



Published in final edited form as:

Nat Genet. 2020 June ; 52(6): 572–581. doi:10.1038/s41588-020-0609-2.

Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses

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

Abstract

Breast cancer susceptibility variants frequently show heterogeneity in associations by tumor subtype^{1–3}. To identify novel loci, we performed a genome-wide association study (GWAS) including 133,384 breast cancer cases and 113,789 controls, plus 18,908 *BRCA1* mutation carriers (9,414 with breast cancer) of European ancestry, using both standard and novel methodologies that account for underlying tumor heterogeneity by estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status and tumor grade. We identified 32 novel susceptibility loci ($P < 5.0 \times 10^{-8}$), 15 of which showed evidence for associations with at least one tumor feature (false discovery rate (FDR) < 0.05). Five loci showed associations ($P < 0.05$) in opposite directions between luminal- and non-luminal subtypes. *In-silico* analyses showed these five loci contained cell-specific enhancers that differed between normal luminal and basal mammary cells. The genetic correlations between five intrinsic-like subtypes ranged from 0.35 to 0.80. The proportion of genome-wide chip heritability explained by all known susceptibility loci was 37.6% for triple-negative and 54.2% for luminal A-like disease. The odds ratios of polygenic risk scores (PRSs), which included 330 variants, for the highest 1% quantiles compared to middle quantiles were 5.63 and 3.02 for luminal A-like and triple-negative disease, respectively. These findings provide an improved understanding of genetic predisposition to breast cancer subtypes and will inform the development of subtype-specific polygenic risk scores.

Based on the largest GWAS to date from the Breast Cancer Association Consortium (BCAC), over 170 independent breast cancer susceptibility variants have been identified. Many of these variants show differential associations by tumor subtypes, particularly ER-positive versus ER-negative or triple-negative disease^{1–3}. However, prior GWAS have not simultaneously accounted for the high correlations between multiple, correlated tumor markers, such as ER, PR, HER2 and grade, to identify specific source(s) of etiologic heterogeneity. We performed a breast cancer GWAS using both standard analyses and a novel two-stage polytomous regression method that efficiently characterizes etiologic heterogeneity while accounting for tumor marker correlations and missing data⁴.

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

Corresponding Author: Nilanjan Chatterjee, 615 N. Wolfe Street, Room E3612, Baltimore, Maryland 21205, nchatte2@jhu.edu.

* Contributed equally

** Jointly supervised this work

Conflicts of interest: None to report

The study populations and genotyping are described elsewhere^{1,2,5,6} and in the Online Methods. Briefly, we analyzed data from 118,474 cases and 96,201 controls of European ancestry participating in 82 studies from the BCAC and 9,414 affected and 9,494 unaffected *BRCA1* mutation carriers from 60 studies from the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA) with genotyping data from one of two Illumina genome-wide custom arrays. In analyses of overall breast cancer, we also included summary level data from 11 other breast cancer GWAS (14,910 cases and 17,588 controls) without subtype information. Our study expands upon previous BCAC GWAS¹ with additional data on 10,407 cases and 7,815 controls, an approximate increase of 10% and 9%, respectively. (Supplementary Tables 1–4).

The statistical methods are further described in the Online Methods and in Extended Data Figure 1. To identify variants for overall breast cancer (invasive, *in situ* or unknown invasiveness) in BCAC, we used standard logistic regression to estimate odds ratios (OR) and 95% confidence-intervals (CI) adjusting for country and principal components (PCs). iCOGS and OncoArray data were evaluated separately and the results were combined with those from the 11 other GWAS using fixed-effects meta-analysis.

To identify breast cancer susceptibility variants displaying evidence of heterogeneity, we used a novel score-test based on a two-stage polytomous model⁴ that allows flexible, yet parsimonious, modelling of associations in the presence of underlying heterogeneity by ER, PR, HER2 and/or grade (Online Methods, Supplementary Note). The model handles missing tumor characteristic data by implementing an efficient Expectation-Maximization algorithm^{4,7}. These analyses were restricted to BCAC controls and invasive cases (Online Methods). We fit an additional two-stage model to estimate case-control ORs and 95% CI between the variants and intrinsic-like subtypes defined by combinations of ER, PR, HER2 and grade⁸ (Online Methods): (1) luminal A-like, (2) luminal B/HER2-negative-like, (3) luminal B-like, (4) HER2-enriched-like and (5) triple-negative or basal-like. We analyzed iCOGS and OncoArray data separately, adjusting for PCs and age, and meta-analyzed the results using a fixed-effects model. We evaluated the effect of country using a leave-one-out sensitivity analysis (Online Methods).

Among *BRCA1* mutation carriers who are prone to develop triple-negative disease⁹, we estimated per-allele hazard ratios (HRs) within a retrospective cohort analysis framework. We assumed estimated ORs for BCAC triple-negative cases and estimated HRs from CIMBA *BRCA1* carriers approximated the same underlying relative risk⁹, and we used a fixed-effect meta-analysis to combine these results (Online Methods). Among all novel variants, we used the two-stage polytomous model to test for heterogeneity in associations across subtypes, globally and by tumor-specific markers (Online Methods).

Overall, we identified 32 novel independent susceptibility loci marked by variants with $P < 5.0 \times 10^{-8}$ (Figure 1, Supplementary Table 5–7, Supplementary Figure 1–5): 22 variants using standard logistic regression, 16 variants using the two-stage polytomous model (eight of which were detected by standard logistic regression) and three variants in the CIMBA/BCAC-triple-negative meta-analysis (rs78378222 was also detected by the two-stage polytomous model in BCAC). Fourteen additional variants ($P < 5.0 \times 10^{-8}$) were excluded,

13 because they lacked evidence of association independent of known susceptibility variants in conditional analyses ($P = 1.0 \times 10^{-6}$; Supplementary Table 8–10), and one (chr22:40042814) for showing a high-degree of sensitivity in the leave-one-out country analysis following exclusion of studies from the USA (Supplementary Figure 6). Supplementary Figures 7–8 and Supplementary Table 11 show associations between all 32 variants and the intrinsic-like subtypes.

Fifteen of the 32 variants showed heterogeneity evidence ($FDR < 0.05$) according to the global heterogeneity test (Figure 2, Supplementary Table 12). ER (7 variants) and grade (7 variants) most often contributed to observed heterogeneity (marker-specific $P < 0.05$), followed by HER2 (4 variants) and PR (2 variants). rs17215231, identified in the CIMBA/BCAC-triple-negative meta-analysis, was the only variant found exclusively associated with triple-negative disease (OR=0.85, 95%CI=0.81–0.89). rs2464195, also identified in the CIMBA/BCAC-triple-negative meta-analysis, was associated with both triple-negative (OR=0.93, 95%CI=0.91–0.96) and luminal B-like subtypes (OR=0.96, 95%CI=0.92–0.99; Supplementary Table 11) and is in linkage disequilibrium (LD; $r^2=0.62$) with rs7953249, which is differentially associated with risk of ovarian cancer subtypes¹⁰. Five variants showed associations with luminal and non-luminal subtypes in opposite directions (Figure 3). Four variants were associated in opposite directions with luminal A-like and triple-negative subtypes (respectively, for rs78378222 OR=1.13, 95%CI=1.05–1.20 vs OR=0.67, 95%CI=0.57–0.80; for rs206435 OR=1.03, 95%CI=1.01–1.05 vs OR=0.95, 95%CI=0.92–0.98; for rs141526427 OR=0.96, 95%CI=0.94–0.98 vs OR=1.04, 95%CI=1.01–1.08; and for rs6065254 OR=0.96, 95%CI=0.94–0.97 vs OR=1.04, 95%CI=1.01–1.07). The tumor-marker heterogeneity test showed associations for rs78378222 with ER ($P_{ER} = 7.0 \times 10^{-6}$) and HER2 ($P_{HER2} = 2.07 \times 10^{-4}$), rs206435 with ER ($P_{ER} = 2.8 \times 10^{-3}$) and grade ($P_{grade} = 2.8 \times 10^{-4}$) and rs141526427 ($P_{ER} = 1.3 \times 10^{-3}$) and rs6065254 ($P_{ER} = 4.3 \times 10^{-3}$) with ER. rs7924772 showed opposite case-control associations between HER2-negative and HER2-positive subtypes and, consistent with these findings, was exclusively associated with HER2 ($P_{HER2} = 1.4 \times 10^{-6}$; Figure 3). rs78378222, located in the 3' UTR of *TP53*, also showed opposite associations with high-grade serous cancers (OR=0.75, $P = 3.7 \times 10^{-4}$) and low-grade serous cancers (OR=1.58, $P = 1.5 \times 10^{-4}$; -). Prior analyses¹¹ did not find rs78378222 associated with breast cancer risk, likely due to its opposite effects between subtypes.

Candidate causal variants were defined (CCVs; Online Methods) for each novel locus and we investigated the CCVs in relation to previously-annotated enhancers in primary breast cells¹². Based on combinations of H3K4me1 and H3K27ac histone modification ChIP-seq signals, putative enhancers in basal cells (BC), luminal progenitor cells (LP) and mature luminal cells (LM) were characterized as “OFF,” “PRIMED,” and “ACTIVE” (Online Methods). We defined “ANYSWITCH” enhancers as those exhibiting different characterizations between cell types. Among the five loci identified with associations in opposite directions between subtypes, at least one CCV per locus overlapped an “ANYSWITCH” enhancer (Figure 4). For example, rs78378222 overlapped an ACTIVE enhancer in basal cells, PRIMED in luminal progenitor cells and OFF in mature luminal cells. In comparison, 63% of the loci with consistent direction of associations across subtypes overlapped with an “ANYSWITCH” enhancer (Supplementary Table 13–14).

These results suggest that some variants may modulate enhancer activity in a cell-type specific manner, thus, differentially influencing risk of tumor subtypes.

We used INQUIST to intersect CCVs with functional annotation data from public databases to identify potential target genes¹ (Supplementary Note, Supplementary Table 15). We predicted 179 unique target genes for 26 of the 32 independent signals. Notably, rs78378222 has been reported associated with *TP53* mRNA levels in blood and adipose tissue¹¹, which we did not replicate in breast tissue. However, our findings of rs78378222 overlapping a cell type-specific regulatory element in breast basal epithelial cells, implicates enhancer function as another potential *TP53* transcriptional control mechanism. Twenty-three target genes in 14 regions were predicted with high confidence (designated “Level 1”), of which 22 target genes in 13 regions were predicted to be distally regulated. Four target genes were previously predicted by INQUISIT^{13,14}, *POLR3C*, *RNF115*, *SOX4* and *TBX3* – a known somatic breast cancer driver gene¹⁵ – and genes implicated by transcriptome-wide association studies (*LINC008861* and *YBEY17*).

We used LD-regression to investigate genetic correlations^{18,19} between subtypes and compare enrichment of genomic features²⁰ between luminal A-like and triple-negative subtypes (Online Methods). All subtypes were moderately- to highly correlated, with luminal A-like and triple-negative having a correlation of 0.46 (SE=0.05). The correlation in breast cancer of *BRCA1* carriers and triple-negative was 0.83 (SE=0.08), suggesting a high-degree of similarity in the genetic basis between these subtypes (Figure 5; Supplementary Table 16). To compare genomic enrichment, we first evaluated 53 annotations and found triple-negative tumors were most enriched for “super-enhancers, extend500bp” (3.04-fold, $P = 3.3 \times 10^{-6}$), and “digital genomic footprint, extend500bp” (from DNase hypersensitive sites) (2.2-fold, $P = 4.0 \times 10^{-4}$); however, no annotations significantly differed between luminal A-like and triple-negative tumors (Supplementary Table 17, Supplementary Figure 9). Investigating cell-specific enrichment of histone markers H3K4me1, H3K3me3, H3K9ac and H3K27ac (Supplementary Note) found both luminal-A and triple-negative subtypes enriched for gastrointestinal cell types and suppression of central nervous system cell types (Supplementary Figure 10).

The proportion of genome-wide chip heritability explained by the 32 novel variants, plus 178 previously identified variants^{1,2,21}, was 54.2%, 37.6% and 26.9% for luminal A-like, triple-negative and *BRCA1* carriers, respectively (Table 1, Supplementary Table 18). These 210 variants explained approximately 18.3% of the two-fold familial relative risk for invasive breast cancer, while all reliably imputable variants on the OncoArray explained 37.1% (Online Methods). The per-standard deviation ORs between PRSs for luminal-A like and triple-negative subtypes (Online Methods), that included 313 published variants²² and 17 novel variants that were independent of the 313 variants (Supplementary Table 19), was 1.83 (95% CI=1.78–1.88) and 1.65 (1.57–1.73), with corresponding area under receiver-operator curves of 66.09 and 63.58, respectively (Extended Data Figure 2–6).

These analyses demonstrate the benefit of combining standard GWAS methods with methods accounting for underlying tumor heterogeneity. Moreover, these methods and results may help clarify mechanisms predisposing to specific molecular subtypes, and

provide precise risk estimates for subtypes to inform development of subtype-specific PRSs²². However, to expand the generalizability of our findings, these analyses should be replicated and expanded in multi-ancestry populations.

