Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer

Manuel A. Ferreira et al.

Genome-wide association studies (GWAS) have identified more than 170 breast cancer susceptibility loci. Here we hypothesize that some risk-associated variants might act in non-breast tissues, specifically adipose tissue and immune cells from blood and spleen. Using expression quantitative trait loci (eQTL) reported in these tissues, we identify 26 previously unreported, likely target genes of overall breast cancer risk variants, and 17 for estrogen receptor (ER)-negative breast cancer, several with a known immune function. We determine the directional effect of gene expression on disease risk measured based on single and multiple eQTL. In addition, using a gene-based test of association that considers eQTL from multiple tissues, we identify seven (and four) regions with variants associated with overall (and ER-negative) breast cancer risk, which were not reported in previous GWAS. Further investigation of the function of the implicated genes in breast and immune cells may provide insights into the etiology of breast cancer.
Breast cancer is the most commonly diagnosed malignancy and most frequent cause of cancer-related mortality in women. Genome-wide association studies (GWAS) have detected more than 170 genomic loci harboring common variants associated with breast cancer risk, including 20 primarily associated with risk of ER-negative disease. Together, these common variants account for 18% of the two-fold familial relative risk of breast cancer.

To translate GWAS findings into an improved understanding of the biology underlying disease risk, it is essential to first identify the target genes of risk-associated variants. This is not straightforward because most risk variants lie in non-coding regions, particularly enhancers, many of which do not target the nearest gene. To help with this task, we recently developed a pipeline that identifies likely target genes of breast cancer risk variants based on breast tissue-specific genomic data, such as promoter–enhancer chromatin interactions and expression quantitative trait loci (eQTL).

Using this approach, called INQUISIT, we identified 689 genes as potential targets of the breast cancer risk variants. However, it is likely that at least some breast cancer risk variants can regulate gene expression in tissues other than breast, which were not considered by INQUISIT; for example, breast cancer risk variants are enriched in histone marks measured in adipose tissue. On the other hand, the immune system also plays a role in the elimination of cancer cells so it is possible that some breast cancer risk variants influence the expression of genes that function in the immune system.

The first aim of this study was to identify additional likely target genes of the breast cancer risk variants identified by the Breast Cancer Association Consortium (i.e. using information on eQTL in multiple relevant tissue types: adipose, breast, immune cells, spleen, and whole-blood). The second aim was to identify previously unreported risk loci for breast cancer by formally integrating eQTL information across tissues with results from the GWAS, using EUGENE, a recently described gene-based test of association, that is, conceptually similar to other transcriptome-wide association study (TWAS) approaches, such as PrediXcan. Gene-based analyses would be expected to identify previously unreported risk loci if, for example, multiple independent eQTL for a given gene are individually associated with disease risk, but not at the genome-wide significance level used for single-variant analyses.

Results

Predicted target genes of overall breast cancer risk variants. Using approximate joint association analysis implemented in GCTA (see Methods), we first identified 212 variants that were independently associated (i.e. with GCTA-COJO joint analysis $P < 5 \times 10^{-8}$) with breast cancer in a GWAS dataset of 122,977 cases and 105,974 controls (Supplementary Data 1). Of note, 20 of these variants reached genome-wide significance in the joint, but not in the original single-variant, association analysis; that is, they represent secondary signals that were masked by the association with other nearby risk variants, as described previously.

We extracted association summary statistics from 117 published eQTL datasets identified in five broad tissue types: adipose, breast, individual immune cell types, spleen and whole-blood (Supplementary Data 2). For each gene and for a given eQTL dataset, we identified cis eQTL (within 1 Mb of gene boundaries) in low linkage disequilibrium (LD; $r^2 < 0.05$) with each other, and with an association with gene expression significant at a conservative significance threshold of $8.9 \times 10^{-10}$. We refer to these as “sentinel eQTL.” The mean number of sentinel eQTL per gene ranged from 1.0 to 2.9 across the 117 eQTL datasets considered, which varied considerably in sample size and number of genes tested (Supplementary Data 2).

When we intersected the list of variants from the joint association analysis and the list of sentinel eQTL from published datasets, we identified 46 sentinel risk variants that were in high LD ($r^2 > 0.8$) with one or more sentinel eQTL, implicating 88 individual genes at 46 loci as likely targets of breast cancer risk variants (Supplementary Data 3 and 4). Twenty-five risk variants had a single predicted target gene, 10 had two, and 11 had three or more (Supplementary Data 5).

Of the 88 genes, 75 (85%) were identified based on eQTL from whole-blood, 10 (11%) from immune cells (PEX14, RNF115, TNNT3, EFEMP2, SDHA, AP4B1, BCL2L15, BTN3A2, HIST1H2BL, SYNE1), and three (4%) exclusively from adipose tissue (ZNF703, HAPLN4, TM6SF2) (Supplementary Data 4). Only four sentinel risk variants were in LD with a sentinel eQTL in breast tissue (for ATG10, PIDD1, RCCD1, and AP0BEC3B); all were also eQTL in whole-blood and immune cells. However, it is noteworthy that an additional 29 sentinel eQTL listed in Supplementary Data 3 had a modest, yet significant association with the expression of the respective target gene in breast tissue (GTEx V7, $n = 251$), suggesting that larger eQTL datasets of this tissue will be informative to identify the target genes of sentinel risk variants.

A total of 62 genes were included in the list of 925 targets predicted in the original GWAS using INQUISIT, while 26 genes represent previously unreported predictions (Supplementary Data 5). Regional association plots for these 26 genes are presented in Supplementary Fig. 1, with three examples shown in Fig. 1.

Directional effect of gene expression on breast cancer risk. For the 88 genes identified as likely targets of breast cancer risk variants, we studied the directional effect of genetically-determined gene expression on disease risk, based on the sentinel eQTL that was in LD with the sentinel risk variant. For each gene, we first determined whether the eQTL allele that was associated with reduced breast cancer risk was associated with higher or lower target gene expression. Of the 77 genes for which this information could be obtained (detailed in Supplementary Data 4), the protective allele was associated with lower expression for 43 genes (e.g. GATAD2A, FAM175A, KCN4, and CTB-161K23.1) and higher expression for 28 genes (e.g. RCCD1, ATG10, ELL, and TLR1) (summarized in Table 1 and Supplementary Data 6). For the remaining six genes (ADCT3, AMFR, AP0BEC3B, CCDC127, HSPA4, and MRPS18C), conflicting directional effects were observed across different tissues, and so the interpretation of results is not straightforward.

Directional effect based on information from multiple eQTL. In the previous analysis, the directional effect of gene expression on disease risk was assessed based on a single eQTL at a time. However, the expression of most genes is associated with multiple independent eQTL, which may not have the same directional effect on disease risk. To address this limitation, we assessed if results from the single QTL analyses above were recapitulated by considering information from multiple eQTL using S-PrediXcan. We applied this approach to the same GWAS results and used transcriptome prediction models from whole-blood, generated based on data from the Depression Genes and Network study ($n = 922$) and GTEx ($n = 369$). We used SNP prediction models for gene expression in whole-blood because most genes (75 of 88) were identified as likely targets based on eQTL information from this tissue.

Results from this analysis are presented in Supplementary Data 7. The predicted directional effect of gene expression on disease risk was available in both the single eQTL and
S-PrediXcan analyses for 48 of the 88 likely target genes. For 42 of those 48 genes (88%) the two predictions matched, supporting a consistent directional effect across multiple eQTL of the same gene. The inconsistent results observed for the remaining six genes were likely caused by technical biases (possible explanations in Supplementary Data 7). Similar findings were obtained when considering whole-blood transcriptome prediction models based on data from the GTEx consortium (Supplementary Data 7). Overall, these results indicate very good agreement between the directional effect of gene expression on disease risk obtained using information from individual or multiple eQTL.
Target gene predictions supported by functional data. The 88 genes identified represent target genes that should be experimentally validated, as outlined previously\cite{12}. To prioritize genes for functional follow-up, we identified a subset for which publicly available functional data supported the presence of either (i) chromatin interactions between an enhancer and the gene promoter\cite{14,15}, or (ii) an association between variation in enhancer epigenetic marks and variation in gene expression levels\cite{16-19}. We only considered enhancers that overlapped a sentinel risk variant (or a proxy with $r^2 > 0.80$) and restricted our analysis to blood cells (Supplementary Data 8), given that most target genes were identified based on eQTL data from whole-blood. We found that 25 (28%) of the 88 target gene predictions were supported by functional data (Supplementary Data 9).

