doi:10.1038/nature24284

Association analysis identifies 65 new breast cancer risk loci

Lists of authors and their affiliations appear in the online version of the paper

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[1]. 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[1]. 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)[2] 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[3–5] 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 6,624 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 $P = 1.6 \times 10^{-10}$; rs373038216: $P = 2.9 \times 10$; 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,

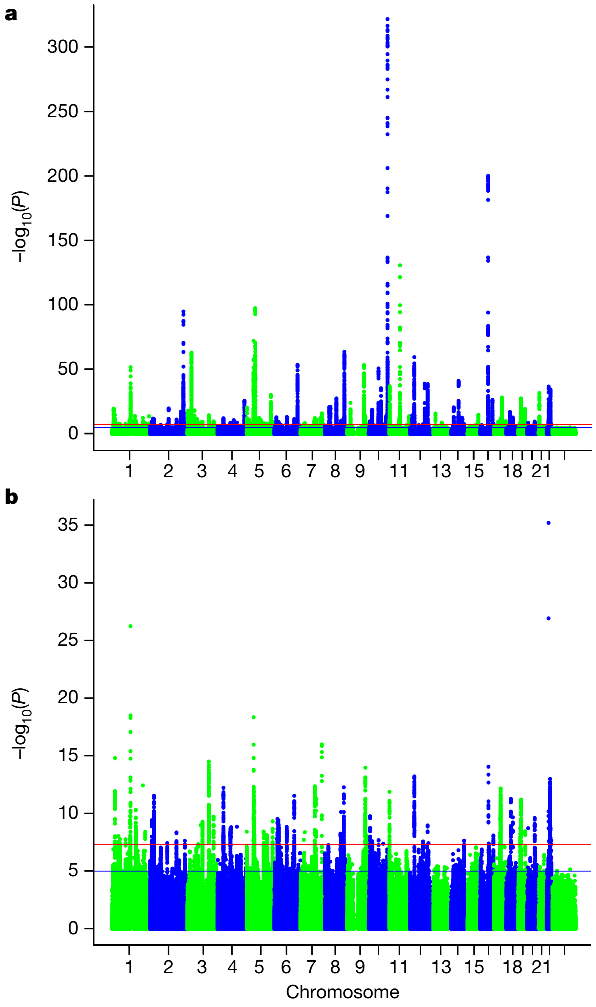


Figure 1 | SNP associations with breast cancer risk. a, Manhattan plot showing $-\log_{10}P$ values for SNP associations with breast cancer risk. **b**, Manhattan plot after excluding previously identified associated regions. The red line denotes 'genome-wide' significance ($P < 5 \times 10^{-8}$); the blue line denotes $P < 10^{-5}$.

nine were associated at $P < 0.05$ and only two showed a difference in the estimated odds ratio between Europeans and Asians ($P < 0.01$; 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 TCF7L2. 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 *PHLDA3*, 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 *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 58 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 ($P = 0.01$; 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 11p15) 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 11p15 (smallest $P = 1.4 \times 10^{-12}$), five were located in the proximal promoter 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 *PRKRIP1* (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 transactivation of a specific promoter and by affecting chromatin looping between regulatory elements and their target genes. These data provide *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 41% of the familial relative risk, assuming a log-additive model (see Methods). Therefore, the identified susceptibility SNPs account for 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 that are 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. *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. *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.* PH domain-only protein PHLDA3 is a p53-regulated repressor of Akt. *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 interactome 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,** 455–461 (2014).
11. Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534,** 47–54 (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,** 506–519 (2015).
14. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat. Commun.* **7,** 11479 (2016).
15. Pers, 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 network-based method for gene-set enrichment visualization and interpretation. *PLoS ONE* **5,** e13984 (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 signaling and breast cancer. *Cancer Biol. Ther.* **3,** 36–41 (2004).
20. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide 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).

Supplementary Information is available in the online version of the paper.

Acknowledgements We 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. Genotyping of the OncoArray was principally funded from three sources: the PERSPECTIVE project, funded by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie, de la Science et de l'Innovation du Québec' through Genome Québec, and the Quebec Breast Cancer Foundation; the NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project (NIH Grants U19 CA148065 and X01HG007492); and Cancer Research UK (C1287/A10118 and C1287/A16563). BCAC is funded by Cancer Research UK (C1287/A16563), by the European Community's Seventh Framework Programme under grant agreement 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the 'CIHR Team in Familial Risks of Breast Cancer' program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec, grant PSR-SIIRI-701. Combining of the GWAS data was supported in part by The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE, part of the GAME-ON initiative). For a full description of funding and acknowledgments, see Supplementary Note.

Author Contributions Writing group: K.Mi., S.Li., J.Bee., S.Hu., S.Ka., P.So., S.L.E., G.D.B., G.C.-T., J.Si., P.K. and D.F.E. Conceived the OncoArray and obtained financial support: C.I.A., J.Si., P.K. and D.F.E. Designed the OncoArray: J.D., E.D., A. Lee, Z.W., A.C.A., C.I.A., S.J.C., P.K. and D.F.E. Led the COGS project: P.Hal. Led the DRIVE project: D.J.H. Led the PERSPECTIVE project: J.Si. Led the working groups of BCAC: A.C.A., I.L.A., P.D.P.P., J.C.-C., R.L.M., M.G.-C., M.K.S. and A.M.D. Data management: J.D., M.K.B., Q.Wan., R.Ke., U.E., S.B., J.C.-C. and M.K.S. Bioinformatics analysis: J.D., J.Bee., A.Lem., P.So., J.A., M.Gh., J.C., A.D., A.E.M.R., S.R.L. Statistical analysis: K.Mi., S.Li., S.Hu., S.Ka., A.Ros., J.T., X.Q.C., L.Fa., X.J., H.Fi., G.D.B., P.K. and D.F.E. Functional analysis: D.G., X.Q.C., J.Bee., J.D.F., K.Mc., S.L.E. and G.C.-T. OncoArray genotyping: M.A., F.B., C.Ba., D.M.C., J.M.C., K.F.D., N.Ha., B.H., K.J., C.L., J.Me., E.P., J.R., G.S., D.C.T., D.V.D.B., D.V., J.V., L.X., B.Z. and A.M.D. Provided DNA samples and/or phenotypic data: M.A.A., H.A., K.A., H.A.-C., N.N.A., V.A., K.J.A., B.A., P.L.A., M.Ba., M.W.B., J.Ben., M.Be., L.Be., C.Bl., N.V.B., S.E.B., B.Bo., A.-L.B.-D., J.S.B., H.Bra., P.Bre., H.Bre., L.Br., P.Bro., I.W.B., A.B., A.B.-W., S.Y.B., T.B., B.Bu., K.B., H.Ca., Q.C., T.C., F.C., A.Ca., B.D.C., J.E.C., T.L.C., T.-Y.D.C., K.S.C., J.-Y.C., H.Ch., C.L.C., M.C., E.C.-D., S.C., A.Co., D.G.C., S.S.C., K.C., M.B.D., P.D., T.D., I.d.-S.-S., M.Du., L.D., M.Dw., D.M.E., A.B.E., A.H.E., C.El., M.El., C.En., M.Er., P.A.F., J.F., D.F.-J., O.F., H.Fl., L.Fr., V.Ga., M.Ga., M.G.-D., Y.-T.G., S.M.G., J.A.G.-S., M.M.G., V.Ge., G.G.G., G.G., M.S.G., D.E.G., A.G.-N., G.I.G.A., M.Gr., J.G., A.G., P.G., L.H., E.H., C.A.H., N.Hå., U.H., S.Ha., P.Har., S.N.H., J.M.H., M.H., A.He., J.H., P.Hi., D.N.H., A.Ho., M.J.H., R.N.H., J.L.H., M.-F.H., C.-N.H., G.H., K.H., J.I., H.It., M.I., H.Iw., A.J., W.J., E.M.J., N.J., M.J., A.J.-V., R.Ka., M.K., K.K., D.K., Y.K., M.J.K., S.Kh., E.K., J.I.K., S.-W.K., J.A.K., V.-M.K., I.M.K., V.N.K., U.K., A.K., D.L., L.L.M., C.N.L., E.L., J.W.L., M.H.L., F.L., J. Li, J.Lil., A.Li., J.Lis., R.L., W.-Y.L., S.Lo., J.Lo., A.Lo., J.Lu., M.P.L., E.S.K.M., R.J.M., T.M., E.M., K.E.M., A.Ma., S.Man., J.E.M., S.Marg., S.Mari., M.E.M., K.Ma., D.M., J.Mc., C.Mc., H.M.-H., A.Me., P.M., U.M., H.M., N.M., K.Mu., A.M.M., C.Mu., S.L.N., H.N., P.N., S.F.N., D.-Y.N., B.G.N., A.N., O.I.O., J.E.O., H.O., C.O., N.O., V.S.P., S.K.P., T.-W.P.-S., J.I.A.P., P.P., J.P., K.-A.P., M.P., D.P.-K., R.P., N.P., D.P., K.P., B.R., P.R., N.R., G.R., H.S.R., V.R., A.Rom., K.J.R., T.R., A.Ru., M.R., E.J.T.R., E.S., D.P.S., S.Sa., E.J.S., D.F.S., R.K.S., A.Sc., M.J.Sc., F.S., P.Sc., C.Sc., R.J.S., S.Se., C.Se., M.S., P.Sh., C.-Y.S., M.E.S., M.J.Sh., X.-O.S., A.Sm., C.So., M.C.S., J.J.S., C.St., S.S.-B., J.St., D.O.S., H.S., A.Sw., N.A.M.T., R.T., J.A.T., M.T., S.H.T., M.B.T., S.Th., K.T., R.A.E.M.T., I.T., L.T., D.T., T.T., C.-C.T., S.Ts., H.-U.U., M.U., G.U., C.V., C.J.v.A., A.M.W.v.d.O., L.v.d.K., R.B.v.d.L., Q.Wai., S.W.-G., C.R.W., C.W., A.S.W., H.W., W.W., R.W., A.W., A.H.W., T.Y., X.R.Y., C.H.Y., K.-Y.Y., J.-C.Y., W.Z., Y.Z., A.Z., E.Z., ABCTB Investigators, kConFab/AOCS Investigators, NBCS Collaborators, A.C.A., I.L.A., F.J.C., P.D.P.P., J.C.-C., P.Hal., D.J.H., R.L.M., M.G.-C., M.K.S., G.D.B., J.Si., P.K. and D.F.E. All authors read and approved the final version of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. Correspondence and requests for materials should be addressed to D.F.E. (dfe20@medschl.cam.ac.uk).

