

A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer

The breast cancer risk variants identified in genome-wide association studies explain only a small fraction of the familial relative risk, and the genes responsible for these associations remain largely unknown. To identify novel risk loci and likely causal genes, we performed a transcriptome-wide association study evaluating associations of genetically predicted gene expression with breast cancer risk in 122,977 cases and 105,974 controls of European ancestry. We used data from the Genotype-Tissue Expression Project to establish genetic models to predict gene expression in breast tissue and evaluated model performance using data from The Cancer Genome Atlas. Of the 8,597 genes evaluated, significant associations were identified for 48 at a Bonferroni-corrected threshold of $P < 5.82 \times 10^{-6}$, including 14 genes at loci not yet reported for breast cancer. We silenced 13 genes and showed an effect for 11 on cell proliferation and/or colony-forming efficiency. Our study provides new insights into breast cancer genetics and biology.

Breast cancer is the most common malignancy among women in many countries¹. Genetic factors play an important role in its etiology^{2–5}. Since 2007, genome-wide association studies (GWAS) have identified approximately 170 genetic loci harboring common, low-penetrance variants for breast cancer^{6–13}, but these variants explain less than 20% of familial relative risk⁷. Most disease-associated risk variants identified by GWAS are located in non-protein-coding regions and are not in linkage disequilibrium with any nonsynonymous coding SNPs¹⁴. Many of these susceptibility variants are located in gene regulatory elements^{15,16}, and it has been hypothesized that many GWAS-identified associations may be driven by the regulatory function of risk variants on the expression of nearby genes. For breast cancer, recent studies have already shown that GWAS-identified associations at more than 15 loci are likely due to the effect of risk variants at these loci on regulating the expression of either nearby or more distal genes^{7,9,10,13,17–22}. However, for the large majority of the GWAS-identified breast cancer risk loci, the genes responsible for the associations remain unknown.

Several studies have reported that regulatory variants may account for a large proportion of disease heritability not yet discovered through GWAS^{23–25}. Many of these variants may have a small effect size, and thus are difficult to identify in individual SNP-based GWAS, even with a large sample size. Applying gene-based approaches that aggregate the effects of multiple variants into a single testing unit may increase study power to identify novel disease-associated loci. Transcriptome-wide association studies (TWAS) systematically investigate the association of genetically predicted gene expression with disease risk, providing an effective approach to identify novel susceptibility genes^{26–29}. A recently performed TWAS including 15,440 cases and 31,159 controls reported significant associations for 5 genes with breast cancer risk³⁰. However, the sample size of that study was relatively small and several reported associations were not significant after Bonferroni correction. Herein, we report results from a larger TWAS of breast cancer that used the MetaXcan method²⁶ to analyze summary statistics data from 122,977 cases and 105,974 controls of European descent from the Breast Cancer Association Consortium (BCAC).

Results

Gene expression prediction models. The study design is shown in Supplementary Fig. 1. We used transcriptome and genotyping data

from 67 women of European descent included in the Genotype-Tissue Expression (GTEx) project to build genetic models to predict RNA expression levels for each gene expressed in normal breast tissues, by applying the elastic net method ($\alpha = 0.5$) with tenfold cross-validation. Genetically regulated expression was estimated using variants within a 2 megabase (Mb) window flanking the respective gene boundaries, inclusive. SNPs with a minor allele frequency of at least 0.05 and included in the HapMap Phase 2 were used for model building. Of the models built for 12,696 genes, 9,109 showed a prediction performance (R^2) of at least 0.01 ($\geq 10\%$ correlation between predicted and observed expression). For genes for which the expression could not be predicted well using this approach, we built models using only SNPs located in the promoter or enhancer regions, as predicted using three breast cell lines in the Roadmap Epigenomics Project/Encyclopedia of DNA Elements Project. This approach leverages information from functional genomics and reduces the number of variants for variable selection, therefore potentially improving statistical power. This enabled us to build genetic models for an additional 3,715 genes with $R^2 \geq 0.01$. Supplementary Table 1 provides detailed information regarding the performance threshold and types of model built. Overall, genes that were predicted with $R^2 \geq 0.01$ in GTEx data were also predicted well in The Cancer Genome Atlas (TCGA) tumor-adjacent normal tissue data (correlation coefficient of 0.55 for R^2 in two data sets; Supplementary Fig. 2). On the basis of model performance in GTEx and TCGA, we prioritized 8,597 genes for analyses of the associations between predicted gene expression and breast cancer risk using the following criteria: genes with a model prediction $R^2 \geq 0.01$ in the GTEx set (10% correlation) and a Spearman's correlation coefficient of ≥ 0.1 in the external validation experiment; genes with a prediction $R^2 \geq 0.09$ (30% correlation) in the GTEx set regardless of their performance in the TCGA set; genes with a prediction $R^2 \geq 0.01$ in the GTEx set (10% correlation) that could not be evaluated in the TCGA set because of a lack of data.

Associations of predicted expression with breast cancer. Using the MetaXcan method²⁶, we performed association analyses to evaluate predicted gene expression and breast cancer risk using the meta-analysis summary statistics of SNPs generated for 122,977 cases and 105,974 controls of European ancestry included in BCAC. For the

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

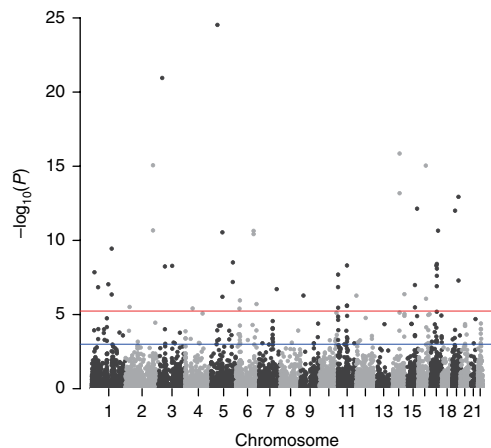


Fig. 1 | A Manhattan plot of the association results from the breast cancer transcriptome-wide association study. The results are based on 122,977 cases and 105,974 controls. The red line represents $P = 5.82 \times 10^{-6}$. The blue line represents $P = 1.00 \times 10^{-3}$.

majority of the tested genes, most of the SNPs selected for prediction models were used for the association analyses (for example, $\geq 80\%$ predicting SNPs used for 95.6% of the tested genes). Lambda 1,000 ($\lambda_{1,000}$), a standardized estimate of the genomic inflation scaling to a study of 1,000 cases and 1,000 controls, was 1.004 in our study (quantile–quantile (QQ) plot presented in Supplementary Fig. 3a). Of the 8,597 genes evaluated, we identified 179 whose predicted expression was associated with breast cancer risk at $P < 1.05 \times 10^{-3}$, a false discovery rate (FDR)-corrected significance level (Fig. 1 and Supplementary Table 2). Of these, 48 showed a significant association at the Bonferroni-corrected threshold of $P \leq 5.82 \times 10^{-6}$ (Fig. 1 and Tables 1–3), including 14 genes located at 11 loci that are 500 kb away from any risk variant identified in previous GWAS (Table 1). An association between lower predicted expression and increased breast cancer risk was detected for *LRRC3B* (3p24.1), *SPATA18* (4q12), *UBD* (6p22.1), *MIR31HG* (9p21.3), *RIC8A* (11p15.5), *B3GNT1* (11q13.2), *GALNT16* (14q24.1) and *MAN2C1* and *CTD-2323K18.1* (15q24.2). Conversely, an association between higher predicted expression and increased breast cancer risk was identified for *ZSWIM5* (1p34.1), *KLHDC10* (7q32.2), *RP11-867G23.10* (11q13.2), *RP11-218M22.1* (12p13.33) and *PLEKHD1* (14q24.1). The remaining 34 associated genes are located at known breast cancer susceptibility loci (Tables 2 and 3). Among them, 23 have not yet been implicated as genes responsible for association signals identified at these loci through expression quantitative trait loci (eQTL) and/or functional studies, and do not harbor GWAS- or fine-mapping-identified risk variants (Table 2), while the other 11 (*KLHDC7A*⁷, *ALS2CR12*³¹, *CASP8*^{31,32}, *ATG10*⁹, *SNX32*³³, *STXBP4*^{34,35}, *ZNF404*⁸, *ATP6AP1L*⁹, *RMND1*¹⁷, *L3MBTL3*³⁶ and *RCCD1*¹⁰) have been reported as potential causal genes at breast cancer susceptibility loci or harbor GWAS- or fine-mapping-identified risk variants (Table 3). Except for *RP11-73O6.3* and *L3MBTL3*, there was no evidence of heterogeneity ($P < 0.2$) across the iCOGS, OncoArray and GWAS data sets included in our analyses (Supplementary Table 3). Overall, we identified 37 novel susceptibility genes for breast cancer and confirmed 11 genes known to potentially play a role in breast cancer susceptibility.

To determine whether the associations between predicted gene expression and breast cancer risk were independent of GWAS-identified association signals, we performed conditional analyses adjusting for the GWAS-identified risk SNPs closest to the TWAS-identified gene (Supplementary Table 4)³⁶. We found that the associations for 11 genes (*LRRC3B*, *SPATA18*, *KLHDC10*, *MIR31HG*,

RIC8A, *B3GNT1*, *RP11-218M22.1*, *MAN2C1*, *CTD-2323K18.1* (Table 1), *ALK*, *CTD-3051D23.1* (Table 2)) remained statistically significant at $P < 5.82 \times 10^{-6}$ (Tables 1–3). This suggests that the expression of these genes may be associated with breast cancer risk independent of the GWAS-identified risk variant(s). For nine of the genes (*SPATA18*, *KLHDC10*, *MIR31HG*, *RIC8A*, *RP11-218M22.1*, *MAN2C1*, *CTD-2323K18.1* (Table 1), *ALK* and *CTD-3051D23.1* (Table 2)), the significance of the association remained essentially unchanged, suggesting that these associations may be entirely independent of GWAS-identified association signals.

Of the 131 genes showing an association at $5.82 \times 10^{-6} < P < 1.05 \times 10^{-3}$ (significant after FDR correction but not Bonferroni correction), 38 are located at GWAS-identified risk loci (Table 4). Except for *RP11-400F19.8*, there was no evidence of heterogeneity in TWAS association ($P < 0.2$) across the iCOGS, OncoArray and GWAS studies (Supplementary Table 3). After adjusting for the risk SNPs, associations for *MTHFD1L*, *PVT1*, *RP11-123K19.1*, *FES*, *RP11-400F19.8*, *CTD-2538G9.5* and *CTD-3216D2.5* remained significant at $P \leq 1.05 \times 10^{-3}$, again suggesting that the association of these genes with breast cancer risk may be independent of the GWAS-identified association signals (Table 4).

For 41 of the 48 associated genes that reached the Bonferroni-corrected significance level, we obtained individual-level data from subjects included in the iCOGS ($n = 84,740$) and OncoArray ($n = 112,133$) data sets, which was 86% of the subjects included in the analysis using summary statistics (Supplementary Table 5). The results from the analysis using individual-level data were very similar to those described above using MetaXcan analyses (Pearson correlation of Z scores was 0.991 for iCOGS data and 0.994 for OncoArray data), although not all associations reached the Bonferroni-corrected significance level, possibly due to a smaller sample size (Supplementary Table 5). Conditional analyses using individual-level data also revealed consistent results compared with analyses using summary data. We found that for several genes within the same genomic region, their predicted expression was correlated with each other (Tables 1–3). The associations between predicted expression of *PLEKHD1* and *ZSWIM5* and breast cancer risk were largely influenced by their corresponding closest risk variants identified in GWAS, although these risk variants are > 500 kb away from these genes (Table 1). There were significant correlations of rs999737 and rs1707302 with genetically predicted expression of *PLEKHD1* ($r = -0.47$ in the OncoArray data set and -0.48 in the iCOGS data set) and *ZSWIM5* ($r = 0.50$ in the OncoArray data set and 0.51 in the iCOGS data set), respectively.