Previously unreported risk loci for breast cancer. The second major goal of this study was to identify previously unreported risk loci for breast cancer using gene-based association analyses. We first used approximate conditional analysis implemented in GCTA\cite{9} to adjust the GWAS results\cite{2} (Fig. 2a) for the effect of the 212 variants that had a significant independent association with overall breast cancer. As expected, in the resulting adjusted GWAS no single variant had a genome-wide significant association (i.e. all had a GCTA-COJO conditional association $P > 5 \times 10^{-8}$; Fig. 2b). We then applied the EUGENE gene-based approach\cite{6,7} to the adjusted GWAS results, considering in a single association analysis cis eQTL identified in five broad tissue types: adipose, breast, immune cells, spleen, and whole-blood (Supplementary Data 10). That is, we did not perform a separate gene-based analysis for each tissue, but rather a single analysis that considers all eQTL reported across the five tissues.

Of the 19,478 genes tested (full results provided as Supplementary Material), 11 had a significant gene-based association after correcting for multiple testing (EUGENE $P < 0.05/19,478 = 2.5 \times 10^{-6}$; Table 2; Fig. 2c). The specific eQTL included in the gene-based test for each of these 11 genes, which were located in six loci >1 Mb apart, are listed in Supplementary Data 10. Regional association plots for the 11 genes are presented in Supplementary Fig. 2, with three examples shown in Fig. 3. Except for the MAN2C1 locus\cite{20}, these loci have not previously been identified by GWAS and thus represent putative breast cancer susceptibility loci.

For most (9 of 11) genes identified, the association $P$-value obtained with the gene-based test was more significant than the $P$-value obtained with the individual eQTL most associated with disease risk, indicating that multiple sentinel eQTL for the same gene were associated with disease risk (range 2–6 associated eQTL per gene; Table 2). For example, the EUGENE gene-based $P$-value for GSTM2 was $6.6 \times 10^{-8}$, while the best individual eQTL showed more moderate association with breast cancer risk (GCTA-COJO conditional association $P = 4.1 \times 10^{-5}$; five of the 14 sentinel eQTL tested for this gene were nominally associated with disease risk (Supplementary Data 11).

We also studied the predicted directional effect of gene expression on disease risk, as described above for target genes of known breast cancer risk variants. When we considered information from all eQTL associated with disease risk for each of the 11 genes (Supplementary Data 11), we found that decreased disease risk was consistently associated with decreased expression for three genes and increased expression for five genes (Table 3 and Supplementary Data 12). For the remaining three genes, inconsistent directional effects were observed across different eQTL.

Lastly, we used EUGENE to determine if any of the 88 target genes of sentinel risk variants identified based on individual eQTL also had a significant gene-based association in the adjusted GWAS results. This would indicate that information from additional breast cancer risk variants (i.e. in low LD with the sentinel risk variants) supported the original target gene prediction, which could be used to prioritize genes for functional follow-up. We found that 11 of the 88 target genes had a nominally significant gene-based association in the adjusted GWAS results (EUGENE $P < 0.05$; Supplementary Data 13), with one remaining significant after correcting for multiple testing: CBX6 (EUGENE $P = 0.0002$).

**Table 1** Directional effect of genetically determined gene expression on disease risk for predicted target genes of breast cancer sentinel risk variants

| Directional effect | Predicted target genes of breast cancer sentinel risk variants |
|--------------------|-------------------------------------------------------------|
| Decreased expression associated with decreased risk | AC007283.5, AHRR, AP006621.5, AP006621.6, AP0BEC3B-AS1, ARRD3C, ASCCC2, BCL2L15, BTN2A1, CCDC170, CCDC91, CDCA7L, CENDI, CES1, COX11, CTB-161K23.1, CTD-2116F7.1, CYP5A1, DDA1, DFFA, EFEMP2, ENPP7, FAM175A, GATA2D2A, HAPLN4, HCG11, HISTH41, KCN5N, LCR2C5, LRRD1, LGFD1, PDD1, PPIF3, PTPN22, RPS23, SIRT5, SMAD5, TGFBR2, TM6SF2, TMEM184B, TNF, ZBTB38, ZNF703 |
| Increased expression associated with decreased risk | AC010335.1, AKA9, AP0BEC3A, AT77TP, ATG10, ATP6AP1L, BTN2A3BP, CBX6, CENP0, CFL1, COQ5, CTD-317HS5.1, DCLRE1B, DNAJC27, ELL, ESRT1, HLF, L3MBTL3, NUDT71B, PGPEP1, RCM1, RHBD3, RNF154, RPI1-486M23.2, SIAV1, SYNE1, TERF, TR1 |
| Ambiguous | ADCY3, AMFR, AP0BEC3B, CCDC127, HSPA4, MRPS18C |
Estrogen receptor (ER)-negative breast cancer. We applied the same analyses described above to results from the Milne et al. GWAS of ER-negative breast cancer, which included data on 21,468 cases and 100,594 controls, combined with 18,908 BRCA1 mutation carriers (9414 with breast cancer)3. Of the 54 sentinel risk variants identified through approximate joint association analysis (Supplementary Data 14), 19 were in LD ($r^2 > 0.8$) with a sentinel eQTL (Supplementary Data 15), implicating 24 genes as likely targets of risk-associated variants for ER-negative breast cancer (Supplementary Data 16). Of these, 13 were also identified as likely targets of variants associated with overall breast cancer risk, while the remaining 11 genes were specific to ER-negative risk variants: ATM, CCNE1, CUL5, MCHR1, MDM4, NPAT, OCEL1, PIK3C2B, RALB, RP5-855D21.3, and WDR43.

Seventeen genes were not highlighted as candidate target genes in the Milne et al. GWAS3 (Supplementary Data 16 and Supplementary Data 17), mostly (15 genes) because they are predicted targets of risk variants identified in previous GWAS, which were not considered by Milne et al.3. The two exceptions were RP5-855D21.3 and CUL5, identified in our study based on eQTL from adipose tissue and whole-blood, respectively. Regional association plots for the 17 genes that represent previously unreported predictions are presented in Supplementary Fig. 3, with three examples shown in Fig. 4.

The disease protective allele was associated with lower gene expression for seven genes and higher gene expression for 11 genes (summary in Table 4 and Supplementary Data 18; detailed information in Supplementary Data 15); for the remaining six genes, directional effect was either not available (ATM, CASP8, OCEL1, PEX14 and WDR43) or inconsistent across tissues (ADCY3).

Of the 24 target gene predictions, 18 were supported by the presence of enhancer–promoter chromatin interactions or an association between enhancer epigenetic marks and gene expression (Supplementary Data 19).

**Fig. 2** Manhattan plots summarizing association results for overall breast cancer. a Association results ($-\log_{10}$ of the $P$-value obtained using an inverse-variance meta-analysis) from the single-variant GWAS originally reported by Michailidou et al2. b Single-variant GWAS adjusted for 212 sentinel risk variants and LD-score intercept; $P$-values were obtained with the GCTA-COJO joint analysis. c Gene-based analysis of adjusted GWAS results; $P$-values were obtained with the EUGENE gene-based test of association.
When we applied EUGENE to the ER-negative GWAS results obtained after conditioning on the 54 sentinel risk variants, we identified four genes in four loci with a significant gene-based association (EUGENE P < 2.5 × 10^{-6}; Table 5, Supplementary Data 20 and Supplementary Fig. 4). Of these, we found that lower disease risk was consistently associated with lower expression for two genes (VPS52, GTF2IRD2B) and higher expression for one gene (INHBB). For the fourth gene (TNFSF10), directional effect was inconsistent across sentinel eQTL (detail and summary in Supplementary Tables 21 and 22, respectively).

Other genes that could be prioritized for functional follow-up include four (of the 24) target genes of sentinel risk variants that had a nominally significant gene-based association in the adjusted GWAS results (EUGENE P < 0.05; Supplementary Data 23): RALB, CCDC170, NPAT, and CASP8.

### Known role of the identified genes in cancer biology
We used OncoScore, a text-mining tool that ranks genes according to their association with cancer based on available biomedical literature, to assess the extent to which each of the breast cancer genes we identified were already known to have a role in cancer. Of the 112 genes we identified across the overall and ER-negative analyses that could be scored by OncoScore, 48 scored below the recommended OncoScore cut-off threshold (21.09) for novelty, including 25 with an OncoScore of 0, indicating no prior evidence for a role in cancer biology (Tables 2 and 5; Supplementary Tables 3 and 16). For the remaining 64 genes there is an extensive literature on their role in cancer, and breast cancer in particular.