Kyriaki Michailidou[1,2]*, Sara Lindström[3,4]*, Joe Dennis[1]*, Jonathan Beesley[5]*, Shirley Hui[6]*, Siddhartha Kar[7]*, Audrey Lemaçon[8], Penny Soucy[8], Dylan Glubb[5], Asha Rostamianfar[6], Manjeet K. Bolla[1], Qin Wang[1], Jonathan Tyrer[7], Ed Dicks[7], Andrew Lee[1], Zhaoming Wang[9,10], Jamie Allen[1], Renske Keeman[11], Ursula Eilber[12], Juliet D. French[5], Xiao Qing Chen[5], Laura Fachal[7], Karen McCue[5], Amy E. McCart Reed[13], Maya Ghoussaini[7], Jason S. Carroll[14], Xia Jiang[4], Hilary Finucane[4,15], Marcia Adams[16], Muriel A. Adank[17], Habibul Ahsan[18], Kristiina Aittomäki[19], Hoda Anton-Culver[20], Natalia N. Antonenkova[21], Volker Arndt[22], Kristan J. Aronson[23], Banu Arun[24], Paul L. Auer[25,26], François Bacot[27], Myrto Barrdahl[12], Caroline Baynes[7], Matthias W. Beckmann[28], Sabine Behrens[12], Javier Benitez[29,30], Marina Bermisheva[31], Leslie Bernstein[32], Carl Blomqvist[33], Natalia V. Bogdanova[21,34,35], Stig E. Bojesen[36,37,38], Bernardo Bonanni[39], Anne-Lise Børresen-Dale[40], Judith S. Brand[41], Hiltrud Brauch[42,43,44], Paul Brennan[45], Hermann Brenner[22,44,46], Louise Brinton[47], Per Broberg[48], Ian W. Brock[49], Annegien Broeks[11], Angela Brooks-Wilson[50,51], Sara Y. Brucker[52], Thomas Brüning[53], Barbara Burwinkel[54,55], Katja Butterbach[22], Qiuyin Cai[56], Hui Cai[56], Trinidad Caldés[57], Federico Canzian[58], Angel Carracedo[59,60], Brian D. Carter[61], Jose E. Castelao[62], Tsun L. Chan[63,64], Ting-Yuan David Cheng[65], Kee Seng Chia[66], Ji-Yeob Choi[67,68], Hans Christiansen[34], Christine L. Clarke[69], NBCS Collaborators†, Margriet Collée[70], Don M. Conroy[7], Emilie Cordina-Duverger[71], Sten Cornelissen[11], David G. Cox[72,73], Angela Cox[49], Simon S. Cross[74], Julie M. Cunningham[75], Kamila Czene[41], Mary B. Daly[76], Peter Devilee[77,78], Kimberly F. Doheny[16], Thilo Dörk[35], Isabel dos-Santos-Silva[79], Martine Dumont[8], Lorraine Durcan[80,81], Miriam Dwek[82], Diana M. Eccles[81], Arif B. Ekici[83], A. Heather Eliassen[84,85], Carolina Ellberg[48,86], Mingajeva Elvira[87], Christoph Engel[88,89], Mikael Eriksson[41], Peter A. Fasching[28,90], Jonine Figueroa[47,91], Dieter Flesch-Janys[92,93], Olivia Fletcher[94], Henrik Flyger[95], Lin Fritschi[96], Valerie Gaborieau[45], Marike Gabrielson[41], Manuela Gago-Dominguez[59,97], Yu-Tang Gao[98], Susan M. Gapstur[61], José A. García-Sáenz[57], Mia M. Gaudet[61], Vassilios Georgoulias[99], Graham G. Giles[100,101], Gord Glendon[102], Mark S. Goldberg[103,104], David E. Goldgar[105], Anna González-Neira[29], Grethe I. Grenaker Alnæs[40], Mervi Grip[106], Jacek Gronwald[107], Anne Grundy[108], Pascal Guénel[71], Lothar Haeberle[28], Eric Hahnen[109,110,111], Christopher A. Haiman[112], Niclas Håkansson[113], Ute Hamann[114], Nathalie Hamel[27], Susan Hankinson[115], Patricia Harrington[7], Steven N. Hart[116], Jaana M. Hartikainen[117,118,119], Mikael Hartman[66,120], Alexander Hein[28], Jane Heyworth[121], Belynda Hicks[10], Peter Hillemanns[35], Dona N. Ho[64], Antoinette Hollestelle[122], Maartje J. Hooning[122], Robert N. Hoover[47], John L. Hopper[101], Ming-Feng Hou[123], Chia-Ni Hsiung[124], Guanmengqian Huang[114], Keith Humphreys[41], Junko Ishiguro[125,126], Hidemi Ito[125,126], Motoki Iwasaki[127], Hiroji Iwata[128], Anna Jakubowska[107], Wolfgang Janni[129], Esther M. John[130,131,132], Nichola Johnson[94], Kristine Jones[10], Michael Jones[133], Arja Jukkola-Vuorinen[134], Rudolf Kaaks[12], Maria Kabisch[114], Katarzyna Kaczmarek[107], Daehee Kang[67,68,135], Yoshio Kasuga[136], Michael J. Kerin[137], Sofia Khan[138], Elza Khusnutdinova[31,87], Johanna I. Kiiski[138], Sung-Won Kim[139], Julia A. Knight[140,141], Veli-Matti Kosma[117,118,119], Vessela N. Kristensen[40,142,143], Ute Krüger[48], Ava Kwong[63,144,145], Diether Lambrechts[146,147], Loic Le Marchand[148], Eunjung Lee[112], Min Hyuk Lee[149], Jong Won Lee[150], Chuen Neng Lee[120,151], Flavio Lejbkowicz[152], Jingmei Li[41], Jenna Lilyquist[116], Annika Lindblom[153], Jolanta Lissowska[154], Wing-Yee Lo[42,43], Sibylle Loibl[155], Jirong Long[56], Artitaya Lophatananon[156,157], Jan Lubinski[107], Craig Luccarini[7], Michael P. Lux[28], Edmond S. K. Ma[63,64], Robert J. MacInnis[100,101], Tom Maishman[80,81], Enes Makalic[101], Kathleen E. Malone[158], Ivana Maleva Kostovska[159], Arto Mannermaa[117,118,119], Siranoush Manoukian[160], JoAnn E. Manson[85,161], Sara Margolin[162], Shivaani Mariapun[163], Maria Elena Martinez[97,164], Keitaro Matsuo[126,165], Dimitrios Mavroudis[99], James McKay[45], Catriona McLean[166], Hanne Meijers-Heijboer[17], Alfons Meindl[167], Primitiva Menéndez[168], Usha Menon[169], Jeffery Meyer[75], Hui Miao[66], Nicola Miller[137], Nur Aishah Mohd Taib[170], Kenneth Muir[156,157], Anna Marie Mulligan[171,172], Claire Mulot[173], Susan L. Neuhausen[32], Heli Nevanlinna[138], Patrick Neven[174], Sune F. Nielsen[36,37], Dong-Young Noh[175], Børge G. Nordestgaard[36,37,38], Aaron Norman[116], Olufunmilayo I. Olopade[176], Janet E. Olson[116], Håkan Olsson[48], Curtis Olswold[116], Nick Orr[94], V. Shane Pankratz[177], Sue K. Park[67,68,135], Tjoung-Won Park-Simon[35], Rachel Lloyd[178], Jose I. A. Perez[179], Paolo Peterlongo[180], Julian Peto[79], Kelly-Anne Phillips[101,181,182,183], Mila Pinchev[152], Dijana Plaseska-Karanfilska[159], Ross Prentice[25], Nadege Presneau[82], Darya Prokofyeva[87], Elizabeth Pugh[16], Katri Pylkäs[184,185], Brigitte Rack[186], Paolo Radice[187], Nazneen Rahman[188], Gadi Rennert[152], Hedy S. Rennert[152], Valerie Rhenius[7], Atocha Romero[57,189], Jane Romm[16], Kathryn J. Ruddy[190], Thomas Rüdiger[191], Anja Rudolph[12], Matthias Ruebner[28], Emiel J. T. Rutgers[192], Emmanouil Saloustros[193], Dale P. Sandler[194], Suleeporn Sangrajrang[195], Elinor J. Sawyer[196], Daniel F. Schmidt[101], Rita K. Schmutzler[109,110,111], Andreas Schneeweiss[54,197], Minouk J. Schoemaker[133], Fredrick Schumacher[198], Peter Schürmann[35], Rodney J. Scott[199,200], Christopher Scott[116], Sheila Seal[188], Caroline Seynaeve[122], Mitul Shah[7], Priyanka Sharma[201], Chen-Yang Shen[202,203], Grace Sheng[112], Mark E. Sherman[204], Martha J. Shrubsole[56], Xiao-Ou Shu[56], Ann Smeets[174], Christof Sohn[197], Melissa C. Southey[205], John J. Spinelli[206,207], Christa Stegmaier[208], Sarah Stewart-Brown[156], Jennifer Stone[178,209], Daniel O. Stram[112], Harald Surowy[54,55], Anthony Swerdlow[133,210], Rulla Tamimi[4,84,85], Jack A. Taylor[194,211], Maria Tengström[117,212,213], Soo H. Teo[163,170], Mary Beth Terry[214], Daniel C. Tessier[27], Somchai Thanasitthichai[215], Kathrin Thöne[93], Rob A. E. M. Tollenaar[216], Ian Tomlinson[217], Ling Tong[18], Diana Torres[114,218], Thérèse Truong[71], Chiu-Chen Tseng[112], Shoichiro Tsugane[219], Hans-Ulrich Ulmer[220], Giske Ursin[221,222], Michael Untch[223], Celine Vachon[116], Christi J. van Asperen[224], David Van Den Berg[112], Ans M. W. van den Ouweland[70], Lizet van der Kolk[225], Rob B. van der Luijt[226], Daniel Vincent[27], Jason Vollenweider[75], Quinten Waisfisz[17], Shan Wang-Gohrke[227], Clarice R. Weinberg[228], Camilla Wendt[162], Alice S. Whittemore[131,132], Hans Wildiers[174], Walter Willett[85,229], Robert Winqvist[184,185], Alicja Wolk[113], Anna H. Wu[112], Lucy Xia[112], Taiki Yamaji[127], Xiaohong R. Yang[47], Cheng Har Yip[230], Keun-Young Yoo[231,232], Jyh-Cherng Yu[233], Wei Zheng[56], Ying Zheng[234], Bin Zhu[10], Argyrios Ziogas[20], Elad Ziv[235], ABCTB Investigators†, kConFab/AOCS Investigators†, Sunil R. Lakhani[13,236], Antonis C. Antoniou[1], Arnaud Droit[8], Irene L. Andrulis[102,237], Christopher I. Amos[238], Fergus J. Couch[75], Paul D. P. Pharoah[1,7], Jenny Chang-Claude[12,239], Per Hall[41,240], David J. Hunter[4,85], Roger L. Milne[100,101], Montserrat García-Closas[47], Marjanka K. Schmidt[11,241], Stephen J. Chanock[47], Alison M. Dunning[7], Stacey L. Edwards[5], Gary D. Bader[6], Georgia Chenevix-Trench[5], Jacques Simard[8]§, Peter Kraft[4,85]§ & Douglas F. Easton[1,7]§