INQUISIT algorithm scores. For the 48 associated genes after Bonferroni correction, we assessed their integrated expression quantitative trait and in silico prediction of GWAS target (INQUISIT) scores⁷ to assess whether there is other evidence beyond the scope of eQTL for supporting our TWAS-identified genes as candidate target genes at GWAS-identified loci. The detailed methodology for INQUISIT scores have been described elsewhere⁷. In brief, a score for each gene–SNP pair is calculated across categories representing potential regulatory mechanisms—distal or proximal gene regulation (promoter). Features contributing to the score are based on functionally important genomic annotations such as chromatin interactions, transcription factor binding and eQTLs. Compared with evidence from eQTL alone, INQUISIT scores incorporate additional lines of evidence, including distal regulations. The INQUISIT scores for our identified genes are shown in Supplementary Table 6. Except for *UBD* with a very low score in the distal regulation category (0.05), none of the genes at novel loci (Table 1) showed evidence of being potential target genes for GWAS-identified breast cancer susceptibility loci. This is interesting and within the expectation since these genes may represent novel association signals. There was evidence

Table 1 | Fourteen expression-trait associations for genes located at genomic loci at least 500 kb away from any GWAS-identified breast cancer risk variants

Region	Gene ^a	Type ^b	Z score	P value ^c	R ^{2c}	Closest risk SNP ^d	Distance to the closest risk SNP (kb)	P value after adjusting for adjacent risk SNPs ^e
1p34.1	ZSWIM5	Protein	5.26	1.43×10^{-7}	0.17	rs1707302	829	0.006
3p24.1	<i>LRRC3B</i>	Protein	-9.57	1.11×10^{-21}	0.17	rs653465	591	1.60×10^{-6}
4q12	<i>SPATA18</i>	Protein	-4.62	3.86×10^{-6}	0.11	rs6815814	14,101	3.98×10^{-6}
6p22.1	<i>UBD</i>	Protein	-4.87	1.10×10^{-6}	0.13	rs9257408	597	0.94
7q32.2	KLHDC10	Protein	5.21	1.92×10^{-7}	0.14	rs4593472	892	2.90×10^{-7}
9p21.3	<i>MIR31HG</i>	lncRNA	-5.02	5.22×10^{-7}	0.12	rs1011970	502	1.23×10^{-7}
11p15.5	<i>RIC8A</i>	Protein	-5.27	1.40×10^{-7}	0.15	rs6597981	588	4.95×10^{-6}
11q13.2	<i>B3GNT1</i>	Protein	-5.85	4.88×10^{-9}	0.09	rs3903072	530	3.50×10^{-6}
11q13.2	<i>RP11-867G23.10</i>	Transcript	4.71	2.49×10^{-6}	0.03	rs3903072	594	2.61×10^{-4}
12p13.33	RP11-218M22.1	lncRNA	5.02	5.27×10^{-7}	0.19	rs12422552	13,641	5.17×10^{-7}
14q24.1	<i>GALNT16</i>	Protein	-8.27	1.38×10^{-16}	0.04	rs999737	691	8.57×10^{-4}
14q24.1	PLEKHD1	Protein	7.50	6.55×10^{-14}	0.02	rs999737	917	0.12
15q24.2	<i>MAN2C1</i> ^f	Protein	-5.32	1.02×10^{-7}	0.39	rs2290203	15,851	9.56×10^{-8}
15q24.2	<i>CTD-2323K18.1</i> ^f	lncRNA	-4.65	3.27×10^{-6}	0.07	rs2290203	15,619	3.16×10^{-6}

^aGenes that were siRNA-silenced for functional assays are shown in bold; SNPs used to predict gene expression are listed in Supplementary Table 13. ^bProtein: protein-coding genes; lncRNA: long non-coding RNAs; transcript: processed transcript. ^cP value: derived from association analyses of 122,977 cases and 105,974 controls; associations with $P \leq 5.82 \times 10^{-6}$ are considered statistically significant on the basis of Bonferroni correction of 8,597 tests (0.05/8,597); ^dR²: prediction performance (R²) derived using GTEx data. ^eRisk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes, are presented in Supplementary Table 4. ^fUse of the COJO method³⁶. ^gPredicted expression of *MAN2C1* and *CTD-2323K18.1* was correlated (Spearman $R=0.76$).

suggesting that *RP11-439A17.7*, *NUDT17*, *ANKRD34A*, *BTN3A2*, *AP006621.6*, *RPLP2*, *LRRC37A2*, *LRRC37A*, *KANSL1-AS1*, *CRHR1* and *HAPLN4* listed in Table 2, and all 11 genes listed in Table 3, may be target genes for risk variants at these loci (Supplementary Table 6). For *NUDT17*, *ANKRD34A*, *RPLP2*, *LRRC37A2*, *LRRC37A*, *KANSL1-AS1*, *CRHR1*, *HAPLN4*, *KLHDC7A*, *ALS2CR12*, *CASP8*, *ATG10*, *ATP6AP1L*, *L3MBTL3*, *RMND1*, *SNX32*, *RCCD1*, *STXBP4* and *ZNF404*, the INQUISIT scores were not derived only from eQTL data, providing orthogonal support for these genes. For these loci, the associations of candidate causal SNPs with breast cancer risk may be mediated through these genes. This is in general consistent with the findings from the conditional analyses.

Pathway enrichment analyses. Ingenuity Pathway Analysis (IPA)³⁷ suggested potential enrichment of cancer-related functions for the identified protein-coding genes (Supplementary Table 7). The top canonical pathways identified included apoptosis-related pathways (granzyme B signaling ($P=0.024$) and cytotoxic T-lymphocyte-mediated apoptosis of target cells ($P=0.046$)), immune system pathways (inflammation pathway ($P=0.030$)) and tumoricidal function of hepatic natural killer cells ($P=0.036$). The identified pathways are largely consistent with previous findings⁷. For the associated long non-coding RNAs (lncRNAs), pathway analysis of their highly co-expressed protein-coding genes also revealed potential over-representation of cancer-related functions (Supplementary Table 7).

In vitro assays of gene functions. To assess the function of genes whose high predicted expression levels were associated with increased breast cancer risk, we selected 13 genes for knockdown experiments in breast cells: *ZSWIM5*, *KLHDC10*, *RP11-218M22.1* and *PLEKHD1* (Table 1), *UBLCP1*, *AP006621.6*, *RP11-467J12.4*, *CTD-3032H12.1* and *RP11-15A1.7* (Table 2), and *ALS2CR12*, *RMND1*, *STXBP4* and *ZNF404* (Table 3). As negative controls, we selected *B2M*, *ARHGDI1A* and *ZAP70* using the criteria: ≥ 2 Mb from any known breast cancer risk locus; not an essential gene in

breast cancer^{38,39}; and not predicted to be a target gene in INQUISIT. In addition, as positive controls, we included *PIDD1* (Table 4)⁷, *NRBF2*²⁰ and *ABHD8*²², which have been functionally validated as target genes at breast cancer risk loci. We performed quantitative PCR (qPCR) on a panel of three 'normal' mammary epithelial and 15 breast cancer cell lines to analyze their expression levels (Supplementary Fig. 4 and Supplementary Table 8). All 19 genes were expressed in the normal mammary epithelial line 184A1⁴⁰ and the luminal breast cancer cell lines, MCF7 and T47D, so we used these cell lines for the proliferation assay, and MCF7 for the colony-formation assay⁴¹. We also evaluated *SNX32*, *ALK* and *BTN3A2* by qPCR, but they were not expressed in T47D and MCF7 cells; therefore, they were not evaluated further. It was difficult to design siRNAs against *RP11-867G23.1* and *RP11-530I9.1* because they both have multiple transcripts with limited, GC-rich regions in common. We did not include *RPLP2* because it is already known to be an essential gene for breast cancer survival³⁹. Knockdown of the 19 tested genes was achieved by siRNA (Supplementary Table 9) and the knockdown efficiency was calculated in 184A1, MCF7 and T47D for each siRNA pair. Robust knockdown of the gene of interests (GOI) was validated by qPCR with the majority of the siRNAs (Supplementary Fig. 5).

To evaluate the survival and proliferation ability of cells following gene interruption, we used an IncuCyte to quantify cell proliferation in real time and quantified the corrected proliferation of cells with knockdown of GOI in comparison to that of cells with non-target control (NTC) siRNA. As expected, knockdown of the three negative control genes (*B2M*, *ARHGDI1A* and *ZAP70*) did not significantly change cell proliferation in any of the three cell lines (Fig. 2a and Supplementary Fig. 6). However, with the exception of *UBLCP1*, *RMND1* and *STXBP4*, knockdown of all other genes (11 TWAS-identified genes along with 2 known genes, *ABHD8* and *NRBF2*) resulted in significantly decreased cell proliferation in 184A1 normal breast cells, with *KLHDC10*, *PLEKHD1*, *RP11-218M22.1*, *AP006621.6*, *ZNF404*, *RP11-467J12.4*, *CTD-3032H12.1* and *STXBP4* showing a similar effect in one or both cancer cell lines.

Table 2 | Twenty-three expression-trait associations for genes located at genomic loci within 500 kb of any previous GWAS-identified breast cancer risk variants but not yet implicated as target genes of risk variants

Region	Gene ^a	Type ^b	Z score	P value ^c	R ^{2c}	Closest risk SNP ^d	Distance to the closest risk SNP (kb)	P value after adjusting for adjacent risk SNPs ^e
1p11.2	<i>RP11-439A17.7</i>	lncRNA	-5.34	9.07×10^{-8}	0.22	rs11249433	442	0.02
1q21.1	<i>NUDT17</i>	Protein	-6.27	3.58×10^{-10}	0.01	rs12405132	56	0.08
1q21.1	<i>ANKRD34A</i>	Protein	-5.05	4.42×10^{-7}	0.01	rs12405132	169	4.28×10^{-5}
2p23.1-2p23.2	<i>ALK</i>	Protein	4.67	3.06×10^{-6}	0.06	rs4577244	295	2.70×10^{-6}
3p21.31	<i>PRSS46</i>	Protein	-5.83	5.68×10^{-9}	0.13	rs6796502	89	0.002
3q12.2	<i>RP11-114I8.4</i>	lncRNA	-5.84	5.19×10^{-9}	0.02	rs9833888	356	0.09
5p12	<i>RP11-53O19.1</i>	lncRNA	10.38	2.94×10^{-25}	0.03	rs10941679	39	7.46×10^{-4}
5q33.3	<i>UBLCP1</i>	Protein	5.93	3.04×10^{-9}	0.07	rs1432679	446	0.37
5q33.3	<i>RP11-32D16.1</i>	lncRNA	-5.41	6.37×10^{-8}	0.09	rs1432679	283	1.32×10^{-4}
6p22.2	<i>BTN3A2</i>	Protein	4.61	3.97×10^{-6}	0.28	rs71557345	229	0.72
6q23.1	<i>RP11-73O6.3^f</i>	lncRNA	-6.61	3.74×10^{-11}	0.11	rs6569648	105	0.41
11p15.5	<i>AP006621.6^g</i>	lncRNA	5.61	2.01×10^{-8}	0.34	rs6597981	21	0.52
11p15.5	<i>RPLP2^h</i>	Protein	4.64	3.46×10^{-6}	0.27	rs6597981	7	0.51
14q32.33	<i>CTD-305I23.1</i>	lncRNA	-5.06	4.21×10^{-7}	0.05	rs10623258	97	7.05×10^{-7}
16q12.2	<i>RP11-467J12.4</i>	lncRNA	8.04	9.02×10^{-16}	0.23	rs3112612	434	0.79
16q12.2	<i>CTD-3032H12.1</i>	lncRNA	4.92	8.58×10^{-7}	0.03	rs28539243	290	0.006
17q21.31	<i>LRRC37A^g</i>	Protein	-5.89	3.85×10^{-9}	0.43	rs2532263	118	0.79
17q21.31	<i>KANSL1-AS1^g</i>	lncRNA	-5.58	2.44×10^{-8}	0.62	rs2532263	18	0.95
17q21.31	<i>CRHR1^g</i>	Protein	-5.29	1.22×10^{-7}	0.22	rs2532263	339	0.99
17q21.31	<i>LINC00671</i>	lncRNA	-5.85	4.95×10^{-9}	0.07	rs72826962	190	0.26
17q21.31	<i>LRRC37A2</i>	Protein	-5.77	7.93×10^{-9}	0.46	rs2532263	336	0.93
19p13.11	<i>HAPLN4</i>	Protein	-7.13	9.88×10^{-13}	0.02	rs2965183	172	0.22
19q13.31	<i>RP11-15A1.7^h</i>	lncRNA	5.45	5.06×10^{-8}	0.02	rs3760982	215	0.28