### Discussion
To predict candidate target genes at breast cancer risk loci, we identified sentinel eQTL in multiple tissues that were in high LD \( (r^2 > 0.8) \) with sentinel risk variants from our recent GWAS. Using this approach, we implicated 88 genes as likely targets of the overall breast cancer risk variants. Because eQTL are widespread, it is possible that some target gene predictions are false-positives due to coincidental overlap between sentinel eQTL and sentinel risk SNPs. At the LD threshold used, statistical methods developed recently to formally test for co-localization between eQTL and risk SNPs are of limited use, due to a high false-positive rate. The 88 genes identified therefore represent target predictions that must be validated by functional studies. Of these, 88, 26 genes had not been predicted as targets using a different approach that considered breast-specific functional annotations and eQTL data, and so were considered previously unreported candidate target genes.

Of the 26 previously unreported target predictions, all but one were identified from eQTL analyses in blood, spleen, or immune cells. They include several genes with a known role in immunity, including: HLF, the expression of which is associated with the extent of lymphocytic infiltration after neo-adjuvant chemotherapy; PTPN22, a shared autoimmunity gene, which encodes a protein tyrosine phosphatase that negatively regulates presentation of immune complex-derived antigens; and RHBDD3, a negative regulator of TLR3-triggered natural killer cell activation and the 112 critical regulator of dendritic cell activation.

In addition, we identified IRF1, which encodes a tumor suppressor and transcriptional regulator serving as an activator of genes involved in both innate and acquired immune response, and ONCO1, as a previously unreported breast cancer risk locus. These results suggest that at least some of the previously unreported predicted target genes play a role in cancer cell elimination or inflammation. However, another possibility is that eQTL detected in well-powered studies of blood are predictors of eQTL in other less accessible tissues, including breast and adipose tissue. Consistent with this possibility, about 50% of the eQTL found to be in LD with a sentinel risk variant for overall breast cancer (and similarly for ER-negative breast cancer) were associated with the expression of the respective target gene in the relatively small TCGA breast tissue dataset, although not at the conservative threshold that we used to define sentinel eQTL. Not only, one previously unreported target was identified through eQTL analyses in adipose tissue: ZNF703. ZNF703 is a known oncogene in breast cancer and has been reported to be associated with breast size, which might suggest a role in adiposity.

Using the same approach, we also identified 24 genes as likely targets of 19 ER-negative risk variants, of which 17 were not proposed as candidate target genes in the original GWAS. Eleven of these 22 genes were unique to ER-negative breast cancer, including for example CUL5, a core component of multiple SCF-like ECS (Elongin-Cullin 2/5-SOCS-box protein) E3 ubiquitin-protein ligase complexes which recognize proteins for degradation and subsequent Class I mediated antigen presentation.

We also identified previously unreported breast cancer risk loci using the recently described EUGENE gene-based association test, which was developed to aggregate evidence for association with a disease or trait across multiple eQTL. Unlike other similar gene-based methods (e.g. S-PrediXcan), EUGENE includes in a single test information from eQTL identified in multiple tissues.

### Table 2 Risk loci for breast cancer identified in the EUGENE gene-based analysis but not in previous GWAS

| Locus index | Gene | Chr | Start | N sentinel eQTL | Gene-based association in the adjusted GWAS | Sentinel eQTL with strongest association in the adjusted GWAS | OncoScore |
|-------------|------|-----|-------|-----------------|---------------------------------------------|-------------------------------------------------------------|------------|
| 1           | GSTM2 | 1   | 110210644 | 14 5 | 6.63E-08       | rs621414       | 4.08E-05 | 38.97       |
| 2           | SEMA44 | 1   | 156171757 | 9 5  | 1.45E-07       | rs887853       | 2.99E-06 | 27.04       |
| 3           | LINC0086 | 3  | 156465135 | 1 1  | 2.34E-06       | rs7641929      | 2.43E-06 | N/A         |
| 4           | ACO34220.3 | 3  | 13164978 | 7 4  | 5.92E-07       | rs1179622      | 0.000314 | N/A         |
| 5           | IRF1  | 5   | 131817301 | 2 2  | 4.99E-07       | rs12548999     | 3.44E-05 | 42.46       |
| 6           | SIK2  | 5   | 111473115 | 2 2  | 4.09E-07       | rs527078       | 3.22E-05 | 39.57       |
| 7           | PPP2R1B | 5  | 111597632 | 8 2  | 2.31E-06       | rs680096       | 2.91E-06 | 56.52       |
| 8           | MAN2C1 | 5   | 75641333 | 14 6 | 1.91E-06       | rs8028277      | 2.16E-06 | 26.66       |
| 9           | RPL11-8(17013.8) | 15 | 7560496 | 4 4  | 3.83E-08       | rs4545784      | 3.85E-06 | N/A         |
| 10          | S10A3 | 5   | 75661720 | 4 4  | 3.83E-08       | rs4545784      | 3.85E-06 | 42.43       |
| 11          | IMP3  | 5   | 75931426 | 1 1  | 2.30E-06       | rs4886708      | 2.30E-06 | 80.41       |

*Gene-based association P-value obtained when the EUGENE gene-based test was applied to the adjusted GWAS results.

P-value obtained when the EUGENE gene-based test was applied to the adjusted GWAS results.
this property is expected to increase power to detect gene associations when multiple cell types/tissues contribute to disease pathophysiology, for two main reasons. First, because tissue-specific eQTL are common, and so a multi-tissue analysis is able to capture the association between all known eQTL and disease risk in a single test. Second, because in single-tissue analyses, one needs to appropriately account for testing multiple tissues, thereby decreasing the significance threshold required for experiment-wide significance, which decreases power. When we applied EUGENE to the overall breast cancer GWAS\textsuperscript{2}, we identified 11 associated genes located in six previously unreported risk loci. For most of these genes, there were multiple sentinel
eQTL associated with overall breast cancer risk. In the analysis of ER-negative breast cancer\(^3\), EUGENE identified four associated genes (INHBB, TNFSF10, VPS52, and GTF2IRD2B) located in four previously unreported risk loci.

Some of the predicted target genes identified are well known to play a role in breast cancer carcinogenesis. For example, the genes identified for ER-negative breast cancer included MDM4, encoding a negative regulator of TP53, which is necessary for normal breast development\(^33\); CCNE1, an important oncogene in breast cancer\(^34,35\); CASP8, encoding a regulator of apoptosis\(^36\); ATM, a known breast cancer susceptibility gene\(^37,38\); and the ER, ESR1, which encodes a critical transcription factor in breast tissue\(^39\). On the other hand, the 11 significant gene-based associations for overall breast cancer included GSTM2, which is part of the mu class of glutathione S-transferases that are involved in increased susceptibility to environmental toxins and carcinogens\(^40\). Other noteworthy gene-based associations included those with: IMP3, which contributes to self-renewal and tumor initiation, properties associated with cancer stem cells\(^41\); PPP2R1B, which encodes the beta isoform of subunit A of Protein Phosphatase 2A, itself a tumor suppressor involved in modulating estrogen and androgen signaling in breast cancer\(^42\); and SEMA4A, recently shown to regulate the migration of cancer cells as well as dendritic cells\(^43\).

Two recent studies reported results from analyses that are similar in scope to those carried out in our study. First, Hoffman et al.\(^44\) reported that genetically determined expression of six genes was associated with risk of breast cancer: three when considering expression in breast tissue (RCCD1, DHODH, and ANKLE1) and three in whole blood (RCCD1, ACAPI, and LRRC25). Of note, RCCD1 and LRRC25 were identified as likely targets of known breast cancer risk variants in our analysis. We also found some support for an association between breast cancer risk and eQTL for ACAPI (EUGENE \(P = 0.003\)) and ANKLE1 (EUGENE \(P = 0.01\)), but not for DHODH (best sentinel eQTL \(P = 0.119\)). Second, we recently applied a different gene-based approach called S-PrediXCan to results from the overall breast cancer GWAS\(^5\), using gene expression levels predicted from breast tissue\(^20\).