[1]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [2]Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [3]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]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. [8]Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, Quebec, Canada. [9]Department of Computational Biology, St Jude Children's Research Hospital, Memphis, Tennessee, USA. [10]Cancer Genomics Research Laboratory (CGR), Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA. [11]Division of Molecular Pathology, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. [12]Division 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]Center for Cancer Epidemiology and Prevention, The University of Chicago, Chicago, Illinois, USA. [19]Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [20]Department of Epidemiology, University of California Irvine, Irvine, California, USA. [21]N. N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus. [22]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. [23]Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, Ontario, 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]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA. [27]McGill University and Génome Québec Innovation Centre, Montréal, 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]Centro 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 Sciences, Beckman Research Institute of City of Hope, Duarte, California, USA. [33]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [34]Department of Radiation Oncology, Hannover Medical School, Hannover, Germany. [35]Gynaecology Research Unit, Hannover Medical School, Hannover, Germany. [36]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [37]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [38]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [39]Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, Milan, Italy. [40]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway. [41]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [42]Dr Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany. [43]University of Tübingen, Tübingen, Germany. [44]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [45]International Agency for Research on Cancer, Lyon, France. [46]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [47]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland, USA. [48]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden. [49]Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK. [50]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [51]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, British Columbia, Canada. [52]Department of Gynecology and Obstetrics, University of Tübingen, Tübingen, Germany. [53]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum, Bochum, Germany. [54]Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, Germany. [55]Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, Germany. [56]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, Tennessee, USA. [57]Medical Oncology Department, Hospital Clínico San Carlos, IdISSC (Centro Investigacion Biomedica en Red), CIBERONC (Instituto de Investigación Sanitaria San Carlos), Madrid, Spain. [58]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. [59]Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago De Compostela, Spain. [60]Centro de Investigación en Red de Enfermedades Raras (CIBERER) y Centro Nacional de Genotipado (CEGEN-PRB2), Universidad de Santiago de Compostela, Santiago De Compostela, Spain. [61]Epidemiology Research Program, American Cancer Society, Atlanta, Georgia, USA. [62]Oncology and Genetics Unit, Instituto de Investigacion Biomedica (IBI) Orense-Pontevedra-Vigo, Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain. [63]Hong Kong Hereditary Breast Cancer Family Registry, Happy Valley, Hong Kong. [64]Department of Pathology, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong. [65]Division of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, New York, USA. [66]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [67]Department of Biomedical Sciences, Seoul National

University Graduate School, Seoul, South Korea. [68]Cancer Research Institute, Seoul National University, Seoul, South Korea. [69]Westmead Institute for Medical Research, University of Sydney, Sydney, Australia. [70]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands. [71]Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France. [72]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [73]INSERM U1052, Cancer Research Center of Lyon, Lyon, France. [74]Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK. [75]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, Minnesota, USA. [76]Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA. [77]Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands. [78]Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. [79]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [80]Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton, Southampton, UK. [81]Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK. [82]Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK. [83]Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany. [84]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [85]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, USA. [86]Oncology and Pathology, Department of Clinical Sciences, Lund University, Lund, Sweden. [87]Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia. [88]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [89]LIFE—Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig, Germany. [90]David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, California, USA. [91]Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, UK. [92]Institute for Medical Biometrics and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [93]Department of Cancer Epidemiology, Clinical Cancer Registry, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [94]Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK. [95]Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [96]School of Public Health, Curtin University, Perth, Australia. [97]Moores Cancer Center, University of California San Diego, La Jolla, California, USA. [98]Department of Epidemiology, Shanghai Cancer Institute, Shanghai, China. [99]Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece. [100]Cancer Epidemiology and Intelligence Division, Cancer Council Victoria, Melbourne, Victoria, Australia. [101]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia. [102]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada. [103]Department of Medicine, McGill University, Montréal, Quebec, Canada. [104]Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, Quebec, Canada. [105]Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah, USA. [106]Department of Surgery, Oulu University Hospital, University of Oulu, Oulu, Finland. [107]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland. [108]Centre de Recherche du Centre Hospitalier de Université de Montréal (CHUM), Université de Montréal, Montréal, Quebec, Canada. [109]Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne, Cologne, Germany. [110]Center for Integrated Oncology (CIO), University Hospital of Cologne, Cologne, Germany. [111]Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. [112]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [113]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [114]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany. [115]Department of Biostatistics & Epidemiology, University of Massachusetts, Amherst, Amherst, Massachusetts, USA. [116]Department of Health Sciences Research, Mayo Clinic, Rochester, Minnesota, USA. [117]Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland. [118]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland. [119]Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland. [120]Department of Surgery, National University Health System, Singapore, Singapore. [121]School of Population Health, University of Western Australia, Perth, Australia. [122]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands. [123]Division of Breast Surgery, Department of Surgery, Kaohsiung Medical University, Kaohsiung, Taiwan. [124]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. [125]Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan. [126]Department of Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan. [127]Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. [128]Department of Breast Oncology, Aichi Cancer Center Hospital, Nagoya, Japan. [129]Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm, Germany. [130]Department of Epidemiology, Cancer Prevention Institute of California, Fremont, California, USA. [131]Department of Health Research and Policy–Epidemiology, Stanford University School of Medicine, Stanford, California, USA. [132]Stanford Cancer Institute, Stanford University School of Medicine, Stanford, California, USA. [133]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. [134]Department of Oncology, Oulu University Hospital, University of Oulu, Oulu, Finland. [135]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, South Korea. [136]Department of Surgery, Nagano Matsushiro General Hospital, Nagano, Japan. [137]School of Medicine, National University of Ireland, Galway, Ireland. [138]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [139]Department of Surgery, Daerim Saint Mary's Hospital, Seoul, South Korea. [140]Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada. [141]Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada. [142]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway. [143]Department of Clinical Molecular Biology, Oslo University Hospital, University of Oslo, Oslo, Norway. [144]Department of Surgery, The University of Hong Kong, Pok Fu Lam, Hong Kong. [145]Department of Surgery, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong. [146]Vesalius Research Center, VIB, Leuven, Belgium. [147]Laboratory for Translational Genetics, Department of Oncology, University of Leuven, Leuven, Belgium. [148]University of Hawaii Cancer Center, Honolulu, Hawaii, USA. [149]Department of Surgery, Soonchunhyang University College of Medicine and Soonchunhyang University Hospital, Seoul, South Korea. [150]Department of Surgery, University of Ulsan College of Medicine and Asan Medical Center, Seoul, South Korea. [151]Department of Cardiac, Thoracic and Vascular Surgery, National University Health System, Singapore, Singapore. [152]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel. [153]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden. [154]Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Memorial Cancer Center & Institute of Oncology, Warsaw, Poland. [155]German Breast Group GmbH, Neu Isenburg, Germany. [156]Division of Health Sciences, Warwick Medical School, Warwick University, Coventry, UK. [157]Institute of Population Health, University of Manchester, Manchester, UK. [158]Division of Public Health Sciences, Epidemiology Program, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [159]Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia. [160]Unit of Medical Genetics, Department of Preventive and Predictive Medicine, Fondazione IRCCS (Istituto Di Ricovero e Cura a Carattere Scientifico) Istituto Nazionale dei Tumori (INT), Milan, Italy. [161]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. [162]Department of Oncology–Pathology, Karolinska Institutet, Stockholm, Sweden. [163]Cancer Research Malaysia, Subang Jaya, Selangor, Malaysia. [164]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, California, USA. [165]Division of Molecular Medicine, Aichi Cancer Center Research Institute, Nagoya, Japan. [166]Anatomical Pathology, The Alfred Hospital, Melbourne, Australia. [167]Division of Gynaecology and Obstetrics, Technische Universität München, Munich, Germany. [168]Servicio de Anatomía Patológica, Hospital Monte Naranco, Oviedo, Spain. [169]Gynaecological Cancer Research Centre, Department of Women's Cancer, Institute for Women's Health, University College London, London, UK. [170]Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia. [171]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [172]Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada. [173]Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris, France. [174]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium. [175]Department of Surgery, Seoul National University College of Medicine, Seoul, South Korea. [176]Center for Clinical Cancer Genetics and Global Health, The University of Chicago, Chicago, Illinois, USA. [177]University of New Mexico Health Sciences Center, University of New Mexico, Albuquerque, New Mexico, USA. [178]The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Australia. [179]Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, Oviedo, Spain. [180]IFOM, The FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan, Italy. [181]Peter MacCallum Cancer Center, Melbourne, Australia. [182]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Australia. [183]Department of Medicine, St Vincent's Hospital, The University of Melbourne, Fitzroy, Australia. [184]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland. [185]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland. [186]Department of Gynecology and Obstetrics, Ludwig-Maximilians University of Munich, Munich, Germany. [187]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS (Istituto Di Ricovero e Cura a Carattere Scientifico) Istituto Nazionale dei Tumori (INT), Milan, Italy. [188]Section of Cancer Genetics, The Institute of Cancer Research, London, UK. [189]Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain. [190]Department of Oncology, Mayo Clinic, Rochester, Minnesota, USA. [191]Institute of Pathology, Staedtisches Klinikum Karlsruhe, Karlsruhe, Germany. [192]Department of Surgery, The Netherlands Cancer Institute, Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. [193]Hereditary Cancer Clinic, University Hospital of Heraklion, Heraklion, Greece. [194]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina, USA. [195]National Cancer Institute, Bangkok, Thailand. [196]Research Oncology, Guy's Hospital, King's College London, London, UK. [197]National Center for Tumor Diseases, University of Heidelberg, Heidelberg, Germany. [198]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, Ohio, USA. [199]Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, Australia. [200]Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Callaghan, Australia. [201]Department of Medicine, University of Kansas Medical Center, Kansas City, Kansas, USA. [202]School of Public Health, China Medical University, Taichung, Taiwan. [203]Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. [204]Division of Cancer Prevention, National Cancer Institute, Rockville, Maryland, USA. [205]Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Australia. [206]Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia, Canada. [207]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada. [208]Saarland Cancer Registry, Saarbrücken, Germany. [209]Department of Obstetrics and Gynaecology, University of Melbourne and the Royal Women's Hospital, Melbourne, Australia. [210]Division of Breast Cancer Research, The Institute of Cancer Research, London, UK. [211]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, North Carolina, USA. [212]Cancer Center, Kuopio University Hospital, Kuopio, Finland. [213]Institute of Clinical Medicine, Oncology, 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]Wellcome 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]Pathology Queensland, The Royal Brisbane and Women's Hospital, Brisbane 4029, Australia. [237]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [238]Center for Genomic Medicine, Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA. [239]University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [240]Department of Oncology, Södersjukhuset, Stockholm, Sweden. [241]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[2,23]. 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[1].