Online Methods

Study populations

The overall breast cancer analyses included women of European ancestry from 82 BCAC studies from over 20 countries, with genotyping data derived from two Illumina genome-wide custom arrays, the iCOGS and OncoArray (Supplementary Table 1). Most of the studies were case-control studies in the general population, or hospital setting, or nested within population-based cohorts, but a subset of studies oversampled cases with a family history of the disease. We included controls and cases of invasive breast cancer, carcinoma *in-situ*, and cases of unknown invasiveness. Information on clinicopathologic characteristics were collected by the individual studies and combined in a central database after quality control checks. We used BCAC database version ‘freeze’ 10 for these analyses. Among a subset of participants (n=16,766) that were genotyped on both the iCOGS and OncoArray arrays, we kept only the OncoArray data. One study (LMBC) contributing to the iCOGS dataset was excluded due to inflation of the test statistics that was not corrected by adjustment for the first ten PCs. We also excluded OncoArray data from Norway (the Norwegian Breast Cancer Study) because there were no controls available from Norway with OncoArray data. All participating studies were approved by their appropriate ethics or institutional review board and all participants provided informed consent. The total sample size for this analysis, including iCOGS, OncoArray and other GWAS data, comprised 133,384 cases and 113,789 controls.

In the GWAS analyses accounting for underlying heterogeneity according to ER, PR, HER2 and grade, we included genotyping data from 81 BCAC studies. These analyses were restricted to controls and cases of invasive breast cancer. We excluded cases of carcinoma *in-situ* and cases with missing information on invasiveness, as ~96% of *in-situ* cases were missing some or all of the tumor markers and *in-situ* cases potentially have different tumor correlations compared to invasive cases, which could potentially bias the estimates from Expectation-Maximization algorithm (Supplementary Table 2). We also excluded all studies from a specific country if there were no controls for that country, or if the tumor marker data were missing on two or more of the tumor marker subtypes (see footnote of Supplementary Table 2 for further explanation of excluded studies). We did not include the summary results from the 14,910 cases and 17,588 controls from the 11 other GWAS in subtype analyses because these studies did not provide data on tumor characteristics. We also excluded invasive cases (n=293) and controls (n=4,285) with missing data on age at diagnosis or age at enrollment, information required by the Expectation-Maximization algorithm to impute missing tumor characteristics. In total, the final sample for the two-stage polytomous logistic regression comprised 106,278 invasive cases and 91,477 controls.

Participants included from CIMBA were women of European ancestry, aged 18 years or older with a pathogenic *BRCA1* variant. Most participants were sampled through cancer genetics clinics. In some instances, multiple members of the same family were enrolled.

OncoArray genotype data was available from 58 studies from 24 countries. Following quality control and removal of participants that overlapped 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 3). We also obtained iCOGS genotype data on 3,342 *BRCA1* mutation carriers (1,630 with breast cancer) from 54 studies through CIMBA. All *BRCA1* mutation carriers provided written informed consent and participated under ethically approved protocols.

Genotyping, quality control, and imputation

Details on genotype calling, quality control and imputation for the OncoArray, iCOGS, and GWAS are described elsewhere^{1,2,5,6}. Genotyped or imputed variants (including bi-allelic and multi-allelic single nucleotide polymorphisms (SNPs) and small indels) marking each of the loci were determined using the iCOGS and the OncoArray genotyping arrays and imputation to the 1000 Genomes Project (Phase 3) reference panel. We included variants, from each component GWAS with an imputation quality score of >0.3 . We restricted analysis to variants with a minor allele frequency >0.005 in the overall breast cancer analysis and >0.01 in the subtype analysis.

Known breast cancer susceptibility variants

Prior studies identified susceptibility variants from genome-wide analyses at a significance level $P < 5.0 \times 10^{-8}$ for all breast cancer types, ER-negative or ER-positive breast cancer, in *BRCA1* or *BRCA2* mutation carriers, or in meta-analyses of these¹⁻³. We defined known breast cancer susceptibility variants as those variants that were identified or replicated in prior BCAC analyses^{1,2}. To help ensure that novel, independent susceptibility variants were identified, we excluded from these analyses variants within 500 kb of a previously published variant. These excluded regions have been subject to a separate, fine-mapping conditional analyses that are focused on identifying additional independent susceptibility variants in these regions¹⁴.

Standard analysis of BCAC data

Logistic regression analyses were conducted separately for the iCOGS and OncoArray datasets, adjusting for country and the array-specific first 10 PCs for ancestry informative variants. The methods for estimating PCs have been described elsewhere^{1,2}. For the remaining GWAS, adjustment for inflation was done by adjusting for up to three PCs and using genomic control adjustment, as previously described¹. We evaluated the associations between approximately 10.8 million variants with imputation quality scores (r^2) ≥ 0.3 and minor allele frequency (MAF) >0.005 . We excluded variants located within ± 500 kb of, or in LD ($r^2 \geq 0.1$) with known susceptibility variants²¹. The association effect size estimates from these, and the previously derived estimates from the 11 other GWAS, were then combined using a fixed effects meta-analysis. Since individual level genotyping data were not available for some previous GWAS, we conservatively approximated the potential overlap between the GWAS and iCOGS and OncoArray datasets, based on the populations contributing to each GWAS (iCOGS/GWAS: 626 controls and 923 cases; OncoArray/GWAS: 20 controls and 990 cases). We then used these adjusted data to estimate the correlation in the effect size

estimates, and incorporated these into the meta-analysis using the method of Lin and Sullivan²³.

Subtypes analysis of BCAC data

We described the two-stage polytomous logistic regression in more detail elsewhere^{4,24} (Supplementary Note). In brief, this method allows for efficient testing of a variant-disease association in the presence of tumor subtype heterogeneity defined by multiple tumor characteristics, while accounting for multiple testing and missing data on tumor characteristics. In the first stage, the model uses a polytomous logistic regression to model case-control ORs between the variants and all possible subtypes that could be of interest, defined by the combination of the tumor markers. For example, in a model fit to evaluate heterogeneity according to ER, PR and HER2 positive/negative status, and grade of differentiation (low, intermediate and high grade), the first stage incorporates case-control ORs for 24 subtypes defined by the cross-classification of these factors. The second stage restructures the first-stage subtype-specific case-control ORs parameters into second-stage parameters through a decomposition procedure resulting in a second-stage baseline parameter that represents a case-control OR of a baseline cancer subtype, and case-case ORs parameters for each individual tumor characteristic. The second-stage case-case parameters can be used to perform heterogeneity tests with respect to each specific tumor marker while adjusting for the other tumor markers in the model. The two-stage model efficiently handles missing data by implementing an Expectation-Maximization algorithm^{4,7} that essentially performs iterative “imputation” of the missing tumor characteristics conditional on available tumor characteristics and baseline covariates based on an underlying two-stage polytomous model. In the two-stage model, the frequency of different tumor subtypes corresponding to different combinations of the tumor characteristics are allowed to vary freely through the model-free specification of the intercepts of the first-stage polytomous model (α_m , see Supplementary Note for details), in other words, the intercepts are kept saturated. As these parameters are estimated from the data itself, the methodology accounts for the correlation among the tumor markers in a robust manner that does not require strong modelling assumptions.

To identify novel susceptibility loci, we used both a fixed-effect two-stage polytomous model and a mixed-effect two-stage polytomous model. The score-test we developed based on the mixed-effect model allows coefficients associated with individual tumor characteristics to enter as either fixed- or random-effect terms. Our previous analyses have shown that incorporation of random effect terms can improve power of the score-test by essentially reducing the effective degrees-of-freedom associated with fixed effects related to exploratory markers (*i.e.*, markers for which there is little prior evidence to suggest that they are a source of heterogeneity)⁴. On the other hand, incorporation of fixed-effect terms can preserve distinct associations of known important tumor characteristics, such as ER. In the mixed-effect two-stage polytomous model, we therefore kept ER as a fixed effect, but modeled PR, HER2 and grade as random effects. We evaluated variants with MAF >0.01 (~10.0 million) and $r^2 \geq 0.3$, and excluded variants within ± 500 kb of, or in LD ($r^2 \geq 0.1$) with known susceptibility variants. A MAF >0.01 was chosen to ensure an adequate sample size

to generate stable estimates. We reported variants that passed the p-value threshold of $P < 5.0 \times 10^{-8}$ in either the fixed- or mixed-effect models.

Both fixed/mixed-effect models adjusted for top ten PCs and age. As age is correlated with the tumor characteristics²⁵, we added age as a covariate to improve the statistical power of Expectation-Maximization (EM) algorithm. Country was not adjusted for in the subtype analyses, since doing so required adequate sample size of each subtype in each country to allow for convergence of the two-stage polytomous model. Instead, we assessed the influence of country on signals identified by the two-stage models by performing a ‘leave one out’ sensitivity analyses in which we reevaluated novel signals after excluding data from each individual country. Data from the OncoArray and iCOGS arrays were analyzed separately and then meta-analyzed using fixed-effects meta-analysis.

Statistical analysis of CIMBA data

We tested for associations between variants and breast cancer risk for *BRCA1* mutation carriers using a score test statistic based on the retrospective likelihood of observing the variant genotypes conditional on breast cancer phenotypes (breast cancer status and censoring time)²⁶. Analyses were performed separately for iCOGS and OncoArray data. To allow for non-independence among related individuals, a kinship-adjusted test was used that accounted for familial correlations²⁷. We stratified analyses by country of residence and, for countries where the strata were sufficiently large (United States and Canada), by Ashkenazi Jewish ancestry. The results from the iCOGS and OncoArray data were then pooled using fixed-effects meta-analysis.

Meta-analysis of BCAC and CIMBA

As the great majority of *BRCA1* related breast cancers are triple-negative²⁸, we performed a meta-analysis with the BCAC triple-negative results to increase the power to detect associations for the triple-negative subtype. We performed a fixed-effects meta-analysis of the results from BCAC triple-negative cases and CIMBA *BRCA1* mutation carriers, using an inverse-variance fixed-effects approach implemented in METAL²⁹. The estimates of association used were the logarithm of the per-allele hazard ratio estimate for association with breast cancer risk for *BRCA1* mutation carriers from CIMBA and the logarithm of the per-allele odds ratio estimate for association with risk of triple-negative breast cancer based on BCAC data.

Conditional analyses

We performed two sets of conditional analyses. First, we investigated for evidence of multiple independent signals in identified loci by performing forward selection logistic regression, in which we adjusted the lead variant and analyzed association for all remaining variants within ± 500 kb of the lead variants, irrespective of LD. Second, we confirmed the independence of 20 variants that were located within ± 2 MB of a known susceptibility region by conditioning the identified signals on the nearby known signal. Since these 20 variants are already genome-wide significant in the original GWAS scan and the conditional analyses restricted to local regions, we therefore used a significance threshold of $P < 1 \times 10^{-6}$ to control for type-one error³⁰.

Heterogeneity analysis of new association signals

We evaluated all novel signals for evidence of heterogeneity using the two-stage polytomous model. We first performed a global test for heterogeneity under the mixed-effect model test to identify variants showing evidence of heterogeneity with respect to any of the underlying tumor markers, ER, PR, HER2 and/or grade. We accounted for multiple testing of the global heterogeneity test using a FDR <0.05 under the Benjamini-Hochberg procedure³¹. Among the variants with observed heterogeneity, we then further used a fixed-effect two-stage model to evaluate influence of specific tumor characteristic(s) driving observed heterogeneity, adjusted for the other markers in the model. We also fit a separate fixed-effect two-stage models to estimate case-control ORs and 95% confidence intervals (CI) for five surrogate intrinsic-like subtypes defined by combinations of ER, PR, HER2 and grade⁸: (1) luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); (2) luminal B/HER2-negative-like (ER+ and/or PR+, HER2-, grade 3); (3) luminal B-like (ER+ and/or PR+, HER2+); (4) HER2-enriched-like (ER- and PR-, HER2+), and (5) triple-negative (ER-, PR-, HER2-). Further, we conducted sensitivity analysis by fitting a standard polytomous model among cases with complete data on the five-intrinsic-like subtypes for the 32 novel variants and compared these results with the results from two-stage polytomous model accounting for missing tumor data.

Candidate causal variants

We defined credible sets of candidate causal variants (CCVs) as variants located within ± 500 kb of the lead variants in each novel region and with *P* values within 100-fold of magnitude of the lead variants. This is approximately equivalent to selecting variants whose posterior probability of causality is within two orders of magnitude of the most significant variant^{32,33}. This approach was applied for detecting a set of potentially causal variants for all 32 identified variants. For the novel variants located within ± 2 Mb of the known signals, we used the conditional *P* values to adjust for the known signals' associations.

Enhancer states analysis in breast sub-populations

We obtained enhancer maps for three enriched primary breast sub-populations (basal, luminal progenitor, and mature luminal) from Pellacani et al.¹². Enhancer annotations were defined as ACTIVE, PRIMED, or OFF based on a combination of H3K27ac and H3K4me1 histone modification CHIP-seq signals using FPKM thresholds as previously described¹². Briefly, genomic regions containing high H3K4me1 signal observed in any cell type were used to define the superset of breast regulatory elements. Sub-population cell type-specific H3K27ac signal (which is characteristic of active elements) within these elements was used as a measure of overall regulatory activity, where "ACTIVE" sites were characterized by H3K4me1-high, H3K27ac-high; "PRIMED" by H3K4me1-high, H3K27ac-low; and "OFF" by H3K4me1-low, H3K27ac-low. This enabled annotation of each enhancer element as either "OFF", "PRIMED" or "ACTIVE" in all cell types. We then defined enhancers which exhibit differing states between at least one cell type as "ANYSWITCH" enhancers.