This study reported significant associations with 46 genes \((P < 5.82 \times 10^{-6})\), including 13 located in 10 regions not yet implicated by GWAS. A major difference between our analyses was that the latter were based on the original GWAS summary statistics, without adjusting for the effects of the sentinel risk variants. This explains why most associated genes in their main analysis were located near known breast cancer risk variants. Of the 13 genes located in previously unreported risk loci, eight were tested in our analysis (which considered eQTL identified in multiple tissues, not just from breast as in ref.\(^20\)), of which four had a nominally significant \((P < 0.05)\) gene-based association: MAN2C1 \((P = 1.9 \times 10^{-6})\), SPATA18 \((P = 0.004)\), B3GNT1 \((P = 0.012)\), and CTD-2323K18.1 \((P = 0.021)\). These results show that at least four of the associations reported by Wu et al.\(^20\), which were based on information from breast eQTL only, are reproducible when a different gene-based approach is applied to the same GWAS results. Conversely, we identified a significant association with 13 genes not reported by Wu et al.\(^20\), all with a gene-based association driven by eQTL identified in non-breast tissues, mostly in immune cells and/or whole-blood. For 78 of the 114 genes that we implicate in breast cancer risk, either through target gene prediction or gene-based analyses, we were able to determine the directional effect of the breast cancer protective alleles on gene expression. In some cases, this was consistent with their known function. For example, ZNF703 is a well-known oncogene in breast cancer\(^30\) and decreased expression was associated with decreased risk. Similarly, oncogenic activity has been reported for PTK2CB2\(^46\), for which we found that decreased expression is associated with decreased risk. Another gene for which decreased expression was associated with decreased risk was PTPN22 which is known to negatively regulate antigen presentation\(^47\) and therefore might suppress immunoeilination. By contrast, CCNE1\(^48\) and APOBEC3A\(^49\) have been reported to have oncogenic roles, but we found that increased expression was associated with decreased risk. We have previously found the same counterintuitive relationship between breast cancer risk alleles and CCND1 expression\(^50\). However, the expression patterns observed in breast tumors may not be relevant to the activity of these genes in the progenitor cells that give rise to breast tumors.

The directional effect of genetically determined gene expression on breast cancer risk is important because drugs that mimic the effect of the protective allele on gene expression might be expected to attenuate (rather than exacerbate) disease risk. For example, decreased risk of ER-negative breast cancer was associated with decreased expression of KCNN4, suggesting that an antagonist that targets this potassium channel and has a good safety profile\(^51\) might reduce disease risk. Given these results, we suggest that KCNN4 should be prioritized for functional and pre-clinical follow up.

### Table 3 Directional effect of genetically determined gene expression on disease risk for genes identified in the gene-based analysis of the adjusted breast cancer GWAS

| Direction of effect | Predicted target genes of breast cancer sentinel risk variants |
|---------------------|---------------------------------------------------------------|
| Decreased expression associated with decreased risk | IMP3, IRF1, SEMA4A |
| Increased expression associated with decreased risk | LINCO00886, MAN2C1, RTN1-1870138, SIK2, SIN3A |
| Ambiguous | AC034220.3, GSTM2, PPP2R1B |
Fig. 4 Examples of previously unreported target gene predictions at known ER-negative breast cancer risk loci. Variants are represented by points colored according to the LD with the sentinel risk variant (red: ≥0.8, orange: 0.6–0.8, green: 0.4–0.6, light blue: 0.2–0.4, and dark blue: <0.2). Sentinel risk variants (triangles) were identified based on joint association analysis. Figure shows on the y-axis the evidence for ER-negative breast cancer association (−log_{10} of the P-value in the original published GWAS results, obtained in that study using an inverse-variance meta-analysis), and on the x-axis chromosomal position. Gene structures from GENCODE v19 gene annotations are shown and the predicted target genes shown in red. The sentinel risk variants are in LD with sentinel eQTL for MDM4 and PIK3C2B (a), ZNF703 (b), and ATM (c; Supplementary Data 17). Regional association plots for the remaining 14 previously unreported target gene predictions are provided in Supplementary Figure 3.
In summary, we have used the largest available GWAS of breast cancer, along with expression data from multiple different tissues, to identify 26 and 17 previously unreported likely target genes of known overall and ER-negative breast cancer risk variants, respectively. We also describe significant gene-based associations at six and four previously unreported risk loci for overall and ER-negative breast cancer, respectively. Further investigation into the function of the genes identified in breast and immune cells, particularly those which have additional support from experimental or computational predictions of chromatin looping, should provide additional insight into the etiology of breast cancer.

**Methods**

**Predicting target genes of breast cancer risk variants.** Recently, Michailidou et al. reported a breast cancer GWAS meta-analysis that combined results from 13 studies: the OncoArray study (61,282 cases and 45,494 controls); the iCOGS study (46,785 cases and 42,892 controls); and 11 other individual GWAS (with a combined 14,910 cases and 17,588 controls). That is, a total of 122,977 cases and 105,974 controls. We refer to these as sentinel risk variants identified in that GWAS.

First, we identified variants associated with variation in gene expression (i.e., eQTL) in published transcriptome studies of five broad tissue types: adipose, breast, immune cells isolated from peripheral blood, spleen and whole-blood. We identified a total of 35 transcriptome studies reporting results from eQTL analyses in any one of those five tissue types (Supplementary Data 2). Some studies included multiple cell types and/or experimental conditions, resulting in a total of 86 separate eQTL datasets. For each eQTL dataset, we then (i) downloaded the original publication tables containing results for the eQTL reported; (ii) extracted the variant identifier, gene name, association P-value and, if available, the effect size (specifically, by “effect size” we mean the beta/z-score and corresponding allele; (iii) excluded eQTL located >1 Mb of the respective gene (i.e., trans eQTL), because often these are thought to be mediated by cis effects; (iv) excluded eQTL with an association P > 8.9 × 10^{-10}, a conservative threshold that corrects for 55,764 transcriptomic genes in human cells; and (v) for each gene, used the -clamp procedure in PLINK to reduce the list of eQTL identified (which often included many correlated variants) to a set of sentinel eQTL, defined as the variants with strongest association with gene expression and in low LD (r^2 < 0.05, linkage disequilibrium (LD) window of 2 Mb) with each other.

Second, we identified variants that were independently associated with breast cancer risk at a rs > 0.5 × 10^{-10} in the GWAS reported by Michailidou et al. 2, which included 122,977 cases and 105,974 controls. We refer to these as sentinel risk variants for breast cancer. To identify independent associations, we first excluded from the original GWAS (which tested 12,396,529 variants) variants with: (i) a sample size < 130,000; (ii) a minor allele frequency < 1%; (iii) not present in, or not polymorphic (Europeans in), or with alleles that did not match, data from the 1000 Genomes project (release 20130502_v5a); and (iv) not present in, or with alleles that did not match, data from the UK Biobank study. After these exclusions, results were available for 8,248,946 variants. Next, we identified sentinel risk variants using the joint association analysis (–cojo-sct) option of GCTA, using imputed data from 5,000 Europeans from the UK Biobank study to calculate LD. These individuals were selected based on the sample IDs (lowest 5,000) from our approved UK Biobank application 25331. Third, we identified genes for which a sentinel eQTL reported in any of the 86 eQTL datasets described above was in high LD (r^2 > 0.8) with a breast cancer sentinel risk variant. That is, we only considered genes for which there was high LD between a sentinel eQTL and a sentinel risk variant, which reduces the chance of spurious co-localization.

Directional effect of gene expression on breast cancer risk. Having identified a list of genes with expression levels correlated with sentinel risk variants, we then studied the directional effect of the breast cancer protective allele on gene expression. For each sentinel eQTL in high LD (r^2 > 0.8) with a sentinel risk variant, we: (i) identified the allele that was associated with reduced breast cancer risk, based on results reported by Michailidou et al. 2; and (ii) determined if that allele was associated with increased or decreased target gene expression in each of the eQTL datasets that reported that eQTL. For many studies, the directional effect of eQTL (i.e., effect allele and beta) was not publicly available, and so for those this analysis could not be performed.

We also assessed whether the directional effect of gene expression on disease risk predicted by the approach described in the previous paragraph, which considered one eQTL at a time (a limitation) but many different eQTL datasets (a strength), would be recapitulated by applying S-PrediXcan to the same breast cancer GWAS using transcriptome information from 922 whole-blood samples studied by Battle et al. 2. S-PrediXcan considers information from multiple eQTL identified for a given gene in a given tissue (e.g., whole-blood) when determining the association between genetically determined gene expression levels and disease risk. Therefore, we reasoned that this approach could be particularly useful for genes with multiple independent eQTL identified in the same tissue. The limitation of this approach is that it first requires the generation of gene expression prediction models based on individual-level variant and transcriptome data, which are not publicly available for most of the 35 transcriptome studies included in our analysis. We used gene expression models generated based on the whole-blood dataset of Battle et al. because (i) most likely target genes were identified in our study based on eQTL information from whole-blood or immune cells isolated from whole-blood; and (ii) this was the largest transcriptome dataset we had access to.