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[1]. 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 applied to 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 >0.4), with reference to the HapMap (v.2) populations, based on the first two principal components. After quality control exclusions and removing overlaps with the previous iCOGS and GWAS genotyping used in the analysis, the final dataset comprized data from 61,282 cases and 45,494 of European ancestry and 7,799 cases and 6,480 controls of Asian ancestry.

We excluded 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 $\frac{(p_1 - p_0)^2}{((p_1 + p_0)(2 - p_1 - p_0))} > 0.007$, where p_0 and p_1 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, using SHAPEIT2 for phasing and IMPUTE version 2 for imputation[24,25]. The imputation was performed in 5-Mb non-overlapping intervals. The subjects were split into subsets of approximately 10,000 samples; where possible subjects from the same study were included in the same subset. The BPC3 and EBCG studies were imputed separately using MACH and Minimac[26,27]. In total, 99.6% of SNPs with frequency >1% were imputable with $r^2 > 0.3$ in the OncoArray dataset and 99.1% in the iCOGS dataset. We generated estimated genotypes for all SNPs that were polymorphic (MAF > 0.1%) in either European or Asian samples (around 21 million SNPs). For the current analysis, however, we restricted our analysis to SNPs with MAF > 0.5% in the European OncoArray dataset (11.8 million SNPs). One-step imputation (without pre-phasing) was performed on 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).

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, using purpose-written software (http://ccge.medschl.cam.ac.uk/software/pccalc). For the main analyses, we used the first ten principal components, as additional components 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.

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 65 novel loci adjusting for study—the effect sizes were essentially identical to those presented. Estimates were derived using ProbAbel for the BPC3 and EBCG studies[28], SNPTEST for the remaining GWAS and purpose-written software for the iCOGS and OncoArray datasets. Odds ratio estimates and standard errors were combined in a fixed-effects, inverse-variance meta-analysis using METAL[29], adjusting the GWAS (but not iCOGS or OncoArray) results for genomic control as described previously[2]. For the GWAS, results were included in the analysis for all SNPs with MAF ≥ 0.01 and imputation $r^2 \geq 0.3$. For iCOGS and OncoArray, we included all SNPs with $r^2 \geq 0.3$ and MAF ≥ 0.005 (11.8 million SNPs in total). We used the tests that were based on the meta-analysis over all stages as the primary tests of association, because this type of analysis has been shown to be more powerful than tests that were based on a test–replication approach[30]. Eight sets of variants were associated with breast cancer at $P < 5 \times 10^{-8}$, but were close to previous susceptibility regions, and these became non-significant after adjustment for the previously identified lead variant. Two SNPs on 22q13.2, rs141447235 and rs73161324, were both associated with overall breast cancer risk but, despite lying >500 kb apart, were strongly correlated with each other ($r^2 = 0.50$) and were therefore considered a single novel signal.

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[31].

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 I^2 statistic and the P value for Cochran's Q statistic (Supplementary Table 28). 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 9).

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 public 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[32], Roadmap Epigenomics Projects[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[34,35]. 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 multiwayvcov.

eQTL analyses. eQTL analyses were performed using data from The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) projects[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 estimate cut-off >95% European)[37]. Germline genotypes were imputed 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[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[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[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[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 CTCF factors were downloaded using the UCSC Table Browser[43]. Hi-C data derived from HMECs were obtained from ref. 44, using 'interaction loops' as defined in ref. 44. Data were reformatted to facilitate intersection of query SNPs using BEDTools 'intersect'[45]. For all interactions, termini were intersected with promoters using GENCODE version 19 (ref. 46) basic gene annotations, for which we defined promoters as −1.0 kb 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[8], IM-PET enhancer–gene predictions in MCF-7, HMEC and HCC1954 cell lines[9]. Enhancer–transcription start site (E–TSS) links were identified from the FANTOM5 Consortium[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 H3K4me3 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, that gene and the corresponding CRV were assigned a point for being an eQTL in INQUISIT.

Topologically associated domain (TAD) boundaries were derived from Hi-C data[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[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) distal gene regulation, (b) proximal gene regulation or (c) impact on protein coding were calculated using the following criteria (see Supplementary Table 17).

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 also overlapped enriched cistromic transcription factor (ESR1, FOXA1, GATA3, TCF7L2, E2F1) ChIP–seq 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 0.05. 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. (c) Intragenic variants were evaluated for their potential to impact protein function using a range of *in silico* prediction tools (CADD[49], FATHMM[50], LRT[51], MutationAssessor[52], Mutation Taster 2 (ref. 53), PolyPhen-2 (ref. 54), PROVEAN[55] and SIFT[56] for missense variants; Human Splicing Finder[57] and MaxEntScan[58] 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 studies[11–14] (considering *ZNF703* and *FGFR1* 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 drivers. Second, we hypothesized that this 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 2, promoter score 3–4, distal score >4; level 2: coding score 1, promoter score 1–2, distal score 1–4; 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 analysis[59]. 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 (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 settings[15]. 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 prediction 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[60]. 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 Syto9, 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 library and, between cell lines, 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/SalI 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 constructs, which had a defined activity of 1. Statistical significance was tested by log transforming the data and performing a two-way ANOVA, followed by Dunnett's multiple comparisons test in GraphPad Prism.

Global genomic enrichment analyses. We performed stratified LD score regression analyses[20] for all breast cancers and for breast cancers stratified by ER status using the summary statistics based on the meta-analyses of the OncoArray, GWAS and iCOGS datasets. We restricted analysis to all SNPs present on the HapMap version 3 dataset that had a MAF > 1% and an imputation quality score $r^2 > 0.3$ in the OncoArray data. LD scores were calculated using the 1000 Genomes Project Phase 3 EUR reference panel.

We first created a 'full baseline model' as previously described that included 24 non-cell-type-specific, publicly available annotations as well as 24 additional annotations that included a 500-bp window around each of the 24 main annotations[20]. Additionally, we included 100-bp windows around ChIP–seq peaks and one annotation containing all SNPs, leading to a total of 53 overlapping annotations.

We subsequently performed analyses using cell-type-specific annotations for four histone markers, H3K4me1, H3K4me3, H3K9ac and H3K27ac, across 27–81 cell types depending on the histone marker[20]. Each cell-type-specific annotation corresponded to a histone marker in a single cell type, and there were 220 annotations in total. We augmented the baseline model by adding these annotations 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: $\sum_i p_i(1-p_i)(\beta_i^2-\tau_i^2)/\ln(\lambda)$ where p_i is the MAF for variant i, β_i is the log(odds ratio) estimate for variant i, τ_i is the standard error of β_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 $h_f^2 = \sum_i 2p_i(1-p_i)\beta_i^2$ where the sum is now over all variants and β_i is the true relative risk conferred by variant i, assuming a log-additive model. We refer to h_f^2 as the frailty scale heritability. We first obtained the estimated observed heritability based on the full set of summary estimates using LD Score Regression[20] and then converted this to an estimate on the frailty scale using the $h_f^2 = h_{obs}^2/P(1-P)$, 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_Genesets, file Human_GOBP_AllPathways_no_GO_iea_April_01_2017_symbol.gmt)[16] from the Bader laboratory dated 1 April 2017 was used in all analyses. This database contains pathways from Reactome[61], NCI Pathway Interaction Database[62], GO (Gene Ontology) biological process[63], HumanCyc[64], MSigdb[65], NetPath[66] and Panther[67]. 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 assigned to genes using the INQUISIT target prediction method described above for all SNPs with $P < 5 \times 10^{-2}$ (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 statistic of the most significant SNP among all SNPs assigned to a gene[68,69]. Because histone genes contained a high number of mapped SNPs, we selected representative SNP associations to avoid pathway enrichments based solely on the increased number of SNPs at these loci (that is, chr6:27657944 for HIST1; chr1:149219841 for HIST2; chr1:228517406 for HIST3; chr12:14871747 for HIST4).

The gene set enrichment analysis algorithm as implemented in the GenGen package[69] was used to perform pathway analysis. Wang *et al.*[70] 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 obtain 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 $P < 5 \times 10^{-8}$ were retained as higher confidence for subsequent analysis. These pathways, together with the genes reaching the significance threshold, are listed in Supplementary Table 22.