Genetic correlation analyses

We used LD score regression^{18–20} to estimate the genetic correlation between five intrinsic-like breast cancer subtypes. The analysis used the summary statistics based on the meta-analysis of the OncoArray, and iCOGS, and CIMBA meta-analysis. The genetic correlation¹⁸ analysis was restricted to the ~1 million variants included in HapMap 3 with MAF > 1% and imputation quality score R²>0.3 in the OncoArray data. Since two-stage polytomous models integrated an imputation algorithm for missing tumor characteristic data, we modified the LD score regression to generate the effective sample size for each variant (Supplementary Note).

Genetic variance explained by identified susceptibility variants and all genome-wide imputable variants

Genetic variance corresponds to heritability on the frailty-scale, which assumes a polygenetic log-additive model as the underlying model. Under the log-additive model, the frailty-scale heritability explained by the identified variants can be estimated by:

$$\sum_{i=1}^n 2p_i(1-p_i)(\hat{\beta}_i^2 - \tau_i^2),$$

where n is the total number of identified variants, p_i is the MAF for i th variant, $\hat{\beta}_i$ is the log odds ratio estimate for the i th variant, and τ_i is the standard error of $\hat{\beta}_i$. To obtain the frailty scale heritability for invasive breast cancer explained by all of the GWAS variants, we used LD score regression to estimate heritability (σ_{GWAS}^2) using the full set of summary statistics from either standard logistic regression for overall invasive breast cancer, the two-stage polytomous regression for the intrinsic-like subtypes, or the CIMBA *BRCA1* analysis for *BRCA1* carriers. σ_{GWAS}^2 is characterized by population variance of the underlying true polygenetic risk scores as $\sigma_{GWAS}^2 = Var(\sum_{m=1}^M \beta_m G_m)$, where G_m is the standardized genotype for m th variant, β_m is the true log odds ratio m th variant, and M are the total number of causal variants among the GWAS variants. Thus, the proportion of heritability explained by identified variants relative to all imputable variants is:

$$\sum_{i=1}^n 2p_i(1-p_i)(\hat{\beta}_i^2 - \tau_i^2) / \sigma_{GWAS}^2.$$

To estimate the proportion of the familial risk of invasive breast cancer that is explained by susceptibility variants, we defined the familial relative risk, λ , as the familial relative risk assuming a polygenic log-additive model that explains all the familial aggregation of the disease³⁴. Under the frailty scale, we define the broad sense heritability³⁵ as σ^2 . The relationship between λ and σ^2 was shown to be $\sigma^2 = 2 * \log(\lambda)$ ³⁴. We assumed $\lambda = 2$ as the overall familial relative risk of invasive breast cancer³⁴ thus $\sigma^2 = 2\log(2)$ and the proportion of the familial relative risk explained by identified susceptibility variants is

$\sum_{i=1}^n 2p_i(1-p_i)(\hat{\beta}_i^2 - \tau_i^2) / \log(2)$, and the proportion of the familial relative risk explained by

GWAS variants is $\sigma_{GWAS}^2/[2 - \log(2)]$. Analysis of heritability and the proportion of explained familial risk were restricted to 106,278 invasive cases and 91,477 controls (Supplementary Table 2). In addition, we compared estimates of GWAS chip heritability across five-intrinsic subtypes using LD-score regression where the summary statistics were derived using either standard polytomous model applied to complete cases or the novel two-stage method that incorporates cases with missing tumor characteristics.

PRSs for five intrinsic-like subtypes

We constructed PRSs for the intrinsic-like subtypes, incorporating the newly identified variants and 313 variants previously reported in the development of PRSs for overall and ER-specific breast cancer²². The 313 SNPs include SNPs that didn't reach genome-wide significance. After excluding variants within 500 kb of the 313 SNPs or $LD \geq 0.1$, 17 out of the 32 novel variants were independent with the 313 SNPs. The BCAC data were split into the training dataset and test dataset with a proportion of 80% and 20%, respectively. Half of the test dataset were five studies nested within prospective cohorts including KARMA, MMHS, PLCO, SISTER, UKBGS (Supplementary Table 2) and the other half was randomly selected among the subjects in OncoArray, excluding studies of bilateral breast cancer, studies or sub studies with oversampling for family history, cases with ambiguous diagnosis, and cases with missing tumor characteristics. We obtained the overall and ER-specific log odds ratios for 313 SNPs by respectively fitting standard and ER-specific logistic regression on the training dataset. We obtained the log odds ratio for 330 SNPs by fitting the fixed-effect two-stage polytomous model for five intrinsic-like subtypes on the training dataset (Supplementary Table 19).

Reporting Summary

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

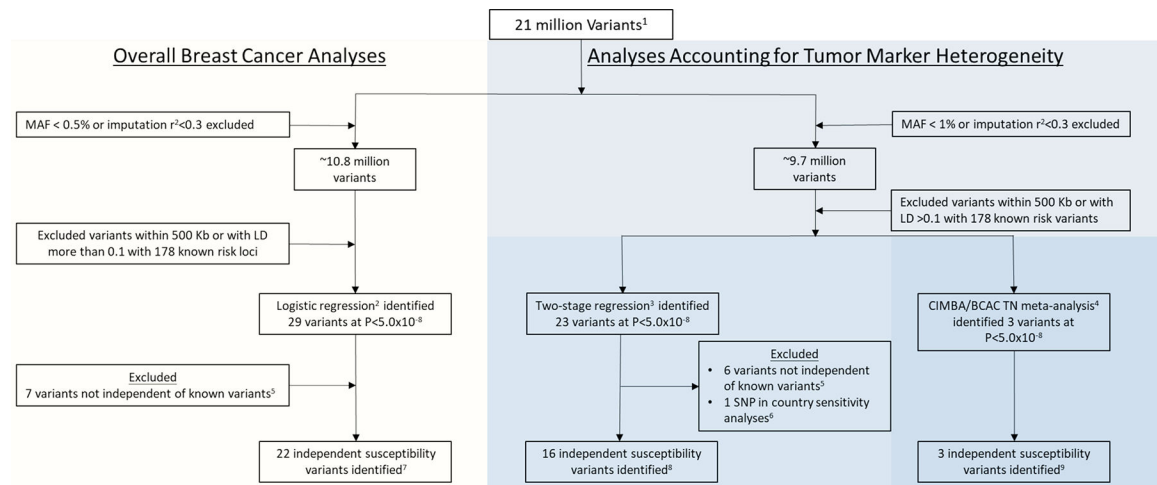
Data Availability Statement

Summary level statistics are available from <http://bcac.ccge.medschl.cam.ac.uk/bcacdata/> and <http://cimba.ccge.medschl.cam.ac.uk/projects/>. Requests for data can be made to the corresponding author or the Data Access Coordination Committees (DACCs) of BCAC (see above URL) and CIMBA (see above URL). BCAC DACC approval is required to access data from the ABCFS, ABCS, ABCTB, BBCC, BBCS, BCEES, BCFR-NY, BCFR-PA, BCFR-UT, BCINIS, BSUCH, CBCS, CECILE, CGPS, CTS, DIETCOMPLYF, ESTHER, GC-HBOC, GENICA, GEPARSIXTO, GESBC, HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KBCP, LMBC, MABCS, MARIE, MBCSG, MCBCS, MISS, MMHS, MTLGEBCS, NC-BCFR, OFBCR, ORIGO, pKARMA, POSH, PREFACE, RBCS, SKKDKFZS, SUCCESSB, SUCCESSC, SZBCS, TNBCC, UCIBCS, UKBGS and UKOPS studies (Supplementary Table 1). CIMBA DACC approval is required to access data from the BCFR-ON, CONSIT TEAM, DKFZ, EMBRACE, FPGMX, GC-HBOC, GEMO, G-FAST, HEBCS, HEBON, IHCC, INHERIT, IOVHBOCS, IPOBCS, MCGILL, MODSQUAD, NAROD, OCGN, OUH and UKGRFOCR studies (Supplementary Table 3).

Code Availability statement

The data analysis code of this paper is available at https://github.com/andrewhaoyu/breast_cancer_data_analysis. The implementation of this two-stage polytomous regression method is available in a R package called TOP (<https://github.com/andrewhaoyu/TOP>) with a detailed tutorial available at <https://github.com/andrewhaoyu/TOP/blob/master/inst/TOP.pdf>.

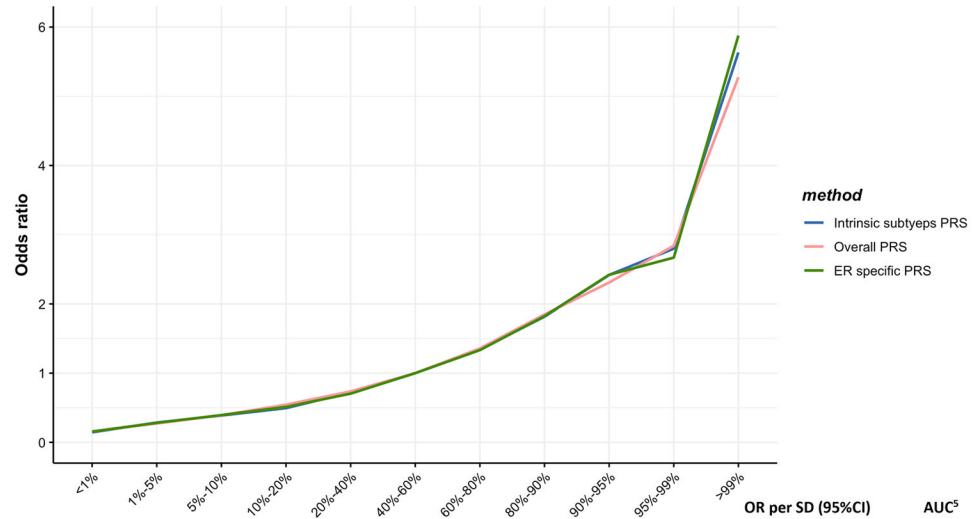
Extended Data



Extended Data Fig. 1. Overview of the analytic strategy and results from the investigation of breast cancer susceptibility variants in women of European descent.

Analyses included investigating for susceptibility variants for overall breast cancer (invasive, in-situ or unknown invasiveness) and for susceptibility variants accounting for tumor heterogeneity according to the estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2), and grade, and specifically investigating for variants that predispose for risk of the triple-negative subtype. 1) Genotyping data from two Illumina genome-wide custom arrays, the iCOGS and Oncoarray, and imputed to the 1000 Genomes Project (Phase 3). (2) Overall breast cancer (invasive, in-situ, or unknown invasiveness) analyses included 82 studies from the Breast Cancer Association Consortium (BCAC; 118,474 cases and 96,201 controls) and summary level data from 11 other breast cancer GWAS (14,910 cases and 17,588 controls; Supplementary Table 1). (3) Analyses accounting for tumor marker heterogeneity according to ER, PR, HER2 and grade included 81 studies from BCAC (106,278 invasive cases and 91,477 controls). (4) Analyses investigating triple-negative susceptibility variants included 91,477 controls and 8,602 triple-negative TN (effective sample, see Supplementary Note) cases from BCAC and 9,414 affected and 9,494 unaffected BRCA1/2 carriers from 60 studies from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA; Supplementary Table 3). (5) Variants excluded following conditional analyses showing the identified variants to not be independent ($P > 1 \times 10^{-6}$) of 178 known susceptibility variants (see Online Methods). (6) See Supplementary Figure 6 for results of country-specific sensitivity analyses. (7) See Supplementary Table 5 for the 22 independent susceptibility variants identified in overall breast cancer analyses. (8) See Supplementary Table 6 for the 16 independent susceptibility

variants identified using two-stage polytomous regression, accounting for tumor markers heterogeneity according to ER, PR, HER2, and grade. Note that 8 of the 16 variants were also detected in the overall breast cancer analysis (9) See Supplementary Table 7 for the 3 independent susceptibility variants identified in the CIMBA / BCAC- triple-negative TN meta-analysis. Note that rs78378222 was detected in both the analyses using the two-stage polytomous regression and in CIMBA / BCAC- triple-negative TN.



| | $\le 1\%$ | 1%-5% | 5%-10% | 10%-20% | 20%-40% | 40%-60% | 60%-80% | 80%-90% | 90%-95% | 95%-99% | >99% | OR per SD (95%CI) | AUC ⁵ |
|---|-----------|-------|--------|---------|---------|---------|---------|---------|---------|---------|------|-------------------|------------------|
| Intrinsic subtypes PRS ORs¹ | 0.14 | 0.29 | 0.39 | 0.49 | 0.72 | 1.00 | 1.34 | 1.82 | 2.42 | 2.80 | 5.63 | 1.83 (1.78-1.88) | 66.09 |
| Overall PRS ORs² | 0.16 | 0.27 | 0.39 | 0.54 | 0.74 | 1.00 | 1.36 | 1.85 | 2.31 | 2.84 | 5.27 | 1.80 (1.75-1.86) | 65.73 |
| ER Specific PRS ORs³ | 0.16 | 0.28 | 0.39 | 0.51 | 0.70 | 1.00 | 1.33 | 1.82 | 2.42 | 2.67 | 5.88 | 1.82 (1.77-1.87) | 65.95 |

