Online Methods

Study subjects

Supplementary Table 1 summarises the studies from the Breast Cancer Association Consortium (BCAC) that contributed data. The majority were case-control studies. Sixty-eight BCAC studies participated in the ER-negative breast cancer component of the OncoArray, contributing 9,655 cases and 45,494 controls. All studies provided core data on disease status and age at diagnosis/observation, and the majority provided information on clinico-pathological and lifestyle factors, which have been curated and incorporated into the BCAC database (version 6). Estrogen receptor status for most (~70%) cases was obtained from clinical records. After removal of overlapping participants, genotype data were also available from eight GWASs[1-5] (4,480 ER-negative cases and 12,632 controls) and 40 studies previously genotyped using the Illumina iCOGS custom array[6] (7,333 ER-negative cases and 42,468 controls).

A total of 21,468 ER-negative cases were included in the combined analyses. Of those 5,793 had tumours that were also negative for progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2) and were defined as triple-negative (TN). PR and HER2 status was also obtained predominantly from clinical records. A further 4,217 were positive for PR or HER and were considered non-TN. The remainder had unknown PR or HER status. All participating studies were approved by their appropriate ethics review boards and all subjects provided informed consent.

Subjects included from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) were women of European ancestry aged 18 years or older with a pathogenic variant in *BRCA1*. The majority of the participants were sampled through cancer genetics clinics. Multiple members of the same families were included in some instances. Fifty-eight studies from 24 countries contributed Oncoarray genotype data. After quality control (see below) and removal of overlapping participants with the BCAC OncoArray study, data were available on 15,566 *BRCA1* mutation carriers, of whom 7,784 were affected with breast cancer (Supplementary Table 2). We also obtained iCOGS genotype data on 3,342 *BRCA1* mutation carriers (1,630 with breast cancer) from 54 studies through CIMBA. All mutation carriers provided written informed consent and participated under ethically approved protocols.

OncoArray SNP selection

Approximately 50% of the SNPs for the OncoArray were selected as a “GWAS backbone” (Illumina HumanCore), which aimed to provide high coverage for the majority of common variants through imputation. The remaining allocation was selected from lists supplied by each of six disease-based consortia, together with a seventh lists of SNPs of interest to multiple disease groups. Approximately 72k SNPs were selected specifically for their relevance to breast cancer, based on prior evidence of association with overall or subtype-specific disease, with breast density or with breast tissue specific gene expression. Lists were merged with lists from the other consortia as described elsewhere[7].

Genotype calling and quality control

Details of the genotype calling and quality control (QC) for the iCOGS and GWAS are described elsewhere[6,8-10].

Of the 568,712 variants selected for genotyping on OncoArray, 533,631 were successfully manufactured on the array (including 778 duplicate probes). OncoArray genotyping of BCAC and CIMBA samples was conducted at six sites. Details of the genotyping calling for the OncoArray are described in more detail elsewhere[7]. Briefly, we developed a single calling pipeline that was applied to more than 500,000 samples. An initial cluster file was generated using from 56,284 samples, selected over all the major genotyping centres and ethnicities, using the Gentrain2 algorithm. Variants likely to have problematic clusters were selected for manual inspection using the following criteria: call rate below 99%, minor allele frequency (MAF) <0.001, poor Illumina intensity and clustering metrics, deviation from the MAF observed in the 1000 Genomes Project using the criterion: $\frac{(|p_1-p_0|-0.01)^2}{((p_1+p_0)(2-p_1-p_0))} > C$, where p_0 and p_1 are the minor frequencies in the 1000 Genome Project and Oncoarray datasets, respectively, and C=0.008. (This latter criterion is approximately equivalent to excluding SNPs on the basis of a Chi-square statistic of 16 for the difference in allele frequencies, assuming 1,000 samples in each group). This resulted in manual adjustment of the cluster file for 3,964 variants, and the exclusion of 16,526 variants. The final cluster file was then applied to the full dataset.