**Target gene predictions supported by functional data.** Sentinels and variants in high LD (r^2 > 0.8) in Europeans of the 1000 Genomes Project, with MAF > 0.01) were queried against the following sources of publicly available data generated from blood-derived samples and cell lines. Computational methods linking regulatory elements with target genes including PreSTIGE, FANTOM5, IMPETE, enhancers and super enhancers from Hnisz et al. 21. Experimental chromatin looping data defined by ChiA-PET and capture Hi-C, and in situ Hi-C were mined to identify physical interactions between query SNPs and target gene promoters. Variants were assigned to potential target genes based on intersection with associated enhancer annotations using BedTools intersect.
Identification of previously unreported risk loci for breast cancer. The second aim of this study was to use a gene-based approach to identify loci containing breast cancer risk variants that were missed by the single-variant analyses reported by Michailidou et al. At least three gene-based approaches have been described recently to combine in a single test the evidence for association with a disease across multiple eQTLs. Of these, we opted to use EUGENE because it is applicable to GWAS summary statistics and combines in the same association test information identified in different tissues and/or transcripts associated with the same tissue (e.g. blood) identify partially (not completely) overlapping lists of eQTL. This might arise, for example, because of differences in sample size, gene expression quantification methods (e.g. microarrays vs. RNA-seq data) or demographics of the ascertained individuals (e.g. age, disease status). Therefore, identifying eQTL based on information from multiple tissues and/or studies is expected to produce a more comprehensive list of regulatory variants that could be relevant to breast cancer pathophysiology. An additional advantage of EUGENE is that it considers in the same test different types of eQTL (e.g. with exon-specific or stimulus-specific effects), thereby increasing the likelihood that causal regulatory variants related to breast cancer are captured in the analysis.

EUGENE requires an input file that lists all eQTL that will be included in the gene-based test for each gene in the genome. First, we took the union of all eQTL reported in the 86 eQTL datasets described above. Second, we used the --clump procedure in PLINK to reduce the list of reported eQTL to a set of ‘sentinel eQTL’, defined as the variants with strongest association with gene expression and in low LD ($r^2 < 0.1$, LD within a block of 1 Mb) with each other. Note that clumping was not performed separately for each tissue or study, but rather applied to the union of eQTL identified across all tissues/studies. If an eQTL was identified in multiple tissues/studies, the clumping procedure was performed using the smallest P-value reported for that eQTL across all tissues/studies. A file (BRCA20170517. eqtl.proxies) containing the sentinel eQTL identified per gene is available at [https://genepi.qimr.edu.au/staff/manuel/eugene/main.html](https://genepi.qimr.edu.au/staff/manuel/eugene/main.html).

For each gene, EUGENE extracts single-variant association results for each sentinel eQTL identified (or, if not available, for a proxy with $r^2 > 0.8$) from the GWAS summary statistics, sums the association chi-square values across those eQTL, and estimates the significance (i.e. P-value) of the resulting sum test statistic using Satterthwaite’s approximation, which accounts for the LD between eQTL. This approximation was originally implemented by Bakshi et al. in the GCTA-fastBAt module and is now also available in EUGENE. LD between eQTL was estimated based on data from 294 Europeans from the 1000 Genomes Project (release 20130502, v5a).

Because our aim was to identify previously unreported breast cancer risk loci, we did not apply EUGENE to the original results reported by Michailidou et al. Had we done so, significant gene-based associations would have been disproporionately local to known breast cancer risk loci; associations driven by previously unreported risk variants would therefore be more difficult to highlight. Instead, we first adjusted the results for the effects of the sentinel risk variants identified (see section above), using the --cojo-cond option of GCTA. In doing so, we obtained 18,908 BRCA1 and BRCA2 mutations across all 1,005,594 controls from the Breast Cancer Association Consortium, combined with 18,908 BRCA1 mutation carriers (9,414 with breast cancer) from the Consortium of Investigators of Modifiers of BRCA1/2, all tested for 17,304,475 variants (9,827,195 after the exclusions described above). The LD-score regression intercept used to correct the single-variant association statistics of this GWAS was 1.0637.

Code availability

The EUGENE gene-based approach is implemented in C++ and is available for download from [https://genepi.qimr.edu.au/staff/manuel/eugene/main.html](https://genepi.qimr.edu.au/staff/manuel/eugene/main.html).

Received: 18 May 2018 Accepted: 14 December 2018

Published online: 15 April 2019

References

1. Ginsburg, O. et al. The global burden of women’s cancers: a grand challenge in global health. Lancet 389, 847–860 (2017).
2. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature 551, 92–94 (2017).
3. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat. Genet. 49, 1767–1778 (2017).
4. Javierre, B. M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. Cell 167, 1369–1384 e19 (2016).
5. Teng, M. W., Swann, J. B., Koebel, C. M., Schreiber, R. D. & Smyth, M. J. Immune-mediated dormancy: an equilibrium with cancer. J. Exp. Med. 84, 988–993 (2008).
6. Ferris, E. A. et al. Genotype-based analysis of regulatory variant identification in putative novel asthma risk genes related to nucleotide synthesis and signaling. J. Allergy Clin. Immunol. 139, 1148–1157 (2017).
7. Ferreira, M. A. R. et al. Eleven loci with new reproductible genetic associations with allergic disease risk. J. Allergy Clin. Immunol. pii: S0091-6749(18)30558-X. [doi:10.1016/j.jaci.2018.03.012] (2018).
8. Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. 47, 1091–1098 (2015).
9. 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). 51-3.
10. Barbera, A. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. bioRxiv https://www.biorxiv.org/content/early/2017/10/03/045260 (2017).
11. Battle, A. et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res. 24, 14–24 (2014).
12. Edwards, S. L., Beesley, J., French, J. D. & Dunning, A. M. Beyond GWAS: illuminating the dark road from association to function. Annu. J. Hum. Genet. 93, 779–797 (2013).
13. Li, G. et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84–98 (2012).
14. Mifsud, B. et al. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat. Genet. 47, 398–406 (2015).
15. 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).
16. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014).
17. 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).
18. 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).
19. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. Cell 155, 934–947 (2013).
20. Wu, L. et al. Identification of novel susceptibility loci and genes for breast cancer risk: a transcriptome-wide association study of 229,000 women of European descent. Nat. Genet. 50, 968–978 (2018).
21. Pizzaro, R. et al. OncoScore: a novel, Internet-based tool to assess the oncogenic potential of genes. Sci. Rep. 7, 46290 (2017).
22. Chuan, S. et al. Limited statistical evidence for shared genetic effects of eQTLs and autoimmune-disease-associated loci in three major immune-cell types. Nat. Genet. 49, 600–605 (2017).
23. Criccietti, C. et al. A gene signature to predict high tumour-infiltrating lymphocytes after neoadjuvant chemotherapy and outcome in patients with triple negative breast cancer. Ann. Oncol. 29, 162–169 (2018).
24. Stanford, S. M. & Bottini, N. PTPN22: the archetypal non-HLA autoimmunity gene. Nat. Rev. Rheumatol. 10, 602–611 (2014).
25. Clarke, F. et al. The protein tyrosine phosphatase PTPN22 negatively regulates presentation of immune complex derived antigens. Sci. Rep. 8, 12692 (2018).

26. Liu, J. et al. Rhomboid domain-containing protein 3 is a negative regulator of TLR3- triggered natural killer cell activation. Proc. Natl Acad. Sci. USA 110, 7814–7819 (2013).

27. Liu, J. et al. Rhombod3 controls autoimmune by suppressing the production of IL-6 by dendritic cells via K27-linked ubiquitination of the regulator NEMO. Nat. Immunol. 15, 612–622 (2014).

28. Miyamoto, M. et al. Regulated expression of a gene encoding a nuclear factor, that differentially regulates luminal and basal progenitors in human mammary epithelium. EMBO Mol. Med. 3, 167–180 (2011).

29. Taniguchi, T., Ogasawara, K., Takaoka, A. & Tanaka, N. IRF family of transcription factors as regulators of host defense. Annu. Rev. Immunol. 19, 623–655 (2001).