Genes presented in table have not been previously reported from eQTL and/or functional studies as target genes of GWAS-identified risk variants and do not harbor GWAS- or fine-mapping-identified risk variants. ^aGenes that were siRNA-silenced for functional assays are shown in bold; SNPs used to predict gene expression are listed in Supplementary Table 13. ^bProtein: protein-coding genes; lncRNA: long non-coding RNAs. ^cP value: nominal P value from association analysis of 122,977 cases and 105,974 controls; the threshold after Bonferroni correction of 8,597 tests ($0.05/8,597 = 5.82 \times 10^{-6}$) was used; ^dR²: prediction performance (R²) derived using GTEx data. ^eRisk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes, are presented in Supplementary Table 4. ^fUse of the **COJO** method²⁶; all index SNPs in the corresponding region were adjusted in the conditional analyses. ^gPredicted expression of *RP11-73O6.3* and *L3MBTL3* was correlated (Spearman R = 0.88). ^hPredicted expression of *AP006621.6* and *RPLP2* was correlated; predicted expression of *LRRC37A*, *KANSL1-AS1* and *CRHR1* was correlated (Spearman R > 0.1). ⁱPredicted expression of *RP11-15A1.7* and *ZNF404* was correlated (Spearman R = 0.64).

Downregulation of three lncRNAs (*RP11-218M22.1*, *RP11-467J12.4* and *CTD-3032H12.1*) resulted in a significant reduction in cell proliferation in all three cell lines. We also evaluated the effect of inhibition of these genes on colony-forming ability in MCF7 cells. Knockdown of the three negative control genes did not significantly affect colony-forming efficiency (CFE). By contrast, knockdown of *PIDD1*, *RP11-15A1.7*, *RP11-218M22.1*, *AP006621.6*, *ZNF404*, *RP11-467J12.4* and *CTD-3032H12.1* resulted in significantly decreased CFE in MCF7 cells compared to the NTC (Fig. 2b and Supplementary Fig. 7).

Discussion

This is the largest study to systematically evaluate associations of genetically predicted gene expression across the human transcriptome with breast cancer risk. We identified 179 genes showing a significant association at the FDR-corrected significance level. Of these, 48 genes showed an association at the Bonferroni-corrected threshold, including 14 at genomic loci that have not previously been implicated for breast cancer risk. Of the 34 genes located at known risk loci, 23 have not previously been shown to be the targets of GWAS-identified risk SNPs at corresponding loci and do not harbor any risk SNPs. Our study provides substantial new information to improve the understanding of genetics and etiology for breast cancer.

It is possible that TWAS-identified genes may be associated with breast cancer through their correlation with disease causal genes. To determine the potential functional significance of TWAS-identified genes and provide evidence for causal inference, we knocked down 13 genes for which high predicted levels of expression were associated with an increased breast cancer risk, in one normal and two breast cancer cell lines, and measured the effect on proliferation and CFE. Although there was some variation between cell lines, knockdown of 11 of the 13 genes showed an effect in at least one cell line, particularly on proliferation in 184A1 normal breast cells; the effects were strongest and most consistent for the lncRNAs *RP11-218M22.1*, *RP11-467J12.4* and *CTD-3032H12.1*. The observation of a more consistent effect in the normal breast cell line compared with the cancer cell lines is not surprising as cancer cell lines have increased capacity to handle gene interference through mutations that enhance cell survival. Rewiring of pathways and compensatory mechanisms is a hallmark of cancer. Knockdown of *PIDD1*, *NRBF2* and *ABHD8*, for which breast cancer risk-associated haplotypes have been shown to be associated with increased expression in reporter assays^{7,20,22}, affected either proliferation or CFE, supporting the results from this study.

Some of the genes with strong functional evidence from our study have been reported to have important roles in carcinogenesis. For example, *RP11-467J12.4* (PR-lncRNA-1) is a p53-regulated

Table 3 | Eleven expression–trait associations for genes previously reported as potential target genes of GWAS-identified breast cancer risk variants or genes harboring risk variants

Region	Gene ^a	Type ^b	Z score	P value ^c	R ^{2c}	Closest risk SNP ^d	Distance to the closest risk SNP (kb)	P value after adjusting for adjacent risk SNPs ^e	Association direction reported previously ^f	Reference
1p36.13	<i>KLHDC7A</i>	Protein	−5.67	1.40 × 10 ^{−8}	0.04	rs2992756	0.085	0.06	−	7
2q33.1	<i>ALS2CR12</i>	Protein	6.70	2.11 × 10 ^{−11}	0.10	rs1830298	Intron of the gene	0.17	NA	31
2q33.1	<i>CASP8</i>	Protein	−8.05	8.51 × 10 ^{−16}	0.22	rs3769821	Intron of the gene	0.16	−	31,32
5q14.1	<i>ATG10</i>	Protein	−6.65	2.85 × 10 ^{−11}	0.51	rs7707921	Intron of the gene	0.21	NA	9
5q14.2	<i>ATP6AP1L</i>	Protein	−4.98	6.32 × 10 ^{−7}	0.63	rs7707921	37	0.98	NA	9
6q23.1	<i>L3MBTL3</i> ^g	Protein	−6.69	2.27 × 10 ^{−11}	0.10	rs6569648	208	0.44	NA	6
6q25.1	<i>RMND1</i>	Protein	4.76	1.95 × 10 ^{−6}	0.13	rs3757322	169	1.11 × 10 ^{−4}	mixed	17
11q13.1	<i>SNX32</i>	Protein	4.70	2.60 × 10 ^{−6}	0.19	rs3903072	18	0.17	NA	33
15q26.1	<i>RCCD1</i>	Protein	−7.18	7.23 × 10 ^{−13}	0.13	rs2290203	6	1.66 × 10 ^{−4}	−	10
17q22	<i>STXBP4</i>	Protein	6.69	2.21 × 10 ^{−11}	0.03	rs6504950	Intron of the gene	0.90	+ in GTEx	34,35
19q13.31	<i>ZNF404</i> ^h	Protein	7.42	1.15 × 10 ^{−13}	0.15	rs3760982	90	0.005	NA	8

^aGenes that were siRNA-silenced for functional assays are shown in bold; SNPs used to predict gene expression are listed in Supplementary Table 13. ^bProtein: protein-coding genes; lncRNA: long non-coding RNAs; NA: not available. ^cP value: nominal P value from association analysis of 122,977 cases and 105,974 controls; the threshold after Bonferroni correction of 8,597 tests (0.05/8,597 = 5.82 × 10^{−6}) was used; R²: prediction performance (R²) derived using GTEx data. ^dRisk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes, are presented in Supplementary Table 4. ^eUse of the COJO method³⁶; all index SNPs in the corresponding region were adjusted for the conditional analyses. ^f−, inverse association; +, positive association; mixed: both inverse and positive associations reported; NA: not available. ^gPredicted expression of *L3MBTL3* and *RP11-7306.3* was correlated (Spearman R = 0.88). ^hPredicted expression of *ZNF404* and *RP11-15A1.7* was correlated (Spearman R = 0.64).

lncRNA that modulates gene expression in response to DNA damage downstream of p53⁴². *STXBP4* encodes syntaxin-binding protein 4, a scaffold protein that can stabilize and prevent degradation of an isoform of p63, a member of the p53 tumor suppressor family⁴³. *KLHDC10* encodes a member of the Kelch superfamily that can activate apoptosis signal-regulating kinase 1, contributing to oxidative stress-induced cell death⁴⁴. Notably, another member of this superfamily, *KLHDC7A*, has recently been identified as the target gene at the 1p36 breast cancer risk locus⁷.

SNX32, *ALK* and *BTN3A2* are also likely susceptibility genes for breast cancer risk. However, their low or absent expression in our chosen breast cell lines prevented further functional analysis. *ALK* (anaplastic lymphoma kinase) copy-number gain and overexpression have been reported in aggressive and metastatic breast cancers⁴⁵. Therapeutic targeting of ALK rearrangement has significantly improved survival in advanced ALK-positive lung cancer⁴⁶, making it an attractive target for breast and other cancers. *BTN3A2* is a member of the B7/butyrophilin-like group of Ig superfamily receptors modulating the function of T lymphocytes. Overexpression of *BTN3A2* in epithelial ovarian cancer is associated with higher infiltrating immune cells and a better prognosis⁴⁷.

Our analyses identified multiple genes with reduced expression associated with increased breast cancer risk. Among them, *LRRC3B* and *CASP8* are putative tumor suppressors in multiple cancers, including breast cancer. Leucine-rich repeat-containing 3B (*LRRC3B*) is a putative LRR-containing transmembrane protein, which is frequently inactivated via promoter hypermethylation leading to inhibition of cancer cell growth, proliferation and invasion⁴⁸. *CASP8* encodes a member of the cysteine-aspartic acid protease family, which play a central role in cell apoptosis. Previous studies have suggested that caspase-8 may act as a tumor suppressor in certain types of lung cancer and neuroblastoma, although this function has not yet been demonstrated in breast cancer. Notably, several large association studies have identified SNPs at the 2q33/*CASP8* locus associated with increased breast cancer risk^{31,49}. Consistent

with our data, eQTL analyses showed that the risk alleles for breast cancer were associated with reduced *CASP8* messenger RNA levels in both peripheral blood lymphocytes and normal breast tissue³¹.

For seven of the genes listed in Tables 1 and 2, we found some evidence from studies using tumor tissues, in vitro or in vivo experiments linking them to cancer risk (Supplementary Table 10), although their association with breast cancer has not been demonstrated in human studies. For five of them (*LRRC3B*, *SPATA18*, *RIC8A*, *ALK* and *CRHR1*), previous in vitro and in vivo experiments and human tissue studies showed a consistent direction of the association as demonstrated in our studies. For two other genes (*UBD* and *MIR31HG*), however, results from previous studies were inconsistent, reporting both potential promoting and inhibiting effects on breast cancer development. Future studies are needed to evaluate the functions of these genes.