We excluded SNPs with a call rate <95% in any consortium, not in Hardy-Weinberg equilibrium ($P<10^{-7}$ in controls, or $P<10^{-12}$ in cases) or with concordance <98% among 5,280 duplicate pairs. For the imputation, we additionally excluded SNPs with a MAF<1% and a call rate <98% in any consortium, SNPs that could not be linked to the 1000 Genomes Project reference, those with MAF for Europeans that differed from that for the 1000 Genomes Project and a further 1,128 SNPs where the cluster plot was judged to be not ideal. Of the 533,631 SNPs which were manufactured on the array, 494,763 passed the initial QC and 469,364 were used in the imputation (see below).

For BCAC, we excluded probable duplicate samples and close relatives within each study, and probable duplicates between studies. These were identified by identity by state (IBS) analysis using a set of approximately 38,000 uncorrelated ($r^2<0.1$) SNPs for OncoArray and iCOGS and 16,000 SNPs for GWAS. Based on inspection of the distribution of IBS values, we identified first-degree relative pairs using the criterion 0.82<IBS<0.90 for OncoArray and 0.85<IBS<0.90 for iCOGS; similar criteria were used for each GWAS (with limits depending on the IBS distribution in that study).

We applied LD score regression to the summary results from GWAS, iCOGS and OncoArray to assess the evidence of overlap in individuals between the three datasets. We conducted three pair-wise cross-trait regression analyses (GWAS-iCOGS, GWAS-OncoArray and iCOGS-OncoArray) and used the intercept from the regression analysis to estimate the amount of overlap[11]. Assuming that the phenotypic correlation is 1 (that is, a case is a case in all datasets and a control is a control in all datasets), we found that for GWAS-iCOGS, the estimated overlap was 1.5% of individuals, for GWAS-OncoArray, the estimated overlap was 3.8% of individuals, and for iCOGS-OncoArray, the estimated overlap was 0.2% of individuals. It is unlikely that this degree of overlap would have influenced the results obtained from our analyses.

We also excluded samples with a call rate <95% and samples with extreme heterozygosity (>4.9 standard deviations from the mean for the reported ethnicity). Ancestry analysis was performed using a standardized approach in which 2,318 ancestry informative markers with minor allele frequencies of 0.05 on a subset of ~66,000 samples including 505 Hapmap 2 samples. The contribution of each of the three major continental ancestry groups (European, Asian and African) was estimated by mapping each individual to regions of a triangle based on the first two principal components, as implemented in the software package FastPop (http://sourceforge.net/projects/fastpop/)[12]. Individuals were thus classified into 4 groups: European (defined as >80% European ancestry), East Asian (>40% Asian ancestry), African (>20% African ancestry) and other (not fulfilling any of the above criteria)[7]. Of the 152,492 samples genotyped, the final dataset consisted of 142,072 samples, of which 9,655 ER-negative cases and 45,494 controls of European origin had not been included in a previous GWAS and had not been genotyped using iCOGS and were included in this analysis.

For the CIMBA samples we excluded individuals of non-European ancestry using multi-dimensional scaling. For this purpose we selected 30,733 uncorrelated autosomal SNPs (pair-wise r^2< 0.10) to compute the genomic kinship between all pairs of *BRCA1* and *BRCA2* carriers, along with 267 HapMap samples (CHB, JPT, YRI and CEU). These were converted to distances and subjected to multidimensional scaling. Using the first two components, we calculated the proportion of European ancestry for each individual and excluded samples with >27% non-European ancestry to ensure that samples of Ashkenazi Jewish ancestry were included in the final sample.