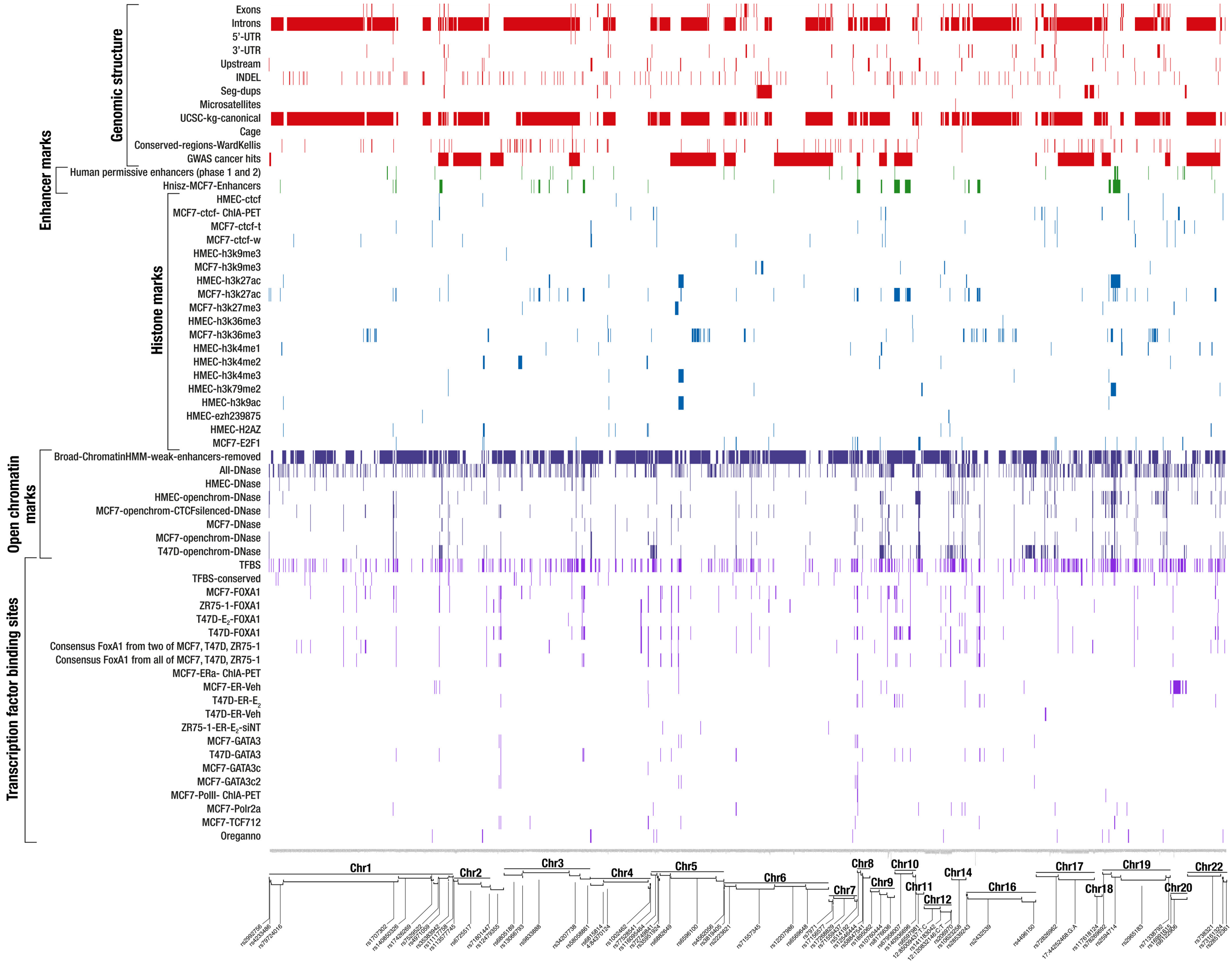
The pathway analysis assigns an enrichment score for each pathway. These values were normalized and *P* values for each pathway were obtained by separately comparing them to null distributions for OncoArray and iCOGS datasets. The null distributions were computed by permuting case/control labels 1,000 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 GWAS[71–75] 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 app[16] 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.

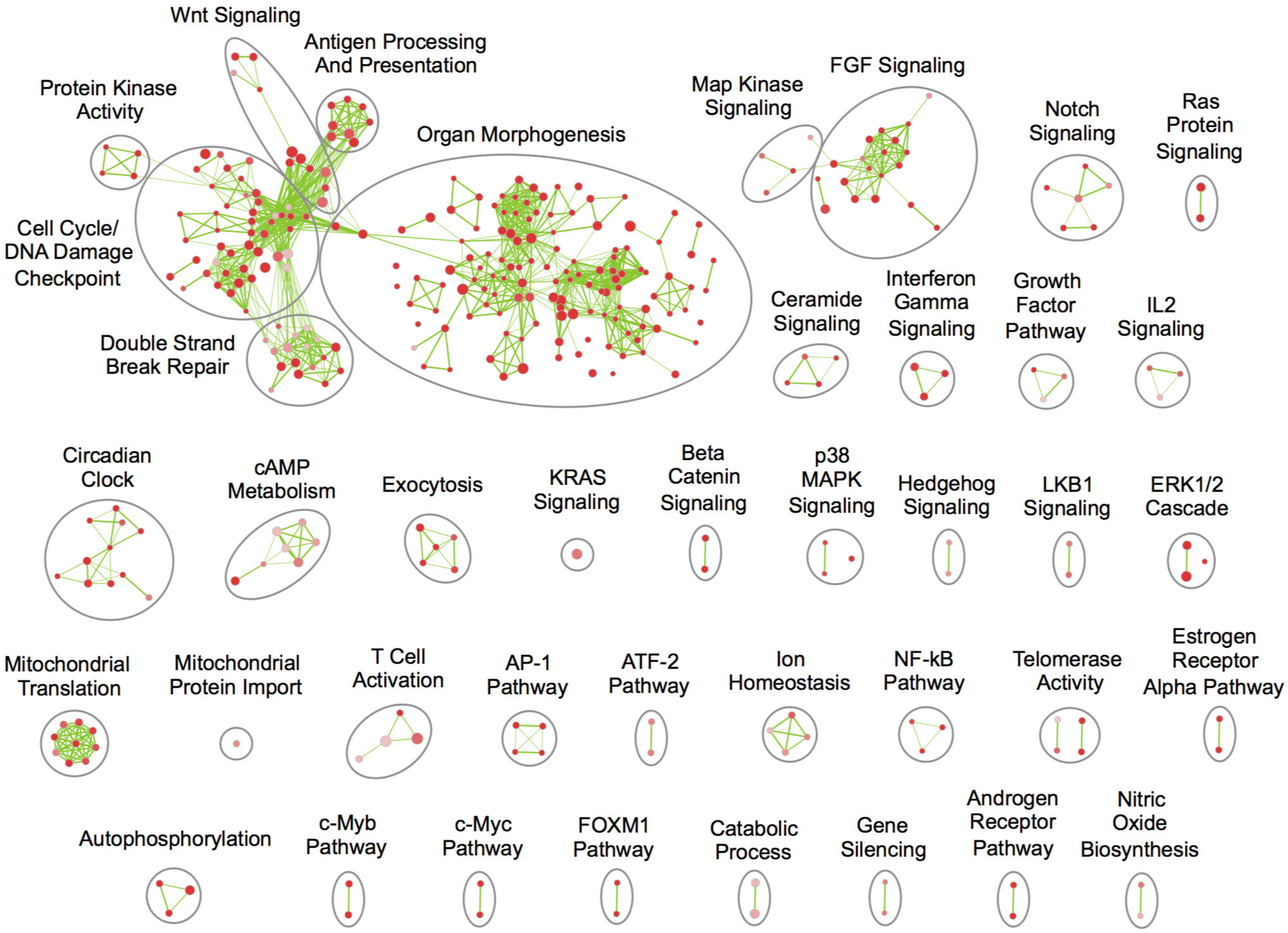
Data availability. A subset of the data that support the findings of this study is publically available via dbGaP (www.ncbi.nlm.nih.gov/gap; accession number phs001265.v1.p1). The complete dataset will not be made publicly available due to restraints imposed by the ethics committees of individual studies; requests for data can be made to the corresponding author or the Data Access Coordination Committee (DACC) of BCAC (http://bcac.ccge.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, DIETCOMPLYF, ESTHER, GC-HBOC, GENICA, GEPARSIXTO, GESBC, HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KBCP, LMBC, MABCS, MARIE, MBCSG, MCBCS, MISS, MMHS, MTLGEBCS, NC-BCFR, OFBCR, ORIGO, pKARMA, POSH, PREFACE, RBCS, SKKDKFZS, SUCCESSB, SUCCESSC, SZBCS, TNBCC, UCIBCS, UKBGS and UKOPS (see Supplementary Table 1). Summary results for all variants are available at http://bcac.ccge.medschl.cam.ac.uk/. Requests for further data should be made through the BCAC DACC (http://bcac.ccge.medschl.cam.ac.uk/).

23. Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47,** 373–380 (2015).
24. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10,** e1004234 (2014).
25. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5,** e1000529 (2009).
26. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44,** 955–959 (2012).
27. Li, Y., Willer, C. J., Ding, J., Scheet, P. & Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.* **34,** 816–834 (2010).
28. Aulchenko, Y. S., Struchalin, M. V. & van Duijn, C. M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11,** 134 (2010).
29. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26,** 2190–2191 (2010).
30. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38,** 209–213 (2006).
31. R Core Team. *R: A Language and Environment for Statistical Computing* https://www.R-project.org (2016).
32. The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.* **9,** e1001046 (2011).
33. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
34. Udler, M. S., Tyrer, J. & Easton, D. F. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet. Epidemiol.* **34,** 463–468 (2010).
35. Maller, J. B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44,** 1294–1301 (2012).
36. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486,** 346–352 (2012).
37. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28,** 1359–1367 (2012).
38. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
39. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).
40. Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12,** R41 (2011).
41. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152,** 633–641 (2013).
42. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28,** 1353–1358 (2012).
43. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32,** D493–D496 (2004).
44. Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159,** 1665–1680 (2014).
45. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
46. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22,** 1760–1774 (2012).
47. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
48. Joly Beauparlant, C. *et al. metagene* profiles analyses reveal regulatory element's factor-specific recruitment patterns. *PLOS Comput. Biol.* **12,** e1004751 (2016).
49. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
50. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34,** 57–65 (2013).
51. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19,** 1553–1561 (2009).
52. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39,** e118 (2011).
53. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11,** 361–362 (2014).
54. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
55. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS ONE* **7,** e46688 (2012).
56. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
57. Desmet, F. O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37,** e67 (2009).
58. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11,** 377–394 (2004).
59. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP–trait associations. *Nucleic Acids Res.* **42,** D1001–D1006 (2014).