We included a large number of cases and controls, providing high statistical power for the association analysis. This large sample size enabled us to identify a large number of candidate breast cancer susceptibility genes, much larger than the number identified in a TWAS study with a sample size of about 20% of ours³⁰. The previous study included subjects of different races, which could affect the results as linkage disequilibrium patterns differ by races. Of the five genes reported in that smaller TWAS that showed a suggestive association with breast cancer risk, the association for the *RCCD1* gene was replicated in our study (Table 3). The other four genes (*ANKLE1*, *DHODH*, *ACAP1* and *LRRC25*) were not evaluated in our study because of unsatisfactory performance of our breast-specific models for these genes that were built using the GTEx reference data set including only female European descendants.

A substantial proportion of SNPs included in OncoArray and iCOGS were selected from breast cancer GWAS and fine-mapping analyses, and thus these arrays were enriched for association signals with breast cancer risk. As a result, the overall λ value for the BCAC association analyses of individual variants is 1.26 after adjusting for population stratifications (QQ plot in Supplementary Fig. 3b)⁷.

Table 4 | Genes at GWAS-identified breast cancer risk loci (± 500 kb of the index SNPs) whose predicted expression levels were associated with breast cancer risk at P values between 5.82×10^{-6} and 1.05×10^{-3} (FDR-corrected P value ≤ 0.05)

Region	Gene	Type ^a	Z score	P value ^b	R ^{2b}	Closest risk SNP ^c	Distance to the closest risk SNP (kb)	P value after adjusting for adjacent risk SNPs ^d
1p34.1	UQCRH	Protein	−3.90	9.51×10^{-5}	0.12	rs1707302	168	0.06
1p22.3	LMO4	Protein	−3.76	1.73×10^{-4}	0.09	rs12118297	15	0.002
2p23.3	DNAJC27-AS1	lncRNA	3.84	1.24×10^{-4}	0.03	rs6725517	65	0.13
4p14	KLHL5	Protein	3.52	4.35×10^{-4}	0.13	rs6815814	230	0.03
5q11.2	AC008391.1	miRNA	−4.03	5.60×10^{-5}	0.13	rs16886113	242	0.76
6p22.1	HCG14	lncRNA	−3.47	5.19×10^{-4}	0.11	rs9257408	61	0.03
6p22.2	TRNAI2	miRNA	−3.71	2.09×10^{-4}	0.02	rs71557345	307	0.007
6q25.1	MTHFD1L	Protein	3.85	1.17×10^{-4}	0.10	rs3757318	491	2.36×10^{-4}
8q24.21	PVT1	Transcript	3.85	1.20×10^{-4}	0.03	rs11780156	81	1.09×10^{-4}
9q33.3	RP11-123K19.1	lncRNA	−4.10	4.05×10^{-5}	0.05	rs10760444	20	1.26×10^{-4}
10q25.2	RP11-57H14.3	lncRNA	3.42	6.16×10^{-4}	0.08	rs7904519	108	0.002
10q26.13	RP11-500G22.2	lncRNA	4.48	7.54×10^{-6}	0.15	rs2981582	336	0.91
11p15.5	PTDSS2	Protein	−3.47	5.16×10^{-4}	0.04	rs6597981	312	0.02
11p15.5	AP006621.5	Protein	4.35	1.37×10^{-5}	0.51	rs6597981	19	0.01
11p15.5	PIDD1	Protein	4.24	2.28×10^{-5}	0.45	rs6597981	Intron of the gene	0.12
11p15.5	MRPL23-AS1	lncRNA	−3.86	1.12×10^{-4}	0.10	rs3817198	95	0.06
11q13.1–11q13.2	PACS1	Protein	−3.59	3.36×10^{-4}	0.06	rs3903072	255	0.001
12p11.22	RP11-860B13.1	lncRNA	3.46	5.42×10^{-4}	0.17	rs10771399	221	0.86
13q22.1	KLF5	Protein	−4.08	4.44×10^{-5}	0.22	rs6562760	306	NA
14q24.1	CTD-2566J3.1	lncRNA	−3.84	1.22×10^{-4}	0.04	rs2588809	64	0.55
14q32.33	C14orf79	Protein	4.37	1.22×10^{-5}	0.11	rs10623258	240	0.91
15q26.1	FES	Protein	4.37	1.26×10^{-5}	0.21	rs2290203	73	3.04×10^{-6}
16q12.2	BBS2	Protein	3.97	7.23×10^{-5}	0.26	rs2432539	80	0.36
16q12.2	CRNDE	lncRNA	3.28	1.05×10^{-3}	0.02	rs28539243	271	0.69
16q24.2	RP11-482M8.1	lncRNA	3.32	9.16×10^{-4}	0.02	rs4496150	441	0.19
17q11.2	GOSR1	Protein	3.79	1.51×10^{-4}	0.10	rs146699004	376	0.04
17q21.2	ATP6V0A1	Protein	3.61	3.02×10^{-4}	0.03	rs72826962	162	0.01
17q21.2	RP11-400F19.8	Transcript	−3.96	7.65×10^{-5}	0.01	rs72826962	122	6.62×10^{-4}
17q21.31	RP11-105N13.4	Transcript	−4.51	6.46×10^{-6}	0.02	rs2532263	359	NA
17q25.3	CBX8	Protein	4.38	1.16×10^{-5}	0.05	rs745570	6	0.99
19p13.11	CTD-2538G9.5	lncRNA	3.56	3.76×10^{-4}	0.01	rs8170	432	4.38×10^{-4}
19p13.11	HOMER3	Protein	−3.87	1.08×10^{-4}	0.10	rs4808801	469	0.18
20q11.22	CTD-3216D2.5	lncRNA	4.03	5.60×10^{-5}	0.16	rs2284378	281	9.24×10^{-4}
22q13.1	TRIOBP	Protein	3.34	8.34×10^{-4}	0.07	rs738321	396	0.003
22q13.1	RP5-1039K5.13	lncRNA	3.73	1.93×10^{-4}	0.01	rs738321	99	0.053
22q13.1	CBY1	Protein	3.91	9.34×10^{-5}	0.05	chr22:39359355	289	0.06
22q13.1	APOBEC3A	Protein	−4.11	3.98×10^{-5}	0.07	chr22:39359355	0.2	0.02
22q13.2	RP1-85F18.6	lncRNA	3.52	4.28×10^{-4}	0.12	rs73161324	460	0.72

^aProtein: protein-coding genes; lncRNA: long non-coding RNAs; transcript: processed transcript. ^bP value: nominal P value from association analysis of 122,977 cases and 105,974 controls; R²: prediction performance derived using GTEx data. ^cRisk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes, are presented in Supplementary Table 4. ^dUse of the COJO method³⁶; all index SNPs in the corresponding region were adjusted for the conditional analyses.

The λ value for the associations of the ~257,000 SNPs included in the gene expression prediction models of the 8,597 genes tested in our association analysis is 1.40 (QQ plot in Supplementary Fig. 3c). This higher λ value is perhaps expected because of a potential further enrichment of breast cancer-associated signals in the set of SNPs selected to predict gene expression. There could be additional gain of power (and thus a higher λ value) in TWAS as it aggregates

the effect of multiple SNPs to predict gene expression and uses genes as the unit for association analyses. The lambda (λ) for our associated analyses of 8,597 genes was 1.51 (QQ plot presented in Supplementary Fig. 3a) likely due to the potential enrichment and power gain as well as our large sample size, and the highly polygenic nature of the disease^{7,50}. Interestingly, high λ values were also found in recent large studies of other polygenic traits, such as body mass

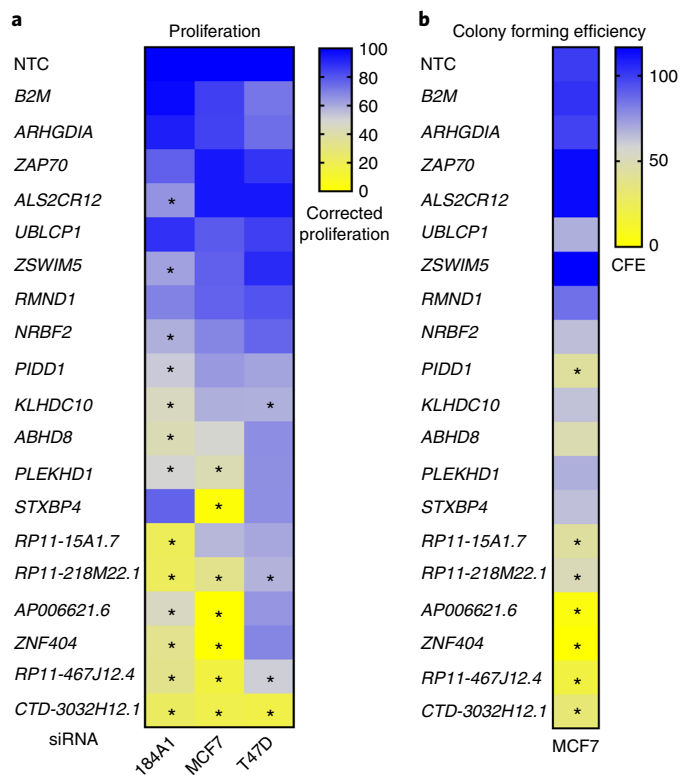


Fig. 2 | Heat maps of proliferation and CFE in breast cells. a, Proliferation efficiency. **b,** CFE. Error bars, s.d. ($N=2$). P values were determined by one-way ANOVA followed by Dunnett's multiple-comparisons test: * P value < 0.05 . NTC: non-target control.

index ($\lambda=1.99$) and height ($\lambda=2.7$)^{51,52}. The $\lambda_{1,000}$, a standardized estimate of the genomic inflation scaling to a study of 1,000 cases and 1,000 controls, is 1.004 in our study.

The statistical power of our study is very high to detect associations for genes with a relatively high *cis*-heritability (h^2) (Supplementary Fig. 8). For example, our study has 80% statistical power to detect an association with breast cancer risk at $P < 5.82 \times 10^{-6}$ with an odds ratio of 1.07 or higher per one standard deviation increase (or decrease) in the expression level of genes with an h^2 of 0.1 or higher. One limitation of our study is the small sample size for building gene expression prediction models, which may have affected the precision of model parameter estimates. We expect that models built with a larger sample size will identify additional association signals. We used samples from women of European origin in model building, given differences in gene expression patterns between males and females and in genetic architecture across ethnicities⁵³. We also used gene expression data of tumor-adjacent normal tissue samples from European descendants in TCGA as an external validation step to prioritize genes for association analyses. Given potential somatic alterations in tumor-adjacent normal tissues, we retained all models showing a prediction R^2 of at least 0.09 in GTEx, regardless of their performance in TCGA. Not all genes have a significant hereditary component in expression regulation, and thus these genes could not be investigated in our study. For example, previous studies have provided strong evidence to support a significant role of the *TERT*, *ESR1*, *CCND1*, *IGFBP5*, *TET2* and *MRPS30* genes in the etiology of breast cancer. However, expression of these genes cannot be predicted well using the data from female European descendants included in the GTEx and thus they were not included in our association analyses. Supplementary Table 11 summarizes the performance of prediction models and association results for breast cancer target genes reported previously at GWAS-identified loci.

In summary, our study has identified multiple gene candidates that can be further functionally characterized. The silencing experiments we performed suggest that many of the genes identified are likely to mediate risk of breast cancer by affecting proliferation or CFE, two hallmarks of cancer. Further investigation of genes identified in our study will provide additional insight into the biology and genetics of breast cancer.

