Association analysis identifies 65 new breast cancer risk loci

Breast cancer risk is influenced by rare coding variants in susceptibility genes, such as BRCA1, and many common, mostly non-coding variants. However, much of the genetic contribution to breast cancer risk remains unknown. Here we report the results of a genome-wide association study of breast cancer in 122,977 cases and 105,974 controls of European ancestry and 14,068 cases and 13,104 controls of East Asian ancestry. We identified 65 new loci that are associated with overall breast cancer risk at $P < 5 \times 10^{-8}$. The majority of credible risk single-nucleotide polymorphisms in these loci fall in distal regulatory elements, and by integrating in silico data to predict target genes in breast cells at each locus, we demonstrate a strong overlap between candidate target genes and somatic driver genes in breast tumours. We also find that heritability of breast cancer due to all single-nucleotide polymorphisms in regulatory features was 2–5-fold enriched relative to the genome-wide average, with strong enrichment for particular transcription factor binding sites. These results provide further insight into genetic susceptibility to breast cancer and will improve the use of genetic risk scores for individualized screening and prevention.

We genotyped 61,282 female cases with breast cancer and 45,494 female controls of European ancestry using the OncoArray. Subjects came from 68 studies collaborating in the Breast Cancer Association Consortium (BCAC) and Discovery, Biology and Risk of Inherited Variants in Breast Cancer Consortium (DRIVE) (Supplementary Table 1). Using the 1000 Genomes Project (phase 3) reference panel, we imputed genotypes for approximately 21 million variants. After filtering for a minor allele frequency of less than 0.5% and imputation quality score of less than 0.3 (see Methods), we assessed the association between breast cancer risk and 11.8 million single-nucleotide polymorphisms (SNPs) adjusting for country and ancestry-informative principal components. We combined these results with results from the Collaborative Oncological Gene-environment Study (iCOGS: 46,785 cases and 42,892 controls) and 11 other breast cancer genome-wide association studies (GWAS; 14,910 cases and 17,588 controls), using a fixed-effect meta-analysis.

Of 102 loci that have previously been associated with breast cancer in Europeans, 49 showed evidence for association with breast cancer in the OncoArray dataset at $P < 5 \times 10^{-8}$ and 94 at $P < 0.05$. Five additional loci that had previously been shown to be associated with breast cancer in Asian women also showed evidence in the European ancestry OncoArray dataset ($P < 0.01$; Supplementary Tables 2–4). We also assessed the association with breast cancer in Asians, including 7,799 cases and 6,480 controls from the OncoArray project and 6,269 cases and 6,269 controls from the iCOGS project. Of the 94 loci that had previously been identified in Europeans that were polymorphic in Asians, 50 showed evidence of association ($P < 0.05$). For the remaining 44, none showed a significant difference in the estimated odds ratio for overall breast cancer risk between Europeans and Asians ($P > 0.01$; Supplementary Table 5). The correlation in effect sizes for all known loci between Europeans and Asians was 0.83, suggesting that the majority of known susceptibility loci are shared between these populations.

To search for additional susceptibility loci, we assessed all SNPs excluding those within 500 kb of a known susceptibility SNP (Fig. 1). This identified 5,969 variants in 65 regions that were associated with overall breast cancer risk at $P < 5 \times 10^{-8}$ (Supplementary Tables 6–8). For two loci (lead SNPs rs58847541 and rs12628403), there was evidence of a second association signal after adjustment for the primary signal (rs13279803: conditional $P = 1.6 \times 10^{-16}$; rs373038216: $P = 2.9 \times 10^{-10}$; Supplementary Table 9). Of the 65 new loci, 21 showed a differential association based on oestrogen receptor (ER) status ($P < 0.05$), with all but two SNPs (rs6725517 and rs6569648) more strongly associated with ER-positive disease (Supplementary Tables 10, 11). Forty-four loci showed evidence of association with ER-negative breast cancer ($P < 0.05$). Of the 51 novel loci that were polymorphic in Asians,
nine were associated at \( P < 0.05 \) and only two showed a difference in the estimated odds ratio between Europeans and Asians \((P < 0.01; \text{Supplementary Table 12}).\)

To define a set of credible risk variants (CRVs) at the new loci, we first selected variants with \( P \) values within two orders of magnitude of the most significant SNPs in each region. Across the 65 novel regions, we identified 2,221 CRVs (Supplementary Table 13), while the 77 previously identified loci contained 2,232 CRVs (Methods and Supplementary Table 14). We examined these CRVs for evidence of enrichment of 67 genomic features, including histone markers and transcription factor binding sites in three breast cancer cell lines (Methods, Extended Data Fig. 1 and Supplementary Tables 15, 16). Thirteen features were significant predictors of CRVs at \( P < 10^{-4} \); the strongest were DNase I hypersensitivity sites in CTCF-silenced MCF7 cells (odds ratio = 2.38, \( P = 4.6 \times 10^{-14} \)). Strong associations were also observed at binding sites for FOXA1, ESR1, GATA3, E2F1 and TCF712. In addition, 7 of the 65 novel loci included only a single CRV (Supplementary Table 6), of which two were non-synonymous. SNP rs16991615 is a missense variant (p.Glu341Lys) in MCM8, which is involved in genome replication and associated with age at natural menopause and impaired DNA repair\(^6\). SNP rs35383942 is a missense variant (p.Arg28Gln) in KLHDC7A, encoding a p53-regulated repressor of AKT\(^7\).

We annotated each CRV with publicly available genomic data from breast cells in order to highlight potentially functional variants, predict target genes and prioritize future experimental validation (Supplementary Tables 7, 13 with UCSC (University of California Santa Cruz) browser links). We developed a heuristic scoring system based on breast-specific genomic data (integrated expression quantitative trait and \textit{in silico} prediction of GWAS targets (INQUISIT)) to rank the target genes at each locus (Supplementary Table 17). Target genes were predicted by combining risk SNP data with multiple sources of genomic information, including chromatin interactions (chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) and genome-wide chromosome conformation capture (Hi-C)), computational enhancer–promoter correlations (PreSTIGE (ref. 8), IM-PET (ref. 9), FANTOM5 (ref. 10) and super-enhancers), results for breast tissue-specific expression quantitative trait loci (eQTLs), transcription factor binding (ENCODE (Encyclopedia of DNA Elements) chromatin immunoprecipitation followed by sequencing (ChIP–seq)), gene expression (ENCODE RNA sequencing (RNA-seq)) and topologically associated domain boundaries (Methods and Supplementary Tables 18–20). Target gene predictions could be made for 38 out of 65 new and 70 out of 77 previously identified loci. Among 689 protein-coding genes predicted by INQUISIT, we found strong enrichment for established breast cancer drivers identified through tumour sequencing (20 out of 147 genes, \( P < 10^{-6} \))\(^{11–14} \), which increased with increasing INQUISIT score \((P = 1.8 \times 10^{-6}) \). We compared INQUISIT with two alternative methods. Firstly, an alternative, published method (Data-driven Expression-Prioritized Integration for Complex Traits (DEPICT), which predicts targets based on shared gene functions between potential targets at other associated loci)\(^{15} \) showed a weaker enrichment of breast cancer driver genes \((P = 0.06 \) after adjusting for the nearest gene, \( P = 0.74 \) after adjusting for INQUISIT score). Secondly, after assigning the association signal to the nearest gene, only a weak enrichment of driver genes after adjusting for the INQUISIT score was found \((P = 0.01; \text{Extended Data Table 1 and Supplementary Table 21}).\)

Notably, most of the 689 putative target genes have no reported involvement in breast tumorigenesis and some may represent additional genes that influence the susceptibility to breast cancer. However, functional assays will be required to confirm whether any of these candidate genes is causally implicated in breast cancer susceptibility.

Having used INQUISIT to predict target genes, we performed pathway gene set enrichment analysis, the results of which are visually summarized as enrichment maps\(^{16} \) (Extended Data Fig. 2 and Supplementary Tables 22, 23). Several growth or development related pathways were enriched, notably the fibroblast growth factor, platelet-derived growth factor and Wnt signalling pathways\(^{17–19} \). Other cancer-related themes included ERK1/2 cascade, immune-response pathways, including interferon signalling, and cell-cycle pathways. Pathways that were not found in earlier breast cancer GWAS include nitric-oxide biosynthesis, AP-1 transcription factor and NF-κB (Supplementary Table 24).

