LETTERS
https://doi.org/10.1038/s41588-020-0609-2

nature genetics

Check for updates

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

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 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, progesterone receptor and human epidermal growth factor receptor 2 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 < 0.05). Five loci showed associations ($P < 0.05$) in opposite directions between luminal and non-luminal subtypes. In silico analyses showed that 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 54.2% for luminal A-like disease and 37.6% for triple-negative disease. The odds ratios of polygenic risk scores, which included 330 variants, for the highest 1% of quantiles compared with 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 genome-wide association study (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 subtype, particularly estrogen-receptor-positive versus estrogen-receptor-negative or triple-negative disease[1–3]. However, previous GWASs have not simultaneously accounted for the high correlations among multiple, correlated tumor markers (such as estrogen receptor, progesterone receptor and human epidermal growth factor receptor 2 (HER2)) and grade, to identify specific source(s) of etiological heterogeneity. We performed a breast cancer GWAS using both standard analyses and a novel two-stage polytomous regression method that efficiently characterizes etiological heterogeneity while accounting for tumor marker correlations and missing data[4].

The study populations and genotyping are described elsewhere[1,2,5,6] and in the Methods. Briefly, we analyzed data from 118,474 cases and 96,201 controls of European ancestry participating in 82 studies from the BCAC, as well as 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 GWASs (14,910 cases and 17,588 controls) without subtype information. Our study expands on previous BCAC GWASs[1], 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 Methods and Extended Data Fig. 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 (ORs) and 95% confidence intervals (95% CIs), adjusting for country and principal components. iCOGS and OncoArray data were evaluated separately and the results were combined with those from the 11 other GWASs 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 model[4] that allows flexible, yet parsimonious, modeling of associations in the presence of underlying heterogeneity by estrogen receptor, progesterone receptor, HER2 and/or grade (Methods and 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 (Methods). We fit an additional two-stage model to estimate case–control ORs and 95% CIs between the variants and intrinsic-like subtypes defined by combinations of estrogen receptor, progesterone receptor, HER2 and grade[8] (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 principal components 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 (Methods).

Among *BRCA1* mutation carriers who are prone to developing triple-negative disease[9], we estimated per-allele hazard ratios within a retrospective cohort analysis framework. We assumed that estimated ORs for BCAC triple-negative cases and estimated hazard ratios from CIMBA *BRCA1* carriers approximated the same underlying relative risk[9], and we used a fixed-effects meta-analysis to combine these results (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 (Methods).

Overall, we identified 32 novel independent susceptibility loci marked by variants with $P < 5.0 \times 10^{-8}$ (Fig. 1, Supplementary Tables 5–7 and Supplementary Figs. 1–5): 22 variants using standard logistic regression, 16 variants using the two-stage