URLs. GTEx protocol, <http://www.gtexportal.org/home/documentationPage>; Gencode V19 annotation file, <http://www.gencodegenes.org/releases/19.html>; HaploReg, <http://archive.broadinstitute.org/mammals/haploreg/data/>; OncoArray, <http://epi.grants.cancer.gov/oncoarray/>.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at <https://doi.org/10.1038/s41588-018-0132-x>.

Received: 21 September 2017; Accepted: 17 April 2018;

Published online: 18 June 2018

References

- Kamangar, F., Dores, G. M. & Anderson, W. F. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J. Clin. Oncol.* **24**, 2137–2150 (2006).
- Beggs, A. D. & Hodgson, S. V. Genomics and breast cancer: the different levels of inherited susceptibility. *Eur. J. Hum. Genet.* **17**, 855–856 (2009).
- Southey, M. C. et al. PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. *J. Med. Genet.* **53**, 800–811 (2016).
- Nathanson, K. L., Wooster, R. & Weber, B. L. Breast cancer genetics: what we know and what we need. *Nat. Med.* **7**, 552–556 (2001).
- Anglian Breast Cancer Study Group. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *Br. J. Cancer* **83**, 1301–1308 (2000).
- Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
- Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
- Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361 (2013).
- 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).
- 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).
- Zheng, W. et al. Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. *Hum. Mol. Genet.* **22**, 2539–2550 (2013).
- Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. *Lancet Oncol.* **12**, 477–488 (2011).
- French, J. D. et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.* **92**, 489–503 (2013).
- Hindorf, L. A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA* **106**, 9362–9367 (2009).
- The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
- Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
- Dunning, A. M. et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat. Genet.* **48**, 374–386 (2016).
- Ghoussaini, M. et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat. Commun.* **4**, 4999 (2014).
- Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
- Darabi, H. et al. Polymorphisms in a putative enhancer at the 10q21.2 breast cancer risk locus regulate NRBF2 expression. *Am. J. Hum. Genet.* **97**, 22–34 (2015).

21. Glubb, D. M. et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am. J. Hum. Genet.* **96**, 5–20 (2015).
22. Lawrenson, K. et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. *Nat. Commun.* **7**, 12675 (2016).
23. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat. Genet.* **47**, 955–961 (2015).
24. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
25. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95**, 535–552 (2014).
26. Barbeira, A.N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nat. Commun.* **9**, 1825 (2018).
27. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).
28. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
29. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).
30. Hoffman, J. D. et al. *Cis*-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. *PLoS Genet.* **13**, e1006690 (2017).
31. Lin, W. Y. et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Hum. Mol. Genet.* **24**, 285–298 (2015).
32. Camp, N. J. et al. Discordant haplotype sequencing identifies functional variants at the 2q33 breast cancer risk locus. *Cancer Res.* **76**, 1916–1925 (2016).
33. Li, Q. et al. Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. *Hum. Mol. Genet.* **23**, 5294–5302 (2014).
34. Caswell, J. L. et al. Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors. *Hum. Mol. Genet.* **24**, 7421–7431 (2015).
35. Darabi, H. et al. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGS). *Sci. Rep.* **6**, 32512 (2016).
36. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–375 (2012).
37. Kramer, A., Green, J., Pollard, J. Jr & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523–530 (2014).
38. Koh, J. L. et al. COLT-Cancer: functional genetic screening resource for essential genes in human cancer cell lines. *Nucleic Acids Res.* **40**, D957–D963 (2012).
39. Marcotte, R. et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer Discov.* **2**, 172–189 (2012).
40. Walen, K. H. & Stampfer, M. R. Chromosome analyses of human mammary epithelial cells at stages of chemical-induced transformation progression to immortality. *Cancer Genet. Cytogenet.* **37**, 249–261 (1989).
41. Treszezamsky, A. D. et al. BRCA1- and BRCA2-deficient cells are sensitive to etoposide-induced DNA double-strand breaks via topoisomerase II. *Cancer Res.* **67**, 7078–7081 (2007).
42. Sanchez, Y. et al. Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. *Nat. Commun.* **5**, 5812 (2014).
43. Li, Y., Peart, M. J. & Prives, C. Stxbp4 regulates DeltaNp63 stability by suppression of RACK1-dependent degradation. *Mol. Cell Biol.* **29**, 3953–3963 (2009).
44. Sekine, Y. et al. The Kelch repeat protein KLHDC10 regulates oxidative stress-induced ASK1 activation by suppressing PP5. *Mol. Cell* **48**, 692–704 (2012).
45. Kim, M. H. et al. Anaplastic lymphoma kinase gene copy number gain in inflammatory breast cancer (IBC): prevalence, clinicopathologic features and prognostic implication. *PLoS One* **10**, e0120320 (2015).
46. Shaw, A.T. et al. Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. *N. Engl. J. Med.* **368**, 2385–2394 (2013).
47. Le Page, C. et al. BTN3A2 expression in epithelial ovarian cancer is associated with higher tumor infiltrating T cells and a better prognosis. *PLoS One* **7**, e38541 (2012).
48. Kan, L. et al. LRRC3B is downregulated in non-small-cell lung cancer and inhibits cancer cell proliferation and invasion. *Tumour Biol.* **37**, 1113–1120 (2016).
49. Cox, A. et al. A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.* **39**, 352–358 (2007).
50. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
51. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
52. Turcot, V. et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. *Nat. Genet.* **50**, 26–41 (2018).
53. Melé, M. et al. The human transcriptome across tissues and individuals. *Science* **348**, 660–665 (2015).

Acknowledgements

The authors thank J. He, W. Wen, A. Giri and T. Edwards of Vanderbilt Epidemiology Center and R. Tao of the Department of Biostatistics, Vanderbilt University Medical Center for their help with the data analysis of this study. The authors would also like to thank all of the individuals for their participation in the parent studies and all of the researchers, clinicians, technicians and administrative staff for their contribution to the studies. We are also grateful to H. K. Im of University of Chicago for her help. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. This project at Vanderbilt University Medical Center was supported in part by grants R01CA158473 and R01CA148677 from the US National Institutes of Health as well as funds from Anne Potter Wilson endowment. L.W. is supported by NCI K99 CA218892 and the Vanderbilt Molecular and Genetic Epidemiology of Cancer (MAGEC) training program (US NCI grant R25 CA160056 awarded to X.-O.S.). 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 Québec Breast Cancer Foundation; the NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and the Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project (National Institutes of Health (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 no. PSR-SIIRI-701. Combining of the GWAS data was supported in part by the NIH Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE, part of the GAME-ON initiative). A full description of funding and acknowledgments for BCAC studies, along with consortium membership, are included in the Supplementary Note.

Author contributions

W.Z. and J. Long conceived the study. L.W. contributed to the study design and performed statistical analyses. L.W., W.Z. and G.C.-T. wrote the manuscript with significant contributions from W.S., J. Long, X.G. and S.L.E. W.S. performed the in vitro experiments. G.C.-T. directed the in vitro experiments. X.G. contributed to the model building and pathway analyses. J.B. contributed to the bioinformatics analyses. F.A.-E., E.R. and S.L.E. contributed to the in vitro experiments. Y.L. and C.Z. contributed to the model building. K.M., M.K.B., X.-O.S., Q.W., J.D., B.L., C.Z., H.F., A.G., R.T.B., A.M.D., P.D.P.P., J.S., R.L.M., P.K. and D.F.E. contributed to manuscript revision, statistical analyses and/or BCAC data management. I.L.A., H.A.-C., V.A., K.J.A., P.L.A., M. Barndahl, C.B., M.W.B., J.B., M. Bermisheva, C.B., N.V.B., S.E.B., H. Brauch, H. Brenner, L.B., P.B., S.Y.B., B.B., Q.C., T.C., F.C., B.D.C., J.E.C., J.C.-C., X.C., T.-Y.D.C., H.C., C.L.C., NBCS Collaborators, M.C., S.C., F.J.C., D.C., A.C., S.S.C., J.M.C., K.C., M.B.D., P.D., K.F.D., T.D., I.D.S.S., M. Dumont, M. Dwek, D.M.E., U.E., H.E., C.E., M.E., L.F., P.A.F., J.F., D.F.-J., O.F., H.F., L.F., M. Gabrielson, M.G.-D., S.M.G., M.G.-C., M.M.G., M. Ghoussaini, G.G.G., M.S.G., D.E.G., A.G.-N., P.G., E. Hahnen, C.A.H., N.H., P. Hall, E. Hallberg, U.H., P. Harrington, A. Hein, B.H., P. Hillemanns, A. Hollestelle, R.N.H., J.L.H., G.H., K.H., D.J.H., A.J., W.J., E.M.J., N.J., K.J., M.E.J., A. Jung, R.K., M.J.K., E.K., V.-M.K., V.N.K., D.L., L.L.M., J. Li, S.L., J. Lissowska, W.-Y.L., S. Loibl, J. Lubinski, C.L., M.P.L., R.J.M., T.M., I.M.K., A. Mannervaa, J.E.M., S.M., D.M., H.M.-H., A. Meindl, U.M., J.M., A.M.M., S.L.N., H.N., P.N., S.F.N., B.G.N., O.I.O., J.E.O., H.O., P.P., J.P., D.P.-K., R.P., N.P., K.P., B.R., P.R., N.R., G.R., H.S.R., V.R., A. Romero, J.R., A. Rudolph, E.S., D.P.S., E.J.S., M.K.S., R.K.S., A.S., R.J.S., C.G.S., S.S., M.S., M.J.S., A.S., M.C.S., J.J.S., J.S., H.S., A.J.S., R.T., W.T., J.A.T., M.B.T., D.C.T., A.T., K.T., R.A.E.M.T., D.T., T.T., M.U., C.V., D.V.D.B., D.V., Q.W., C.R.W., C.W., A.S.W., H.W., W.C.W., R.W., A.W., L.X., X.R.Y., A.Z., E.Z. and kConFab/AOCS Investigators contributed to the collection of the data and biological samples for the original BCAC studies. All authors have reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information








Supplementary information is available for this paper at <https://doi.org/10.1038/s41588-018-0132-x>.

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

Correspondence and requests for materials should be addressed to G.C. or W.Z.