30. Holland, D. G. et al. ZNF703 is a common Luminal B breast cancer oncogene. Breast Cancer Res. Treat. 131, 53 (2012).

31. Eriksson, N. et al. Genetic variants associated with breast size also influence breast cancer risk. BMC Med. Genet. 13, 53 (2012).

32. Okumura, F., Joo-Oekumura, A., Nakatsukasa, K. & Kamura, T. The role of cullin 5-containing ubiquitin ligases. Cell Div. 11, 1 (2016).

33. Haupt, S. et al. The role of MDM2 and MDM4 in breast cancer development and prevention. J. Mol. Cell Biol. 9, 53–67 (2011).

34. Caelles, C. E., Sestori, N., Sutherland, R. L., Regev, A. E. Differences in degradation lead to asynchronous expression of cyclin E1 and cyclin E2 in cancer cells. Cell Cycle 12, 596–605 (2013).

35. Rogers, S. et al. Cyclin E2 is the predominant E-cyclin associated with NPAT in breast cancer cells. Cell Div. 10, 1 (2015).

36. Shenlin, S., Dorstyn, L., Davar, S. & Kumar, S. Old, new and emerging functions of caspases. Cell Death Differ. 22, 526–539 (2015).

37. Chenexvix-Trench, G. et al. Dominant negative ATM mutations in breast cancer families. J. Natl. Cancer Inst. 94, 205–215 (2002).

38. Goldgar, D. E. et al. Rare variants in the ATM gene and risk of breast cancer. Breast Cancer Res. 13, R73 (2011).

39. Magnani, L. & Lupien, M. Chromatin and epigenetic determinants of estrogen receptor alpha (ESR1) signaling. Mol. Cell Endocrinol. 382, 633–641 (2014).

40. Raza, H. Dual localization of glutathione S-transferase in the cytosol and mitochondria: implications in oxidative stress, toxicity and disease. FEBS J. 278, 4243–4251 (2011).

41. Samanta, S. et al. IMP3 promotes stem-like properties in triple-negative breast cancer by regulating Slug. Oncogene 35, 1111–1121 (2016).

42. Esplin, E. D. et al. The glycine 90 to aspartate alteration in the Abeta subunit of PP2A (PP2R1B) associates with breast cancer and causes a deficit in protein function. Genes Chromosomes Cancer 45, 182–190 (2006).

43. Petrovic, A. et al. Structure of the Msi12 complex and molecular basis of its interaction with CENP-C at human kinetochores. Cell 167, 1028–1040 e15 (2016).

44. Sun, T. et al. A reverse signaling pathway downstream of Sema4A controls cell migration via Scrib. J. Cell Biol. 216, 199–215 (2017).

45. Hoffman, J. D. et al. Cis-eQTL-based trans-ethnic meta-analysis reveals functional variation in humans. Cell 165, 32894 (2016).

46. Chikh, A. et al. Class II phosphoinositide 3-kinase C2beta regulates a novel interaction with CENP-C at human kinetochores. Cell Death Differ., 1–13 (2015).

47. Yu, X. et al. Structure, inhibitor, and regulatory mechanism of Lyp, a cell disease: a phase III randomized, placebo-controlled, double-blind study of the Gardos channel blocker senicapoc (ICA-17043). J. Clin. Invest. 104 (2011).

48. French, J. D. et al. Functional variants at the 11q13 risk locus for breast cancer do not correlate with acute vaso-occlusive crises in patients with sickle cell disease: a phase III randomized, placebo-controlled, double-blind study of the Gardos channel blocker senicapoc (ICA-17043). Br. J. Haematol. 153, 92–104 (2011).

49. Navaratnam, N. & Sarwar, R. An overview of cytidine deaminases. Nat. Immunol. 18, 195–203 (2017).

50. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. 47, 291–295 (2015).

Acknowledgements
Full details of the Acknowledgements are provided in Supplementary Note 1.

Author contributions
Writing Group: M.A.F., F. Al-E., W.Z., D.F.E., R.L.M., J.B., G.C.-T.; Statistical analysis: M.F. and E.R.G.; Provided DNA samples and/or phenotypic data: K.A., I.A., H.A.-C., A.C.A, A.A., Y.A., K.J.A., B.K.A., E.A., J.A., J.B., D.R.R., D.B., M.W.R., S.R.J.B., M.B., C.B., N.V.B., S.E.B., M.K.B., A.B., H.B., H.R., A.B., R.B., T.C., M.A.C., I.C., F.C., J.C., B.D.C., E.C., J.G.-C., D.C., H.C., W.K.C., K.B.M.C., C.L.C., F.J.C., A.C., S.J.C., S.S.C., K.C., M.B.D., M. de la H, J.D., P.D., O.D., T.D., A.M.D., M.D.E., B.L., C.E., C.E., M.E., P.A.F., O.F., H.E., F.F., D.F., M.G., M.G.-D., P.A.G., S.M.G., J.G., M.G.-J., A.S.-M., M.G.G., G.G.G., A.K.G., M.S.G., D.E.G., A.G.N., M.H.G., J.G., P.G., C.A.H., P.H., U.H., W.H., J.H., F.B.L.H., A.H., R.N.H., H.E., P.H., K.H., E.N.I., C.J., M.I., A.I., P.I., R.I., R.C.I., E.M.I., N.I., J.E.I., V.I., K.B., Y.K., T.K., E.K., J.J.K., Y.-D.K., I.K., P.N.K., V.N.K., Y.L., D.L., C.L., G.L., I.L., F.L., S.L., J.L., J.T.L., I.L., E.M., A.M., S.M., D.M., L.M.A.M., U.M., K.M., A.M., M.M., F.M., L.M., A.M.M., K.L.N., S.N., H.N., I.N., F.C.N., I.N.-Z., K.O., E.O., O.L.O., H.O., A.O., I.P., T.-W.S.-P., M.T.P., I.S.P., A.P.P.P., P.D.P.D.P., D.P.K., B.P., N.P., P.R., J.R., G.R., H.A.R., E.S., K.S., I.E.S., M.K.S., R.K.S., P.S., X.O., S.I., F.C.S., P.S., C.M.S., J.C.S., A.B.S., A.J.S., J.B., J.P., B.K.T., I.A.T., J.M.T., M.B.T., A.T.e., M.T., K.T., D.L.T., M.T., A.E.T., D.T., M.T., N.T., C.M.V., R.L.N., C.J.A.V.A, M.A.W. van den O., E.J.V.R., A.V., A.V., W.R., J.N.W., C.W., R.W., X.R.Y., D.A., A.M., G.M.O., G.M.R. Study Collaborators, G.-C.G.-D. study Collaborators, EMBRACE Collaborators, HBOC Investigators, BCIRG Investigators and ABCTB Investigators.

Additional information
Supplementary information accompanies this paper at https://doi.org/10.1038/s41467-018-08053-5.

Competing interests: The authors declare no competing interests.

Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

Journal peer review information: Nature Communications thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