To explore more globally the genomic features that contribute to breast cancer risk, we estimated the proportion of genome-wide SNP heritability attributable to 53 publicly available annotations\(^{20} \). We observed the largest enrichment in heritability (5.2-fold, \( P = 8.5 \times 10^{-5} \)) of transcription factor binding sites, followed by a fourfold \((P = 0.0006)\) enrichment of histone marker H3K4me3 (marking promoters). By contrast, we observed a significant depletion (0.27, \( P = 0.0007 \)) in repressed regions (Supplementary Table 25). We conducted cell-type-specific enrichment analysis for four histone markers and observed significant enrichments in several tissue types (Extended Data Figs 3–7 and Supplementary Table 26, 27), including a 6.7-fold enrichment of H3K4me1 in breast myoepithelial tissue \((P = 7.9 \times 10^{-5}) \). We compared the cell-type-specific enrichments for all, ER-positive and ER-negative breast cancers to the enrichments for 16 other complex traits (Extended Data Figs 3–7). Breast cancer showed enrichment in adipose and epithelial cell types (including breast epithelial cells). By contrast, psychiatric diseases showed enrichment specific to cell types of the central nervous system and autoimmune disorders showed enrichment in immune cells.

We selected four loci for further evaluation to represent those that are predicted to act through proximal regulation (1p36 and 1p15) and distal regulation (1p34 and 7q22), because they had a relatively small number of CRVs. Firstly, the only CRV at 1p36, rs2992756 \((P = 1.6 \times 10^{-15}) \), is located 84 bp from the transcription start site of KLHDC7A. Secondly, of the 19 CRVs at 1p15 (smallest \( P = 1.4 \times 10^{-12} \)), five were located in the proximal promotor of PIDD1, which is implicated in DNA-damage-induced apoptosis and tumorigenesis\(^{21} \). INQUISIT predicted KLHDC7A and PIDD1 to be target genes and these genes had the highest score for the likelihood of promoter regulation (Supplementary Table 19). Using reporter assays, we showed that the KLHDC7A promoter construct containing the risk T-allele of rs2992756 has significantly lower activity than the reference construct, while the PIDD1 promoter construct containing the risk haplotype significantly increased PIDD1 promoter activity (Extended Data Fig. 8).

Thirdly, the 1p34 locus included four CRVs (smallest \( P = 9.1 \times 10^{-9} \)) that are within two putative regulatory elements (PREs) and are predicted by INQUISIT to regulate CITED4 (Extended Data Fig. 8). CITED4 encodes a transcriptional coactivator that interacts with CBP, p300 and TFAP2 and can inhibit hypoxia-activated transcription in cancer cells\(^{22} \). Chromatin conformation capture assays confirmed that the PREs physically interacted with the CITED4 promoter (Extended Data Fig. 8). Subsequent reporter assays showed that the PRE1 reference construct reduced CITED4 promoter activity, whereas the risk T-allele of SNP rs4233486 located in PRE1 negates this effect.

Finally, the 7q22 risk locus contained six CRVs (smallest \( P = 5.1 \times 10^{-12} \)) that were found to be in several PREs spanning around 40 kb of intron 1 of CUX1. Chromatin interactions were identified between PRE1 (containing SNP rs6979850) and the promoters of CUX1 and RASA4 and between PRE2 (containing SNP rs71559437) and the promoters of RASA4 and PRRKIP1 (Extended Data Fig. 9). Allele-specific chromatin conformation capture assays in heterozygous MBA-MB-231 cells showed that the risk haplotype was associated with chromatin looping, suggesting that the protective allele abrogates looping between the PREs and target genes (Extended Data Fig. 9). These results identify two mechanisms by which CRVs may affect target gene expression: through transcriptionactivation of a specific promoter and by affecting chromatin looping between regulatory elements and their target genes. These data provide \textit{in vitro} evidence of target identification.
and regulation; however further studies that include genome editing, oncogenic assays and/or animal models will be required to fully elucidate disease-related gene function.

We estimate that the newly identified susceptibility loci explain around 4% of the twofold familial relative risk of breast cancer and that in total, common susceptibility variants identified through GWAS explain 18% of the familial relative risk. Furthermore, we estimate that variants that can be reliably imputed using the OncoArray explain around 44% (18% out of 41%) of the familial relative risk that can be explained by all imputable SNPs. The identified SNPs will be incorporated into risk prediction models, which can be used to improve the identification of women at high or low risk of breast cancer: for example, using a polygenic risk score based on the variants that have been identified to date, women in the highest 1% of the distribution have a 3.5-fold greater risk of breast cancer than the population average. Such risk prediction can inform targeted early detection and prevention.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 June 2016; accepted 17 September 2017.

Published online 23 October 2017.