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

Lang Wu^{1,159}, Wei Shi^{2,159}, Jirong Long¹, Xingyi Guo¹, Kyriaki Michailidou^{3,4}, Jonathan Beesley², Manjeet K. Bolla³, Xiao-Ou Shu¹, Yingchang Lu¹, Qiuyin Cai¹, Fares Al-Ejeh², Esdy Rozali², Qin Wang³, Joe Dennis³, Bingshan Li⁵, Chenjie Zeng¹, Helian Feng^{6,7}, Alexander Gusev^{8,9,10}, Richard T. Barfield⁶, Irene L. Andrulis^{11,12}, Hoda Anton-Culver¹³, Volker Arndt¹⁴, Kristan J. Aronson¹⁵, Paul L. Auer^{16,17}, Myrto Barrdahl¹⁸, Caroline Baynes¹⁹, Matthias W. Beckmann²⁰, Javier Benitez^{21,22}, Marina Bermisheva^{23,24}, Carl Blomqvist^{25,26}, Natalia V. Bogdanova^{24,27,28}, Stig E. Bojesen^{29,30,31}, Hiltrud Brauch^{32,33,34}, Hermann Brenner^{14,34,35}, Louise Brinton³⁶, Per Broberg³⁷, Sara Y. Brucker³⁸, Barbara Burwinkel^{39,40}, Trinidad Caldés⁴¹, Federico Canzian⁴², Brian D. Carter⁴³, J. Esteban Castela⁴⁴, Jenny Chang-Claude^{18,45}, Xiaoqing Chen², Ting-Yuan David Cheng⁴⁶, Hans Christiansen²⁷, Christine L. Clarke⁴⁷, NBCS Collaborators⁴⁸, Margriet Collée⁴⁹, Sten Cornelissen⁵⁰, Fergus J. Couch⁵¹, David Cox^{52,53}, Angela Cox⁵⁴, Simon S. Cross⁵⁵, Julie M. Cunningham⁵¹, Kamila Czene⁵⁶, Mary B. Daly⁵⁷, Peter Devilee^{58,59}, Kimberly F. Doheny⁶⁰, Thilo Dörk²⁴, Isabel dos-Santos-Silva⁶¹, Martine Dumont⁶², Miriam Dwek⁶³, Diana M. Eccles⁶⁴, Ursula Eilber¹⁸, A. Heather Eliassen^{7,65}, Christoph Engel⁶⁶, Mikael Eriksson⁵⁶, Laura Fachal¹⁹, Peter A. Fasching^{20,67}, Jonine Figueroa^{36,68}, Dieter Flesch-Janys^{69,70}, Olivia Fletcher⁷¹, Henrik Flyger⁷², Lin Fritschi⁷³, Marike Gabrielson⁵⁶, Manuela Gago-Dominguez^{74,75}, Susan M. Gapstur⁴³, Montserrat García-Closas³⁶, Mia M. Gaudet⁴³, Maya Ghoussaini¹⁹, Graham G. Giles^{76,77}, Mark S. Goldberg^{78,79}, David E. Goldgar⁸⁰, Anna González-Neira²¹, Pascal Guénel⁸¹, Eric Hahnen^{82,83,84}, Christopher A. Haiman⁸⁵, Niclas Håkansson⁸⁶, Per Hall^{56,87}, Emily Hallberg⁸⁸, Ute Hamann⁸⁹, Patricia Harrington¹⁹, Alexander Hein²⁰, Belynda Hicks⁹⁰, Peter Hillemanns²⁴, Antoinette Hollestelle⁹¹, Robert N. Hoover³⁶, John L. Hopper⁷⁷, Guanmengqian Huang⁸⁹, Keith Humphreys⁵⁶, David J. Hunter^{7,92}, Anna Jakubowska^{93,94}, Wolfgang Janni⁹⁵, Esther M. John^{96,97,98}, Nichola Johnson⁷¹, Kristine Jones⁹⁰, Michael E. Jones⁹⁹, Audrey Jung¹⁸, Rudolf Kaaks¹⁸, Michael J. Kerin¹⁰⁰, Elza Khusnutdinova^{123,101}, Veli-Matti Kosma^{102,103,104}, Vessela N. Kristensen^{105,106,107}, Diether Lambrechts^{108,109}, Loïc Le Marchand¹¹⁰, Jingmei Li¹¹¹, Sara Lindström^{112,113}, Jolanta Lissowska¹¹⁴, Wing-Yee Lo^{32,33}, Sibylle Loibl¹¹⁵, Jan Lubinski⁹³, Craig Luccarini¹⁹, Michael P. Lux²⁰, Robert J. MacInnis^{76,77}, Tom Maishman¹¹⁶, Ivana Maleva Kostovska^{24,117}, Arto Mannermaa^{102,103,104}, JoAnn E. Manson^{7,118}, Sara Margolin¹¹⁹, Dimitrios Mavroudis¹²⁰, Hanne Meijers-Heijboer¹²¹, Alfons Meindl¹²², Usha Menon¹²³, Jeffery Meyer⁵¹, Anna Marie Mulligan^{124,125}, Susan L. Neuhausen¹²⁶, Heli Nevanlinna¹²⁷, Patrick Neven¹²⁸, Sune F. Nielsen^{29,30}, Børge G. Nordestgaard^{29,30,31}, Olufunmilayo I. Olopade¹²⁹, Janet E. Olson⁸⁸, Håkan Olsson³⁷, Paolo Peterlongo¹³⁰, Julian Peto⁶¹, Dijana Plaseska-Karanfilska¹¹⁷, Ross Prentice¹⁶, Nadege Presneau⁶³, Katri Pylkäs^{131,132}, Brigitte Rack⁹⁵, Paolo Radice¹³³, Nazneen Rahman¹³⁴, Gad Rennert¹³⁵, Hedy S. Rennert¹³⁵, Valerie Rhenius¹⁹, Atocha Romero^{41,136}, Jane Romm⁶⁰, Anja Rudolph¹⁸, Emmanouil Saloustros¹³⁷, Dale P. Sandler¹³⁸, Elinor J. Sawyer¹³⁹, Marjanka K. Schmidt^{50,140}, Rita K. Schmutzler^{82,83,84}, Andreas Schneeweiss^{39,141}, Rodney J. Scott^{142,143}, Christopher G. Scott⁸⁸, Sheila Seal¹³⁴, Mitul Shah¹⁹, Martha J. Shrubsole¹, Ann Smeets¹²⁸, Melissa C. Southey¹⁴⁴, John J. Spinelli^{145,146}, Jennifer Stone^{147,148}, Harald Surowy^{39,40}, Anthony J. Swerdlow^{99,149}, Rulla M. Tamimi^{6,7,65}, William Tapper⁶⁴, Jack A. Taylor^{138,150}, Mary Beth Terry¹⁵¹, Daniel C. Tessier¹⁵², Abigail Thomas⁸⁸, Kathrin Thöne⁴⁵, Rob A. E. M. Tollenaar¹⁵³, Diana Torres^{89,154}, Thérèse Truong⁸¹, Michael Untch¹⁵⁵, Celine Vachon⁸⁸, David Van Den Berg⁸⁵, Daniel Vincent¹⁵², Quinten Waisfisz¹³², Clarice R. Weinberg¹⁵⁶, Camilla Wendt¹¹⁹, Alice S. Whittemore^{97,98}, Hans Wildiers¹²⁸, Walter C. Willett^{7,65,157}, Robert Winqvist^{131,132},

Alicja Wolk¹ , Lucy Xia², Xiaohong R. Yang³, Argýrios Ziogas⁴ , Elad Ziv⁵ , kConFab/AOCS Investigators⁴⁸, Alison M. Dunning¹⁹, Paul D. P. Pharoah^{13,19} , Jacques Simard⁶², Roger L. Milne^{76,77} , Stacey L. Edwards², Peter Kraft^{6,7}, Douglas F. Easton^{13,19} , Georgia Chenevix-Trench^{2*} and Wei Zheng^{1*} 

¹Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA. ²Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. ³Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. ⁴Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. ⁵Department of Molecular Physiology & Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA. ⁶Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ⁷Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ⁸Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA, USA. ⁹Department of Medicine, Harvard Medical School, Boston, MA, USA. ¹⁰Division of Genetics, Brigham and Women's Hospital, Boston, MA, USA. ¹¹Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, Ontario, Canada. ¹²Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. ¹³Department of Epidemiology, University of California Irvine, Irvine, CA, USA. ¹⁴Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹⁵Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, Ontario, Canada. ¹⁶Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. ¹⁷Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA. ¹⁸Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. ¹⁹Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. ²⁰Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany. ²¹Human Cancer Genetics Program, Spanish National Cancer Research Centre, Madrid, Spain. ²²Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain. ²³Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa, Russia. ²⁴Gynaecology Research Unit, Hannover Medical School, Hannover, Germany. ²⁵Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. ²⁶Department of Oncology, University of Örebro, Örebro, Sweden. ²⁷Department of Radiation Oncology, Hannover Medical School, Hannover, Germany. ²⁸N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus. ²⁹Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. ³⁰Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. ³¹Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ³²Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany. ³³University of Tübingen, Tübingen, Germany. ³⁴German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. ³⁵Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. ³⁶Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA. ³⁷Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden. ³⁸Department of Gynecology and Obstetrics, University of Tübingen, Tübingen, Germany. ³⁹Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, Germany. ⁴⁰Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, Germany. ⁴¹Medical Oncology Department, CIBERONC Hospital Clínico San Carlos, Madrid, Spain. ⁴²Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. ⁴³Epidemiology Research Program, American Cancer Society, Atlanta, GA, USA. ⁴⁴Oncology and Genetics Unit, Instituto de Investigación Biomedica Galicia Sur (IISGS), Xerencia de Xestión Integrada de Vigo-SERGAS, Vigo, Spain. ⁴⁵University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ⁴⁶Department of Epidemiology, University of Florida, Gainesville, FL, USA. ⁴⁷Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia. ⁴⁸A list of NBCS Collaborators and kConFab/AOCS Investigators appears in the Supplementary Note. ⁴⁹Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands. ⁵⁰Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. ⁵¹Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA. ⁵²Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. ⁵³INSERM U1052, Cancer Research Center of Lyon, Lyon, France. ⁵⁴Sheffield Institute for Nucleic Acids, Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK. ⁵⁵Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK. ⁵⁶Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. ⁵⁷Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA. ⁵⁸Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands. ⁵⁹Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands. ⁶⁰Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. ⁶¹Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. ⁶²Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Québec City, QC, Canada. ⁶³Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK. ⁶⁴Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK. ⁶⁵Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ⁶⁶Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. ⁶⁷David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, USA. ⁶⁸Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, UK. ⁶⁹Institute for Medical Biometrics and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ⁷⁰Department of Cancer Epidemiology, Clinical Cancer Registry, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. ⁷¹The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK. ⁷²Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. ⁷³School of Public Health, Curtin University, Perth, Western Australia, Australia. ⁷⁴Genomic 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. ⁷⁵Moore's Cancer Center, University of California San Diego, La Jolla, CA, USA. ⁷⁶Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Victoria, Australia. ⁷⁷Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia. ⁷⁸Department of Medicine, McGill University, Montréal, Quebec, Canada. ⁷⁹Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, Quebec, Canada. ⁸⁰Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, USA. ⁸¹Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France. ⁸²Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne, Cologne, Germany. ⁸³Center for Integrated Oncology (CIO), University Hospital of Cologne, Cologne, Germany. ⁸⁴Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany. ⁸⁵Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. ⁸⁶Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. ⁸⁷Department of Oncology, Södersjukhuset, Stockholm, Sweden. ⁸⁸Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. ⁸⁹Molecular Genetics of Breast Cancer, German Cancer Research Center