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

© The Author(s) 2019

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.
Manuel A. Ferreira1, Eric R. Gamazon2,3, Fares Al-Ejeh1, Kristiina Aittomäki4, Irene L. Andrilis5,6, Hoda Anton-Culver7, Adalgeir Arason8,9, Volker Arndt10, Kristian J. Aronson11, Banu K. Arun12, Ella Asseryanis13, Jacopo Azzollini14, Judith Balmaña15,16, Daniel R. Barnes17, Daniel Barrowdale17, Matthias W. Beckmann18, Sabine Behrens19, Javier Benitez20,21, Marina Bermisheva22, Katarzyna Bialkowska23, Carl Blomqvist24,25, Natalia V. Bogdanova26,27,28, Stig E. Bojesen29,30,31, Manjeet K. Bolla17, Ake Borg32, Hiltrud Brauch33,34,35, Hermann Brenner36,37, Annegien Broeks38,39, Barbara Burwinkel38,39, Trinidad Caldés40, Maria A. Caligo41, Daniele Campa19,42, Ian Campbell43,44, Federico Canzian45, Jonathan Carter46, Brian D. Carter47, Jose E. Castelao48, Jenny Chang-Claude19,49, Stephen J. Chanock50, Hans Christiansen26, Wendy K. Chung51, Kathleen B.M. Claes52, Christine L. Clarke53, GC-HBOC Study Collaborators, GEMO Study Collaborators, EMBRACE Collaborators, Fergus J. Couch54, Angela Cox55, Simon S. Cross56, Kamila Czene57, Mary B. Daly58, Miguel de la Hoya40, Joe Dennis17, Peter Devileee59,60, Orland Diez15,61, Thilo Dörk27, Alison M. Dunning62, Miriam Dwek63, Diana M. Eccles64, Bent Ejertsen65, Carolina Ellberg16, Christoph Engel67, Mikael Eriksson57, Peter A. Fasching18,68, Olivia Fletcher69, Henrik Flyger70, Eitan Friedman71,72, Debra Frost17, Marike Gabrielsson57, Manuela Gago-Dominguez73,74, Patricia A. Gan5, Susan M. Gapstur47, Judy Garber76, Montserrat García-Closs50,77, José A. García-Sáenz78, Mia M. Gaudet47, Graham G. Giles79,80,81, Gord Glendon5, Andrew K. Godwin82, Mark S. Goldberg83,84, David E. Goldgar85, Anna González-Neira21, Mark H. Greene86, Jacek Gronwald23, Pascal Guénel17, Christopher A. Haiman88, Per Hall57,89, Ute Hamann90, Wei He57, Jane Heyworth91, Frans B.L. Hogervorst92, Antoinne Hollestelle93, Robert N. Hoover50, John L. Hopper80, Peter J. Hulick94,95, Keith Humphreys57, Evgeny N. Imyanitov96, HEBON Investigators, BCFR Investigators, ABCTB Investigators, Claudine Isaacs97, Milena Jakimovska98,99, Anna Jakubowska23,99, Paul A. James100,44, Ramunas Janavicius101, Rachel C. Jankowitz102, Esther M. John103, Nichola Johnson69, Vijai Joseph104, Beth Y. Karlan105, Elza Khusnutdinova22,106, Johanna I. Kiiski107, Yon-Dschun Ko108, Michael E. Jones109, Irene Konstantopoulou110, Vesselka N. Kristensen111,112, Yael Laitman21, Diether Lambrechts113,114, Conxi Lazaro15, Goska Leslie17, Jenny Lester105, Fabienne Lesueur116,117,118, Sara Lindström119,120, Jirong Long121, Jennifer T. Loud86, Jan Lubinski23, Enes Makalic80, Arto Mannermaa122,123,124, Mehdi Manoochehri90, Sara Margolin89,125, Tabea Maurer49, Dimitrios Mavroudis126, Lesley McGuffog17, Alfonso Meind127, Usha Menon128, Kyriaki Michailidou17,129, Austin Miller130, Marco Montagna131, Fernando Moreno78, Lidia Moserle131, Anna Marie Mulligan132,133, Katherine L. Nathanson134, Susan L. Neuhausen135, Helen Nevanlinna107, Ines Nevelsteen136, Finn C. Nielsen137, Liene Nikitina-Zake138, Robert L. Nussbaum139, Kenneth Offit104,140, Edith Olah141, Olufunmilayo I. Olopade142, Håkan Olsson66, Ana Osorio20,21, Janos Papp141, Tjoung-Won Park-Simon27, Michael T. Parsons16, Inge Sokilde Pedersen143,144,145, Ana Peixoto146, Paolo Peterlongo147, Paul D.P. Pharoah17,62, Dijana Plaseska-Karanfiloka,98, Bruce Poppe52, Nadege Presneau63, Paolo Radice148, Johanna Rantalala149, Gad Rennert150, Harvey A. Risch151, Emmanouil Saloustros152, Kristin Sanden153, Elinor J. Sawyer154, Marjanka K. Schmidt155, Rita K. Schmutzler156,157, Priyanka Sharma158, Xiaow-Ou Shu121, Jacques Simard159, Christian F. Singel13, Penny Soucy159, Melissa C. Southey160,161, John J. Spinelli162,163, Amanda B. Spurdle1, Jennifer Stone80,164, Anthony J. Swerdlow109,165, William J. Tapper164, Jack A. Taylor166,167, Manuel R. Teixeira146,168, Mary Beth Terry169, Alex Teulé170, Mads Thomassen171, Kathrin Thöne49, Darcy L. Thull172, Marc Tischkowitz73,174, Amanda E. Toland175, Diana Torres90,176, Thérèse Truong87, Nadine Tung177, Celine M. Vachon178, Christi J. van Asperen179, Ans M.W. van den Ouweland180,
GC-HBCO Study Collaborators

Norbert Arnold221,222, Bernd Auber223, Nadja Bogdanova-Markov224, Julika Borde156,157,225, Almuth Caliebe226, Nina Ditsch127, Bernd Dworniczak224, Stefanie Engert227, Ulrike Faust228, Andrea Gehrig229, Eric Hahnen230, Jan Hauke156,157, Julia Hentschel230, Natalie Herold156,157,225, Ellen Honisch231, Walter Just232, Karin Kast233, Mirjam Larsen156,157, Johannes Lemke230, Huu Phuc Nguyen228, Dieter Niederacher231, Claus-Eric Ott234, Konrad Platzer230, Esther Pohl-Rescigno156,157, Juliane Ramser227, Kerstin Rhiem156,157, Doris Steinemann223, Christian Sutter235, Raymonda Varon-Mateeva234, Shan Wang-Gohrke236 & Bernhard H.F. Weber237

GEMO Study Collaborators

Fabienne Prieur238, Pascal Pujol239, Charlotte Sagne240, Nicolas Sevenet241, Hagay Sobol242,243, Johanna Sokolowska244, Dominique Stoppa-Lyonnet245,246,247 & Laurence Venat-Bouvet248

221Department of Gynaecology and Obstetrics, University Hospital of Schleswig-Holstein, Campus Kiel, Christian-Albrechts University Kiel, Kiel 24118, Germany. 222Institute of Clinical Molecular Biology, University Hospital of Schleswig-Holstein, Campus Kiel, Christian-Albrechts University Kiel, Kiel 24118, Germany. 223Institute of Human Genetics, Hannover Medical School, 30625 Hannover, Germany. 224Institute of Human Genetics, University of Münster, Münster 48149, Germany. 225Center for Molecular Medicine Cologne (CMMC), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne 50931, Germany. 226Institute of Human Genetics, University Hospital of Schleswig-Holstein, Campus Kiel, Christian-Albrechts University Kiel, Kiel 24118, Germany. 227Division of Gynaecology and Obstetrics, Klinikum rechts der Isar der Technischen Universität München, Munich 80333, Germany. 228Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen 72074, Germany. 229Department of Human Genetics, University Würzburg, Würzburg 97074, Germany. 230Institute of Human Genetics, University Hospital Leipzig, Leipzig 04103, Germany. 231Department of Gynecology and Obstetrics, University Hospital Düsseldorf, Heinrich-Heine University Düsseldorf, Düsseldorf 40225, Germany. 232Institute of Human Genetics, University Hospital Ulm, Ulm 89075, Germany. 233Department of Gynecology and Obstetrics, Technical University of Dresden, Dresden 01307, Germany. 234Institute of Human Genetics, Campus Virchow Klinikum, Charte, Berlin 13353, Germany. 235Institute of Human Genetics, University Hospital Heidelberg, 69120 Heidelberg, Germany. 236Department of Gynaecology and Obstetrics, University Hospital Ulm, 89075 Ulm, Germany. 237Institute of Human Genetics, University Regensburg, 93053 Regensburg, Germany

238Service de Génétique Clinique Chromosomique et Moléculaire, Hôpital Nord, CHU Saint Etienne, 42270 St. Etienne, France. 239Unité d’Oncogénétique, CHU Arnaud de Villeneuve, 34295 Montpellier, France. 240Bâtiment Cheney D, Centre Léon Bérard, Lyon 69373, France. 241Unicité d’Oncogénétique, Institut Bergonié, Bordeaux 33076, France. 242Département Oncologie Génétique, Prévention et Dépistage, Institut Paoli-Calmettes, Marseille 13009, France. 243Marseille Medical School, Aix-Marseille University, Marseille 13007, France. 244Laboratoire de génétique médicale, Nancy Université, Centre Hospitalier Régional et Universitaire, Vandoeuvre-les-Nancy 54511, France. 245Service de Génétique, Institut Curie, Paris 75005, France. 246Department of Tumour Biology, INSERM U830, 75005 Paris, France. 247Université Paris Descartes, 75006 Paris, France. 248Department of Medical Oncology, CHU Dupuytren, Limoges 87042, France.
EMBRACE Collaborators