Imputation

Genotypes for ~21M SNPs were imputed for all samples using the October 2014 (Phase 3) release of the 1000 Genomes Project data as the reference panel and Nhap=800. The iCOGS, OncoArray and six of the GWAS datasets were imputed using a two-stage imputation approach, using SHAPEIT[13] for phasing and IMPUTEv2[14] for imputation. The imputation was performed in 5Mb non-overlapping intervals. All subjects were split into subsets of ~10,000 samples, with subjects from the same grouped in the subset. The Breast and Prostate Cancer Cohort Consortium (BPC3) and Breast Cancer Family Registry (BCFR) GWAS performed the imputation separately using MACH and Minimac[15,16]. We imputed genotypes for all SNPs that were polymorphic (MAF>0.1%) in either European or Asian samples. For the BCAC GWAS, data were included in the analysis for all SNPs with MAF>0.01 and imputation r^2>0.3. For iCOGS and OncoArray we included data for all SNPs with imputation r^2>0.3 and MAF>0.005.

Statistical analyses of BCAC data

Per-allele odds ratios and standard errors were generated for the Oncoarray, iCOGS and each GWAS, adjusting for principal components using logistic regression. The Oncorray and iCOGS analyses were additionally adjusted for country and study, respectively. For the OncoArray dataset, principal components analysis was performed using data for 33,661 SNPs (which included the 2,318 markers of continental ancestry) with a MAF≥0.05 and maximum correlation of 0.1, using purpose-written software to allow standard calculations to be performed sufficiently rapidly on a very large dataset (http://ccge.medschl.cam.ac.uk/software/pccalc/). We used the first 10 principal components, as additional components did not further reduce inflation in the test statistics. We used nine principal components for the iCOGS and up to 10 principal components for the other GWAS, where this was found to reduce inflation.

OR estimates were derived using MACH for the BCFR GWAS, ProbABEL[17] for the BPC3 GWAS, SNPTEST (https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html) for the remaining GWAS and purpose written software for the iCOGS and Oncoarray datasets. OR estimates and standard errors were combined by a fixed effects inverse variance meta-analysis using METAL[18]. This was first done across the eight GWAS, applying genomic control, as described previously[6]. It was then applied (without genomic control) to combine findings from the three BCAC genotyping initiatives (GWAS, iCOGS, OncoArray).

The independence of signals from two variants at 11q22.3 was by fitting the logistic regression models described above with both variants as covariates. This was done separately for iCOGS and OncoArray data and results for each variant combined by meta-analysis.

For selected SNPs we estimated per-allele ORs by ER-status using all available BCAC data for 82,263 cases with known ER status and 87,962 controls from the iCOGS and OncoArray studies. We also estimated the per-allele ORs by TN status (TN versus other ER-negative subtypes) and tumour grade, using available BCAC data for ER-negative cases and corresponding controls. Tests for heterogeneity by subtype were derived by applying logistic regression to cases only. This was done separately for the iCOGS and Oncoarray datasets, adjusted as before, and then combined in a fixed-effects meta-analysis. Multinomial regression was applied to cases only to test a linear trend for grade, with the model constrained so that the difference between grade 1 and 3 was double that for the difference between grade 2 and 3; this method was also used to test for a linear trend with age with ordinal values 1, 2, 3 and 4 representing ages <40, 40-49, 50-59 and ≥60, respectively.

Statistical analyses of CIMBA data

Associations between genotypes and breast cancer risk for *BRCA1* mutation carriers were evaluated using a 1*df* per allele trend-test (*P*-trend), based on modeling the retrospective likelihood of the observed genotypes conditional on breast cancer phenotypes[19]. This was done separately for iCOGS and OncoArray data. To allow for the non-independence among related individuals, an adjusted test statistic was used which took into account the correlation in genotypes[20]. All analyses were stratified by country of residence and, for countries where strata were sufficiently large (USA and Canada), by Ashkenazi Jewish ancestry. The results from the iCOGS and OncoArray datasets were then pooled using fixed effects meta-analysis. We repeated these analyses modelling ovarian cancer as a competing risk and observed no substantial difference in the results obtained.

The independence of signals from two variants at 11q22.3 was assessed using OncoArray data only, fitting a Cox regression model with per-allele effects for both variants, adjusting for birth cohort, stratified by country of residence and using robust standard errors and clustered observations for relatives. This approach provides valid significance tests of associations, although the HR estimates can be biased[21].