(DKFZ), Heidelberg, Germany. ⁹⁰Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD, USA. ⁹¹Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands. ⁹²Nuffield Department of Population Health, University of Oxford, Big Data Institute, Oxford, UK. ⁹³Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland. ⁹⁴Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland. ⁹⁵Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm, Germany. ⁹⁶Department of Epidemiology, Cancer Prevention Institute of California, Fremont, CA, USA. ⁹⁷Department of Health Research and Policy - Epidemiology, Stanford University School of Medicine, Stanford, CA, USA. ⁹⁸Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. ⁹⁹Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. ¹⁰⁰School of Medicine, National University of Ireland, Galway, Ireland. ¹⁰¹Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia. ¹⁰²Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland. ¹⁰³Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland. ¹⁰⁴Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland. ¹⁰⁵Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway. ¹⁰⁶Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway. ¹⁰⁷Department of Clinical Molecular Biology, Oslo University Hospital, University of Oslo, Oslo, Norway. ¹⁰⁸VIB KULeuven Center for Cancer Biology, VIB, Leuven, Belgium. ¹⁰⁹Laboratory for Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium. ¹¹⁰Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA. ¹¹¹Human Genetics, Genome Institute of Singapore, Singapore, Singapore. ¹¹²Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA. ¹¹³Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. ¹¹⁴Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Institute - Oncology Center, Warsaw, Poland. ¹¹⁵German Breast Group, GmbH, Neu Isenburg, Germany. ¹¹⁶Southampton Clinical Trials Unit, University of Southampton, Southampton, UK. ¹¹⁷Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", Macedonian Academy of Sciences and Arts, Skopje, Macedonia. ¹¹⁸Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ¹¹⁹Department of Oncology - Pathology, Karolinska Institutet, Stockholm, Sweden. ¹²⁰Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece. ¹²¹Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands. ¹²²Division of Gynaecology and Obstetrics, Technische Universität München, Munich, Germany. ¹²³Gynaecological Cancer Research Centre, Women's Cancer, Institute for Women's Health, University College London, London, UK. ¹²⁴Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. ¹²⁵Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada. ¹²⁶Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA. ¹²⁷Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. ¹²⁸Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium. ¹²⁹Center for Clinical Cancer Genetics and Global Health, The University of Chicago, Chicago, IL, USA. ¹³⁰IFOM, The FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan, Italy. ¹³¹Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland. ¹³²Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland. ¹³³Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS (Istituto Di Ricovero e Cura a Carattere Scientifico) Istituto Nazionale dei Tumori (INT), Milan, Italy. ¹³⁴Section of Cancer Genetics, The Institute of Cancer Research, London, UK. ¹³⁵Department of Community Medicine and Epidemiology, Carmel Medical Center, Ruth and Bruce Rappaport Faculty of Medicine, Technion-Israel Institute of Technology and Clalit National Cancer Control Center, Haifa, Israel. ¹³⁶Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain. ¹³⁷Hereditary Cancer Clinic, University Hospital of Heraklion, Heraklion, Greece. ¹³⁸Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA. ¹³⁹Research Oncology, Guy's Hospital, King's College London, London, UK. ¹⁴⁰Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands. ¹⁴¹National Center for Tumor Diseases, University of Heidelberg, Heidelberg, Germany. ¹⁴²Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, New South Wales, Australia. ¹⁴³Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Newcastle, New South Wales, Australia. ¹⁴⁴Department of Pathology, The University of Melbourne, Melbourne, Victoria, Australia. ¹⁴⁵Cancer Control Research, BC Cancer Agency, Vancouver, British Columbia, Canada. ¹⁴⁶School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada. ¹⁴⁷The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Western Australia, Australia. ¹⁴⁸Department of Obstetrics and Gynaecology, University of Melbourne and the Royal Women's Hospital, Melbourne, Victoria, Australia. ¹⁴⁹Division of Breast Cancer Research, The Institute of Cancer Research, London, UK. ¹⁵⁰Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA. ¹⁵¹Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA. ¹⁵²McGill University and Génome Québec Innovation Centre, Montréal, Quebec, Canada. ¹⁵³Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands. ¹⁵⁴Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia. ¹⁵⁵Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany. ¹⁵⁶Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA. ¹⁵⁷Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA. ¹⁵⁸Department of Medicine, Institute for Human Genetics, UCSF Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA. ¹⁵⁹These authors contributed equally: Lang Wu, Wei Shi. *e-mail: Georgia.Trench@qimrberghofer.edu.au; wei.zheng@vanderbilt.edu

Methods

Building of gene expression prediction models. We used transcriptome and high-density genotyping data from the GTEx study to establish prediction models for genes expressed in normal breast tissues. Details of the GTEx have been described elsewhere⁵⁴. Genomic DNA samples obtained from study subjects included in the GTEx were genotyped using Illumina OMNI 5M or 2.5M SNP Array and RNA samples from 51 tissue sites were sequenced to generate transcriptome profiling data. Genotype data were processed according to the GTEx protocol (see the URLs section). SNPs with a call rate <98%, with differential missingness between the two array experiments (5M/2.5M Arrays), with Hardy-Weinberg equilibrium P value < 10^{-6} (among subjects of European ancestry) or showing batch effects were excluded. One Klinefelter individual, three related individuals and a chromosome 17 trisomy individual were also excluded. The genotype data were imputed to the Haplotype Reference Consortium reference panel⁵⁵ using Minimac3 for imputation and SHAPEIT for prephasing^{56,57}. SNPs with high imputation quality ($r^2 \geq 0.8$), minor allele frequency ≥ 0.05 and included in the HapMap Phase 2 version were used to build expression prediction models. For gene expression data, we used reads per kilobase per million (RPKM) units from RNA-SeQC⁵⁸. Genes with a median expression level of 0 RPKM across samples were removed, and the RPKM values of each gene were log₂ transformed. We performed quantile normalization to bring the expression profile of each sample to the same scale, and performed inverse quantile normalization for each gene to map each set of expression values to a standard normal. We adjusted for the top ten principal components derived from genotype data and the top 15 probabilistic estimation of expression residual factors to correct for batch effects and experimental confounders in model building⁵⁹. Genetic and transcriptome data from 67 female subjects of European descent without a prior breast cancer diagnosis were used to build gene expression prediction models for this study.

We built an expression prediction model for each gene by using the elastic net method as implemented in the glmnet R package, with $\alpha = 0.5$, as recommended earlier²⁷. The genetically regulated expression for each gene was estimated by including variants within a 2 Mb window flanking the respective gene boundaries, inclusive. Expression prediction models were built for protein-coding genes, lncRNAs, microRNAs (miRNAs), processed transcripts, immunoglobulin genes and T-cell receptor genes, according to categories described in the Gencode V19 annotation file (see the URLs section). Pseudogenes were not included in the present study because of potential concerns of inaccurate calling⁶⁰. Tenfold cross-validation was used to validate the models internally. Prediction R^2 values (the square of the correlation between predicted and observed expression) were generated to estimate the prediction performance of each of the gene prediction models established.

For genes that cannot be predicted well using the above approach, we built models using only SNPs located in predicted promoter or enhancer regions in breast cell lines. This approach reduces the number of variants for model building, and thus potentially improves model accuracy, by increasing the ratio of sample size to effective degrees of freedom. SNP-level annotation data in three breast cell lines, namely breast myoepithelial primary cells (E027), breast variant human mammary epithelial cells (vHMEC) (E028) and HMEC mammary epithelial primary cells (E119) in the Roadmap Epigenomics Project/Encyclopedia of DNA Elements Project¹⁶, were downloaded from HaploReg (Version 4.0, assessed on December 6, 2016) (see the URLs section). SNPs in regions classified as promoters (TssA, TssAFlnk), enhancers (Enh, EnhG) or regions with both promoter and enhancer signatures (ExFlnk) according to the core 15 chromatin state model¹⁶ in at least one of the cell lines were retained as input SNPs for model building.

Evaluating performance of gene expression prediction models using TCGA data. To assess further the validity of the models, we performed external validation using data generated in tumor-adjacent normal breast tissue samples obtained from 86 European-ancestry female breast cancer patients included in the TCGA. Genotype data were imputed using the same approach as described for GTEx data. Expression data were processed and normalized using a similar approach as described above. The predicted expression level for each gene was calculated using the model established using GTEx data and then compared with the observed level of that gene using the Spearman's correlation.

Evaluating statistical power for association tests. We conducted a simulation analysis to assess the power of our TWAS analysis. Specifically, we set the number of cases and controls to be 122,977 and 105,974, respectively, and generated the gene expression levels from the empirical distribution of predicted gene expression levels in the BCAC. We calculated statistical power at $P < 5.82 \times 10^{-6}$ (the significance level used in our TWAS) according to *cis*-heritability (h^2), which we aim to capture using gene expression prediction models (R^2). The results based on 1,000 replicates are summarized in Supplementary Fig. 8. On the basis of the power calculation, our TWAS analysis has 80% power to detect a minimum odds ratio of 1.11, 1.07, 1.05, 1.04 or 1.03 for breast cancer risk per one standard deviation increase (or decrease) in the expression level of a gene whose *cis*-heritability is 5%, 10%, 20%, 40% or 60%, respectively.

Association analyses of predicted gene expression with breast cancer risk. We used the following criteria to select genes for the association analysis: with a model

prediction R^2 of ≥ 0.01 in GTEx and a Spearman's correlation coefficient of ≥ 0.1 in TCGA; with a prediction R^2 of ≥ 0.09 in GTEx regardless of the performance in TCGA; with a prediction R^2 of ≥ 0.01 in GTEx but unable to be evaluated in TCGA. The second group of genes was selected because some gene expression levels might have changed in TCGA tumor-adjacent normal tissues, and thus it is anticipated that some genes may show low prediction performance in TCGA data due to the influence of tumor growth^{61,62}. Overall, a total of 8,597 genes met the criteria and were evaluated for their expression-trait associations.

To identify novel breast cancer susceptibility loci and genes, the MetaXcan method, as described elsewhere, was used for the association analyses²⁶. Briefly, the formula:

$$Z_g \approx \sum_{l \in \text{Model}_g} w_{lg} \frac{\hat{\beta}_l}{\hat{\sigma}_{\hat{\beta}_l}}$$

was used to estimate the Z score of the association between predicted expression and breast cancer risk. Here w_{lg} is the weight of SNP l for predicting the expression of gene g , $\hat{\beta}_l$ and $\text{se}(\hat{\beta}_l)$ are the GWAS association regression coefficient and its standard error for SNP l , and $\hat{\sigma}_l$ and $\hat{\sigma}_g$ are the estimated variances of SNP l and the predicted expression of gene g , respectively. Therefore, the weights for predicting gene expression, GWAS summary statistics results, and correlations between SNPs included in the prediction models are the input variables for the MetaXcan analyses. For this study, we estimated correlations between SNPs included in the prediction models using the phase 3, 1000 Genomes Project data focusing on European population.

For the association analysis, we used the summary statistics data for genetic variants associated with breast cancer risk generated in 122,977 breast cancer patients and 105,974 controls of European ancestry from the BCAC. The details of the BCAC have been described elsewhere^{7,9,13,63,64}. Briefly, 46,785 breast cancer cases and 42,892 controls of European ancestry were genotyped using a custom Illumina iSelect genotyping array (iCOGS) containing ~211,155 variants. A further 61,282 cases and 45,494 controls of European ancestry were genotyped using the OncoArray including 570,000 SNPs (see the URLs section). Also included in this analysis were data from 9 GWAS studies including 14,910 breast cancer cases and 17,588 controls of European ancestry. Genotype data from iCOGS, OncoArray and GWAS were imputed using the October 2014 release of the 1000 Genomes Project data as a reference. Genetic association results for breast cancer risk were combined using inverse-variance fixed-effect meta-analyses⁷. For our study, only SNPs with imputation $r^2 \geq 0.3$ were used. All participating BCAC studies were approved by their appropriate ethics review boards. Relevant ethical regulations were complied with. This study was approved by the BCAC Data Access Coordination Committee.

Lambda 1,000 ($\lambda_{1,000}$) was calculated to represent a standardized estimate of the genomic inflation scaling to a study of 1,000 cases and 1,000 controls, using the following formula: $\lambda_{1,000} = 1 + (\lambda_{\text{obs}} - 1) \times (1/n_{\text{cases}} + 1/n_{\text{controls}}) / (1/1,000_{\text{cases}} + 1/1,000_{\text{controls}})$ (refs ^{65,66}). We used a Bonferroni-corrected P threshold of 5.82×10^{-6} ($0.05/8,597$) to determine a statistically significant association for the primary analyses. To identify additional gene candidates at previously identified susceptibility loci, we also used an FDR-corrected P threshold of 1.05×10^{-3} ($\text{FDR} \leq 0.05$) to determine a significant association. Associated genes with an expression of >0.1 RPKM in fewer than 10 individuals in GTEx data were excluded as the corresponding prediction models may not be stable.