1. Amos, C. I. et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. Cancer Epidemiol. Biomarkers Prev. 26, 126–135 (2017).
2. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. 45, 353–361 (2013).
3. Long, J. J. et al. Genome-wide association study in East Asians identifies novel susceptibility loci for breast cancer. PLoS Genet. 8, e1002532 (2012).
4. 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 (2014).
5. Long, J. J. et al. A common deletion in the APOBEC3 genes and breast cancer risk. J. Natl Cancer Inst. 105, 573–579 (2013).
6. He, C. et al. Genome-wide association studies identify loci associated with age at menarche and age at natural menopause. Nat. Genet. 41, 724–728 (2009).
7. Kawase, T. et al. PHLD3 is a p53-regulated repressor of Act. Cell 136, 535–550 (2009).
8. Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res. 24, 1–13 (2014).
9. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer–promoter interactions in human cells. Proc. Natl Acad. Sci. USA 111, E2191–E2199 (2014).
10. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 459–461 (2014).
11. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature 534, 47–56 (2016).
12. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70 (2012).
13. Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. Cell 163, 505–519 (2015).
14. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptionic landscapes. Nat. Commun. 7, 11479 (2016).
15. Pets, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. Nat. Commun. 6, 5890 (2015).
16. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G. D. Enrichment map: a gene interaction network interactome in human cells. Proc. Natl Acad. Sci. USA 107, 1926–1931 (2010).
17. Turner, N. & Grose, R. Fibroblast growth factor signalling: from development to cancer. Nat. Rev. Cancer 10, 116–129 (2010).
18. Heldin, C. H. Targeting the PDGF signaling pathway in tumor treatment. Cell Commun. Signal. 11, 97 (2013).
19. Howe, L. R. & Brown, A. M. Wnt signalling and breast cancer. Cancer Biol. Ther. 3, 36–41 (2004).
20. Finucane, H. K. et al. Partitioning heritability by functional annotation using genomewide association summary statistics. Nat. Genet. 47, 1228–1235 (2015).
21. Lin, Y., Ma, W. & Benchimol, S. Pidd, a new death-domain-containing protein, is induced by p53 and promotes apoptosis. Nat. Genet. 26, 122–127 (2000).
22. Fox, S. B. et al. CITED4 inhibits hypoxia-activated transcription in cancer cells, and its cytoplasmic location in breast cancer is associated with elevated expression of tumor cell hypoxia-inducible factor 1α. Cancer Res. 64, 6075–6081 (2004).
Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. 1 Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. 2 Department of Epidemiology, University of Washington School of Public Health, Seattle, Washington, USA. 4 Program in Genetic Epidemiology and Statistical Genetics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA. 5 Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Australia. 6 The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada. 7 Department of Epidemiology, University of Cambridge, Cambridge, UK. 8 Genomics Centre, Centre Hospitalier Universitaire de Quebec Research Center, Laval University, Quebec City, Canada. 9 Department of Computational Biology, St. Jude Children’s Research Hospital, Memphis, Tennessee, USA. 10 Cancer Genomics Research Laboratory (CGR), Division of Cancer Genetics and Genomics, National Cancer Institute, Rockville, Maryland, USA. 11 Division of Molecular Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. 12 Department of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. 13 UQ Centre for Clinical Research, The University of Queensland, Brisbane, Australia. 14 Cancer Research UK Cambridge Research Institute, University of Cambridge, Li Ka Shing Centre, Cambridge, UK. 15 Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. 16 Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. 17 Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands. 18 Centre for Epidemiology and Prevention, The University of Chicago, Chicago, Illinois, USA. 19 Department of Clinical Genetics, Helsinki University Central Hospital, Helsinki, Finland. 20 Department of Epidemiology, University of California Irvine, Irvine, California, USA. 21 A. N. Alexander Research Institute of Oncology and Medical Radiology, Mirsk, Belarus. 22 Division of Clinical Cancer Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. 23 Department of Research, Genomics Center, Centre Hospitalier Universitaire de Quebec Research Center, Laval University, Quebec City, Canada. 24 Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA. 25 Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. 26 Ziller School of Public Health University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA. 27 McGill University and Genome Quebec Innovation Centre, Montreal, Quebec, Canada. 28 Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany. 29 Human Cancer Genetics Program, Spanish National Cancer Research Centre, Madrid, Spain. 30 Centre de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain. 31 Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa, Russia. 32 Department of Population Epidemiology and Health Research Institute of Hope, DuPage, California, USA. 33 Department of Public Health Sciences, and Cancer Research Institute, Queen’s University, Kingston, Ontario, Canada. 34 Department of Radiation Oncology, Hannover Medical School, Hannover, Germany. 35 Gynakologische Ressourcenverband Ten, Hannover, Germany. 36 Cancer Research Unit, Hannover Medical School, Hannover, Germany. 37 Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. 38 Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. 39 Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. 40 Division of Cancer Prevention and Genetics, Istituto Europeo di OncoLogia, Milan, Italy. 41 Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway. 42 Department of Medical and Molecular Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. 43 Genome Sciences Centre, University of Cambridge, Cambridge, UK. 44 Department of Oncology, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany. 45 Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. 46 Institute of Pathology, Karolinska University Hospital, Munkbro, Stockholm, Sweden. 47 Department of Radiology, VU University Medical Center, Amsterdam, The Netherlands. 48 Department of Molecular Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. 49 Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Molecular Biology, University of Sheffield, Sheffield, UK. 50 Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. 51 Department of Biomedical Physics and Kinesiology, Simon Fraser University, Burnaby, British Columbia, Canada. 52 Department of Gynaecology and Obstetrics, University of Tubingen, Tubingen, Germany. 53 Institute for Prevention and Occupational Health of the German Social Accident Insurance, Institute of the Ruhr University Bochum, Bochum, Germany. 54 Department of Obstetrics and Gynaecology, University of Heidelberg, Heidelberg, Germany. 55 Molecular Epidemiology Group, Dutch Cancer Institute, Amsterdam, The Netherlands. 56 Department of Epidemiology, American Cancer Society, Atlanta, Georgia, USA. 57 Oncogenics and Genetics Unit, Instituto de Investigación Biomédica (IIB) Ourense-Pontevedra-Vigo, Xerencia de Xestión Integrada de Vigo-SERGAS, Vigo, Spain. 58 Hong Kong Hereditary Breast Cancer Family Registry, Happy Valley, Hong Kong. 59 Department of Pathology, Hong Kong Sanatorium and Hospital, Hong Kong, China. 60 Department of Cancer Prevention and Control, National Cancer Institute, Buffalo, New York, USA. 61 Saw Sée Hock School of Public Health, National University of Singapore, Singapore, Singapore. 62 Department of Biomedical Sciences, Seu National
University of Eastern Finland, Kuopio, Finland. 214 Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, New York, USA. 215 National Cancer Institute, Ministry of Public Health, Nonthaburi, Thailand. 216 Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands. 217 Welcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK. 218 Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia. 219 Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. 220 Frauenklinik der Stadtklinik Baden-Baden, Baden-Baden, Germany. 221 Cancer Registry of Norway, Oslo, Norway. 222 Department of Nutrition, Institute of Basic Medical Sciences, University of Oslo, Oslo, Norway. 223 Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany. 224 Department of Clinical Genetics, Leiden University Medical Center, Leiden, The Netherlands. 225 Family Cancer Clinic, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. 226 Division of Biomedical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands. 227 Department of Gynaecology and Obstetrics, University of Ulm, Ulm, Germany. 228 Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina, USA. 229 Department of Nutrition, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA. 230 Subang Jaya Medical Centre, Subang Jaya, Selangor, Malaysia. 231 Seoul National University College of Medicine, Seoul, South Korea. 232 Armed Forces Capital Hospital, Seongnam, South Korea. 233 Department of Surgery, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan. 234 Shanghai Municipal Center for Disease Control and Prevention, Shanghai, China. 235 Department of Medicine, Institute for Human Genetics, UCSF Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, USA. 236 University of California San Francisco, San Francisco, California, USA. 237 Pathology Queensland, The Royal Brisbane and Women’s Hospital, Brisbane 4029, Australia. 238 Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. 239 Center for Genomic Medicine, Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA. 240 University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. 241 Department of Oncology, Södersjukhuset, Stockholm, Sweden. 242 Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands.

† Lists of participants and their affiliations appear in the Supplementary Information.
*These authors contributed equally to this work.
§ These authors jointly supervised this work.
METHODS

Details of the studies and genotype calling and quality control for the iCOGS and eleven other GWAS are described elsewhere. Seventy-eight studies participated in the breast cancer component of the OncoArray, of which 67 studies contributed European ancestry data and 12 contributed Asian ancestry data (one study, NBCS, was excluded as there were no controls from Norway) (Supplementary Table 1). The majority of studies were population-based case–control studies, or case–control studies nested within population-based cohorts, but a subset of studies oversampled cases with a family history of the disease. All studies provided core data on disease status and age at diagnosis or observation, and the majority provided additional data on clinico-pathological and lifestyle factors, which have been curated and incorporated into the BCAC database (version 6). All participating studies were approved by their appropriate ethics review board and all subjects provided informed consent.

OncoArray SNP selection. Approximately 50% of the SNPs for the OncoArray were selected as a ‘GWAS backbone’ (Illumina HumanCore), which aimed to provide high coverage for the majority of common variants through imputation. The remaining SNPs were selected from lists supplied by each of six disease-based consortia, together with a seventh list of SNPs of interest to multiple disease-focused groups. Approximately 72 thousand SNPs were selected specifically for their relevance to breast cancer. These included: (a) SNPs showing evidence of association from previous genotype data, based on a combined analysis of eleven existing GWAS together with the data from the iCOGS experiment; (b) SNPs showing evidence of association with ER-negative disease (through a combined analysis with the CIMBA consortium), triple-negative disease, breast cancer diagnosed before the age of 40 years, high-grade disease, node-positive disease or ductal carcinoma in situ; (c) SNPs potentially associated with breast cancer survival; (d) SNPs selected for fine-mapping of 55 regions showing evidence of breast cancer association at genome-wide significance; (e) rare variants showing evidence of association through exome sequencing in multiple case families, whole-genome sequencing in high-risk cases (DRIVE) or analysis of the ExomeChip (BCAC); (f) specific follow-up of regions of interest from breast cancer GWAS in Asian, Latina and African/African–American women; (g) SNPs associated with breast density, selected from GWAS conducted by the MODE consortium; (h) breast tissue-specific eQTLs; (i) lists of functional candidates from >30 groups. Lists were merged with lists from the other consortia as described elsewhere.

OncoArray calling and quality control. Of the 568,712 variants selected for genotyping, 533,631 were successfully manufactured on the array (including 778 duplicate probes). Genotyping for the breast cancer component of the OncoArray, which included 152,492 samples, was conducted at six sites. Details of the genotyping calling for the OncoArray are described in more detail elsewhere. In brief, we developed a single calling pipeline that was applied to more than 500,000 samples. An initial cluster file was generated using data from 56,284 samples, selected to cover all the major genotyping centres and ethnicities, using the Gentrain2 algorithm. Variants that were likely to have problematic clusters were selected for manual inspection using the following criteria: call rate below 99%, variants with minor allele frequency (MAF) < 0.001, poor Illumina intensity and clustering metrics, or deviation from the expected frequency as observed in the 1000 Genomes Project. This resulted in manual adjustment of the cluster file for 3,964 variants, and the exclusion of 16,526 variants. The final cluster file was then used to apply the full dataset. We excluded probable duplicates and close relatives within each study, and probable duplicates across studies. We excluded samples with a call rate < 95% or samples with extreme heterozygosity (4.89 s.d. from the mean for the ethnicity). Ancestry was computed using a principal component analysis, applied to the full OncoArray dataset, using 2,318 informative markers on a subset of around 47,000 samples. The analysis presented here was restricted to women of European ancestry, defined as individuals with an estimated proportion of European ancestry >0.8, and women of East Asian ancestry (estimated proportion of Asian ancestry >0.8). We excluded SNPs with a call rate < 98% in any consortium, SNPs that could not be linked to the 1000 Genomes Project reference or were contained within population-based cohorts, or were contained within a seventh list of SNPs of interest to multiple disease-focused groups. Approximately 72 thousand SNPs were selected specifically for their relevance to breast cancer. These included: (a) SNPs showing evidence of association from previous genotype data, based on a combined analysis of eleven existing GWAS together with the data from the iCOGS experiment; (b) SNPs showing evidence of association with ER-negative disease (through a combined analysis with the CIMBA consortium), triple-negative disease, breast cancer diagnosed before the age of 40 years, high-grade disease, node-positive disease or ductal carcinoma in situ; (c) SNPs potentially associated with breast cancer survival; (d) SNPs selected for fine-mapping of 55 regions showing evidence of breast cancer association at genome-wide significance; (e) rare variants showing evidence of association through exome sequencing in multiple case families, whole-genome sequencing in high-risk cases (DRIVE) or analysis of the ExomeChip (BCAC); (f) specific follow-up of regions of interest from breast cancer GWAS in Asian, Latina and African/African–American women; (g) SNPs associated with breast density, selected from GWAS conducted by the MODE consortium; (h) breast tissue-specific eQTLs; (i) lists of functional candidates from >30 groups. Lists were merged with lists from the other consortia as described elsewhere.