¹ Intrinsic-like subtypes PRS based on 330 SNPs (Online Methods, Supplementary Table 19)

² Overall breast cancer PRS with 313 SNPs previously reported²²

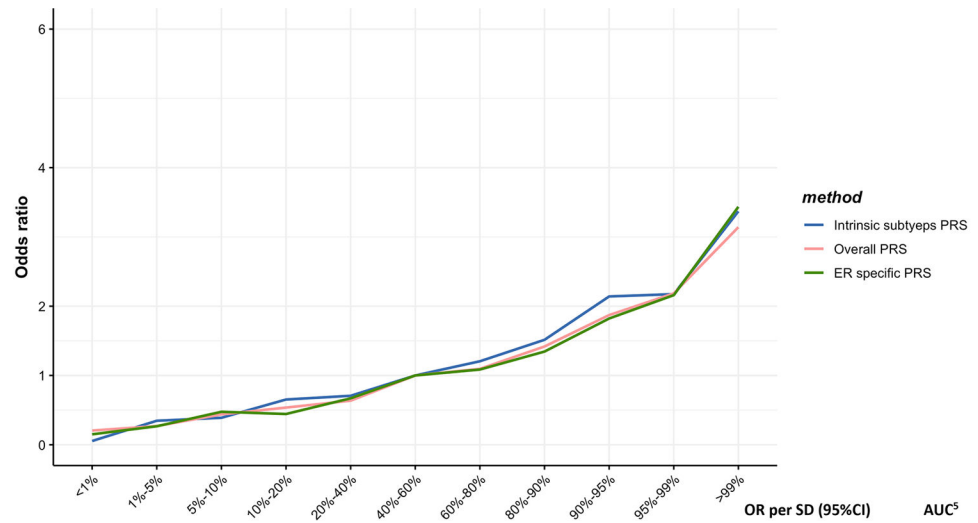
³ ER-specific PRS with 313 SNPs previously reported²²

⁴ Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2).

⁵ Area under the curve

Extended Data Fig. 2. Associations between three different polygenetic risk scores^{1,2,3} and luminal A-like⁴ risk in the test dataset.

Odds ratios for different quantiles of the PRS against the middle quantile of the PRS. The odds ratios were estimated using the test dataset like (n = 7,325 Luminal-A like cases, n = 20,815 controls).



| | | OR per SD (95%CI) | AUC ⁵ | | | | | | | | | | |
|---|------|-------------------|------------------|------|------|------|------|------|------|------|------|------------------|-------|
| Intrinsic subtypes PRS ORs¹ | 0.05 | 0.35 | 0.39 | 0.65 | 0.71 | 1.00 | 1.20 | 1.51 | 2.14 | 2.17 | 3.37 | 1.62 (1.54-1.70) | 63.30 |
| Overall PRS ORs² | 0.20 | 0.27 | 0.44 | 0.54 | 0.64 | 1.00 | 1.10 | 1.42 | 1.87 | 2.19 | 3.14 | 1.62 (1.55-1.71) | 63.36 |
| ER Specific PRS ORs³ | 0.15 | 0.27 | 0.48 | 0.44 | 0.67 | 1.00 | 1.09 | 1.35 | 1.82 | 2.16 | 3.43 | 1.62 (1.54-1.70) | 63.23 |

¹ Intrinsic-like subtypes PRS based on 330 SNPs (**Online Methods, Supplementary Table 19**)

² Overall breast cancer PRS with 313 SNPs previously reported²²

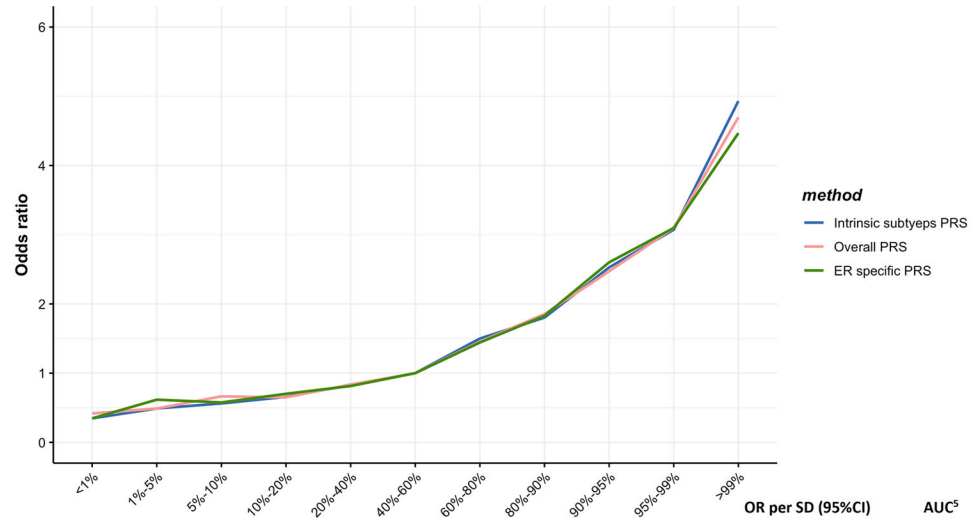
³ ER-specific PRS with 313 SNPs previously reported²²

⁴ Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2).

⁵ Area under the curve

Extended Data Fig. 3. Associations between three different polygenetic risk scores^{1,2,3} and luminal B/HER2-negative-like⁴ risk in the test dataset.

Odds ratios for different quantiles of the PRS against the middle quantile of the PRS. The odds ratios were estimated using the test dataset like (n= 1,779 Luminal B/HER2-negative-like cases, n = 20,815 controls).



| | $\le 1\%$ | 1%-5% | 5%-10% | 10%-20% | 20%-40% | 40%-60% | 60%-80% | 80%-90% | 90%-95% | 95%-99% | >99% | OR per SD (95%CI) | AUC ⁵ |
|---|-----------|-------|--------|---------|---------|---------|---------|---------|---------|---------|------|-------------------|------------------|
| Intrinsic subtypes PRS ORs¹ | 0.35 | 0.49 | 0.56 | 0.66 | 0.83 | 1.00 | 1.50 | 1.80 | 2.53 | 3.07 | 4.93 | 1.69 (1.61-1.78) | 64.31 |
| Overall PRS ORs² | 0.42 | 0.49 | 0.66 | 0.65 | 0.84 | 1.00 | 1.45 | 1.85 | 2.47 | 3.10 | 4.70 | 1.68 (1.60-1.77) | 64.32 |
| ER Specific PRS ORs³ | 0.35 | 0.62 | 0.58 | 0.70 | 0.81 | 1.00 | 1.44 | 1.83 | 2.60 | 3.10 | 4.47 | 1.66 (1.58-1.75) | 64.00 |

¹ Intrinsic-like subtypes PRS based on 330 SNPs (**Online Methods, Supplementary Table 19**)

² Overall breast cancer PRS with 313 SNPs previously reported²²

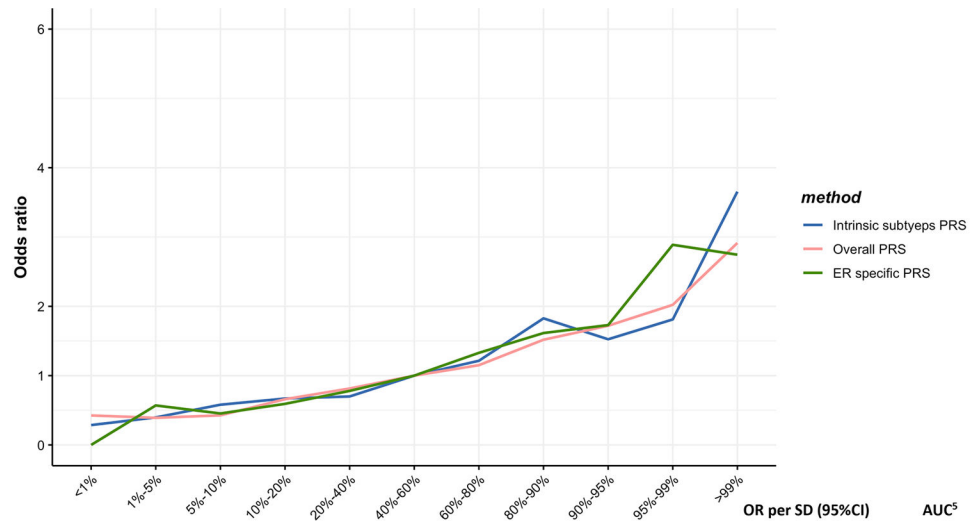
³ ER-specific PRS with 313 SNPs previously reported²²

⁴ Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2).

⁵ Area under the curve

Extended Data Fig. 4. Associations between three different polygenetic risk scores^{1,2,3} and luminal B-like⁴ risk in the test dataset.

Odds ratios for different quantiles of the PRS against the middle quantile of the PRS. The odds ratios were estimated using the test dataset like (n = 1,682 Luminal B-like cases, n = 20,815 controls).



| | <1% | 1%-5% | 5%-10% | 10%-20% | 20%-40% | 40%-60% | 60%-80% | 80%-90% | 90%-95% | 95%-99% | >99% | OR per SD (95%CI) | AUC ⁵ |
|---|------|-------|--------|---------|---------|---------|---------|---------|---------|---------|------|-------------------|------------------|
| Intrinsic subtypes PRS ORs ¹ | 0.29 | 0.40 | 0.58 | 0.67 | 0.70 | 1.00 | 1.21 | 1.83 | 1.52 | 1.81 | 3.65 | 1.53 (1.42-1.65) | 62.08 |
| Overall PRS ORs ² | 0.43 | 0.39 | 0.43 | 0.66 | 0.81 | 1.00 | 1.15 | 1.52 | 1.72 | 2.02 | 2.91 | 1.49 (1.38-1.60) | 60.93 |
| ER Specific PRS ORs ³ | 0.00 | 0.57 | 0.45 | 0.59 | 0.78 | 1.00 | 1.33 | 1.61 | 1.73 | 2.89 | 2.74 | 1.59 (1.48-1.71) | 62.91 |

¹ Intrinsic-like subtypes PRS based on 330 SNPs (Online Methods, Supplementary Table 19)

² Overall breast cancer PRS with 313 SNPs previously reported²²

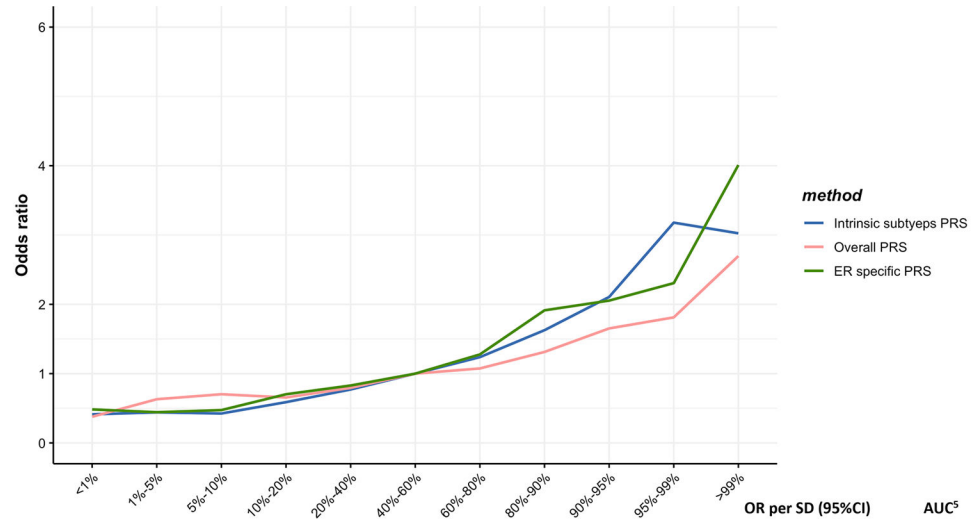
³ ER-specific PRS with 313 SNPs previously reported²²

⁴ Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2).

⁵ Area under the curve

Extended Data Fig. 5. Associations between three different polygenetic risk scores^{1,2,3} and HER2-enriched-like⁴ risk in the test dataset.