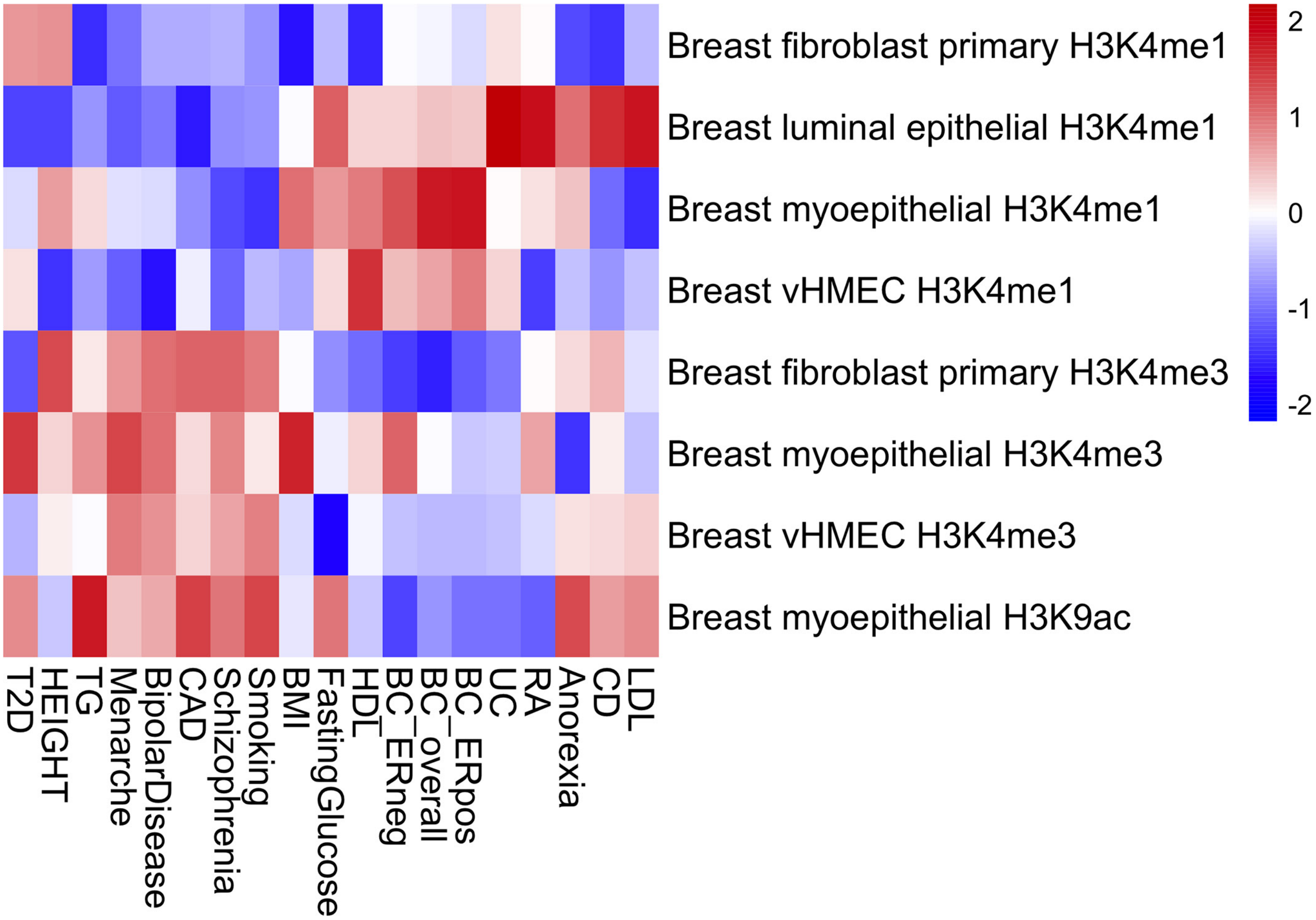
60. Ghoussaini, M. *et al.* Evidence that breast cancer risk at the 2q35 locus is mediated through *IGFBP5* regulation. *Nat. Commun.* **4,** 4999 (2014).
61. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*. **33,** D428–D432 (2005).
62. Schaefer, C. F. *et 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 genomewide 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).



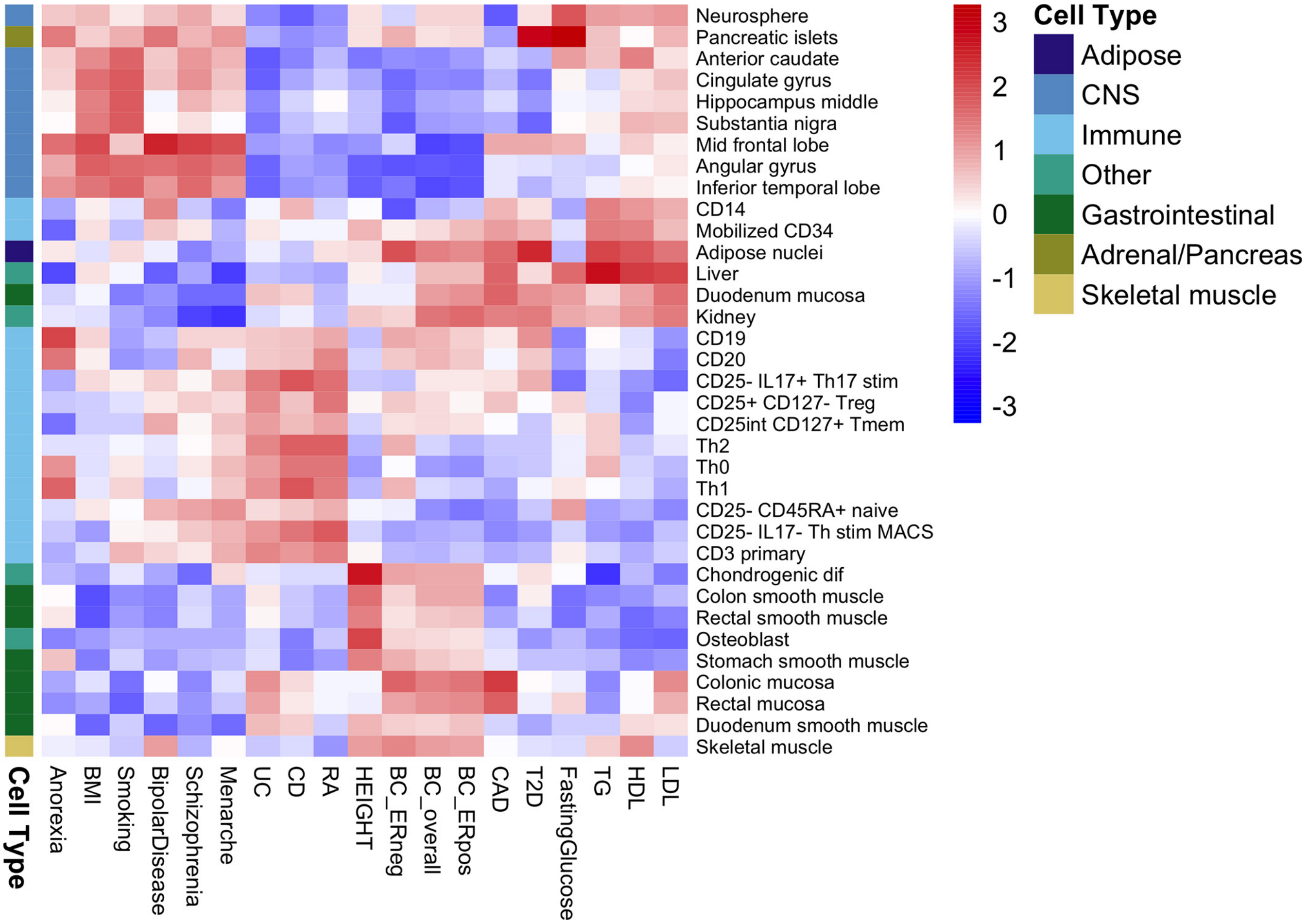
Extended Data Figure 1 | Global mapping of biofeatures across novel loci associated with overall breast cancer risk. The overlaps between potential genomic predictors in relevant breast cell lines and credible risk variants (CRVs) within each locus. On the *x* axis, each column represents a CRV (see 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).