SNPs with a call rate < 95% in any consortium, SNPs not in Hardy–Weinberg equilibrium ($P < 10^{-7}$ in controls or $P < 10^{-12}$ in cases) and SNPs with concordance < 98% among 5,280 duplicate sample pairs. For the imputation, we additionally excluded SNPs with a MAF < 1% and a call rate < 98% in any consortium, SNPs that could not be linked to the 1000 Genomes Project reference or differed significantly in frequency from the 1000 Genomes Project dataset (using the criterion $|p_2 - p_3| > 0.007$, where $p_2$ and $p_3$ are the MAFs in the 1000 Genomes Project and OncoArray European datasets, respectively). A further 1,128 SNPs for which the cluster plots were judged, on visual inspection, to be too poor to call genotypes reliably were excluded. Of the 533,631 SNPs that were manufactured on the array, 494,763 SNPs passed the initial quality control and 469,364 SNPs were used in the imputation.

Genotype imputation. All samples were imputed using the October 2014 (version 3) release of the 1000 Genomes Project dataset as the reference panel and the number of sampled haplotypes per individual ($n_{hap}$) = 800. The iCOGS, OncoArray and nine of the GWAS datasets were imputed using a two-stage imputation approach, first using IMPUTE version 2 for imputation and then the IMPUTE version 2 quality score in the OncoArray dataset, was >0.80 for all but one locus (rs72749841, quality score = 0.65; see Supplementary Table 28).

Principal components analysis. To adjust for potential (intra-continental) population stratification in the OncoArray dataset, principal components analysis was performed using data from 33,661 uncorrelated SNPs (which included 2,318 SNPs specifically selected on informativeness for determining continental ancestry) with a MAF of at least 0.05 and maximum correlation of 0.1 in the OncoArray dataset. The number of principal components for the iCOGS and OncoArray datasets, as a quality control step for those associated loci where the imputation quality score was < 0.9. Imputation quality for the lead variants, as assessed by the IMPUTE version 2 quality score in the OncoArray dataset, was >0.80 for all but one locus (rs72749841, quality score = 0.65; see Supplementary Table 28).

Statistical analyses. No statistical methods were used to predetermine sample size. Genotyping was conducted without knowledge of the disease phenotypes. Per-allele odds ratios and standard errors were generated for the OncoArray, iCOGS and each GWAS, adjusting for principal components using logistic regression. The OncoArray and iCOGS analyses were additionally adjusted for country and study, respectively. For the OncoArray analysis, we adjusted for country and 10 principal components. Adjustment for country rather than study was used to improve power because some studies had few or no controls. We evaluated the adequacy of this approach by comparing the inflation in the test statistic with that obtained in corresponding analysis in which we adjusted for study—the inflation was very similar ($\lambda = 1.15$ versus 1.17, based on the backbone SNPs, equivalent to $\lambda_{1,000} = 1.003$, for a study of 1,000 cases and 1,000 controls, in both cases). As an additional sensitivity analysis, we computed the effect sizes for the SNPs that were discovered in the discovery stage (i.e. not included in the previous analyses) and those presented. Estimates were derived using ProAbel for the BCAC and EBCG studies, SNPTEST for the remaining GWAS and purpose-written software (http://ccge.medschl.cam.ac.uk/software/pcalc). For the main analyses, we used the first ten principal components as additional adjustments did not further reduce inflation in the test statistics. We used nine principal components for the iCOGS and up to ten principal components for the other GWAS, where this was found to reduce inflation.