Odds ratios for different quantiles of the PRS against the middle quantile of the PRS. The odds ratios were estimated using the test dataset like (n = 718 HER2-enriched- like, n = 20,815 controls).



| | <1% | 1%-5% | 5%-10% | 10%-20% | 20%-40% | 40%-60% | 60%-80% | 80%-90% | 90%-95% | 95%-99% | >99% | OR per SD (95%CI) | AUC ⁵ |
|---|------|-------|--------|---------|---------|---------|---------|---------|---------|---------|------|-------------------|------------------|
| Intrinsic subtypes PRS ORs ¹ | 0.41 | 0.44 | 0.42 | 0.59 | 0.77 | 1.00 | 1.24 | 1.63 | 2.11 | 3.18 | 3.02 | 1.65 (1.57-1.73) | 63.58 |
| Overall PRS ORs ² | 0.38 | 0.63 | 0.70 | 0.66 | 0.80 | 1.00 | 1.07 | 1.31 | 1.65 | 1.81 | 2.70 | 1.38 (1.31-1.44) | 58.77 |
| ER Specific PRS ORs ³ | 0.48 | 0.44 | 0.47 | 0.70 | 0.83 | 1.00 | 1.28 | 1.91 | 2.05 | 2.31 | 4.01 | 1.59 (1.51-1.66) | 62.76 |

¹ Intrinsic-like subtypes PRS based on 330 SNPs (Online Methods, Supplementary Table 19)

² Overall breast cancer PRS with 313 SNPs previously reported²²

³ ER-specific PRS with 313 SNPs previously reported²²

⁴ Luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2).

⁵ Area under the curve

Extended Data Fig. 6. Associations between three different polygenetic risk scores^{1,2,3} and triple-negative⁴ risk in the test dataset.

Odds ratios for different quantiles of the PRS against the middle quantile of the PRS. The odds ratios were estimated using the test dataset like (n = 2,006 triple-negative cases, n = 20,815 controls).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Haoyu Zhang^{1,2,*}, Thomas U. Ahearn^{1,*}, Julie Lecarpentier³, Daniel Barnes³, Jonathan Beesley⁴, Guanghao Qi², Xia Jiang⁵, Tracy A. O'Mara⁴, Ni Zhao², Manjeet K. Bolla⁶, Alison M. Dunning³, Joe Dennis⁶, Qin Wang⁶, Zumuruda Abu Ful⁷, Kristiina Aittomäki⁸, Irene L. Andrulis⁹, Hoda Anton-Culver¹⁰, Volker Arndt¹¹, Kristan J. Aronson¹², Banu K. Arun¹³, Paul L. Auer^{14,15}, Jacopo Azzollini¹⁶, Daniel Barrowdale¹⁷, Heiko Becher¹⁸, Matthias W. Beckmann¹⁹, Sabine Behrens²⁰, Javier Benitez²¹, Marina Bermisheva²², Katarzyna Bialkowska²³, Ana Blanco^{24,25,26}, Carl Blomqvist^{27,28}, Natalia V. Bogdanova^{29,30,31}, Stig E. Bojesen^{32,33,34,35}, Bernardo Bonanni³⁶, Davide Bondavalli³⁶, Ake Borg³⁷, Hiltrud Brauch^{38,39,40}, Hermann Brenner^{11,40,41}, Ignacio Briceno⁴², Annegien Broeks⁴³, Sara Y. Brucker⁴⁴, Thomas Brüning⁴⁵, Barbara Burwinkel^{46,47}, Sandra S. Buys⁴⁸, Helen Byers⁴⁹, Trinidad Caldés⁵⁰, Maria A. Caligo⁵¹, Mariarosaria Calvella³⁶, Daniele Campa^{20,52}, Jose E. Castelao⁵³, Jenny Chang-Claude^{20,54}, Stephen J. Chanock¹, Melissa Christiaens⁵⁵,

Hans Christiansen³¹, Wendy K. Chung⁵⁶, Kathleen B.M. Claes⁵⁷, Christine L. Clarke⁵⁸, Sten Cornelissen⁴³, Fergus J. Couch⁵⁹, Angela Cox⁶⁰, Simon S. Cross⁶¹, Kamila Czene⁶², Mary B. Daly⁶³, Peter Devilee⁶⁴, Orland Diez⁶⁵, Susan M. Domchek⁶⁶, Thilo Dörk³⁰, Miriam Dwek⁶⁷, Diana M. Eccles⁶⁸, Arif B. Ekici⁶⁹, D.Gareth Evans^{70,49}, Peter A. Fasching^{71,19}, Jonine Figueroa⁷², Lenka Foretova⁷³, Florentia Fostira⁷⁴, Eitan Friedman⁷⁵, Debra Frost¹⁷, Manuela Gago-Dominguez^{76,77}, Susan M. Gapstur⁷⁸, Judy Garber⁷⁹, José A. García-Sáenz⁵⁰, Mia M. Gaudet⁷⁸, Simon A. Gayther⁸⁰, Graham G. Giles^{81,82,83}, Andrew K. Godwin⁸⁴, Mark S. Goldberg^{85,86,87}, David E. Goldgar⁸⁸, Anna González-Neira³⁵, Mark H. Greene⁸⁹, Jacek Gronwald²³, Pascal Guénel⁹⁰, Lothar Häberle⁹¹, Eric Hahnen⁹², Christopher A. Haiman⁹³, Christopher R. Hake⁹⁴, Per Hall^{62,95}, Ute Hamann⁹⁶, Elaine F. Harkness^{97,98}, Bernadette A.M. Heemskerk-Gerritsen⁹⁹, Peter Hillemanns³⁰, Frans B.L. Hogervorst¹⁰⁰, Bernd Holleczeck¹⁰¹, Antoinette Hollestelle⁹⁹, Maartje J. Hooning⁹⁹, Robert N. Hoover¹, John L. Hopper⁸², Anthony Howell¹⁰², Hanna Huebner¹⁹, Peter J. Hulick¹⁰³, Evgeny N. Imyanitov¹⁰⁴, kConFab Investigators¹⁰⁵, ABCTB Investigators¹⁰⁵, Claudine Isaacs¹⁰⁶, Louise Izatt¹⁰⁷, Agnes Jager⁹⁹, Milena Jakimovska¹⁰⁸, Anna Jakubowska^{23,109}, Paul James¹¹⁰, Ramunas Janavicius^{111,112}, Wolfgang Janni¹¹³, Esther M. John¹¹⁴, Michael E. Jones¹¹⁵, Audrey Jung²⁰, Rudolf Kaaks²⁰, Pooja Middha Kapoor^{20,116}, Beth Y. Karlan¹¹⁷, Renske Keeman⁴³, Sofia Khan¹¹⁸, Elza Khusnutdinova^{22,119}, Cari M. Kitahara¹²⁰, Yon-Dschun Ko¹²¹, Irene Konstantopoulou⁷⁴, Linetta B. Koppert¹²², Stella Koutros¹, Vessela N. Kristensen^{123,124}, Anne-Vibeke Laenholm¹²⁵, Diether Lambrechts^{126,127}, Susanna C. Larsson^{128,129}, Pierre Laurent-Puig¹³⁰, Conxi Lazaro¹³¹, Emilija Lazarova¹³², Flavio Lejbkovicz⁷, Goska Leslie⁶, Fabienne Lesueur¹³³, Annika Lindblom^{134,135}, Jolanta Lissowska¹³⁶, Wing-Yee Lo^{38,137}, Jennifer T. Loud⁸⁹, Jan Lubinski²³, Alicja Lukomska²³, Robert J. MacInnis^{81,82}, Arto Mannermaa^{138,139,140}, Mehdi Manoochehri⁹⁶, Siranoush Manoukian¹⁶, Sara Margolin^{95,141}, Maria Elena Martinez^{77,142}, Laura Matricardi¹⁴³, Lesley McGuffog⁶, Catriona McLean¹⁴⁴, Noura Mebirouk¹⁴⁵, Alfons Meindl¹⁴⁶, Usha Menon¹⁴⁷, Austin Miller¹⁴⁸, Elvira Mingazheva¹⁴⁹, Marco Montagna¹⁴³, Anna Marie Mulligan^{150,151}, Claire Mulot¹³⁰, Taru A. Murañen¹¹⁸, Katherine L. Nathanson⁶⁶, Susan L. Neuhausen¹⁵², Heli Nevanlinna¹¹⁸, Patrick Neven⁵⁵, William G. Newman^{49,70}, Finn C. Nielsen¹⁵³, Liene Nikitina-Zake¹⁵⁴, Jesse Nodora^{155,156}, Kenneth Offit¹⁵⁷, Edith Olah¹⁵⁸, Olufunmilayo I. Olopade^{159,160}, Håkan Olsson^{161,162}, Nick Orr¹⁶³, Laura Papi¹⁶⁴, Janos Papp¹⁵⁸, Tjung-Won Park-Simon³⁰, Michael T. Parsons¹⁶⁵, Bernard Peissel¹⁶, Ana Peixoto¹⁶⁶, Beth Peshkin¹⁶⁷, Paolo Peterlongo¹⁶⁸, Julian Peto^{169,6}, Kelly-Anne Phillips^{82,170,171}, Marion Piedmonte¹⁴⁸, Dijana Plaseska-Karanfilska¹⁰⁸, Karolina Prajzencanc²³, Ross Prentice¹⁴, Darya Prokofyeva¹¹⁹, Brigitte Rack¹¹³, Paolo Radice¹⁷², Susan J. Ramus^{173,174,175}, Johanna Rantala¹⁷⁶, Muhammad U. Rashid^{96,177}, Gad Rennert⁷, Hedy S. Rennert⁷, Harvey A. Risch¹⁷⁸, Atocha Romero^{179,180}, Matti A. Rookus¹⁸¹, Matthias Rübner⁹¹, Thomas Rüdiger¹⁸², Emmanouil Saloustros¹⁸³, Sarah Sampson¹⁸⁴, Dale P. Sandler¹⁸⁵, Elinor J. Sawyer¹⁸⁶, Maren T. Scheuner¹⁸⁷, Rita K. Schmutzler⁹², Andreas Schneeweiss^{47,188}, Minouk J. Schoemaker¹¹⁵, Ben Schöttker¹¹, Peter Schürmann³⁰, Leigha Senter¹⁸⁹, Priyanka Sharma¹⁹⁰, Mark E. Sherman¹⁹¹, Xiao-

Ou Shu¹⁹², Christian F. Singer¹⁹³, Snezhana Smichkoska¹³², Penny Soucy¹⁹⁴, Melissa C. Southey⁸³, John J. Spinelli^{195,196}, Jennifer Stone^{82,197}, Dominique Stoppa-Lyonnet¹⁹⁸, EMBRACE Study¹⁰⁵, GEMO Study Collaborators¹⁰⁵, Anthony J. Swerdlow^{115,199}, Csilla I. Szabo²⁰⁰, Rulla M. Tamimi^{5,201,202}, William J. Tapper²⁰³, Jack A. Taylor^{185,204}, Manuel R. Teixeira^{166,180}, MaryBeth Terry²⁰⁵, Mads Thomassen²⁰⁶, Darcy L. Thull²⁰⁷, Marc Tischkowitz^{208,209}, Amanda E. Toland²¹⁰, Rob A.E.M. Tollenaar²¹¹, Ian Tomlinson^{212,213}, Diana Torres^{96,214}, Melissa A. Troester²¹⁵, Th  r  se Truong⁹⁰, Nadine Tung²¹⁶, Michael Untch²¹⁷, Celine M. Vachon²¹⁸, Ans M.W. van den Ouweland²¹⁹, Lizet E. van der Kolk¹⁰⁰, Elke M. van Veen^{49,70}, Elizabeth J. vanRensburg²²⁰, Ana Vega^{24,25,26}, Barbara Wappenschmidt⁹², Clarice R. Weinberg²²¹, Jeffrey N. Weitzel²²², Hans Wildiers⁵⁵, Robert Winqvist^{223,224,225,226}, Alicja Wolk^{109,128,129}, Xiaohong R. Yang¹, Drakoulis Yannoukakos⁷⁴, Wei Zheng¹⁹², Kristin K. Zorn²²⁷, Roger L. Milne^{81,82,83}, Peter Kraft^{5,202}, Jacques Simard¹⁹⁴, Paul D.P. Pharoah^{3,6}, Kyriaki Michailidou^{6,228,229}, Antonis C. Antoniou⁶, Marjanka K. Schmidt^{43,230}, Georgia Chenevix-Trench⁴, Douglas F. Easton³, Nilanjan Chatterjee^{2,231,**}, Montserrat Garc  a-Closas^{1,**}

Affiliations

¹Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, USA, ²Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, ³Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK, ⁴Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia, ⁵Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA, ⁶Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, ⁷Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel, ⁸Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, ⁹Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada, ¹⁰Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA, USA, ¹¹Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany, ¹²Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON, Canada, ¹³Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA, ¹⁴Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, ¹⁵Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA, ¹⁶Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy, ¹⁷Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK, ¹⁸Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany, ¹⁹Department of Gynecology and Obstetrics, Comprehensive