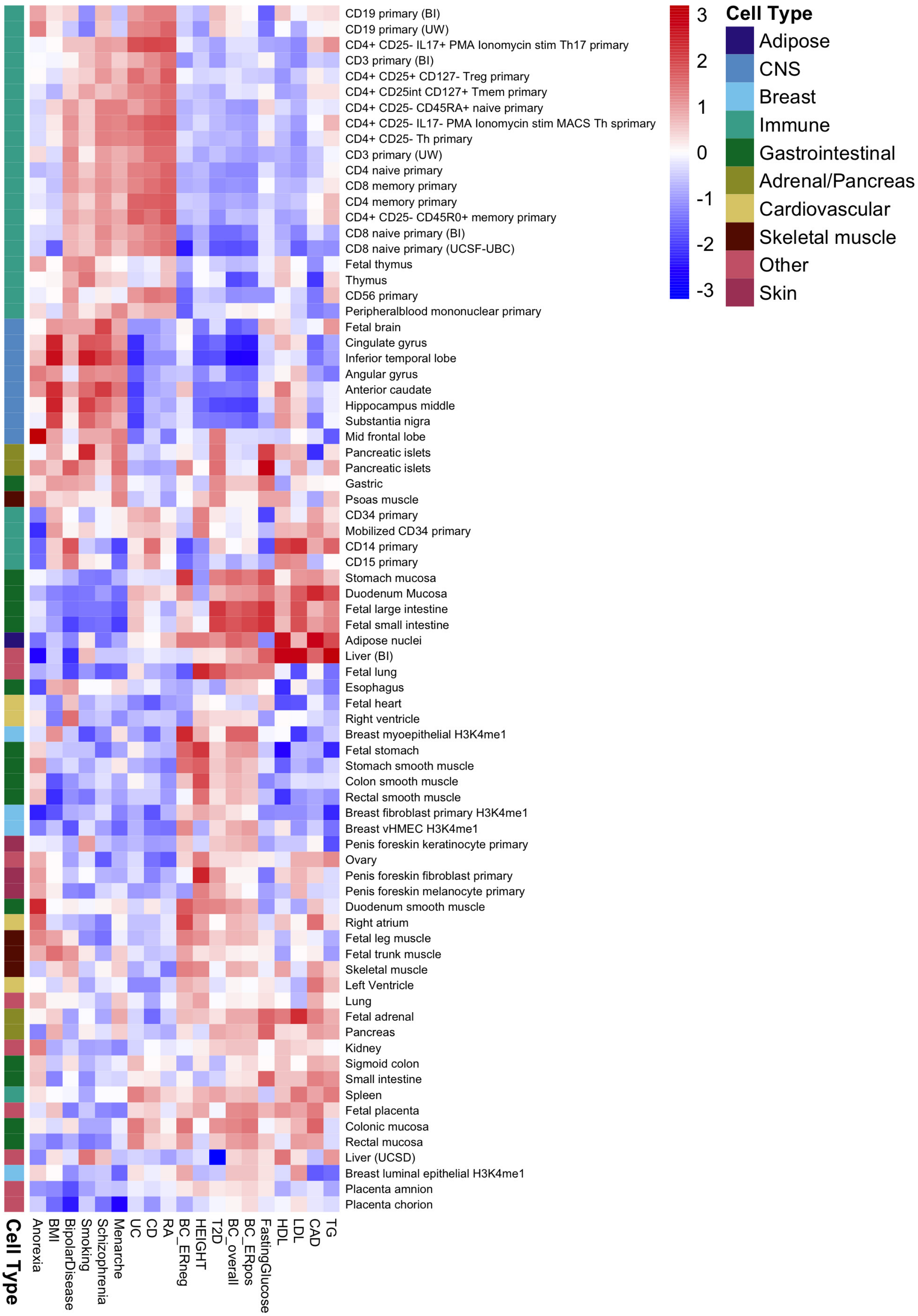
Extended Data Figure 2 | Pathway enrichment map for susceptibility loci based on summary association statistics. Each coloured circle (node) represents a pathway (gene set), coloured by enrichment score where redder nodes indicate lower FDRs. Larger nodes indicate pathways with more genes. Green lines connect pathways with overlapping genes (minimum overlap 0.55). Pathways are grouped by similarity and organized into major themes (large labelled circles).



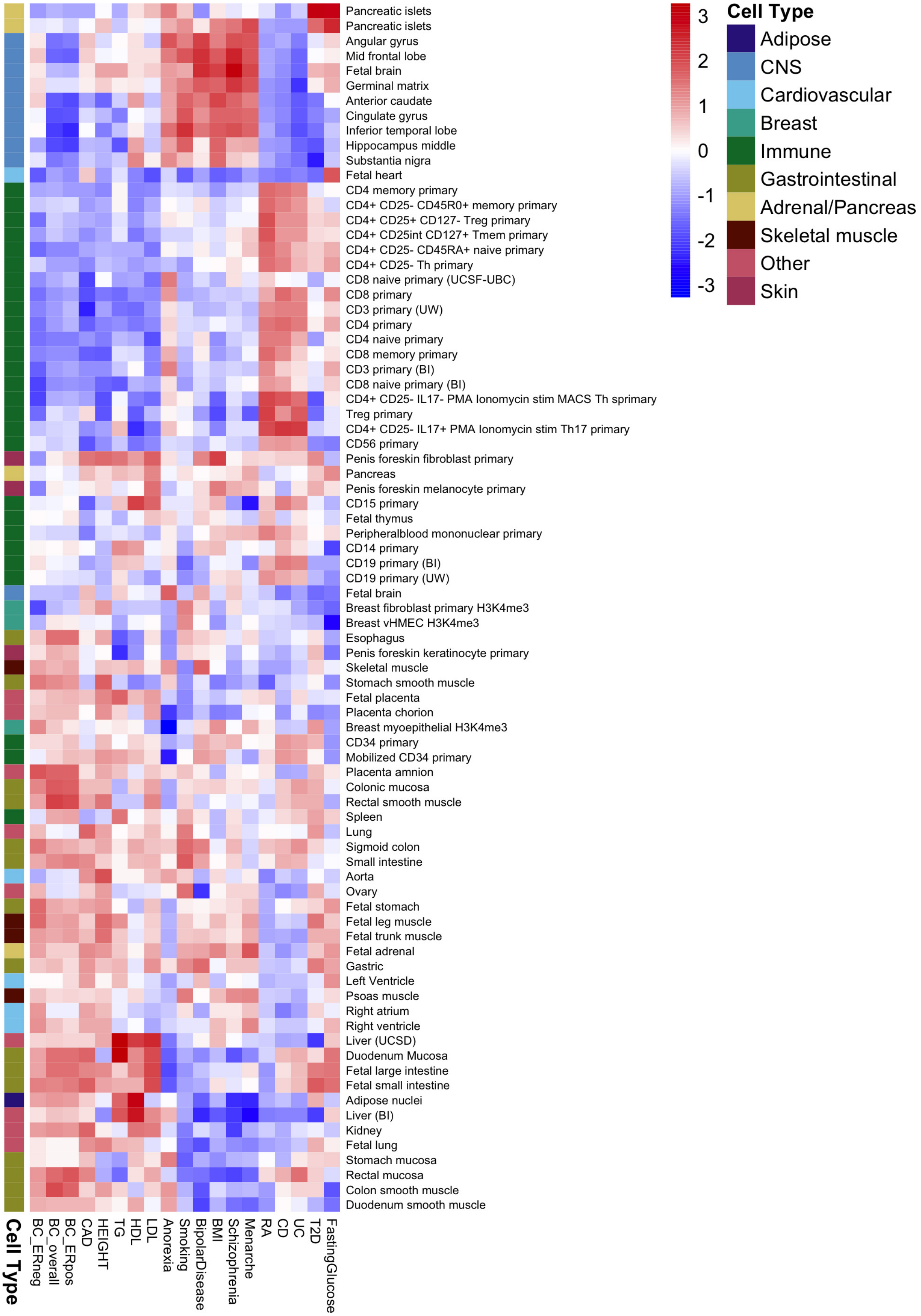
Extended Data Figure 3 | Heat map showing patterns of cell-type-specific enrichments for breast tissue across three histone marks (H3K4me1, H3K4me3 and H3K9ac) for all breast cancer types, ER-positive breast cancer and ER-negative breast cancer as well as 16 other traits. BC_ERneg, ER-negative breast 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.



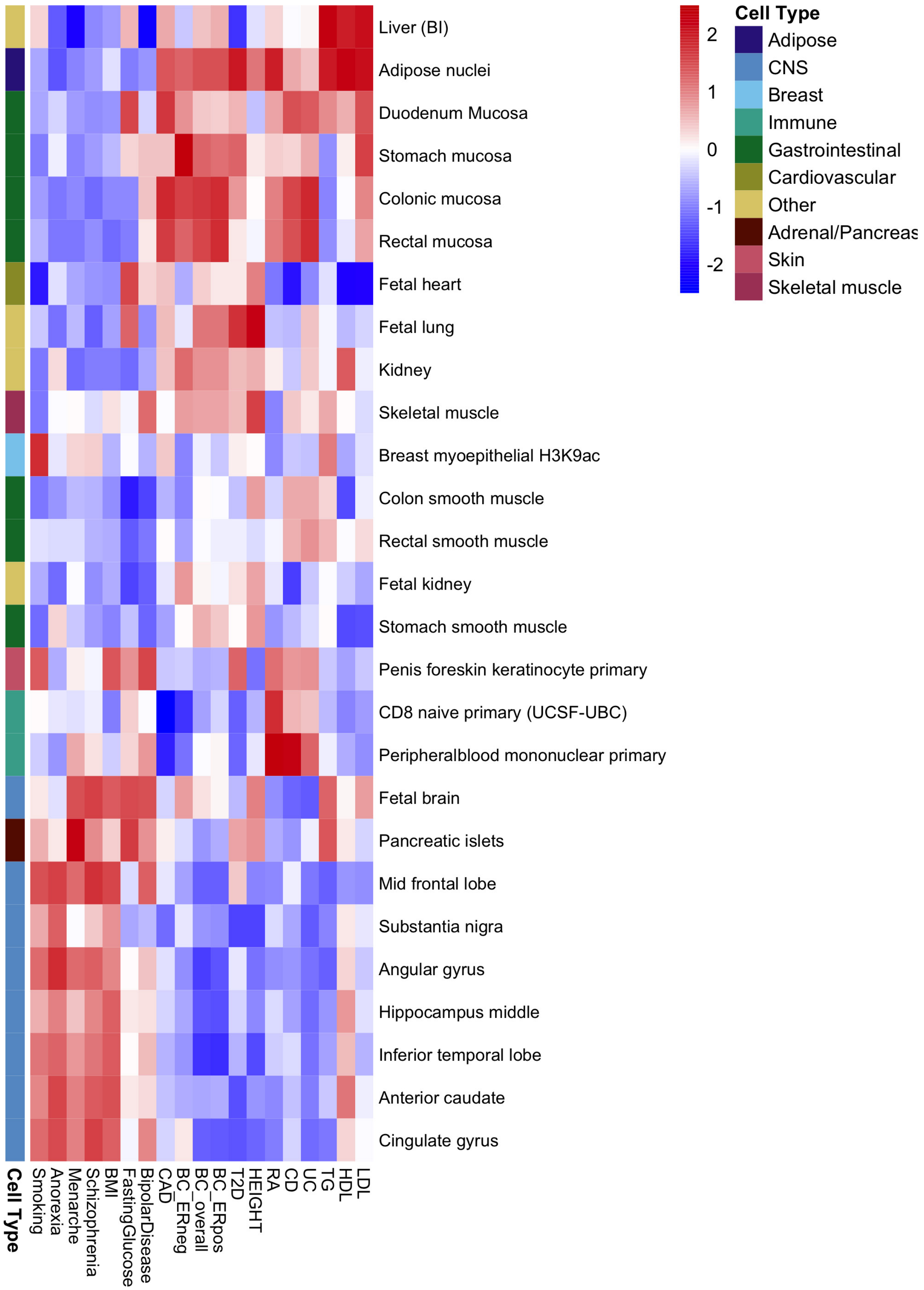
Extended Data Figure 4 | Heat map showing patterns of cell-type-specific enrichments for histone mark H3K27ac in all breast cancer types, ER-positive and ER-negative breast cancer as well as 16 other traits.