For SNPs showing evidence of association, we additionally computed genotype-specific odds ratios for the iCOGS and OncoArray dataset, and per-allele odds ratios for ER-negative and ER-positive disease. Departures from a log-additive model were evaluated using a one degree of freedom likelihood ratio test, comparing the log-additive model (genotypes parameterized as the number of rare alleles carried) with the general model estimating odds ratios for each genotype. The genotype-specific risks for all variants were consistent with a log-additive model ($P > 0.01$; Supplementary Table 29). Tests for differences in the odds ratio by
ER-status were derived using case-only analyses, in which estimates were derived by logistic regression separately in the iCOGS and OncoArray datasets, adjusted as before, and then combined in a fixed-effects meta-analysis. These analyses were performed in R.  
We assessed heterogeneity in the odds ratio estimates among studies within each of the OncoArray, iCOGS and GWAS components, and between the (combined) estimates for the three components, using both the I² statistic and the P value for Cochran’s Q statistic (Supplementary Table S8). There was no evidence of heterogeneity among studies in the odds ratios for any of the loci in the OncoArray, but three loci showed some evidence of heterogeneity in the odds ratios among the GWAS, iCOGS and OncoArray datasets.  
To determine whether there were multiple independent signals in a given region, we performed multiple logistic regression analysis using SNPs within 500 kb of each lead SNP, adjusting for the lead SNP. We used the genotypes derived by one-step imputation, performed the analyses separately in the iCOGS and OncoArray datasets and combined the results (adjusted effect sizes and standard errors) using a fixed-effects meta-analysis. For one of the two loci for which there was an additional signal significant at \( P < 5 \times 10^{-8} \), the lead SNP from the one-step imputation differed from the lead SNP in the overall analysis, but was strongly correlated with it (Supplementary Table S9).  
**Definition of known hits.** We attempted to identify all associations previously reported from genome-wide or candidate analysis at a significance level \( P < 5 \times 10^{-8} \) for all breast cancer types, ER-negative or ER-positive breast cancer, in BRCA1 or BRCA2 carriers, or in meta-analyses of these categories. Where multiple studies reported associations in the same region, we used the first reported association unless later studies identified a variant that was clearly more strongly associated. We only included one SNP per 500 kb interval, unless joint analysis provided clear evidence \( (P < 5 \times 10^{-8}) \) of more than one independent signal. For the analysis of CRVs, we restricted our analysis to regions in which the most significant signal had a \( P \) value < 10^{-7} in Europeans (77 regions). To avoid complications with defining CRVs for secondary signals, we considered only the primary signal and defined CRVs as those for which the \( P \) value was within two orders of magnitude of the most significant \( P \) value.  
**In silico analysis of CRVs.** We combined multiple sources of in silico functional annotation from databases to help to identify potential functional SNPs and target genes. To investigate functional elements enriched across the region encompassing the strongest CRVs, we analysed data of chromatin biofeatures from the Encyclopedia of DNA Elements (ENCODE) Project\(^\text{12}\), Roadmap Epigenomics Projects\(^\text{33}\) and other data obtained through the National Center for Biotechnology Information (NCBI) gene expression omnibus (GEO) namely: chromatin state segmentation by hidden Markov Models (chromHMM), DNase I hypersensitive and histone modifications of epigenetic markers H3K4, H3K9 and H3K27 in human mammary epithelial (HMEC) and myoepithelial cells, T47D and MCF7 breast cancer cells and transcription factor ChIP–seq in a range of breast cell lines (Supplementary Table 13).  
**Association of genomic features with CRVs.** We first defined credible candidate variants as those located within 500 kb of the most significant SNP in each region, and with \( P \) values within two orders of magnitude of the most significant SNPs. This is approximately equivalent to flagging variants whose posterior probability of causality is within two orders of magnitude of the value of the most significant SNP\(^\text{4,35}\). We then selected 800 random 1-Mb control regions that were at least 1 Mb from each other and from the intervals defined by the associated SNPs. The association with each feature was then evaluated using logistic regression, with being a CRV as the outcome, and adjusting for the dependence due to linkage disequilibrium using robust variance estimation, clustering on region, using the R package multitypanic2.  
**eQTL analyses.** eQTL analyses were performed using data from The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) projects\(^\text{12,36}\). The TCGA eQTL analysis was based on 458 breast tumours that had matched gene expression, copy number and methylation profiles together with the corresponding germline genotypes available. All 458 individuals were of European ancestry as ascertained using the genotype data and the Local Ancestry in admixed Populations (LAMP) software package (LAMP estimator \( g = 99\% \) European)\(^\text{37}\). Germline SNPs into the 1000 Genomes Project reference panel (October 2014 release) using IMPUTE version 2 (refs 26, 38). Gene expression had been measured on the Illumina HiSeq 2000 RNA-Seq platform (gene-level RSEM normalized counts\(^\text{39}\)), copy-number estimates were derived from the Affymetrix SNP 6.0 (somatic copy number alteration minus germline copy number variation called using the GISTIC2 algorithm\(^\text{40}\)), and methylation beta values measured on the Illumina Infinium HumanMethylation450. eQTL analysis focused on all variants within 500 kb of the most significantly associated risk SNP in 142 genomic regions (each 2-Mb wide) containing at least one previously identified or new overall breast cancer risk locus confirmed at genome-wide significance in the current meta-analysis. Each variant was evaluated for its association with the expression of every gene within 2 Mb that had been profiled for each of the three data types. The effects of tumour copy number and methylation on gene expression were first regressed out using a method described previously\(^\text{41}\). eQTL analysis was performed by linear regression, with residual gene expression as outcome, germline SNP genotype dosage as the covariate of interest and \( ESR1 \) expression and age as additional covariates, using the R package Matrix eQTL\(^\text{42}\).  
The METABRIC eQTL analysis was based on 138 normal breast tissue samples resected from breast cancer patients of European ancestry. Germline genotyping for the METABRIC study was also done on the Affymetrix SNP 6.0 array, and gene expression in the METABRIC study was measured using the Illumina HT12 microarray platform (probe-level estimates). No adjustment was implemented for somatic copy number and methylation status since we were evaluating eQTLs in normal breast tissue. All other steps were identical to the TCGA eQTL analysis described above.  
**INQUISIT.** We developed a computational pipeline, the integrated expression quantitative trait and *in silico* prediction of GWAS targets (INQUISIT), to interrogate publically available data for the prioritization of candidate target genes.  
**Data used for INQUISIT.** Chromatin interaction data from ENCODE CHIA-PET analysis in MCF-7 cells for RNApolII, ERα and CTFC factors were downloaded using the UCSC Table Browser\(^\text{43}\). Hi-C data from HMECs were obtained from ref. 44, using ‘interaction loops’ as defined in ref. 44. Data were reformatiod to facilitate intersection of query SNPs using BEDTools ‘intersect’\(^\text{45}\). For all interactions, termini were intersected with promoters using GENCODE version 19 (ref. 46) basic gene annotations, for which we defined promoters as –1 to 0.1 kb surrounding a transcription start site.  
Enhancer–target gene predictions of several computational algorithms were collected. Each of these datasets assigned genes to enhancers. We used all MCF-7 and HMEC enhancer predictions (low and high stringency) made by PreSTIGE\(^\text{6}\), IM-PET enhancer–gene predictions in MCF-7, HMEC and HCC1954 cell lines\(^\text{8}\). Enhancer–transcription start site (E–TSS) links were identified from the FANTOMS Consortium\(^\text{10}\), and enhancers detected in mammary epithelial cells were intersected with E–TSS links. We also collected typical and super-enhancers in MCF-7, HMEC and HCC1954 cells defined in ref. 47.  
ChIP–seq peak data for transcription factors ESR1, FOXA1, GATA3, TCF7L2 and E2F1 from MCF-7, T47D and MCF-10A cells were downloaded in narrowPeak format from ENCODE. ChIP–seq peak data for H3K-4me3 and H3K9ac (characteristic of promoters) histone modification for all breast cells were obtained from ENCODE and Roadmap Epigenomics Project. ChromHMM data for breast cell samples (HMEC and myoepithelial: E027, E028 and E119) were downloaded from Roadmap Epigenomics.  
eQTL analyses were conducted as described above. For the interpretation of the eQTL results for INQUISIT (and in general), we focused on the overlap between the CRVs (risk signal) and the top eQTL variants for a given gene (eQTL signal). If the eQTL \( P \) value for a CRV was the same as, or within 1/100th of the eQTL \( P \) value of the SNP most significantly associated with expression of a particular gene, then the corresponding CRV were assigned a point for being an eQTL in INQUISIT.  
Topologically associated domain (TAD) boundaries were derived from Hi-c data\(^\text{44}\). Genomic intervals corresponding to ‘contact domains’ from eight human cell types were merged using BEDTools ‘merge’ resulting in annotation of regions that were most likely to encompass TAD units. Inter-TAD boundaries were identified using BEDTools ‘complement’.  
Gene level RNA-seq expression data generated under multiple experimental conditions in MCF-7 and normal mammary epithelial cells were downloaded from ENCODE. The FPKM (fragments per kilobase of exon per million fragments mapped) values for each gene were extracted using the metagene R package\(^\text{48}\) and averaged across all experiments to give an approximation of expression in breast cells. Accession numbers of previously published data are given in Supplementary Table 30.  
**Inquisit pipeline.** Candidate target genes were evaluated by assessing the potential impact of each CRV on regulatory or coding features. Scores categorized by (a) conserved regulatory, (b) proximal gene regulation or (c) impact on protein coding were calculated using the following criteria (see Supplementary Table 13).  
Genomic annotation data for target gene predictions (chromatin interaction and computational enhancer–promoter assignment), ChIP–seq, histone modification and chromHMM were curated into a BED–formatted database. We intersected the chromosomal positions of CRVs with each category of genomic annotation data using BEDTools ‘intersect’ (minimum 1-bp overlap), resulting in annotation of SNP–gene pairs with presence or absence of multiple classes of genomic data. Each gene was scored using a custom R script on the basis of the following criteria. (a) For distally regulated genes, a candidate gene was given two points if a CRV
fell in an element that revealed long-range ChIA-PET or Hi-C interactions with the promoter of that gene. One point was added to a gene’s score in the case of enhancers predicted by computational methods to target that gene (in addition to experimental interactions if also observed). If the distal elements containing the SNPs overlapped enriched cistromic transcription factor (ESR1, FOXA1, GATA3, TCF7L2, ZNF703, or PIDD1) peaks, an additional point was given when an overlap between a SNP, enhancer and ChIP-seq peak occurred, and two points were added when there were multiple transcription factor binding sites that were overlapping with SNPs in distinct interactions or enhancers (see Supplementary Table 17 for details). One point was given to significant eSNP–eGENE pairs. Predicted distal target genes that were among the list of breast cancer driver genes were up-weighted with a further point (except for the analysis of driver gene enrichment). Information regarding TAD boundaries was used to down-weight genes: genes that were separated from CRVs by a TAD boundary were down-weighted by multiplying their scores by 0.50. Scores for genes exhibiting no expression in MCF7 or HMEC (mean FPKM = 0) were multiplied by 0.1. This resulted in scores for each candidate target gene ranging from 0 to 8. (b) Variants were treated as potentially affecting proximal promoter regulation if they resided between −1.0 and +0.1 kb surrounding a transcription start site. Additional points were awarded to genes when variants overlapped promoter H3K4me3 or H3K9ac histone-modification peaks, intersected with ESR1, FOXA1, GATA3, TCF7L2 or E2F1 transcription factor binding sites, were significant eSNP–eGENE pairs, and if the gene was annotated as a breast cancer driver gene. Gene scores were down-weighted (by a factor of 0.1) if they lacked expression in MCF-7 or HMEC samples. Resultant scores ranged from 0 to 5.1 Intergenic variants were weighted with their potential to impact protein function using a range of in silico prediction tools (CADD59, FATHMM60, LRT51, MutationAssessor52, Mutation Taster 2 (ref. 53), Polyphen-2 (ref. 54), PROVEAN55 and SIFT66 for missense variants; Human Splicing Finder57 and MaxEntScan68 for splice variants). We scored genes with missense and nonsense variants predicted to be functionally deleterious, and points for genes containing variants predicted to alter splicing. Genes could therefore carry SNPs that affect coding and splicing and receive increased scores. Additional points were given to genes that were breast cancer driver genes. We multiplied scores by 0.1 when genes showed a lack of expression in breast cells. Possible coding scores ranged from 0 to 4.