Julian Adlard195, Munaza Ahmed196, Julian Barwell197, Angela Brady198, Carole Brewer199, Jackie Cook200, Rosemarie Davidson201, Alan Donaldson202, Jacqueline Eason203, Ros Eeles204, D. Gareth Evans205,206, Helen Gregory207, Helen Hanson208, Alex Henderson209, Shirley Hodgson210, Louise Izatt211, M. John Kennedy212, Fiona Lalloo206, Clare Miller213, Patrick J. Morrison214, Kai-ren Ong215, Jo Perkins17, Mary E. Porteous216, Mark T. Rogers217, Lucy E. Side218, Katie Snape219, Lisa Walker220 & Patricia A. Harrington62

195Yorkshire Regional Genetics Service, Chapel Allerton Hospital, Leeds LS7 4SA, UK. 196North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Trust, London WC1N 3JH, UK. 197Leicestershire Clinical Genetics Service, University Hospitals of Leicester NHS Trust, Leicester LE1 5WW, UK. 198North West Thames Regional Genetics Service, Kennedy Galton Centre, The North West London Hospitals NHS Trust, Middlesex HA1 3UJ, UK. 199Department of Clinical Genetics, Royal Devon & Exeter Hospital, Exeter EX2 5DW, UK. 200Sheffield Clinical Genetics Service, Sheffield Children’s Hospital, Sheffield S10 2TH, UK. 201Department of Clinical Genetics, South Glasgow University Hospitals, Glasgow G51 4TF, UK. 202Clinical Genetics Department, St. Michael’s Hospital, Bristol BS2 8EG, UK. 203Nottingham Clinical Genetics Service, Nottingham University Hospitals NHS Trust, Nottingham NG5 1PB, UK. 204Oncogenetics Team, The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, London SM2 5NG, UK. 205Genomic Medicine, Division of Evolution and Genomic Sciences, The University of Manchester, Manchester Academic Health Science Centre, Manchester Universities Foundation Trust, St. Mary’s Hospital, Manchester M13 9WL, UK. 206Genomic Medicine, North West Genomics hub, Manchester Academic Health Science Centre, Manchester Universities Foundation Trust, St. Mary’s Hospital, Manchester M13 9WL, UK. 207North of Scotland Regional Genetics Service, NHS Grampian & University of Aberdeen, Foresterhill, Aberdeen, UK. 208Southwest Thames Regional Genetics Service, St. George’s Hospital, London SW17 0QT, UK. 209Institute of Genetic Medicine, Centre for Life, Newcastle upon Tyne Hospitals NHS Trust, Newcastle upon Tyne NE1 3BZ, UK. 210Department of Clinical Genetics, St George’s, University of London, London SW17 0RE, UK. 211Clinical Genetics, Guy’s and St Thomas’ NHS Foundation Trust, London SE1 9RT, UK. 212Academic Unit of Clinical and Molecular Oncology, Trinity College Dublin and St. James’s Hospital, Dublin, Eire, Ireland. 213Department of Clinical Genetics, Alder Hey Hospital, Liverpool L12 2AP, UK. 214Northern Ireland Regional Genetics Centre, Belfast City Hospital, Belfast BT9 7AB, UK. 215West Midlands Regional Genetics Service, Birmingham Women’s Hospital Healthcare NHS Trust, Birmingham B15 2TG, UK. 216South East of Scotland Regional Genetics Service, Western General Hospital, Edinburgh EH4 2XU, UK. 217All Wales Medical Genetics Services, University Hospital of Wales, Cardiff CF14 4XW, UK. 218Princess Anne Hospital, Southampton SO16 5YA, UK. 219Medical Genetics Unit, St. George’s, University of London, London SW17 0RE, UK. 220Oxford Regional Genetics Service, Churchill Hospital, Oxford OX3 7LJ, UK.

HEBON Investigators

Bernadette A.M. Heemskerk-Gerritsen93, Matti A. Rookus249, Caroline M. Seynaeve93, Frederieke H. van der Baan249, Annemieke H. van der Hout250, Lizet E. van der Kolk92, Rob B. van der Luijt251, Carolien H.M. van deurzen252, Helena C. van Doorn253, Klaartje van Engelen254, Liselotte van Hest255, Theo A.M. van Os256, Senno Verhoef92, Maartje J. Vogel257 & Juul T. Wijnen258

249Department of Epidemiology, The Netherlands Cancer Institute, Amsterdam 1066 CX, The Netherlands. 250Department of Genetics, University Medical Center Groningen, University Groningen, Groningen 9713 GZ, The Netherlands. 251Department of Medical Genetics, University Medical Center, Utrecht 3594 CX, The Netherlands. 252Department of Pathology, Erasmus University Medical Center, Rotterdam 3015 CN, The Netherlands. 253Department of Gynaecology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam 3015 CE, The Netherlands. 254Department of Clinical Genetics, VU University Medical Center, Amsterdam 1105 AZ, The Netherlands. 255Clinical Genetics, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam 1007 MB, The Netherlands. 256Department of Clinical Genetics, Amsterdam UMC, location AMC, Amsterdam 1100 DD, The Netherlands. 257Department of Pathology, Netherlands Cancer Institute, Amsterdam 1006 BE, The Netherlands. 258Department of Human Genetics and Department of Clinical Genetics, Leiden University Medical Center, Leiden 2333 ZA, The Netherlands.

BCFR Investigators

Alexander Miron189,190, Miroslav Kapuscinski80, Anita Bane191, Eric Ross192, Saundra S. Buys193 & Thomas A. Conner194

189Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA. 190Department of Surgery, Harvard Medical School, Boston, MA 02215, USA. 191Department of Pathology & Molecular Medicine, Juravinski Hospital and Cancer Centre, McMaster University, Hamilton, ON L8V 1C3, Canada. 192Population Studies Facility, Fox Chase Cancer Center, Philadelphia, PA 19111, USA. 193Department of Medicine, Huntsman Cancer Institute, Salt Lake City, UT 84112, USA. 194Huntsman Cancer Institute, Salt Lake City, UT 84112, USA.
**ABCTB Investigators**

Rosemary Balleine\textsuperscript{259}, Robert Baxter\textsuperscript{260}, Stephen Braye\textsuperscript{261}, Jane Carpenter\textsuperscript{262}, Jane Dahlstrom\textsuperscript{263}, John Forbes\textsuperscript{264}, Soon C Lee\textsuperscript{265}, Deborah Marsh\textsuperscript{260}, Adrienne Morey\textsuperscript{266}, Nirmala Pathmanathan\textsuperscript{267}, Peter Simpson\textsuperscript{268}, Allan Spigelman\textsuperscript{269}, Nicholas Wilcken\textsuperscript{270} & Desmond Yip\textsuperscript{271}

\textsuperscript{259}Pathology West ICPMR, Westmead, NSW 2145, Australia. \textsuperscript{260}Kolling Institute of Medical Research, University of Sydney, Royal North Shore Hospital, Sydney, NSW 2065, Australia. \textsuperscript{261}Pathology North, John Hunter Hospital, Newcastle, NSW 2305, Australia. \textsuperscript{262}Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research, University of Sydney, Sydney, NSW 2145, Australia. \textsuperscript{263}Department of Anatomical Pathology, ACT Pathology, Canberra Hospital and ANU Medical School, Australian National University, Canberra, ACT 2605, Australia. \textsuperscript{264}Department of Surgical Oncology, Calvary Mater Newcastle Hospital, Australian New Zealand Breast Cancer Trials Group, and School of Medicine and Public Health, University of Newcastle, Newcastle, NSW 2035, Australia. \textsuperscript{265}School of Science and Health, The University of Western Sydney, Sydney, NSW 2650, Australia. \textsuperscript{266}SydPath St. Vincent's Hospital, Sydney, NSW 2010, Australia. \textsuperscript{267}Department of Tissue Pathology and Diagnostic Oncology, Pathology West, Westmead Breast Cancer Institute, Westmead Hospital, Sydney, NSW 2145, Australia. \textsuperscript{268}UQ Centre for Clinical Research and School of Medicine, The University of Queensland, Brisbane, QLD 4072, Australia. \textsuperscript{269}Hereditary Cancer Clinic, St. Vincent's Hospital, The Kinghorn Cancer Centre, Sydney, NSW 2010, Australia. \textsuperscript{270}Crown Princess Mary Cancer Centre, Westmead Hospital, Sydney, NSW 2145, Australia. \textsuperscript{271}Department of Medical Oncology, The Canberra Hospital, Canberra, ACT 2605, Australia