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

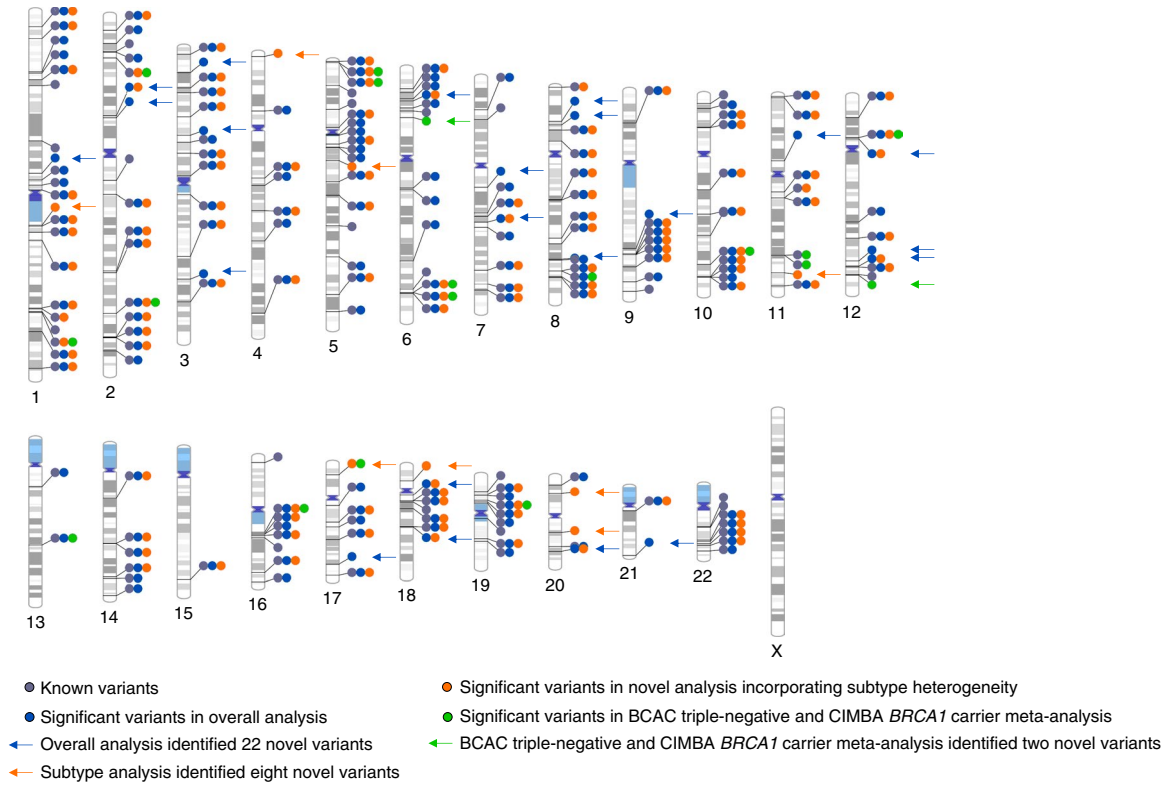


Fig. 1 | Ideogram of all of the independent genome-wide-significant breast cancer susceptibility variants in overall, subtype, BCAC triple-negative and CIMBA *BRCA1* carrier meta-analyses. The 32 novel variants are labeled with arrows. The other significant variants are within ±500 or linkage disequilibrium > 0.3 with previously reported variants.

polytomous model (eight of which were not 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 ($P \geq 1.0 \times 10^{-6}$; Supplementary Tables 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 United States (Supplementary Fig. 6). Supplementary Figs. 7 and 8 and Supplementary Table 11 show associations between all 32 variants and the intrinsic-like subtypes.

Fifteen of the 32 variants showed heterogeneity evidence (false discovery rate (FDR) < 0.05) according to the global heterogeneity test (Fig. 2 and Supplementary Table 12). Estrogen receptor (seven variants) and grade (seven variants) most often contributed to observed heterogeneity (marker-specific $P < 0.05$), followed by HER2 (four variants) and progesterone receptor (two variants). rs17215231, which was identified in the CIMBA/BCAC triple-negative meta-analysis, was the only variant found to be exclusively associated with triple-negative disease (OR = 0.85; 95% CI = 0.81–0.89). rs2464195, which was also identified as associated 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 (coefficient of determination (r^2) = 0.62) with