.K.D., V.A. and C. Stemmayer coordinated ESTHER. A. Meindl, R.K.S., C. Sutter and R.Y. coordinated GC-HBOC. H. Brauch, U.H. and T.B. coordinated GENICA. H.N., T.A.M., K. Aittomäki, C. Blomqvist, K. Aaltonen and S.K. coordinated HEBCS. K. Matsuo, H. Ito, H. Itawa and K.T. coordinated HERPACC. T.D. and N.V. coordinated HMBCS. A. Lindblom and S. Margolin coordinated KARBC. A. Manne B. V.K. and J.M.H. coordinated KBGP. G.C.-T. and J. Beesley coordinated kConFab/AOCS. A.H.W., C. Tseng, D.V.D.R. and D.O.S. coordinated LAABC. D.L., P.N., H.W. and E.V.L. coordinated LMBC. J.C.-C., D.F.-J., U.E. and S.B.
and AR coordinated MARIE. Pr Peterlongo, S Manoukian and L Bernard coordinated RBCS. PR, PR Peterlongo, S Manoukian and L Bernard coordinated MCCS. CAH, BEH, ES and LLM coordinated MEC. JS, MSG, EL and MD coordinated MTDLGBCS. SHT, CAY, NAMT and GH coordinated MYBRA. VNK, LGIA and S.N coordinated NBCS. WZ, SLH, M Shrubsole and J Long coordinated NBHS. RW, KP, AJ-V and MG coordinated OBcs. ILA, JAK, G.G and A.MM coordinated OFBCR. PD, RAEEMT, C Seynaeve and CJ.VA coordinated ORIGO. MG-C, JF, SJC and L Brinton coordinated PBCS. KC, HD, ME and J.S.B coordinated pKARMA. MJH, A Holleste, J.EWM, and IMM coordinated MBCS. PH, LL, JI, LH and KP coordinated SARSAS- XOS, -WLI, -WY, T.G and HG coordinated SBCBs. AC, S.SC and M.WRR coordinated SBCs. WB, LBS and QC coordinated SCCS. MH and B.JP coordinated SEARCH. D.K, J- YC, SPK and K-YY coordinated SEBCS. MHH, KSC and CW coordinated SGBC. U.K, and M.D Torres coordinated SKKDVFZ. SJ, LBubinski, KJ and TH coordinated SBZCS. S Sangrajrang, VG, PB and JM coordinated TBCS. FG, GS, Slager, A.E, CA, B.A and D.Y coordinated the TNBCC. CYS, C-NH, P-EW and M-FH coordinated the TNBCS. AJS, A.A, N.O and MJS coordinated UKBGs. HAA, MGK, A.SW, EMJ, KEM, DRG, RSM, GU, EM, DES and GC coordinated EBCG GWAS. QW, Waisfisz, HM-H, MAa and RBvdL coordinated DFBBCS GWAS. DFE, NR, and C Turnbull coordinated UK2 GWAS. F.C, DTrichopoulos, P. Peeters, EL, M. Sund, K-TX, MJC, D.P, LD, J-MH and LMM coordinated EPIC. All authors provided critical review of the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Kamangar, F., Dorrs, G.M & Andrus, W.F: Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J. Clin. Oncol. 24, 2137–2150 (2006).
2. Easton, D.F & et al: Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447, 1087–1093 (2007).
3. Hunter, DJ & et al: A genome-wide association study identifies alleles in FGF2 associated with risk of sporadic postmenopausal breast cancer. Nat. Genet. 39, 870–874 (2007).
4. Stacey, S.N & et al: Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-negative breast cancer. Nat. Genet. 39, 865–869 (2007).
5. Ahmed, S. & et al: Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat. Genet. 41, 585–590 (2009).
6. Zheng, W. & et al: Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat. Genet. 41, 324–328 (2009).
7. Thomas, G. et al: Early-onset breast cancer: genome-wide association study identifies two new risk alleles at 10p11.2 and 14q24.1 (RAD51L1). Nat. Genet. 41, 579–584 (2009).
8. Turnbull, C. & et al: Genome-wide association study identifies five new breast cancer susceptibility loci. Nat. Genet. 42, 703–706 (2008).
9. 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 (2010).
10. Fletcher, C. et al: Novel breast cancer susceptibility loci on 9q31.2: results of a genome-wide association study. J. Natl. Cancer Inst. 103, 425–435 (2011).

Kyriaki Michailidou1, Jonathan Beesley2, Sara Lindstrom3, Sander Canisius4, Joe Dennis1, Michael J Lush1, Mel J Maranian5, Manjeet K Bolla1, Qin Wang1, Mitul Shah5, Barbara J Perkins5, Kamila Czene6, Mikael Eriksson6, Hafte Darabi6, Judith S Brand6, Stig E Bojesen7,9, Borge G Nordestgaard7,9, Henrik Flyer10, Sune F Nielsen7,8, Nazneen Rahman11, Clare Turnbull11, BOCS12, Olivia Fletcher13, Julian Peto14, Lorna Gibson14, Isabel dos Santos-Silva14, Jenny Chang-Claude15, Dieter Flesch-Jayns16,17, Anja Rudolph15, Ursula Eilber15, Sabine Behrens15, Heli Nevanlinna18, Taru A Muranen18, Kristina Aittomäki19, Carl Blomqvist20, Sofia Khan18, Kirstimari Aaltonen18, Habibul Ahsan21,25, Muhammad G Kibria21,22, Alice S Whittemore26,27, Esther M John26,28, Kathleen E Malone29, Marilie D Gammon30, Regina M Santella31, Giske Ursin32, Enes Makalic33, Daniel F Schmidt33, Graham Casey34, David J Hunter3, Susan M Gapstur30, Mia M Gaudet35, W Ryan DiVer35, Christopher A Haiman34, Fredrick Schumacher34, Brian E Henderson34, Loic Le Marchand36, Christine D Berg37, Stephen J Chanock38, Jonine Figueroa38, Robert N Hoover38, Diether Lambrechts39,40, Patrick Neven41, Harold Wilsers41, Erik van Limbergen41, Marjanka K Schmidt42, David J Hunter34, Jonine Figueroa38, Robert N Hoover38, Diether Lambrechts39,40, Patrick Neven41, Harold Wilsers41, Erik van Limbergen41, Marjanka K Schmidt42.
LETTERS