**Enrichment of somatic breast cancer driver genes in INQUISIT target gene predictions.** We listed 147 unique protein-coding driver genes for breast cancer identified from four recent tumour genome- and exome-sequencing studies11–14 (considering ZNF703 and FGRF1 as independent genes; Supplementary Table 31). First, we examined overlap between this list of 147 genes and the total set of unique target genes predicted by INQUISIT (n = 689) by one or more of the three regulatory mechanisms (distal, promoter and coding). The significance of this overlap was assessed by randomly drawing (without replacement) 689 genes from the set of all protein-coding genes (GENCODE release 19, n = 20,243) one million times and calculating the probability of observing the same (or stronger) overlap with the list of 147 driver genes that was observed. The enrichment would be stronger with progressively higher INQUISIT scores. We categorized all 20,243 protein-coding genes into four levels based on their INQUISIT scores: level 1: coding score 1; level 2: coding score 1–2; level 3: any score >0 but <1; level 4: score 0 that is not a predicted target). The gene nearest to a risk locus is frequently assigned as a candidate target gene in GWAS in the absence of additional functional analysis59. We observed that 7 of the 147 drivers were among the genes nearest to a previously or newly identified breast cancer risk locus. Therefore, we used logistic regression, including data for all target genes predicted by INQUISIT, with driver status as outcome, and evaluated INQUISIT score level and nearest gene status as potential predictors of driver status (see Supplementary Table 21).

Lead SNPs at 142 breast cancer risk-associated loci were used as input into DEPICT, which was then run using the default settings53. We examined the relative performance of INQUISIT and DEPICT in predicting driver gene status using logistic regression models as above (Supplementary Table 21), adding DEPICT covariance as a covariate.

**Chromatin conformation capture (3C).** MCF7 (ATCC HTB22) and MDA-MB-231 (ATCC HTB26) breast cancer cell lines were grown in RPMI medium with 10% FCS and antibiotics. Bre–80 normal breast epithelial cells (provided as a gift from R. Reddel) were grown in DMEM/F12 medium with 5% horse serum, 10 μg ml−1 insulin, 0.5 μg ml−1 hydrocortisone, 20 ng ml−1 epidermal growth factor, 100 ng ml−1 cholera toxin and antibiotics. Cell lines were maintained under standard conditions, routinely tested for *Mycoplasma* and short tandem repeat profiled to confirm cell line identity. 3C libraries were generated using EcoRI as described previously.3C interactions were quantified by real-time PCR (qPCR) using primers designed within restriction fragments (Supplementary Table 32). qPCR was performed on a RotorGene 6000 using MyTaq HS DNA polymerase (Bioline) with the addition of 5 mM of 9097999, annealing temperature of 66 °C and extension of 30 s. 3C analyses were performed in three independent 3C libraries from each cell line with each experiment quantified in duplicate. BAC clones covering each region were used to create artificial libraries of ligation products in order to normalize for PCR efficiency. Data were normalized to the signal from the BAC clone. TAD boundaries, lines of replication and, where available, by reference to a region within GAPDH. All qPCR products were electrophoresed on 2% agarose gels, gel purified and sequenced to verify the 3C product.

**Plasmid construction and reporter assays.** Promoter-driven luciferase reporter constructs were generated by insertion of PCR amplified fragments or synthesized gBlocks (Integrated DNA Technologies), containing the KLHDC7A, PIDD1 or *CITED4* promoters, into the KpnI/HindIII sites of pGL3-Basic. For the 1p34 locus, a 1.169-bp putative regulatory element (PRE1) or 951-bp PRE2 were synthesized as gBlocks and cloned into the BamHI/Sall sites of the *CITED4*-promoter construct. The minor alleles of SNPs were introduced into promoter or PRE sequences by overlap extension PCR or gBlocks. Sequencing of all constructs confirmed variant incorporation (AGRF). MCF7 or Bre-80 cells were transfected with equimolar amounts of luciferase reporter plasmids and 50 ng of pRLTK transfection control plasmid with Lipofectamine 2000. The total amount of transfected DNA was kept constant at 600 ng for each construct by the addition of pUC19 as a carrier plasmid. Luciferase activity was measured 24 h after transfection by the Dual-Glo Luciferase Assay System. To correct for any differences in transfection efficiency or cell-lysate preparation, Firefly luciferase activity was normalized to Renilla luciferase, and the activity of each construct was measured relative to the reference promoter construct individually, creating 220 separate models, each with 54 annotations (53 + 1). This procedure controls for the overlap with the 53 functional categories in the full baseline model but not with the 219 other cell-type-specific annotations.

We further tested the differences in functional enrichment between ER-positive and ER-negative subsets through a Wald test, using the regression coefficients and standard errors for the two subsets based on the models described above.

**Contribution of identified variants to the familial relative risk of breast cancer.** We estimated the proportion of the familial risk of breast cancer due to the identified variants, under a log-additive model, using the formula:

\[
\beta_{\text{risk}} = \frac{1}{2(1)l \ln(1)} \sum \left( \frac{1}{P_i(1 - P_i)} \right) \left( \frac{1}{P_i(1 - P_i)} \right) \ln \left( \frac{1}{P_i(1 - P_i)} \right)
\]

where \( P_i \) is the MAF for variant \( i \), \( \beta \) is the log(odds ratio) estimate for variant \( i \), and \( \lambda = 2 \) is the assumed overall familial relative risk (FRR). To compute the corresponding estimate for the FRR due to all variants, we wish to estimate \( \beta = \frac{1}{2} \sum \left( \frac{1}{P_i(1 - P_i)} \right) \left( \frac{1}{P_i(1 - P_i)} \right) \ln \left( \frac{1}{P_i(1 - P_i)} \right) \), where the sum is now over all variants and \( \beta \) is the true relative risk conferred by variant \( i \), assuming a log-additive model. We refer to \( \beta \) as the frailty scale heritability. We first obtained the estimated observed heritability based on the full set of summary estimates using LD Score Regression20 and then converted this to an estimate on the frailty scale using the formula:

\[
\beta_{\text{FRR}} = \frac{1}{\sum \left( \frac{1}{P_i(1 - P_i)} \right) \left( \frac{1}{P_i(1 - P_i)} \right) \ln \left( \frac{1}{P_i(1 - P_i)} \right)}
\]

where \( P \) is the proportion of samples in the population that are cases.

**Pathway analyses.** The pathway gene set database (http://download.baderlab.org/ EM_Genestes, file Human_GOBP_AllPathways_no.GO.iea_April_01_2017_symbol.gmt)74 was used for the Bader laboratory dated 1 April 2017 was used in all analyses. This database contains pathways from Reactome60, NCI Pathway Interaction Database62, GO (Gene Ontology) biological process63, HumanCyc64, MSigdb65, NetPath66 and Panther67. For GO, terms inferred from electronic annotation were excluded from our analyses. The same pathway may be defined in two or more databases with potentially different sets of genes. All versions of such ‘duplicate’ pathways were included. To provide more biologically meaningful results
and reduce false positives, only pathways that contained between 10 and 200 genes were used. Pathway size was determined by the total number of genes in the pathway that could also be mapped to the genes included in the GWAS dataset (actual pathway size may be larger).

SNPs were also filtered using the INQUSIT target prediction method described above for all SNPs with \( p < 5 \times 10^{-6} \) (around 1.25 million associations). This cut-off was chosen based on a threshold analysis that showed that 19 of the 20 pathway themes found using all SNP associations (around 16 million) and a simple distance-based SNP-to-gene mapping method could be recovered using this smaller subset of associations. More stringent cut-offs resulted in fewer themes being covered (for example, three themes found using SNPs with \( p < 5 \times 10^{-6} \) or around 33 thousand SNP associations). Gene significance was calculated by assigning the statistically significant SNP data from all SNPs assigned to a gene.25–27 Because histone gene loci contained a high number of mapped SNPs, we selected significant SNP associations to avoid pathway enrichments based solely on the increased number of SNPs at these loci (that is, chr2:27657944 for HIST; chr1:149219841 for HIST2; chr1:228514706 for HIST3; chr1:124871747 for HIST4).