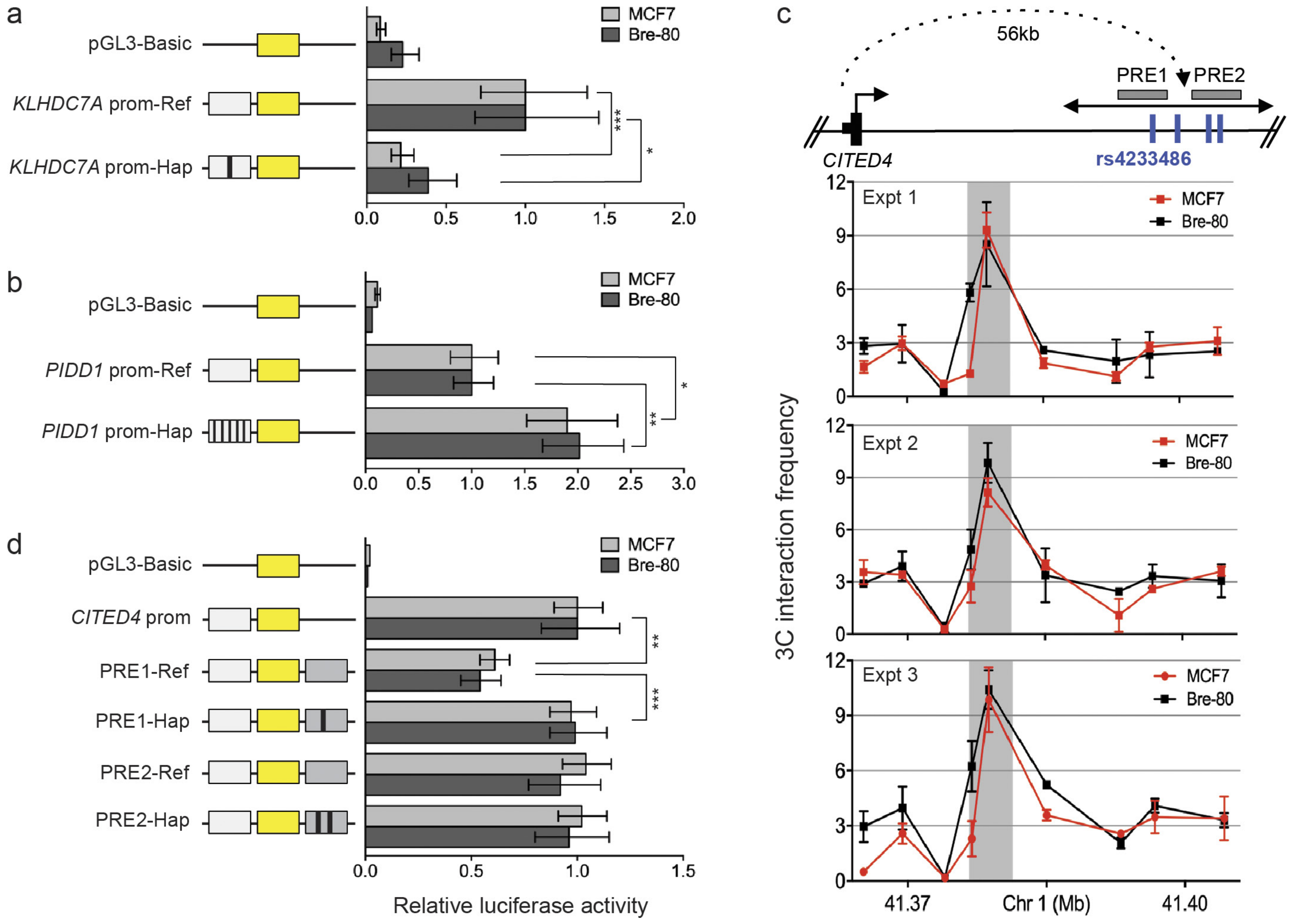
Extended Data Figure 5 | Heat map showing patterns of cell-type-specific enrichments for histone mark H3K4me1 in all breast cancer types, ER-positive and ER-negative breast cancer as well as 16 other traits.



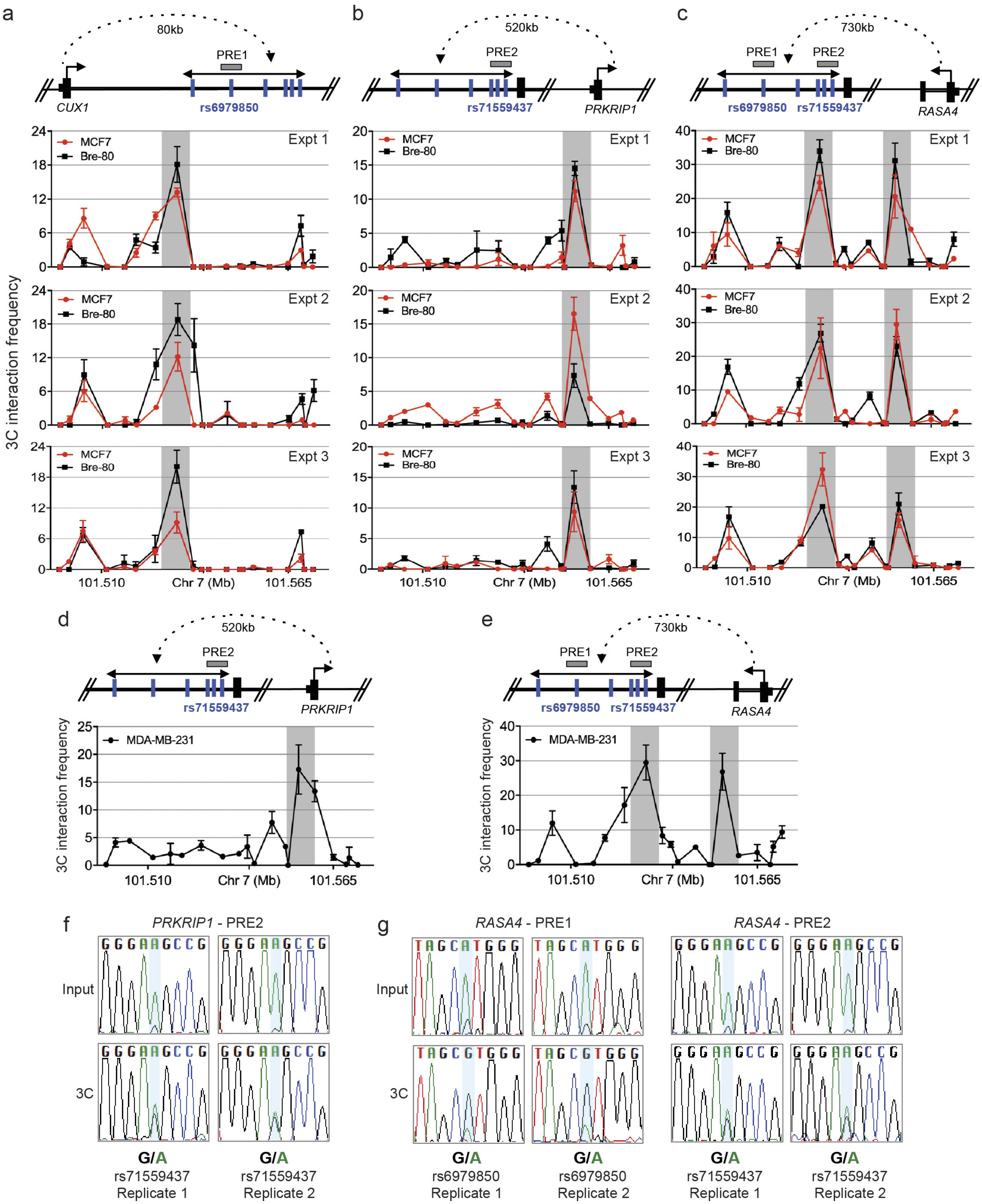
Extended Data Figure 6 | Heat map showing patterns of cell-type-specific enrichments for histone mark H3K4me3 in breast cancer overall, ER+ and ER- breast cancer as well as 16 other traits.



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 *KLHDC7A* (**a**) or *PIDD1* (**b**) promoter regions, containing the reference (prom-Ref) or risk alleles (prom-Hap), were cloned upstream of the 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 (*$P < 0.05$, **$P < 0.01$, ***$P < 0.001$). **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 *CITED4* promoter-driven luciferase construct (*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**), *PRKRIP1* (**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 *PRKRIP1*-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).

nature research

Corresponding author(s): Douglas Easton

□ Initial submission ☐ Revised version ☒ Final submission

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.

▸ Experimental design

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.

Software

Policy information about availability of computer code

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.

Materials and reagents

Policy information about availability of materials

8. Materials availability

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

a. State the source of each eukaryotic cell line used.	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.
b. Describe the method of cell line authentication used.	All cell lines were short tandem repeat (STR) profiled.
c. Report whether the cell lines were tested for mycoplasma contamination.	All cell lines tested negative for mycoplasma contamination.
d. 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.	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.