School of Medicine, Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland. Cancer Center of Eastern Finland, University of Eastern Finland, Kuopio, Finland. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland. Cancer Center, Kuopio University Hospital, Kuopio, Finland. Central Finland Hospital District, Jyväskylä Central Hospital, Jyväskylä, Finland. Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands. Department of Surgical Oncology, Leiden University Medical Center, Leiden, the Netherlands. Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, the Netherlands. Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland. National Cancer Institute of Thailand, Bangkok, Thailand. International Agency for Research on Cancer, Lyon, France. Department of Molecular Virology, Immunology and Medical Genetics, Ohio State University, Columbus, Ohio, USA. Roswell Park Cancer Institute, Buffalo, New York, USA. Molecular Diagnostics Laboratory, Institute of Radioisotopes and Radiodiagnostic Products (IRRP), National Centre for Scientific Research ‘Demokritos’, Aghia Paraskevi Attikis, Athens, Greece. Institute of Human Genetics, Pontificia Universidad Javeriana, Bogotá, Colombia. Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, California, USA. Department of Epidemiology, University of California–Irvine, Irvine, California, USA. McGill University and Génome Québec Innovation Centre, Montreal, Quebec, Canada. Correspondence should be addressed to D.F.E. (dfe20@medschl.cam.ac.uk).
ONLINE METHODS

Study overview. Details of the subjects, genotyping and quality control measures for the GWAS and iCOGS data are described elsewhere12,14,16,36,37. All participating studies were approved by their appropriate ethics review boards, 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 TNBC GWAS nominated additional SNPs with evidence for association with ER-negative or triple-negative (ER-, progesterone receptor (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 2-stage procedure, using SHAPEIT to derive phased genotypes and IMPUTEv2 to perform imputation on the phased data22. We performed imputation using 5-Mb non-overlapping intervals for the whole genome. OR estimates and standard errors were obtained using logistic regression with SNPTEST21. For two of the studies, we adjusted for the three leading principal components, as this adjustment was found to materially reduce 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 Minimac23. Genomic control adjustment was applied to each GWAS as previously described16. The iCOGS data were also imputed in a two-stage procedure using SHAPEIT and IMPUTEv2, again using 5-Mb non-overlapping intervals. We split the ~90,000 samples into 10 subsets, keeping subjects from the same study in the same subset where possible. We obtained OR estimates and standard errors using logistic regression adjusting for study and nine principal components. For regions showing evidence of association, we repeated imputation in iCOGS, using IMPUTEv2 but without prephasing in SHAPEIT to improve imputation accuracy. We also increased the number of Markov chain Monte Carlo (MCMC) iterations from 30 to 90 and increased the buffer region from 250 kb to 500 kb.

Meta-analysis. OR estimates and standard errors were combined in a fixed-effects inverse variance meta-analysis using METAL23. For the GWAS, results were included in the analysis for all SNPs with MAF >0.01 and imputation r² >0.3, except for the triple-negative 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 reimputation and meta-analysis was genotyped in 4,123 samples from SEARCH, using TaqMan assays according to the manufacturer’s instructions. Squared correlations between the observed genotypes and the genotypes estimated by imputation are shown in Supplementary Table 5. For all imputed SNPs, the squared correlations were >0.7, the call rates were ≥0.98 and there was no evidence of departure of genotype frequencies from those expected under Hardy–Weinberg equilibrium (P > 0.1).

eQTL analyses. Germline genotype, mRNA expression and somatic copy number data for samples taken from breast tumors and tumor-adjacent normal tissue were obtained from the TCGA38. Copy number and genotype data were obtained using the Affymetrix Genome-Wide Human SNP 6.0 platform. For mRNA expression data, we used the expression profiles obtained with the Agilent G4502A-07-3 microarray. Genotype data were subjected to the following quality control filters. SNPs were excluded in case of low frequency (MAF <1%) or high heterozygosity (false discovery rate (FDR) < 1%). Furthermore, individuals were also excluded in case of non-European ancestry or male sex. Quality control and intersection with the other genomic data types resulted in 380 tumor samples and 56 normal samples. Genotype data were imputed as described above. eQTL analysis was performed using linear regression with SNPTEST, regressing the mRNA expression levels of selected candidate genes on the imputed genotype. For each gene, we performed eQTL analysis against every microarray probe that uniquely mapped to that gene. We adjusted the analyses for the somatic copy number of the gene and for SNPs that intersected the probe sequence, provided that the MAF for these SNPs exceeded 1% in individuals of European ancestry in the 1000 Genomes Project 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 MCF-7, HMEC and HCC1954 were intersected with candidate causal variants using Galaxy. ENCODE ChIA-PET chromatin interaction data from MCF-7 cells (mediated by RNA polymerase 2 and ERα) were downloaded using the UCSC Table browser. Galaxy was used to identify the 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).

36. 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 (2016).
37. Stevens, K.N. et al. 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. Cancer Res. 72, 1795–1803 (2012).
38. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).