Cancer Center ER-EMN, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, Germany, ²⁰Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany, ²¹Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain, ²²Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia, ²³Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland, ²⁴Molecular Medicine Unit, Fundación Pública Galega de Medicina Xenómica, Santiago de Compostela, Spain, ²⁵Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain, ²⁶Centro de Investigación en Red de Enfermedades Raras (CIBERER), Santiago de Compostela, Spain, ²⁷Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, ²⁸Department of Oncology, Örebro University Hospital, Örebro, Sweden, ²⁹N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus, ³⁰Gynaecology Research Unit, Hannover Medical School, Hannover, Germany, ³¹Department of Radiation Oncology, Hannover Medical School, Hannover, Germany, ³²Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark, ³³Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark, ³⁴Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark, ³⁵Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, ³⁶Division of Cancer Prevention and Genetics, IEO, European Institute of Oncology IRCCS, Milan, Italy, ³⁷Department of Oncology, Lund University and Skåne University Hospital, Lund, Sweden, ³⁸Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany, ³⁹iFIT-Cluster of Excellence, University of Tübingen, Tübingen, Germany, ⁴⁰German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany, ⁴¹Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany, ⁴²Bioscience Department, Faculty of Medicine, Universidad de la Sabana, Chia, Colombia, ⁴³Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands, ⁴⁴Department of Women's Health, University of Tübingen, Tübingen, Germany, ⁴⁵Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany, ⁴⁶Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, Germany, ⁴⁷Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg, University of Heidelberg, Heidelberg, Germany, ⁴⁸Department of Medicine, Huntsman Cancer Institute, Salt Lake City, UT, USA, ⁴⁹Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester NIHR Biomedical Research Centre, Manchester University Hospitals NHS, Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK, ⁵⁰Medical Oncology Department, Hospital Clínico San Carlos, Instituto de

Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain, ⁵¹Section of Molecular Genetics, Dept. of Laboratory Medicine, University Hospital of Pisa, Pisa, Italy, ⁵²Department of Biology, University of Pisa, Pisa, Italy, ⁵³Oncology and Genetics Unit, Instituto de Investigación Sanitaria Galicia Sur (IISGS), Xerencia de Xestión Integrada de Vigo-SERGAS, Vigo, Spain, ⁵⁴Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany, ⁵⁵Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium, ⁵⁶Departments of Pediatrics and Medicine, Columbia University, New York, NY, USA, ⁵⁷Centre for Medical Genetics, Ghent University, Ghent, Belgium, ⁵⁸Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia, ⁵⁹Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA, ⁶⁰Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK, ⁶¹Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK, ⁶²Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden, ⁶³Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA, ⁶⁴Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands, ⁶⁵Oncogenetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain, ⁶⁶Department of Medicine, Abramson Cancer Center, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA, ⁶⁷Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK, ⁶⁸Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK, ⁶⁹Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany, ⁷⁰Division of Evolution and Genomic Medicine, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK, ⁷¹David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, USA, ⁷²Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, UK, ⁷³Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic, ⁷⁴Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research "Demokritos", Athens, Greece, ⁷⁵The Susanne Levy Gertner Oncogenetics Unit, Chaim Sheba Medical Center, Ramat Gan, Israel, ⁷⁶Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain, ⁷⁷Moore's Cancer Center, University of California San Diego, La Jolla, CA, USA, ⁷⁸Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA, ⁷⁹Cancer Risk and Prevention Clinic, Dana-Farber Cancer Institute, Boston, MA,

USA, ⁸⁰Center for Bioinformatics and Functional Genomics and the Cedars Sinai Genomics Core, Cedars-Sinai Medical Center, Los Angeles, CA, USA, ⁸¹Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia, ⁸²Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia, ⁸³Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, Australia, ⁸⁴Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Kansas City, KS, USA, ⁸⁵Department of Medicine, McGill University, Montréal, QC, Canada, ⁸⁶Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, QC, Canada, ⁸⁷Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia, ⁸⁸Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, USA, ⁸⁹Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, ⁹⁰Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France, ⁹¹Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany, ⁹²Center for Familial Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, ⁹³Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, ⁹⁴Waukesha Memorial Hospital-Pro Health Care, Waukesha, WI, USA, ⁹⁵Department of Oncology, Södersjukhuset, Stockholm, Sweden, ⁹⁶Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany, ⁹⁷Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK, ⁹⁸Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, UK, ⁹⁹Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands, ¹⁰⁰Family Cancer Clinic, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands, ¹⁰¹Saarland Cancer Registry, Saarbrücken, Germany, ¹⁰²Division of Cancer Sciences, University of Manchester, Manchester, UK, ¹⁰³Center for Medical Genetics, NorthShore University HealthSystem, Evanston, IL, USA, ¹⁰⁴N.N. Petrov Institute of Oncology, St. Petersburg, Russia, ¹⁰⁵A full list of authors can be found in the Supplementary Note, ¹⁰⁶Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA, ¹⁰⁷Clinical Genetics, Guy's and St. Thomas' NHS Foundation Trust, London, UK, ¹⁰⁸Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov", Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia, ¹⁰⁹Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland, ¹¹⁰Parkville Familial Cancer Centre, Peter MacCallum Cancer Center, Melbourne,

Victoria, Australia, ¹¹¹Hematology, oncology and transfusion medicine center, Dept. of Molecular and Regenerative Medicine, Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania, ¹¹²4/30/01, Vilnius, Lithuania, ¹¹³Department of Gynaecology and Obstetrics, University Hospital Ulm, Ulm, Germany, ¹¹⁴Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA, ¹¹⁵Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK, ¹¹⁶Faculty of Medicine University of Heidelberg, Heidelberg, Germany, ¹¹⁷David Geffen School of Medicine, Department of Obstetrics and Gynecology, University of California at Los Angeles, Los Angeles, CA, USA, ¹¹⁸Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland, ¹¹⁹Department of Genetics and Fundamental Medicine, Bashkir State Medical University, Ufa, Russia, ¹²⁰Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA, ¹²¹Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany, ¹²²Department of Surgical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands, ¹²³Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway, ¹²⁴Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway, ¹²⁵Department of Surgical Pathology, Zealand University Hospital, Slagelse, Denmark, ¹²⁶VIB Center for Cancer Biology, VIB, Leuven, Belgium, ¹²⁷Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium, ¹²⁸Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden, ¹²⁹Department of Surgical Sciences, Uppsala University, Uppsala, Sweden, ¹³⁰Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris, France, ¹³¹Molecular Diagnostic Unit, Hereditary Cancer Program, ICO-IDIBELL (Bellvitge Biomedical Research Institute, Catalan Institute of Oncology), CIBERONC, Barcelona, Spain, ¹³²Ss. Cyril and Methodius University in Skopje, Medical Faculty, University Clinic of Radiotherapy and Oncology, Skopje, Republic of North Macedonia, ¹³³Genetic Epidemiology of Cancer team, Inserm U900, Paris, France, ¹³⁴Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden, ¹³⁵Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden, ¹³⁶Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Cancer Center, Oncology Institute, Warsaw, Poland, ¹³⁷University of Tübingen, Tübingen, Germany, ¹³⁸Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland, ¹³⁹Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland, ¹⁴⁰Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland, ¹⁴¹Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden, ¹⁴²Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, USA, ¹⁴³Immunology and Molecular Oncology Unit, Veneto Institute of Oncology IOV - IRCCS, Padua, Italy, ¹⁴⁴Department of Anatomical Pathology, The Alfred Hospital, Prahran, Victoria, Australia, ¹⁴⁵Genetic

Epidemiology of Cancer team, Inserm U900, Institut Curie, PSL University, Mines ParisTech, Paris, France, ¹⁴⁶Department of Gynecology and Obstetrics, Ludwig Maximilian University of Munich, Munich, Germany, ¹⁴⁷MRC Clinical Trials Unit at UCL, Institute of Clinical Trials & Methodology, University College London, London, UK, ¹⁴⁸NRG Oncology, Statistics and Data Management Center, Roswell Park Cancer Institute, Buffalo, NY, USA, ¹⁴⁹Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia, ¹⁵⁰Laboratory Medicine Program, University Health Network, Toronto, ON, Canada, ¹⁵¹Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada, ¹⁵²Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA, ¹⁵³Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark, ¹⁵⁴Latvian Biomedical Research and Study Centre, Riga, Latvia, ¹⁵⁵Moore's Cancer Center, University of California, San Diego, La Jolla, CA, USA, ¹⁵⁶Department of Family Medicine and Public Health, School of Medicine, University of California, San Diego, La Jolla, CA, USA, ¹⁵⁷Clinical Genetics Research Lab, Department of Cancer Biology and Genetics, Memorial Sloan-Kettering Cancer Center, New York, NY, USA, ¹⁵⁸Department of Molecular Genetics, National Institute of Oncology, Budapest, Hungary, ¹⁵⁹Center for Clinical Cancer Genetics, The University of Chicago, Chicago, IL, USA, ¹⁶⁰Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia, ¹⁶¹Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden, ¹⁶²Clinical Genetics Service, Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, NY, USA, ¹⁶³Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, Ireland, UK, ¹⁶⁴Unit of Medical Genetics, Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy, ¹⁶⁵Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia, ¹⁶⁶Department of Genetics, Portuguese Oncology Institute, Porto, Portugal, ¹⁶⁷Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA, ¹⁶⁸Genome Diagnostics Program, IFOM, The FIRC Institute of Molecular Oncology, Milan, Italy, ¹⁶⁹Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK, ¹⁷⁰Peter MacCallum Cancer Center, Melbourne, Victoria, Australia, ¹⁷¹Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia, ¹⁷²Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy, ¹⁷³Adult Cancer Program, Lowy Cancer Research Centre, University of NSW Sydney, Sydney, New South Wales, Australia, ¹⁷⁴School of Women's and Children's Health, Faculty of Medicine, University of NSW Sydney, Sydney, New South Wales, Australia, ¹⁷⁵The Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, Australia, ¹⁷⁶Clinical Genetics, Karolinska Institutet, Stockholm, Sweden, ¹⁷⁷Department of Basic Sciences, Shaikat Khanum Memorial Cancer Hospital and Research Centre (SKMCH & RC), Lahore, Pakistan, ¹⁷⁸Chronic Disease

Epidemiology, Yale School of Public Health, New Haven, CT, USA, ¹⁷⁹Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain, ¹⁸⁰Biomedical Sciences Institute (ICBAS), University of Porto, Porto, Portugal, ¹⁸¹Department of Epidemiology, The Netherlands Cancer Institute, Amsterdam, The Netherlands, ¹⁸²Institute of Pathology, Staedtisches Klinikum Karlsruhe, Karlsruhe, Germany, ¹⁸³Department of Oncology, University Hospital of Larissa, Larissa, Greece, ¹⁸⁴Prevent Breast Cancer Centre and Nightingale Breast Screening Centre, Manchester University NHS Foundation Trust, Manchester, UK, ¹⁸⁵Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA, ¹⁸⁶Research Oncology, Guy's Hospital, King's College London, London, UK, ¹⁸⁷Cancer Genetics and Prevention Program, University of California San Francisco, San Francisco, CA, USA, ¹⁸⁸National Center for Tumor Diseases, University Hospital and German Cancer Research Center, Heidelberg, Germany, ¹⁸⁹Clinical Cancer Genetics Program, Division of Human Genetics, Department of Internal Medicine, The Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA, ¹⁹⁰Department of Internal Medicine, Division of Oncology, University of Kansas Medical Center, Westwood, KS, USA, ¹⁹¹Department of Health Sciences Research, Mayo Clinic College of Medicine, Jacksonville, FL, USA, ¹⁹²Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA, ¹⁹³Dept of OB/GYN and Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria, ¹⁹⁴Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval, Research Center, Québec City, QC, Canada, ¹⁹⁵Population Oncology, BC Cancer, Vancouver, BC, Canada, ¹⁹⁶School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada, ¹⁹⁷The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Western Australia, Australia, ¹⁹⁸Department of Genetics, Inserm U830, Institut Curie, Paris Descartes Sorbonne-Paris-Cité University, Paris, France, ¹⁹⁹Division of Breast Cancer Research, The Institute of Cancer Research, London, UK, ²⁰⁰National Human Genome Research Institute, National Cancer Institute, Bethesda, MD, USA, ²⁰¹Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA, ²⁰²Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA, ²⁰³Faculty of Medicine, University of Southampton, Southampton, UK, ²⁰⁴Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA, ²⁰⁵Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA, ²⁰⁶Department of Clinical Genetics, Odense University Hospital, Odense C, Denmark, ²⁰⁷Department of Medicine, Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA, ²⁰⁸Program in Cancer Genetics, Departments of Human Genetics and Oncology, McGill University, Montréal, QC, Canada, ²⁰⁹Department of Medical Genetics, National Institute for Health Research Cambridge Biomedical Research Centre,

University of Cambridge, Cambridge, UK, ²¹⁰Department of Cancer Biology and Genetics, The Ohio State University, Columbus, OH, USA, ²¹¹Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands, ²¹²Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK, ²¹³Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK, ²¹⁴Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia, ²¹⁵Department of Epidemiology, Gillings School of Global Public Health and UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, ²¹⁶Department of Medical Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA, ²¹⁷Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany, ²¹⁸Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA, ²¹⁹Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands, ²²⁰Department of Genetics, University of Pretoria, Arcadia, South Africa, ²²¹Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA, ²²²Clinical Cancer Genomics, City of Hope, Duarte, CA, USA, ²²³Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland, ²²⁴Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland, ²²⁵Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada, ²²⁶Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands, ²²⁷Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA, ²²⁸Biostatistics Unit, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus, ²²⁹Cyprus School of Molecular Medicine, Nicosia, Cyprus, ²³⁰Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands, ²³¹Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA

Acknowledgments and Funding

We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out.