Meta-analysis of BCAC and CIMBA

A fixed effects meta-analysis of results from BCAC and CIMBA was conducted using an inverse variance approach assuming fixed effects, as implemented in METAL[18]. The effect estimates used were the logarithm of the per-allele hazard ratio (HR) estimate for the association with breast cancer risk in *BRCA1* mutation carriers from CIMBA and the logarithm of the per-allele OR estimate for the association with risk of ER-negative breast cancer based on BCAC data, both of which were assumed to approximate the same relative risk. We assessed genomic inflation using common (MAF>1%) GWAS backbone variants. As lambda is influenced by sample size, we calculated lambda1000 to be comparable with other studies.

All statistical tests conducted were two-sided.

Definition of known hits

We identified all associations previously reported from genome-wide or candidate analysis at a significance level $P<5\text{x}10^{-8}$ for overall breast cancer, ER-negative or ER-positive breast cancer, in *BRCA1* or *BRCA2* carriers, or in meta-analyses of these categories. We included only one SNP in any 500kb interval, unless joint analysis provided genome-wide significant evidence (conditional $P<5\text{x}10^{-8}$) of more than one independent signal. Where multiple studies reported associations in the same region, we considered the first reported association unless a later study identified a different variant in the same region that was more strongly associated with breast cancer risk. One hundred and seven previously reported hits were identified, 11 of these through GWAS of ER-negative disease or of breast cancer in *BRCA1* mutation carriers, or reported as more strongly associated with ER-negative breast cancer. These are listed in Table 2. The other 96 previously reported hits are listed in Supplementary Table 10.

Definition of new hits

To search for novel loci, we assessed all SNPs excluding those within 500kb of a known hit. This identified 206 SNPs in nine regions that were associated with disease risk at $P<5\text{x}10^{-8}$ in the meta-analysis of BCAC ER-negative breast cancer and CIMBA *BRCA1* mutation carriers. The SNP with lowest p-value from this analysis was considered the lead SNP. No additional loci were detected from the analysis of BCAC data only. Imputation quality, as assessed by the IMPUTE2 imputation r^2 in the Oncoarray dataset, was ≥0.89 for the 10 lead SNPs reported (Supplementary Table 3).

Candidate causal SNPs

To define the set of potentially causal variants at each of the novel susceptibility loci, we selected all variants with p-values within two orders of magnitude of the most significant SNP at each of the 10 novel loci. This is approximately equivalent to selecting variants whose posterior probability of causality is within two orders of magnitude of the most significant SNP[22,23]. This approach was applied to identify potentially causal variants for the signal given by the more frequent lead SNP at 11q22.3 (rs11374964). A similar approach was applied for the rarer lead SNP at this locus (rs74911261), but based on p-values from analyses adjusted for rs11374964.

Proportion of familial risk explained

The relative risk of ER-negative breast cancer for the first degree female relative of a woman with ER-negative disease has not been estimated. We therefore assumed that the 2-fold risk observed for overall disease also applied to ER-negative disease. In order to estimate the proportion of this explained by the 125 variants associated with ER-negative disease, we used minor allele frequency and OR estimates from the OncoArray-based genotype data and applied the formula: $\sum_i p_i(1-p_i)(\beta_i^2-\tau_i^2)/\ln(\lambda))$, where p_i is the minor allele frequency for variant *i*, β_i is the log(OR) estimate for variant *i*, τ_i is the standard error of β_i and λ=2 is the assumed overall familial relative risk.

The corresponding estimate for the FRR due to all variants is the *frailty scale* heritability, defined as $h_f^2 = \sum_i 2p_i(1-p_i)\gamma_i^2$, where the sum over all variants and γ_i is the true relative risk conferred by variant *i*, assuming a log-additive model. We first obtained the estimated heritability based on the full set of summary estimates using LD Score Regression[24], which derives a heritability estimate on the observed scale. We then converted this to an estimate on the friaility scale using the formula $h_f^2 = {h_{obs}^2}/{P(1-P)}$, where *P* is the proportion of samples in the population that are cases.