To determine whether the predicted expression-trait associations were independent of the top signals identified in previous GWAS, we performed GCTA-COJO analyses developed in an earlier study³⁶ to calculate association betas and standard errors of variants with breast cancer risk after adjusting for the index SNPs of interest. We then re-ran the MetaXcan analyses using the association statistics after conditioning on the index SNPs. This information was used to determine whether the detected expression-trait associations remained significant after adjusting for the index SNPs.

For 41 identified associated genes at the Bonferroni-corrected threshold, we also performed analyses using individual-level data in the iCOGS ($n = 84,740$) and OncoArray ($n = 112,133$) data sets. We generated predicted gene expression using predicting SNPs (Supplementary Table 12), and then assessed the association between predicted gene expression and breast cancer risk adjusting for study and nine principal components in the iCOGS data set, and country and the first ten principal components in the OncoArray data set. Conditional analyses adjusting for index SNPs were performed to assess the potential influence of reported index SNPs on the association between predicted gene expression and breast cancer risk. Furthermore, we evaluated whether the predicted expression levels of genes within a same genomic region were correlated with each other by using the OncoArray data.

INQUISIT algorithm scores for TWAS-identified genes. To evaluate whether there are additional lines of evidence supporting the identified genes as putative target genes of GWAS-identified risk SNPs beyond the scope of eQTL, we assessed their INQUISIT algorithm scores, which have been described elsewhere⁷. Briefly, this approach evaluates chromatin interactions between distal and proximal regulatory transcription-factor-binding sites and the promoters at the risk regions

using Hi-C data generated in HMECs⁶⁷ and chromatin interaction analysis by paired end tag in MCF7 cells. This could detect genome-wide interactions brought about by, or associated with, CCCTC-binding factor (CTCF), DNA polymerase II (POL2) and estrogen receptor (ER), all involved in transcriptional regulation⁶⁷. Annotation of predicted target genes used the Integrated Method for Predicting Enhancer Targets (IM-PET)⁶⁸, the Predicting Specific Tissue Interactions of Genes and Enhancers (PreSTIGE) algorithm⁶⁹, Hnisz⁷⁰ and FANTOM⁷¹. Features contributing to the scores are based on functionally important genomic annotations such as chromatin interactions, transcription factor binding and eQTLs. The detailed information for the INQUISIT pipeline and scoring strategy has been included in a previous publication⁷. In brief, besides assigning integral points according to different features, we also set up weighting and down-weighting criteria according to breast cancer driver genes, topologically associated domain boundaries and gene expression levels in relevant breast cell lines. Scores in the distal regulation category range from 0 to 7, and in the promoter category from 0 to 4. A score of 0 represents that no evidence was found for regulation of the corresponding gene.

Functional enrichment analysis using IPA. We performed functional enrichment analysis for the identified protein-coding genes reaching the Bonferroni-corrected association threshold. To assess the potential functionality of the identified lncRNAs, we examined their co-expressed protein-coding genes determined using expression data for normal breast tissue of European females in GTEx. Spearman's correlations between protein-coding genes and identified lncRNAs of ≥ 0.4 or ≤ -0.4 were used to indicate a high co-expression. Canonical pathways, top associated diseases and biofunctions, and top networks associated with genes of interest were estimated using IPA software⁷².

Gene expression in breast cell lines. Total RNA was isolated using the RNeasy Mini Kit (Qiagen) from 18 cell lines (Supplementary Table 8) that were authenticated using Promega's Geneprint 10 kit that conforms with ATCC standard ASN-0002-2011, and verified as free from viable Mycoplasma by using Lonza's Mycoalert kit. cDNA was synthesized using SuperScript III (Invitrogen) and amplified using the Platinum SYBR Green qPCR SuperMix-UDG cocktail (Invitrogen). Two or three primer pairs were used for each gene and the mRNA levels for each sample were measured in technical triplicates for each primer set. The primer sequences are listed in Supplementary Table 13. Experiments were performed using an ABI ViiA(TM) 7 System (Applied Biosystems), and data processing was performed using ABI QuantStudio Software V1.1 (Applied Biosystems). The average of Ct from all the primer pairs for each gene was used to calculate ΔC_t . The relative quantification of each mRNA normalized to that in 184A1 was performed using the comparative Ct method ($\Delta\Delta C_t$) and is summarized in Supplementary Fig. 4.

siRNA silencing. 184A1, MCF7 and T47D cells obtained from the American Type Culture Collection were reverse-transfected with siRNAs targeting GOI or a NTC siRNA (consi; Shanghai Genepharma) with RNAiMAX (Invitrogen) according to the manufacturer's protocol. Verification of siRNA knockdown of gene expression by qPCR was performed 36 h after transfection.

Proliferation and colony-formation assays. For proliferation assays, MCF7 and T47D cells were trypsinized at 16 h post-transfection and seeded into 24-well plates to achieve ~10% confluency. Phase-contrast images were collected with IncuCyte ZOOM (Essen Bioscience) for seven days. Duplicate samples were assessed for cells treated with siRNAs against each GOI along with cells treated with NTC siRNA in the same plate. 184A1 cells were reverse-transfected in 96-well plates to achieve 50% confluency at 8 h after transfection. Two independent experiments were carried out for all siRNAs in all three cell lines. Each cell proliferation time course was normalized to the baseline confluency and analyzed in GraphPad Prism. The area under the curve was calculated for each concentration ($n=4$) and used to calculate corrected proliferation (corrected proliferation (%) = $100 \pm$ (relative proliferation in indicated siRNA – proliferation in NTC siRNA)/knockdown efficiency ('+' if the GOI promotes proliferation and '-' if it inhibits proliferation)). For each gene, results from two siRNAs in two independent experiments were averaged and are summarized in Fig. 2 and Supplementary Fig. 6. For colony-formation assays, the same number of GOI siRNA-transfected MCF7 cells was seeded in 6-well plates at 16 h after transfection to assay CFE at two weeks. All siRNA-treated cells were seeded in duplicate. Colonies (defined to consist of at least 50 cells) were fixed with methanol, stained with crystal violet (0.5% w/v), scanned and counted using ImageJ as batch analysis by a self-defined plug-in Macro. Corrected CFE (%) = $100 \pm$ (relative CFE in indicated siRNA – CFE in NTC siRNA)/knockdown efficiency ('+' if the GOI promotes colony formation

and '-' if it inhibits colony formation). For each gene, results from two siRNAs in two independent experiments were averaged and are summarized in Fig. 2 and Supplementary Fig. 7. *P* values were determined by one-way analysis of variance (ANOVA) followed by Dunnett's multiple comparisons test.

Reporting Summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

Code availability. The computer codes used in our study are available upon reasonable request.

Data availability. The GTEx data are publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap; dbGaP Study Accession: [PHS000424.v6.p1](https://www.ncbi.nlm.nih.gov/gap/study/PHS000424.v6.p1)). TCGA data are publicly available via the National Cancer Institute's Genomic Data Commons Data Portal (<https://gdc.cancer.gov/>). A subset of the BCAC data that support the findings of this study is publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap; accession number [PHS001265.v1.p1](https://www.ncbi.nlm.nih.gov/gap/study/PHS001265.v1.p1)). Most of the BCAC data used in this study are or will be publicly available via dbGaP. Data from some BCAC studies are not publicly available due to restraints imposed by the ethics committees of individual studies; requests for further data can be made to the BCAC (<http://bcac.ccge.medschl.cam.ac.uk/>) Data Access Coordination Committee (DACC). BCAC DACC approval is required to access data from the studies ABCFS, ABCS, ABCTB, BBCC, BBSCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UT, BCINIS, BSUCH, CBCS, CECILE, CGPS, CTS, DIETCOMPLYE, ESTHER, GC-HBOC, GENICA, GEPARSIXTO, GESBC, HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KBCC, LMBC, MABCS, MARIE, MBCSG, MCBCS, MISS, MMHS, MTLGEBSCS, NC-BCFR, OFBCR, ORIGO, pKARMA, POSH, PREFACE, RBSCS, SKKDKFZS, SUCCESSB, SUCCESSC, SZBCS, TNBCC, UCIBCS, UKBGS and UKOPS.

References

- The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
- McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
- Delaneau, O., Marchini, J. & Zagury, J. F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012).
- 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).
- DeLuca, D. S. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics* **28**, 1530–1532 (2012).
- Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
- Guo, X., Lin, M., Rockowitz, S., Lachman, H. M. & Zheng, D. Characterization of human pseudogene-derived non-coding RNAs for functional potential. *PLoS One* **9**, e93972 (2014).
- Casbas-Hernandez, P. et al. Tumor intrinsic subtype is reflected in cancer-adjacent tissue. *Cancer Epidemiol. Biomark. Prev.* **24**, 406–414 (2015).
- Huang, X., Stern, D. F. & Zhao, H. Transcriptional profiles from paired normal samples offer complementary information on cancer patient survival – Evidence from TCGA pan-cancer data. *Sci. Rep.* **6**, 20567 (2016).
- Ghoussaini, M. et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nat. Genet.* **44**, 312–318 (2012).
- Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat. Genet.* **45**, 392–398 (2013).
- Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
- Freedman, M. L. et al. Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
- 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).
- 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).
- 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).
- Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
- The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).

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.

We did not use statistical methods to predetermine sample size. Rather, to maximize the statistical power for this study, for gene expression prediction model building and validation, we used the latest version of GTEx and TCGA data with the largest sample size at the time of study conduction. We used data from 122,977 cases and 105,974 controls of European ancestry, the largest available sample size to date (essentially all available GWAS data on breast cancer) for association analyses of predicted gene expression with breast cancer risk.

2. Data exclusions

Describe any data exclusions.

We excluded samples from non-European descendants as this study focus on European descendants. We used standard quality control protocols to exclude samples with poor data quality.

3. Replication

Describe whether the experimental findings were reliably reproduced.

For gene expression prediction models, both internal (cross validations) and external validations were performed. Bonferroni correction was used to reduce type 1 errors in association analyses. For identified significant associations with breast cancer risk, we performed stratified analyses in the subsets of iCOGS, OncoArray and GWAS sets. Finally, in vitro assays were performed to evaluate the function of selected genes identified in the association analyses. The functional experiments for prioritized genes were replicated as described in the methods section. All siRNAs were dissolved and aliquoted to avoid more than one freeze-thaw cycle. All attempts at replication were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

This is not relevant to our study, as our study is a genetic epidemiological study but not a randomized clinical trial.

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

For the in vitro functional experiments, the investigators were blinded to group allocation during data collection and analyses.

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.

R version 3.1.2; Plink v1.07; GCTA Version 1.26.0; IPA software (version 42012434); MetaXcan (version 0.2.5); Minimac3; SHAPEIT; ABI QuantStudio™ Software V1.1; GraphPad Prism, ImageJ

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

9. Antibodies

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

No antibodies were used

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

MCF7, T-47D and 184A1 cells were obtained from the American Type Culture Collection.

b. Describe the method of cell line authentication used.

Human cell lines were authenticated using Promega's Genepoint 10 kit that conforms with ATCC standard ASN-0002-2011.

c. Report whether the cell lines were tested for mycoplasma contamination.

All cell lines were verified as free from viable Mycoplasma by the using Lonza's Mycoalert kit.

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 commonly misidentified cell lines were used.

► 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.

No animals were used

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.

For GTEx project, data of female European subjects were used; For the BCAC studies, participants were female breast cancer patients or healthy female controls of European descent. Detailed relevant information for the subjects included in the BCAC has been described in a published study by Michailidou et al, Nature, 2017 (PMID:29059683) and referenced in the manuscript.