Genotyping for the OncoArray was funded by the government of Canada through Genome Canada and the Canadian Institutes of Health Research (GPH-129344), the Ministère de l'Économie, de la Science et de l'Innovation du Québec through Génome Québec, the Quebec Breast Cancer Foundation for the PERSPECTIVE project, the US National Institutes of Health (NIH) (1 U19 CA 148065 for the Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project and X01HG007492 to the Center for Inherited Disease Research (CIDR) under contract HHSN268201200008I), Cancer Research UK (C1287/A16563), the Odense University Hospital Research Foundation (Denmark), the National R&D Program for Cancer Control—Ministry of Health and Welfare (Republic of Korea) (1420190), the Italian Association for Cancer Research (AIRC; IG16933), the Breast Cancer Research Foundation, the National Health and Medical Research Council (Australia) and German Cancer Aid (110837).

Genotyping for the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710, C1287/A10118 and C12292/A11174), NIH grants (CA128978, CA116167 and

CA176785) and the Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 (GAME-ON initiative)), an NCI Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), the Canadian Institutes of Health Research (CIHR) for the CIHR Team in Familial Risks of Breast Cancer, the Ministère de l'Économie, Innovation et Exportation du Québec (PSR-SIIRI-701), the Komen Foundation for the Cure, the Breast Cancer Research Foundation and the Ovarian Cancer Research Fund.

Combination of the GWAS data was supported in part by the NIH Cancer Post-Cancer GWAS initiative (1 U19 CA 148065) (DRIVE, part of the GAME-ON initiative). LD score regression analysis was supported by grant CA194393.

BCAC was funded by Cancer Research UK (C1287/A16563) and by the European Union via its Seventh Framework Programme (HEALTH-F2-2009-223175, COGS) and the Horizon 2020 Research and Innovation Programme (633784, B-CAST; 634935, BRIDGES). CIMBA was funded by Cancer Research UK (C12292/A20861 and C12292/A11174).

Dr. Nilanjan Chatterjee's was funded by NHGRI (1R01 HG010480-01).

For a full description of funding and acknowledgments, see the Supplementary Note.

References

1. Michailidou K et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* 551, 92–94 (2017). [PubMed: 29059683]
2. Milne RL et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet* 49, 1767–1778 (2017). [PubMed: 29058716]
3. Garcia-Closas M et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet* 45, 392–8, 398e1–2 (2013). [PubMed: 23535733]
4. Zhang H et al. A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *Biostatistics*, Doi: 10.1093/biostatistics/kxz065 (2020).
5. Michailidou K et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet* 45, 353–61, 361e1–2 (2013). [PubMed: 23535729]
6. 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). [PubMed: 25751625]
7. Dempster AP, Laird NM & Rubin DB Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological* 39, 1–38 (1977).
8. Curigliano G et al. De-escalating and escalating treatments for early-stage breast cancer: the St. Gallen International Expert Consensus Conference on the Primary Therapy of Early Breast Cancer 2017. *Ann Oncol* 28, 1700–1712 (2017). [PubMed: 28838210]
9. Spurdle AB et al. Refined histopathological predictors of BRCA1 and BRCA2 mutation status: a large-scale analysis of breast cancer characteristics from the BCAC, CIMBA, and ENIGMA consortia. *Breast Cancer Res* 16, 3419 (2014). [PubMed: 25857409]
10. Phelan CM et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat Genet* 49, 680–691 (2017). [PubMed: 28346442]
11. Stacey SN et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat Genet* 43, 1098–103 (2011). [PubMed: 21946351]
12. Pellacani D et al. Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. *Cell Rep* 17, 2060–2074 (2016). [PubMed: 27851968]
13. Beesley J et al. Chromatin interactome mapping at 139 independent breast cancer risk signals. *bioRxiv*, 520916 (2019).
14. Fachal L et al. Fine-mapping of 150 breast cancer risk regions identifies 178 high confidence target genes. *Nat Genet* 52, 56–73 (2020). [PubMed: 31911677]
15. Nik-Zainal S et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 534, 47–54 (2016). [PubMed: 27135926]
16. Ferreira MA et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat Commun* 10, 1741 (2019). [PubMed: 30988301]

17. Wu L et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* 50, 968–978 (2018). [PubMed: 29915430]
18. Bulik-Sullivan B et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* 47, 1236–41 (2015). [PubMed: 26414676]
19. Bulik-Sullivan BK et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291–5 (2015). [PubMed: 25642630]
20. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* 47, 1228–35 (2015). [PubMed: 26414678]
21. Ahearn TU et al. Common breast cancer risk loci predispose to distinct tumor subtypes. *bioRxiv*, 733402 (2019).
22. Mavaddat N et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *Am J Hum Genet* 104, 21–34 (2019). [PubMed: 30554720]
23. Lin DY & Sullivan PF Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* 85, 862–72 (2009). [PubMed: 20004761]
24. Chatterjee N A Two-Stage Regression Model for Epidemiological Studies with Multivariate Disease Classification Data. *Journal of the American Statistical Association* 99, 127–138 (2004).
25. Anderson WF, Rosenberg PS, Prat A, Perou CM & Sherman ME How many etiological subtypes of breast cancer: two, three, four, or more? *J Natl Cancer Inst* 106(2014).
26. Barnes DR et al. Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations. *Genet Epidemiol* 36, 274–91 (2012). [PubMed: 22714938]
27. Antoniou AC 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). [PubMed: 20852631]
28. Mavaddat N et al. Pathology of breast and ovarian cancers among BRCA1 and BRCA2 mutation carriers: results from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). *Cancer Epidemiol Biomarkers Prev* 21, 134–47 (2012). [PubMed: 22144499]
29. Willer CJ, Li Y & Abecasis GR METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–1 (2010). [PubMed: 20616382]
30. Hendricks AE, Dupuis J, Logue MW, Myers RH & Lunetta KL Correction for multiple testing in a gene region. *Eur J Hum Genet* 22, 414–8 (2014). [PubMed: 23838599]
31. Benjamini Y & Hochberg Y Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 289–300 (1995).
32. Udler MS, Tyrer J & Easton DF Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet Epidemiol* 34, 463–8 (2010). [PubMed: 20583289]
33. Wellcome Trust Case Control, C. et al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 44, 1294–301 (2012). [PubMed: 23104008]
34. Pharoah PD et al. Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 31, 33–6 (2002). [PubMed: 11984562]
35. Visscher PM, Hill WG & Wray NR Heritability in the genomics era--concepts and misconceptions. *Nat Rev Genet* 9, 255–66 (2008). [PubMed: 18319743]

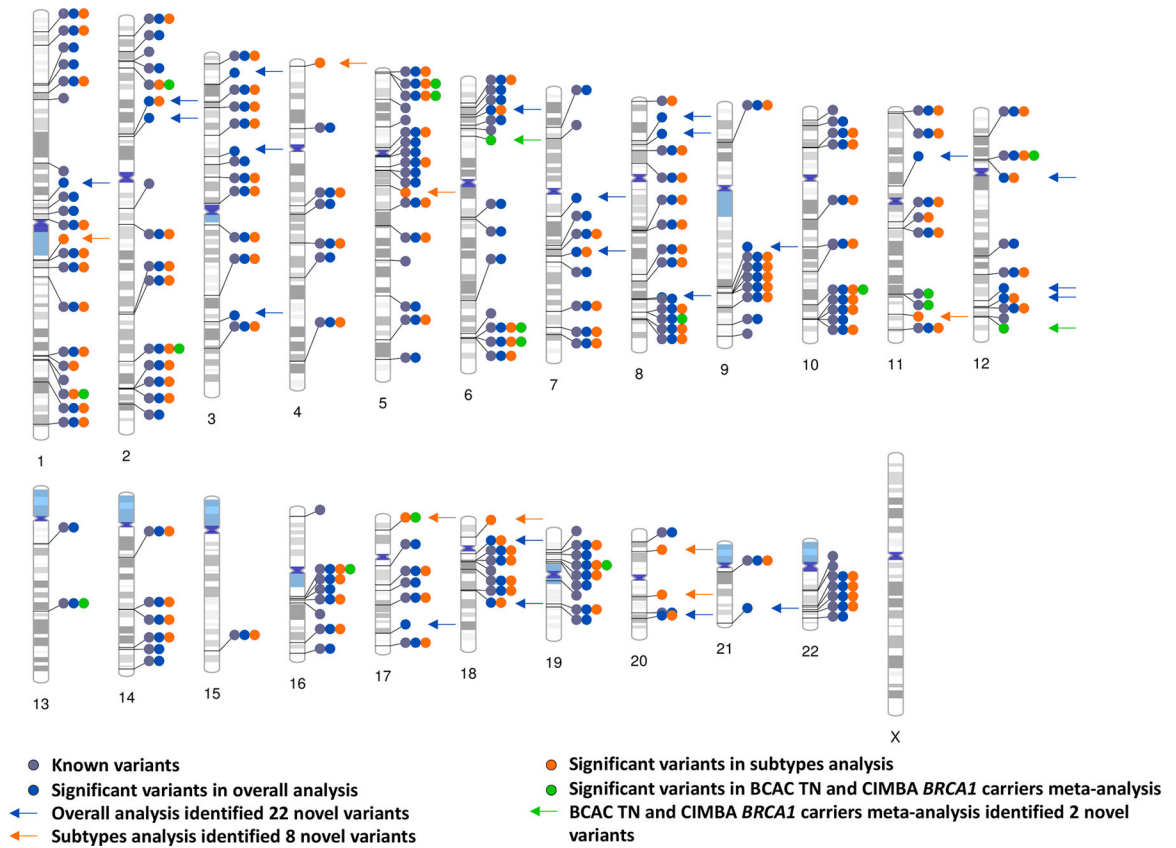


Figure 1.

Ideogram of all the independent genome-wide significant breast cancer susceptibility variants in overall, subtypes, BCAC triple-negative (TN) and CIMBA *BRCA1* carriers meta-analysis. The 32 novel variants are labeled with arrows. The other significant variants are within ± 500 or LD > 0.3 with previously reported variants.

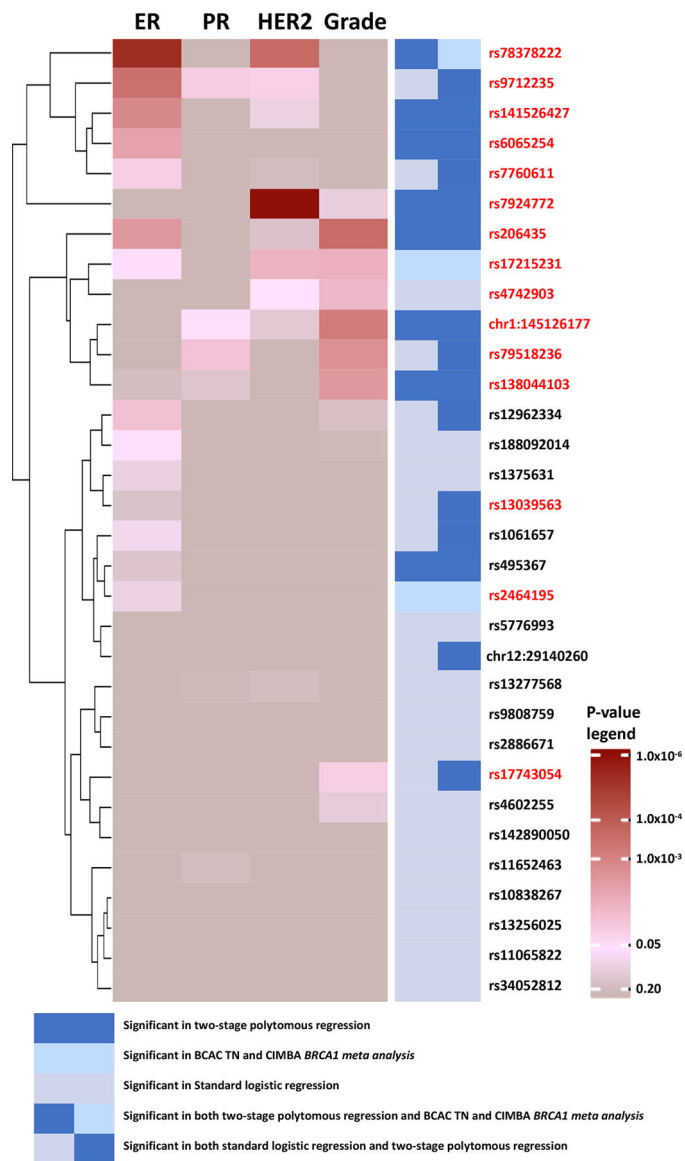
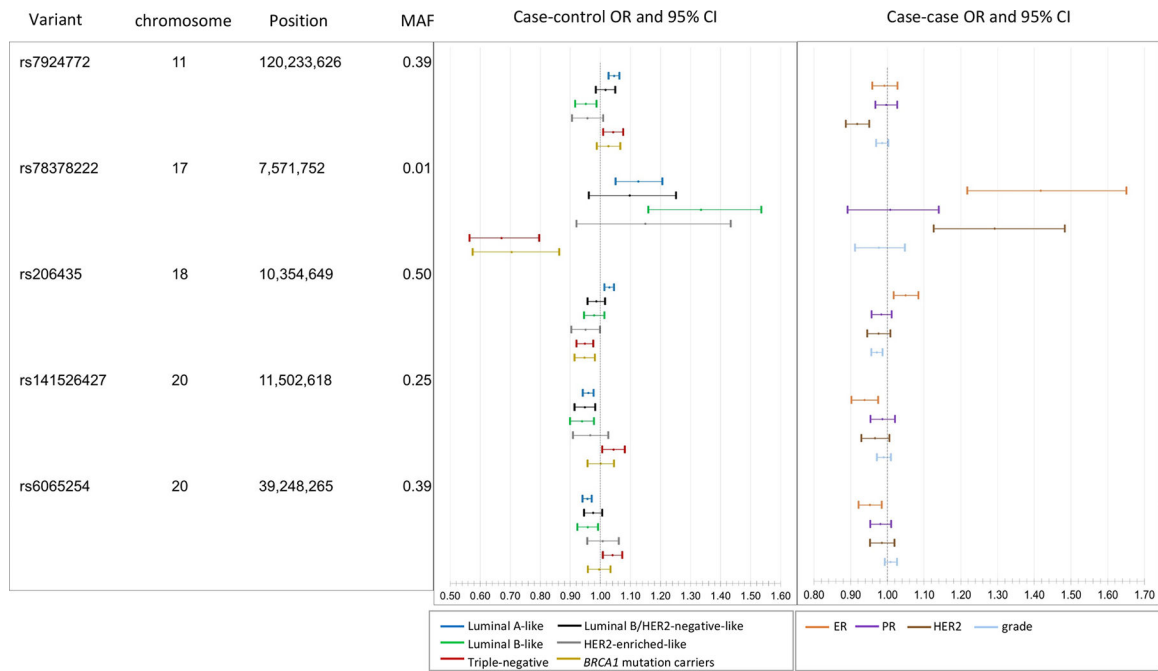


Figure 2.