Proportion of polygenic risk-modifying variance explained for *BRCA1* carriers.

The proportion of the variance in the polygenic frailty modifying risk in BRCA1 carriers explained by the set of associated SNPs was estimated by $\sum_i \ln c_i/\sigma^2$, where c_i is the squared estimated coefficient of variation in incidences associated with SNP_i[25] and σ^2 is the total polygenic variance, estimated from segregation data[26].

In Silico Annotation of Candidate Causal variants

We combined multiple sources of *in silico* functional annotation from public databases to help identify potential functional SNPs and target genes, based on previous observations that breast cancer susceptibility alleles are enriched in *cis*-regulatory elements and alter transcriptional activity[27-30]. The influence of candidate causal variants on transcription factor binding sites was determined using the ENCODE-Motifs resource[31]. To investigate functional elements enriched across the region encompassing the strongest candidate causal SNPs, we analysed chromatin biofeatures data from the Encyclopedia of DNA Elements (ENCODE) Project[32], Roadmap Epigenomics Projects[33] and other data obtained through the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) namely: Chromatin State Segmentation by Hidden Markov Models (chromHMM), DNase I hypersensitive and histone modifications of epigenetic markers H3K4, H3K9, and H3K27 in Human Mammary Epithelial (HMEC) and myoepithelial (MYO) cells, T47D and MCF7 breast cancer cells and transcription factor ChIP-seq in a range of breast cell lines (Supplementary Table 6). To identify the SNPs most likely to be functional we used RegulomeDB[34], and to identify putative target genes, we examined potential functional chromatin interactions between distal and proximal regulatory transcription-factor binding sites and the promoters at the risk regions, using Hi-C data generated in HMECs[35] and Chromatin Interaction Analysis by Paired End Tag (ChiA-PET) in MCF7 cells. This detects 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 regulation[35]. Annotation of putative *cis*-regulatory regions and predicted target genes used the Integrated Method for Predicting Enhancer Targets (IM-PET)[36], the "Predicting Specific Tissue Interactions of Genes and Enhancers" (PreSTIGE) algorithm[37], Hnisz[38] and FANTOM[39]. Intersections between candidate causal variants and regulatory elements were identified using Galaxy, BedTools v2.24 and HaploReg v4.1, and visualised in the UCSC Genome Browser. Publically available eQTL databases including Gene-Tissue Expression (GTEx;[40] version 6, multiple tissues) and Westra[41] (blood), were queried for candidate causal variants.

<u>eQTL analyses</u>

Expression quantitative trait loci (eQTL) analyses were performed using data from The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) projects[42,43].

The TCGA eQTL analysis was based on 79 ER-negative breast tumors that had matched gene expression, copy number, and methylation profiles together with the corresponding germline genotypes available. All 79 individuals were of European ancestry as ascertained using the genotype data and the Local Ancestry in adMixed Populations (LAMP) software package (LAMP estimate cut-off >95% European)[44]. Germline genotypes were imputed into the 1000 Genomes reference panel (October 2014 release) using IMPUTE2[15,45]. Gene expression had been measured on the Illumina HiSeq 2000 RNA-Seq platform (gene-level RSEM normalized counts[46]), copy number estimates were derived from the Affymetrix SNP 6.0 (somatic copy number alteration minus germline copy number variation called using the GISTIC2 algorithm[47]), and methylation beta values measured on the Illumina Infinium HumanMethylation450, as previously described[42]. Primary TCGA eQTL analysis focused on all potentially causal variants in the 10 new regions associated with breast cancer risk in the meta-analysis of ER-negative cases and controls from BCAC and *BRCA1* mutation carriers from CIMBA. We considered all genes located up to 1 Mb on either side of each of these variants. The effects of tumor copy number and methylation on gene expression were first removed using a method described previously[48], and eQTL analysis was performed by linear regression as implemented in the R package Matrix eQTL[49].