The gene set enrichment analysis algorithm as implemented in the GenGen package28 was used to perform pathway analysis. Wang et al.29 modified the original gene set enrichment analysis algorithm to work with GWAS datasets, using SNP significance and SNP-to-gene mapping instead of gene expression data. In brief, the algorithm calculates an enrichment score for each pathway based on a weighted Kolmogorov–Smirnov statistic (refer to ref. 70 for more details). Pathways that have most of their genes at the top of the ranked list of genes obtained higher enrichment scores. Note that only the largest positive enrichment score was considered as opposed to largest absolute enrichment score (that is, the largest deviation from zero). This modification (recommended by the GenGen authors for GWAS analysis) was performed to include only pathways that are significantly affected between cases and controls and ignore those with significant negative enrichment scores (this may happen if a pathway is significantly less altered than expected by chance). Only pathways containing more than 10 genes with at least one of these genes with \( p < 5 \times 10^{-6} \) were retained as higher confidence for subsequent analysis. These pathways, together with the genes reaching the significance thresholds, were listed in the supplementary material.

The pathway analysis assigns an enrichment score for each pathway. These values were normalized and \( P \) values for each pathway were obtained by separately comparing them to null distributions for OncoArray and iCOGS datasets. The null distributions were computed by permuting case/control labels 1,000 times (keeping the number of cases and controls the same in each iteration) and recomputing all enrichment statistics. FDR values were computed using the statistics from the null distributions and all pathways with FDR < 0.05 in either OncoArray or iCOGS distributions were considered further. Pathway findings were further considered if they contained more than one significant gene and if they could be confirmed to be involved in breast cancer as reported in at least one of five published large-scale breast cancer GWAS1–6 or reported elsewhere in the literature. Furthermore, themes that were weakly associated with breast cancer (based on a literature search) were only included if they had a FDR < 0.05 and at least four novel genes (that is, was not found among the genes from mapped themes containing pathways known to be involved in breast cancer) (Extended Data Fig. 2). Pathways related to ‘sensory perception of smell’ were removed as there is no evidence in the literature for their involvement in breast cancer and because they contain genes close to each other on chromosome 6 that are frequently correlated.

An enrichment map was created using the Enrichment Map version 2.1.0 app16 in Cytoscape version 3.3 (ref. 76). Pathway nodes were laid out using a force-directed layout and nodes with gene set overlap of over 0.55 were connected by edges. Related pathway nodes were manually clustered and labelled as themes.

Data availability. A subset of the data that support the findings of this study is publically available via dbGaP (http://www.ncbi.nlm.nih.gov/gap; accession number phs001265.v1.p1). The complete dataset will not be made publicly available due to restraints imposed by the ethics committees of individual studies; requests for data can be made to the corresponding author or the Data Access Coordination Committee (DACC) of BCAC (http://bcac.cge.medschl.cam.ac.uk/). BCAC DACC approval is required to access data from studies ABCFS, ABCS, ABCTB, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UT, BCINIS, BSUCH, CBCS, CECILE, CGPS, CTS, DIETCOMPLY, ESTHER, GC-HBOC, GENICA, GEPARISIXTO, GESBC, HABCS, HCSC, HEBCS, HBUCS, KBARBC, KBCP, LMBC, MBACS, MARIE, MBCGS, MCBS, MISS, MMTLGBECS, NC-BCFR, OFBCR, ORIGO, pKARMA, POSH, PREFACE, RBCS, SKDKFZS, SUCCESB, SUCCSESS, SZBSC, TNNBC, UCIBCS, UKGBS and UKOPS (see Supplementary Table 1). Summary results for all variants are available at http://bcac.cge.medschl.cam.ac.uk/). Requests for further data should be made through the BCAC DACC (http://bcac.cge.medschl.cam.ac.uk/).

An enrichment map was created using the Enrichment Map version 2.1.0 app16 in Cytoscape version 3.3 (ref. 76). Pathway nodes were laid out using a force-directed layout and nodes with gene set overlap of over 0.55 were connected by edges. Related pathway nodes were manually clustered and labelled as themes.

Data availability. A subset of the data that support the findings of this study is publically available via dbGaP (http://www.ncbi.nlm.nih.gov/gap; accession number phs001265.v1.p1). The complete dataset will not be made publicly available due to restraints imposed by the ethics committees of individual studies; requests for data can be made to the corresponding author or the Data Access Coordination Committee (DACC) of BCAC (http://bcac.cge.medschl.cam.ac.uk/). BCAC DACC approval is required to access data from studies ABCFS, ABCS, ABCTB, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UT, BCINIS, BSUCH, CBCS, CECILE, CGPS, CTS, DIETCOMPLY, ESTHER, GC-HBOC, GENICA, GEPARISIXTO, GESBC, HABCS, HCSC, HEBCS, HBUCS, KBARBC, KBCP, LMBC, MBACS, MARIE, MBCGS, MCBS, MISS, MMTLGBECS, NC-BCFR, OFBCR, ORIGO, pKARMA, POSH, PREFACE, RBCS, SKDKFZS, SUCCESB, SUCCSESS, SZBSC, TNNBC, UCIBCS, UKGBS and UKOPS (see Supplementary Table 1). Summary results for all variants are available at http://bcac.cge.medschl.cam.ac.uk/). Requests for further data should be made through the BCAC DACC (http://bcac.cge.medschl.cam.ac.uk/).
60. Ghoussaini, M. et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nat. Commun. 4, 4999 (2014).
61. Joshi-Tope, G. et al. Reactome: a knowledgebase of biological pathways. Nucleic Acids Res. 33, D428–D432 (2005).
62. Schaefer, C. F. et al. PID: the pathway interaction database. Nucleic Acids Res. 37, D674–D679 (2009).
63. Ashburner, M. et al. Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000).
64. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 6, R2 (2004).
65. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl Acad. Sci. USA 102, 15545–15550 (2005).
66. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 11, R3 (2010).
67. Thomas, P. D. et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 13, 2129–2141 (2003).
68. Wang, L., Jia, P., Wolfinger, R. D., Chen, X. & Zhao, Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. Genomics 98, 1–8 (2011).
69. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. Nat. Rev. Genet. 11, 843–854 (2010).
70. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genome-wide association studies. Am. J. Hum. Genet. 81, 1278–1283 (2007).
71. Mogushi, K. & Tanaka, H. PathAct: a novel method for pathway analysis using gene expression profiles. Bioinformation 9, 394–400 (2013).
72. Medina, I. et al. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res. 37, W340–W344 (2009).
73. Lee, Y. H., Kim, J. H. & Song, G. G. Genome-wide pathway analysis of breast cancer. Tumour Biol. 35, 7699–7705 (2014).
74. Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. Bioinformatics 27, 95–102 (2011).
75. Braun, R. & Buetow, K. Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. PLoS Genet. 7, e1002101 (2011).
76. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13, 2498–2504 (2003).
Extended Data Figure 1 | Global mapping of biofeatures across novel loci associated with overall breast cancer risk. The overlaps between potential genomic predictors in relevant breast cell lines and credible risk variants (CRVs) within each locus. On the x axis, each column represents a CRV (see Methods). The most significant SNPs are identified in each region. On the y axis, biofeatures are grouped into five functional categories: genomic structure (red), enhancer markers (dark green), histone markers (blue), open chromatin markers (dark blue) and transcription factor binding sites (dark violet). Coloured elements indicate SNPs for which the feature is present. For data sources, see Methods (In silico analysis of CRVs).
Extended Data Figure 2 | Pathway enrichment map for susceptibility loci based on summary association statistics. Each coloured circle (node) represents a pathway (gene set), coloured by enrichment score where redder nodes indicate lower FDRs. Larger nodes indicate pathways with more genes. Green lines connect pathways with overlapping genes (minimum overlap 0.55). Pathways are grouped by similarity and organized into major themes (large labelled circles).
Extended Data Figure 3 | Heat map showing patterns of cell-type-specific enrichments for breast tissue across three histone marks (H3K4me1, H3K4me3 and H3K9ac) for all breast cancer types, ER-positive breast cancer and ER-negative breast cancer as well as 16 other traits. BC_ERneg, ER-negative breast cancer; BC_ERpos, ER-positive breast cancer; BC_overall, all breast cancer types; BMI, body mass index; CAD, cardiovascular disease; CD, Crohn's disease; HDL, high-density lipoprotein; LDL, low-density lipoprotein; RA, rheumatoid arthritis; T2D, type 2 diabetes; TG, triglycerides; UC, ulcerative colitis; vHMEC, variant human mammary epithelial cells.
Extended Data Figure 4 | Heat map showing patterns of cell-type-specific enrichments for histone mark H3K27ac in all breast cancer types, ER-positive and ER-negative breast cancer as well as 16 other traits.
Extended Data Figure 5 | Heat map showing patterns of cell-type-specific enrichments for histone mark H3K4me1 in all breast cancer types, ER-positive and ER-negative breast cancer as well as 16 other traits.
Extended Data Figure 6 | Heat map showing patterns of cell-type-specific enrichments for histone mark H3K4me3 in breast cancer overall, ER+ and ER- breast cancer as well as 16 other traits.
Extended Data Figure 7 | Heat map showing patterns of cell-type-specific enrichments for histone marker H3K9ac in all breast cancer types, ER-positive and ER-negative breast cancer as well as 16 other traits.
Extended Data Figure 8 | Functional assessment of regulatory variants at 1p36, 11p15 and 1p34 risk loci. a, b, The \textit{KLHDC7A} (a) or \textit{PIDD1} (b) promoter regions, containing the reference (prom-Ref) or risk alleles (prom-Hap), were cloned upstream of the \textit{pGL3} luciferase reporter gene. MCF7 or Bre-80 cells were transfected with constructs and assayed for luciferase activity after 24 h. The means and 95% confidence intervals are shown. (n = 3). P values were determined by two-way ANOVA followed by Dunnett’s multiple comparisons test (**P < 0.01, ***P < 0.001). c, 3C assays. Top, a physical map of the region analysed by 3C. Grey boxes depict the PREs, blue vertical lines indicate the risk-associated SNPs and the black dotted line represents chromatin looping. Bottom, graphs representing three independent 3C interaction profiles. 3C libraries were generated with EcoRI, grey vertical boxes indicate the interacting restriction fragment (containing PRE1 and PRE2). Means and standard deviations are shown. d, PRE1 or PRE2 containing the reference (PRE-ref) or risk (PRE-Hap) haplotypes were cloned downstream of a \textit{CITED4} promoter-driven luciferase construct (\textit{CITED4} prom). MCF7 or Bre-80 cells were transfected with constructs and assayed for luciferase activity after 24 h. Error bars denote 95% CI (n = 3). P values were determined by two-way ANOVA followed by Dunnett’s multiple comparisons test (**P < 0.01, ***P < 0.001).
Extended Data Figure 9 | Functional assessment of regulatory variants at the 7q22 risk locus. a–e, 3C assays. Top, a physical map of the region interrogated by 3C. Grey horizontal boxes depict the putative regulatory elements (PREs), blue vertical lines indicate the risk-associated SNPs and the black dotted line represents chromatin looping. Bottom, graphs represent three independent 3C interaction profiles between the CUX1 (a), PRKRI1 (b, d) or RASA4 (c, e) promoter regions and PREs. 3C libraries were generated with EcoRI, grey vertical boxes indicate the interacting restriction fragment (containing PRE1 and/or PRE2). Means and standard deviations are shown. f, g, Allele-specific 3C. 3C followed by Sanger sequencing for the PRKRI1-PRE2 (f) or RASA4-PRE1 or -PRE2 (g) in heterozygous MDA-MB-231 breast cancer cells.
Extended Data Table 1 | INQUISIT, DEPICT and the nearest gene as predictors of driver status