rs7953249, which is differentially associated with the risk of ovarian cancer subtypes[10]. Five variants showed associations with luminal and non-luminal subtypes in opposite directions (Fig. 3). Four variants were associated in opposite directions with luminal A-like and triple-negative subtypes (rs78378222: OR = 1.13 and 95% CI = 1.05–1.20 versus OR = 0.67 and 95% CI = 0.57–0.80; rs206435: OR = 1.03 and 95% CI = 1.01–1.05 versus OR = 0.95 and 95% CI = 0.92–0.98; rs141526427: OR = 0.96 and 95% CI = 0.94–0.98 versus OR = 1.04 and 95% CI = 1.01–1.08; rs6065254: OR = 0.96 and 95% CI = 0.94–0.97 versus OR = 1.04 and 95% CI = 1.01–1.07). The tumor marker heterogeneity test showed associations for rs78378222 with estrogen receptor ($P_{ER} = 7.0 \times 10^{-6}$) and HER2 ($P_{HER2} = 2.07 \times 10^{-4}$), for rs206435 with estrogen receptor ($P_{ER} = 2.8 \times 10^{-3}$) and grade ($P_{grade} = 2.8 \times 10^{-4}$) and for rs141526427 ($P_{ER} = 1.3 \times 10^{-3}$) and rs6065254 ($P_{ER} = 4.3 \times 10^{-3}$) with only estrogen receptor. rs7924772 showed opposite case–control associations between HER2-negative and HER2-positive subtypes and, in agreement with these findings, was exclusively associated with HER2 ($P_{HER2} = 1.4 \times 10^{-6}$; Fig. 3). rs78378222, which is located in the 3′ untranslated region 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; $P = 1.5 \times 10^{-4}$). Previous analyses[11] did not find rs78378222 to be associated with breast cancer risk, probably due to its opposite effects between subtypes.