The METABRIC eQTL analysis was based on 135 normal breast tissue samples resected from breast cancer patients of European ancestry. Germline genotyping for the METABRIC study was also done on the Affymetrix SNP 6.0, and ancestry estimation and imputation for this data set was conducted as described for TCGA. Gene expression in the METABRIC study had been measured using the Illumina HT12 microarray platform and we used probe-level estimates. As for TCGA, we considered all genes in 10 regions using Matrix eQTL.

We also performed additional eQTL analyses using the METABRIC data set for all variants within 1 Mb of *L3MBTL3* and *CDH2* and the expression of these specific genes.

Global Genomic Enrichment Analyses

We performed stratified LD score regression analyses[24] for ER- breast cancer using the summary statistics based on the meta-analyses of OncoArray, GWAS, iCOGS and CIMBA. We used all SNPs in the 1000 Genomes Project phase 1 v3 release that had a minor allele frequency > 1% and an imputation quality score R^2>0.3 in the OncoArray data. LD scores were calculated using the 1000 Genomes Project Phase 1 v3 EUR panel.

We created a "full baseline model" as previously described[24] that included 52 "baseline" genomic features (24 non-cell-type specific publicly available annotations, a 500-bp window around each of the 24 annotations and a 100-bp window around each of four ChIP-seq peaks) and one category containing all SNPs. We estimated the enrichment for these 53 functional categories in a single multivariable LD score regression analysis.

We subsequently performed analyses using cell-type specific annotations for the four histone marks H3K4me1, H3K4me3, H3K9ac and H3K27ac across 27-81 cell types, depending on histone mark, giving a total of 220 cell-type specific marks[24]. We estimated the enrichment for each of these marks after adjusting for the baseline annotations by running 220 LD score regressions, each adding a different histone mark to the baseline model. We observed no associations after adjusting for 220 tests

We tested the differences in functional enrichment between ER-positive and ER-negative subsets for individual features through a Wald test, using the regression coefficients and standard errors for the two subsets based on the models described above. Finally, we assessed the heritability due to genotyped and imputed SNPs[50] and estimated the genetic correlation between ER-positive and ER-negative breast cancer[11]. The genetic correlation analysis was restricted to the ~1M SNPs included in HapMap 3.

Pathway Enrichment Analyses

Pathway enrichment analysis was performed to identify pathways associated with ER-negative breast cancer risk, pointing to biological hypotheses that can be further tested experimentally.

The pathway gene set database Human_GOBP_AllPathways_no_GO_iea_January_19_2016_symbol.gmt (http://baderlab.org/GeneSets)[51], was used for all analyses. This database contains pathway gene sets from Reactome[52], NCI Pathway Interaction Database[53], GO (Gene Ontology) biological process[54], HumanCyc[55], MSigdb[56], NetPath[57] and Panther[58]. GO pathways inferred from electronic annotation terms were excluded. Some manual annotation was performed on the pathway gene set database where annotation errors from public data were discovered. In particular, in several pathways, the PDPK1 gene was mistakenly entered as PDK1 gene and was manually corrected. The same pathway (e.g. apoptosis) may be defined in two or more databases with potentially different sets of genes, and all versions of these duplicate/overlapping pathways were included. Pathway size was determined by the total number of genes in the pathway to which SNPs in the imputed GWAS dataset could be mapped. To provide more biologically meaningful results, and reduce false positives, only pathways that contained between 10 and 200 genes were considered.

Gene information (hg19) was downloaded from the ANNOVAR[59] website (http://www.openbioinformatics.org/annovar/). SNPs were mapped to the nearest gene within 500kb; those that were further than 500 kb away from any gene were excluded. Gene significance was calculated by assigning the lowest p-value observed across all SNPs assigned to a gene[60,61], based on the meta-analysis of BCAC and CIMBA data described above. Some pathways include genes that are also grouped closely together in the genome and are thus are likely to share the significance of a single SNP, which would artificially increase the pathway significance in our analysis. This was the case for pathways including histone genes. Thus, we selected representative SNP-gene associations to control for this effect (chr6:26055031 for HIST1, chr1:120904839, 149864043 for HIST2, chr1: 228615251 for HIST3 and chr12: 14919727 for HIST4).