Heatmap and clustering of p-values from marker specific heterogeneity test for 32 breast cancer susceptibility loci ($n = 106,278$ invasive cases, $n = 91,477$ controls). P-values are for associations between the most significant variants marking each loci and estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) or grade, adjusting for top ten principal components and age. P-values are raw p-values from two-tailed z-test statistics. Fifteen variants in red color were significant according to the global heterogeneity tests ($FDR < 0.05$), of which 14 were identified by methods accounting for tumor heterogeneity. TN, triple negative.



¹ Per-minor allele odds ratio and 95% confidence limits

² luminal A-like (ER+ and/or PR+, HER2-, grade 1 & 2); luminal B/HER2-negative-like (ER+ and/or PR+, HER2-, grade 3); luminal B-like (ER+ and/or PR+, HER2+); (4) HER2-enriched-like (ER- and PR-, HER2+), and triple-negative (ER-, PR-, HER2-)

Figure 3.

Susceptibility variants with associations in opposite direction across subtypes. The case-control odds ratios (OR) and 95% confidence intervals (95% CI)¹ (left panel) are for associations of each of the five variants and risk for breast cancer intrinsic-like subtypes² estimated from the first-stage of the two-stage polytomous regression fixed-effects model (n = 106,278 invasive cases, n = 91,477 controls). The case-case ORs 95%CI (right panel) are estimated from the second stage parameters of a fixed effect two-stage polytomous models testing for heterogeneity between the five variants and estrogen receptor (ER), progesterone receptor (PR), human epidermal growth factor receptor 2 (HER2) and grade, where ER, PR, HER2, and grade are mutually adjusted for each other. MAF, minor allele frequency.

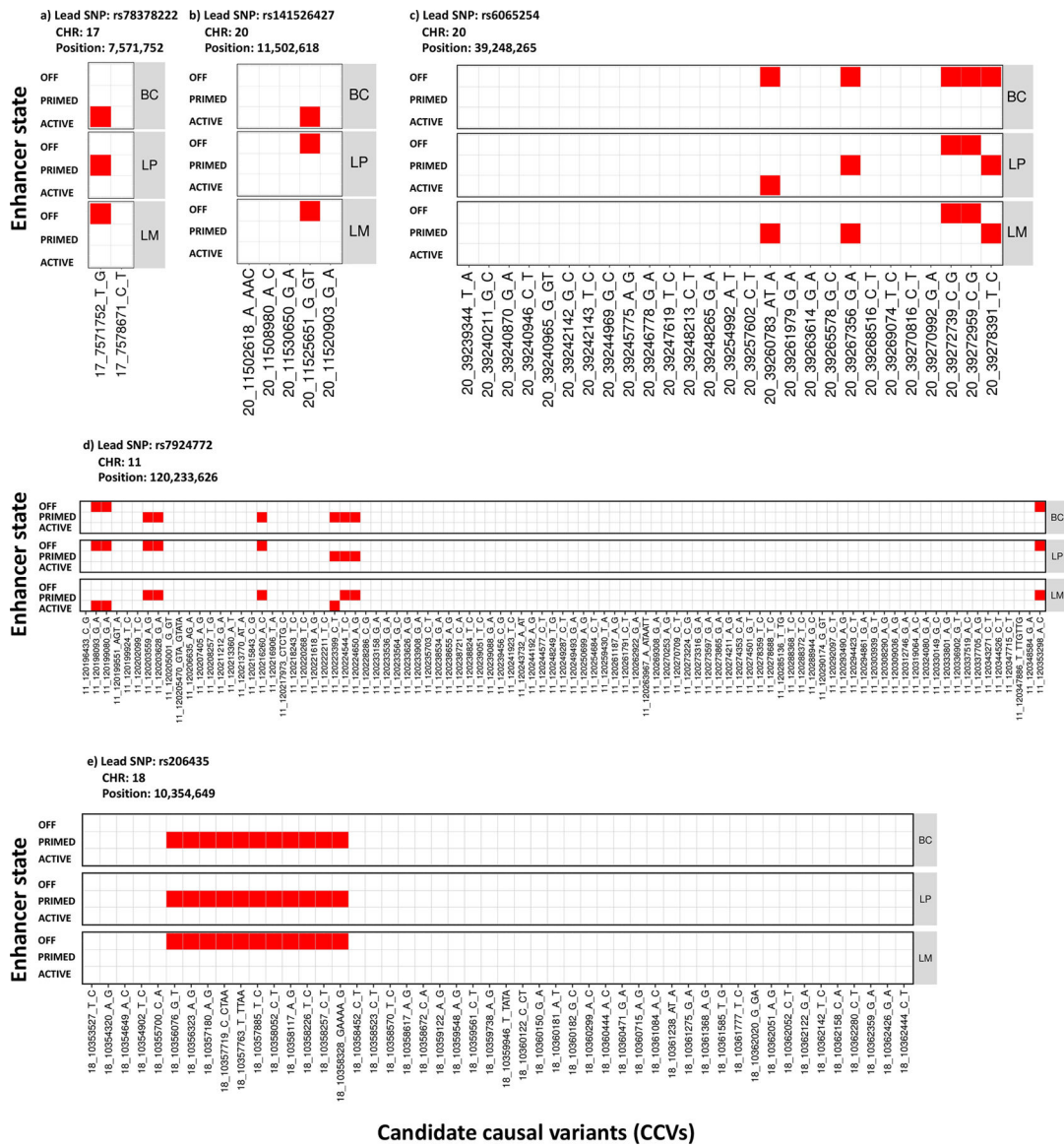


Figure 4. Heatmap of candidate causal variants (CCVs) overlapping with enhancer states in primary breast subpopulations for five variants with associations in opposite direction across subtypes. Three different breast subpopulations were considered: basal cells (BC), luminal progenitor (LP) and luminal cells mature (LM). Based on a combination of H3K4me1 and H3K27ac histone modification ChiP-seq signals, putative enhancers in BC, LP, and LM were characterized as “OFF”, “PRIMED” and “ACTIVE” (Online Methods). The CCVs overlapping with enhancers were colored as red, otherwise were white.

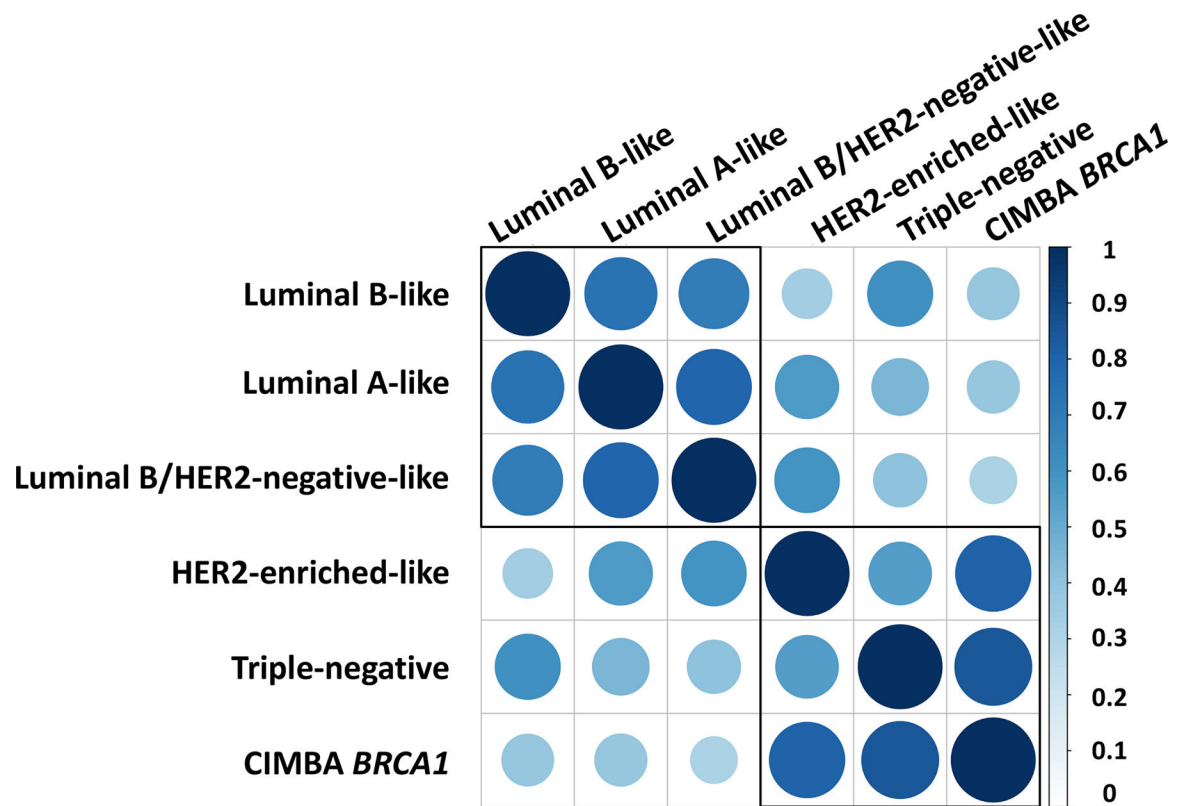


Figure 5.

Genetic correlation between the five intrinsic-like breast cancer subtypes and *BRCA1* mutation carriers estimated through LD score regression. See Supplementary Table 16 for further details. Both the color and size of the circles reflect the strength of the genetic correlations.

Genetic variance of invasive breast cancer explained by identified susceptibility variants and all reliably genome-wide imputable variants¹

Table 1.

| Phenotype | Genetic variance for 210 identified susceptibility variants ² | Genetic variance for 32 newly identified variants ² | Genetic variance for all GWAS variants ³ | Proportion of genetic variance explained by identified susceptibility loci ⁴ |
|-------------------------------------|--|--|---|---|
| Invasive breast cancer ⁵ | 0.253 | 0.016 | 0.515 | 45.51% |
| Luminal A-like | 0.336 | 0.022 | 0.620 | 54.22% |
| Luminal B/HER2-negative-like | 0.233 | 0.018 | 0.597 | 38.95% |
| Luminal B-like | 0.270 | 0.020 | 0.740 | 36.46% |
| HER2-enriched-like | 0.200 | 0.011 | 0.689 | 29.05% |
| Triple negative | 0.185 | 0.025 | 0.492 | 37.63% |
| CIMBA BRCA1 carriers | 0.083 | 0.016 | 0.309 | 26.86% |

¹ Genetic variance corresponds to heritability on the frailty-scale, which assumes the polygenetic log-additive model as the underlying model.

² Susceptibility variants included 178 variants previously identified or replicated^{1,2} and 32 newly identified variants in this paper.

³ Genetic variance of all reliably genome-wide imputable variants was estimated through LD-score regression described in Nat Genet 47, 291–5 (2015), and Nat Genet 47, 1236–41 (2015). Under the frailty-scale, the genetic variance for all GWAS variants is characterized by population variance of the underlying true polygenic risk score as $\sigma_{GWAS}^2 = Var\left(\sum_{m=1}^M \beta_m G_m\right)$, where G_m is the standardized genotype for the m th variant, β_m is the true log odds ratio for the m th variant and M are the total number of causal variants among the GWAS variants. (Online Methods).

⁴ Proportion of genetic variance explained by 210 identified GWAS significant variants over the genetic variance explained by all GWAS variants.

⁵ Invasive breast cancer summary level statistics were generated from 106,278 invasive cases and 91,477 controls, which were the same samples used in subtypes analyses (Supplementary Table 2).