Candidate causal variants (CCVs) were defined (Methods) for each novel locus and we investigated the CCVs in relation to previously annotated enhancers in primary breast cells[12]. Based on combinations of H3K4me1 and H3K27ac histone modification chromatin immunoprecipitation sequencing (ChIP–seq) signals, putative enhancers in basal cells, luminal progenitor cells and mature luminal cells were characterized as off, primed or active (Methods). We defined switch 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 a switch enhancer (Fig. 4). For example, rs78378222 overlapped an active enhancer in basal cells, a primed enhancer in luminal progenitor cells and an off

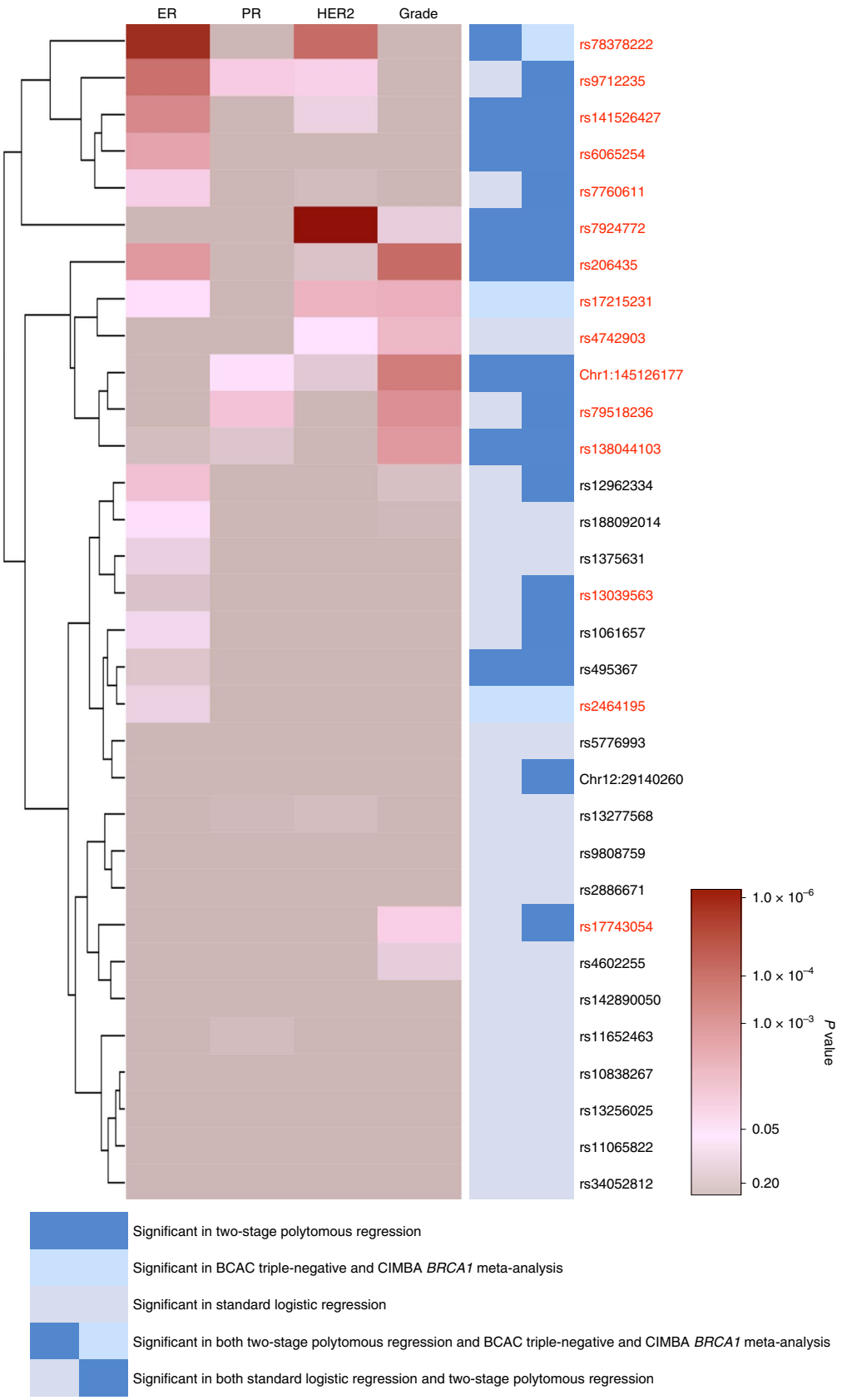


Fig. 2 | Heatmap and clustering of *P* values from a marker-specific heterogeneity test for 32 breast cancer susceptibility loci. *P* values are for associations between the most significant variants marking each locus ($n = 106,278$ invasive cases; $n = 91,477$ controls) and estrogen receptor (ER), progesterone receptor (PR), HER2 or grade, adjusting for the top ten principal components and age. *P* values are raw *P* values from two-tailed *z*-test statistics. The 15 variants in red were significant according to the global heterogeneity tests (FDR < 0.05), of which 14 were found to be genome-wide significant by methods accounting for tumor heterogeneity. The blue color scale indicates variants significantly ($P < 5.0 \times 10^{-8}$) associated with risk in different models.