The gene set enrichment analysis (GSEA)[51] algorithm, as implemented in the GenGen package (http://gengen.openbioinformatics.org/en/latest/)[61,62] was used to perform pathway analysis. Although there are several methods for pathway enrichment analysis, we chose the GSEA approach as it is one of the most established methods that is threshold free; many other methods such as SRT, ALIGATOR and Plink set-based test require an arbitrary p-value threshold to be defined for SNPs and applied before pathway analysis. Briefly, the algorithm calculates an enrichment score (ES) for each pathway based on a weighted Kolmogorov-Smirnov statistic[62]. Pathways that have most of their genes at the top of the ranked list of genes obtain higher ES values.

To focus on pathway enrichment analysis results about which we were most confident, we implemented a number of filters. First, only pathways with positive ES and containing at least one gene linked to a significant SNP ($P<5x10^{-8}$) were retained for subsequent analysis. Second, we defined an ES threshold (ES≥0.4086) based on a comparison with a gold standard pathway enrichment analysis we previously performed on the iCOGS data alone and where we were able to analytically compute FDR values by shuffling case/control labels (this was not computationally feasible with the more complex meta-analysis scheme used in this paper). This ES threshold was chosen to yield a true-positive rate (TPR) > 0.20 and a false-positive rate (FPR) < 0.15, with true-positive pathways defined as those observed with false discovery rate (FDR)<0.05 in a prior analysis carried out using the analytic approach defined above applied to iCOGS data for ER-negative disease.

We chose the true positive rate (TPR) threshold by varying the TPR in steps of 0.1 and observing how the FPR changed. A TPR of 0.1 resulted in a very low FPR (0.02), but we considered this to be unduly conservative as it resulted in a small number of pathways (40, clustered into 9 themes) and excluded many pathways known to be involved in breast cancer. A TPR of 0.2 (FPR = 0.15) gave a reasonable balance between the true and false positive rates, while including pathways known to be involved in breast cancer. Thus this threshold was chosen for this study. A TPR of 0.3 gave an FPR of 0.28, which we considered high; further, the resulting additional pathways included (in addition to those included at TPR=0.2) were weaker (i.e. they had worse enrichment scores [ES<0.4086] and had relatively very few genes included) than pathways appearing at lower FPRs (and TPRs). We rejected TPR thresholds >0.3 because each gave an FPR that was larger than the TPR.

Finally, we performed an in depth literature search on all resulting pathways to confirm their relevance to breast cancer biology, applying the following criteria: 1) reported in at least one of five published breast cancer pathway analyses[63-67]; or 2) reported elsewhere in the literature to be involved in breast cancer. We also removed pathways that were significant due to incorrect gene function annotation.

To visualize the pathway enrichment analysis results, an enrichment map was created using the Enrichment Map (EM) v 2.1.0 app[51] in Cytoscape v3.30[68], applying an edge-weighted force directed layout. To measure the contribution of each gene to enriched pathways and annotate the map, we reran the pathway enrichment analysis multiple times, each time excluding one gene. A gene was considered to drive the enrichment if the ES dropped to zero or less (pathway enrichment driver) after it was excluded. Pathways were grouped in the map if they shared >70% of their genes or their enrichment was driven by a shared gene.

REFERENCES

1. Siddiq, A. *et al.* A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum Mol Genet* **21**, 5373-84 (2012).
2. Haiman, C.A. *et al.* A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. *Nat Genet* **43**, 1210-4 (2011).
3. Ahsan, H. *et al.* A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. *Cancer Epidemiol Biomarkers Prev* **23**, 658-69 (2014).
4. Stevens, K.N. *et al.* 19p13.1 is a triple-negative-specific breast cancer susceptibility locus. *Cancer Res* **72**, 1795-803 (2012).
5. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat Genet* **42**, 504-7 (2010).
6. Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* **45**, 353-61 (2013).
7. Amos, C.I. *et al.* The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* (in press).
8. 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**, 373-80 (2015).
9. Couch, F.J. *et al.* Identification of four novel susceptibility loci for estrogen receptor negative breast cancer. *Nat Commun* (in press).