| Variable                        | Coefficient | Standard Error | Z-value | P-value |
|---------------------------------|-------------|----------------|---------|---------|
| **Multivariable logistic regression model with INQUISIT, DEPICT, Nearest gene** |             |                |         |         |
| INQUISIT                        | -0.61       | 0.14           | -4.31   | 1.6E-05 |
| DEPICT                          | -0.17       | 0.50           | -0.33   | 0.74    |
| Nearest gene                    | 1.12        | 0.59           | 1.88    | 0.06    |
| **Multivariable logistic regression model with INQUISIT and Nearest gene** |             |                |         |         |
| INQUISIT                        | -0.63       | 0.13           | -4.77   | 1.8E-06 |
| Nearest gene                    | 1.23        | 0.48           | 2.56    | 0.01    |
| **Multivariable logistic regression model with DEPICT and Nearest gene** |             |                |         |         |
| DEPICT                          | -0.89       | 0.49           | -1.82   | 0.07    |
| Nearest gene                    | 1.61        | 0.63           | 2.57    | 0.01    |

Scores were converted into levels for analysis. For INQUISIT: level 1 (‘coding score of 2’ or ‘promoter score of 3 or 4’ or ‘distal score >4’), level 2 (‘coding score of 1’ or ‘promoter score of 1 or 2’ or ‘distal score of 1, 2, 3, or 4’), level 3 (coding/promoter/distal scores >0 but <1) and level 4 (not predicted to be a target gene by INQUISIT). For DEPICT: level 1 (DEPICT predicted target gene at $P \leq 0.05$), level 2 (DEPICT predicted target gene, but with $P > 0.05$) and level 3 (not predicted to be a target gene by DEPICT).
Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see Reporting Life Sciences Research. For further information on Nature Research policies, including our data availability policy, see Authors & Referees and the Editorial Policy Checklist.

1. Sample size
   Describe how sample size was determined. The dataset included essentially all available GWAS data on breast cancer.

2. Data exclusions
   Describe any data exclusions. We excluded samples, and variants, according to predefined filtering criteria to remove unreliable genotype calls, as described in the statistical methods. These criteria included, for variants: low call rate, low minor allele frequency, deviation from Hardy-Weinberg and deviation in frequency from the 1000 genomes reference panel. We also excluded variants judged to be poor on manual inspection of intensity cluster plots. Sample exclusions included low call rate, unusually high or low heterozygosity, and samples not of European or Asian ancestry, as appropriate, based on principal components analysis. We also excluded probable duplicates and close relatives, and samples overlaps between the contributing GWAS.

3. Replication
   Describe whether the experimental findings were reliably reproduced. N/A. This was an observational study - analyses were based on all available data.

4. Randomization
   Describe how samples/organisms/participants were allocated into experimental groups. This was an observational genetic association study, hence randomisation was not relevant. The analyses were adjusted for country and ancestry informative principal components, as described in the methods.

5. Blinding
   Describe whether the investigators were blinded to group allocation during data collection and/or analysis. The laboratories conducting the genotyping did not have access to the phenotypic data (i.e. were blinded). Moreover, genotype calling was automated. The phenotype and genotype data were only combined during the statistical analysis.

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.
6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

| n/a | Confirmed |
|-----|-----------|
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |
| ☑   | ☑         |

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on statistics for biologists for further resources and guidance.

7. Software

Describe the software used to analyze the data in this study.

Analyses were mostly conducted using standard software. These include ShapeIT2, IMPUTE v2, Mach, Minimac, R, and METAL. We used two purpose written software for principal components analysis, pccalc (http://ccge.medschl.cam.ac.uk/software/pccalc) and for fast logistic logistic regression, mlogit (http://ccge.medschl.cam.ac.uk/software/mlogit/)

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). Nature Methods guidance for providing algorithms and software for publication provides further information on this topic.

8. Materials and reagents

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

No unique materials were used (with the exception of the patient DNA samples). The Oncoarray is available for purchase from Illumina.

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

N/A

10. Eukaryotic cell lines

- State the source of each eukaryotic cell line used.
- Describe the method of cell line authentication used.
- Report whether the cell lines were tested for mycoplasma contamination.
- If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.

MCF7 and MDA-MB-231 breast cancer cells were purchased from ATCC, Bre-80 normal breast cells were provided as a gift from Prof Roger Reddel, CMRI, Sydney.

All cell lines were short tandem repeat (STR) profiled.

All cell lines tested negative for mycoplasma contamination.

No cell lines are on the database of commonly misidentified lines.
Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals
   Provide details on animals and/or animal-derived materials used in the study.
   N/A

Policy information about studies involving human research participants

12. Description of human research participants
   Describe the covariate-relevant population characteristics of the human research participants.
   Participants were female breast cancer patients or healthy female controls.