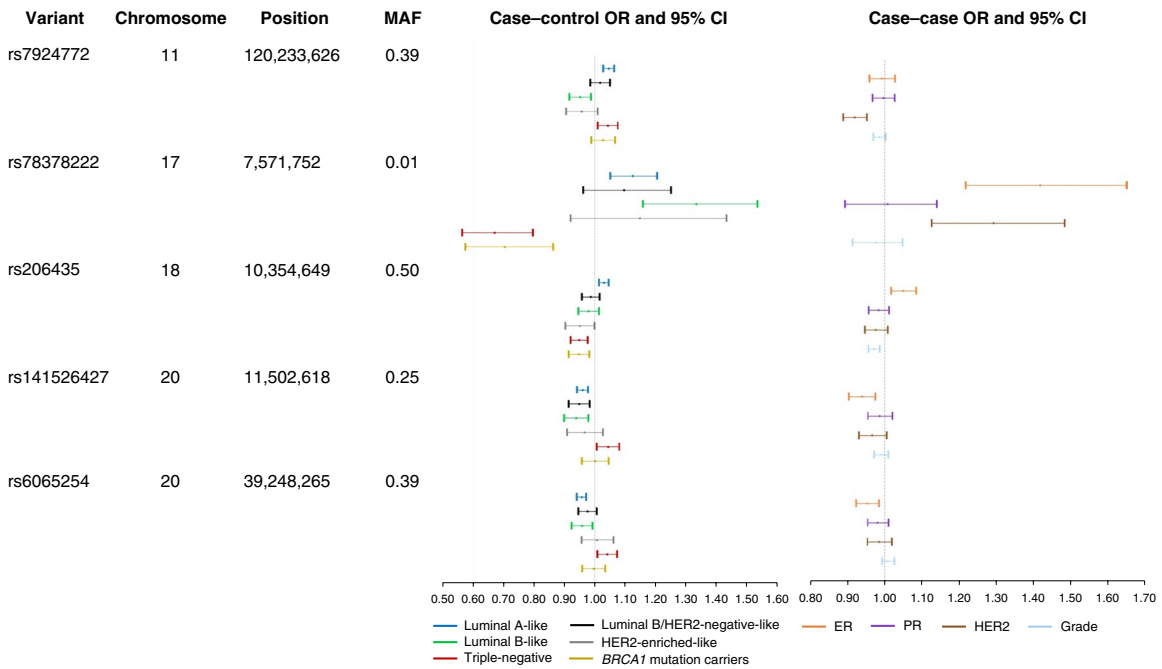


Fig. 3 | Susceptibility variants with associations in opposite directions across subtypes. The case–control ORs and 95% CIs (left) are for associations of each of the five variants and risk for breast cancer intrinsic-like subtypes, estimated from the two-stage polytomous regression fixed-effects model ($n = 106,278$ invasive cases; $n = 91,477$ controls). The case–case ORs and 95% CIs (right) were estimated from the same two-stage polytomous model, and the parameter for each tumor characteristic was adjusted for the others. In both plots, the data points represent the per-minor-allele OR and the error bars show the 95% CI. Luminal A-like: ER^+ and/or PR^+, $HER2^-$, grades 1 and 2; luminal B/HER2-negative-like: ER^+ and/or PR^+, $HER2^-$, grade 3; luminal B-like: ER^+ and/or PR^+, $HER2^+$; HER2-enriched-like: ER^- and PR^-, $HER2^+$; and triple-negative: ER^-, PR^-, $HER2^-$.

Table 1 | Genetic variance of invasive breast cancer explained by identified susceptibility variants and all reliably genome-wide imputable variants[a]

Phenotype	Genetic variance for 210 identified susceptibility variants[b]	Genetic variance for 32 newly identified variants[b]	Genetic variance for all GWAS variants[c]	Proportion of genetic variance explained by identified susceptibility loci[d]
Invasive breast cancer[e]	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%

[a]Genetic variance corresponds to heritability on the frailty scale, which assumes the polygenetic log-additive model as the underlying model. [b]Susceptibility variants included 178 previoulsy identified variants[1,2] and 32 variants newly identified in this paper. [c]The genetic variance of all reliably genome-wide imputable variants was estimated through linkage disequilibrium score regression, as described in refs. [18,19]. Under the frailty scale, the genetic variance for all GWAS variants was characterized by population variance of the underlying true PRS as $\sigma^2_{GWAS} = \mathrm{Var}\left(\sum_{m=1}^{M} \beta_m G_m\right)$, where G_m is the standardized genotype for the mth variant, β_m is the true log[odds ratio] for the mth variant and M is the total number of causal variants among the GWAS variants (Methods). [d]Proportion of genetic variance explained by 210 identified GWAS significant variants over the genetic variance explained by all GWAS variants. [e]Invasive breast cancer summary-level statistics were generated from 106,278 invasive cases and 91,477 controls, which were the same samples used in the subtype analyses (Supplementary Table 2).

enhancer in mature luminal cells. In comparison, 63% of the loci with consistent direction of associations across subtypes also overlapped with a switch enhancer (Supplementary Tables 13 and 14). These results suggest that some variants may modulate enhancer activity in a cell type-specific manner, thus differentially influencing the risk of tumor subtypes.

We used INQUISIT to intersect CCVs with functional annotation data from public databases to identify potential target genes[1] (Supplementary Note and Supplementary Table 15). We predicted 179 unique target genes for 26 of the 32 independent signals. Notably, rs78378222 has been reported to be associated with *TP53* messenger RNA levels in blood and adipose tissue[11], 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] (that is, *POLR3C*, *RNF115*, *SOX4* and *TBX3*