10. Garcia-Closas, M. *et al.* Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* **45**, 392-8 (2013).

11. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).

12. Li, Y. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).

13. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2012).

14. 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).

15. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955-9 (2012).

16. Li, Y., Willer, C.J., Ding, J., Scheet, P. & Abecasis, G.R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816-34 (2010).

17. Aulchenko, Y.S., Struchalin, M.V. & van Duijn, C.M. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics* **11**, 134 (2010).

18. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).

19. Barnes, D.R., Lee, A., Easton, D.F. & Antoniou, A.C. Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations. *Genet Epidemiol* **36**, 274-91 (2012).

20. Antoniou, A.C. *et al.* A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat Genet* **42**, 885-92 (2010).

21. Antoniou, A.C. *et al.* A weighted cohort approach for analysing factors modifying disease risks in carriers of high-risk susceptibility genes. *Genet Epidemiol* **29**, 1-11 (2005).

22. Udler, M.S., Tyrer, J. & Easton, D.F. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet Epidemiol* **34**, 463-8 (2010).

23. Maller, J.B. *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294-301 (2012).

24. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228-35 (2015).

25. Antoniou, A.C. & Easton, D.F. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol* **25**, 190-202 (2003).

26. Antoniou, A.C. *et al.* The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions. *Br J Cancer* **98**, 1457-66 (2008).

27. 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).

28. 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).
29. Ghoussaini, M. *et al.* Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat Commun* **4**, 4999 (2014).
30. 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).
31. Kheradpour, P. & Kellis, M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. *Nucleic Acids Res* **42**, 2976-87 (2014).
32. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9**, e1001046 (2011).
33. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-30 (2015).
34. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790-7 (2012).
35. Rao, S.S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-80 (2014).
36. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. *Proc Natl Acad Sci U S A* **111**, E2191-9 (2014).
37. 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).
38. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934-47 (2013).
39. Forrest, A.R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
40. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648-60 (2015).
41. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-43 (2013).
42. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70 (2012).
43. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346-52 (2012).
44. Baran, Y. *et al.* Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359-67 (2012).
45. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
46. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
47. Mermel, C.H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41 (2011).
48. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
49. Shabalin, A.A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353-8 (2012).

50. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
51. Merico, D., Isserlin, R., Stueker, O., Emili, A. & Bader, G.D. Enrichment map: a network-based method for gene-set enrichment visualization and interpretation. *PLoS One* **5**, e13984 (2010).
52. Joshi-Tope, G. *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* **33**, D428-32 (2005).
53. Schaefer, C.F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res* **37**, D674-9 (2009).
54. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9 (2000).
55. Romero, P. *et al.* Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* **6**, R2 (2005).
56. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
57. Kandasamy, K. *et al.* NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).
58. Thomas, P.D. *et al.* PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**, 2129-41 (2003).
59. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
60. Wang, L., Jia, P., Wolfinger, R.D., Chen, X. & Zhao, Z. Gene set analysis of genome-wide association studies: methodological issues and perspectives. *Genomics* **98**, 1-8 (2011).
61. Wang, K., Li, M. & Hakonarson, H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet* **11**, 843-54 (2010).
62. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81**, 1278-83 (2007).
63. Kar, S.P. *et al.* Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types. *Cancer Discov* **6**, 1052-67 (2016).
64. Braun, R. & Buetow, K. Pathways of distinction analysis: a new technique for multi-SNP analysis of GWAS data. *PLoS Genet* **7**, e1002101 (2011).
65. Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics* **27**, 95-102 (2011).
66. Mogushi, K. & Tanaka, H. PathAct: a novel method for pathway analysis using gene expression profiles. *Bioinformation* **9**, 394-400 (2013).
67. Medina, I. *et al.* Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res* **37**, W340-4 (2009).
68. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).