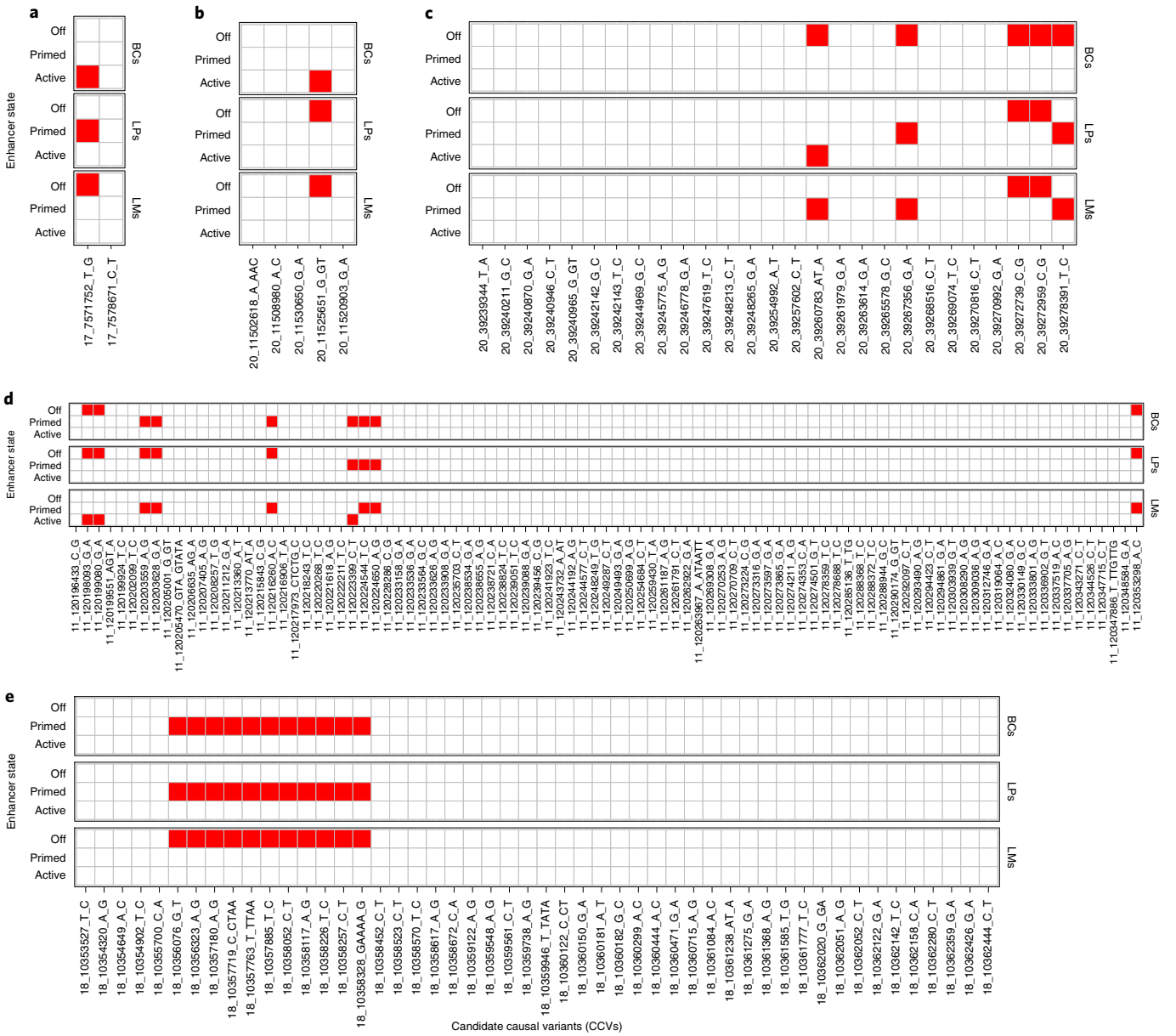


Fig. 4 | Heatmap of CCVs overlapping with enhancer states in primary breast subpopulations for five variants with associations in the opposite direction across subtypes. a–e, The lead variants, chromosomes and positions for each of the five variants were, respectively: rs78378222, 17 and 7,571,752 (**a**), rs141526427, 20 and 11,502,618 (**b**), rs6065254, 20 and 39,248,265 (**c**), rs7924772, 11 and 120,233,626 (**d**) and rs206435, 18 and 10,354,649 (**e**). Three different breast subpopulations were considered: basal cells (BCs), luminal progenitor cells (LPs) and mature luminal cells (LMs). Based on a combination of H3K4me1 and H3K27ac histone modification ChiP-seq signals, putative enhancers in basal cells, luminal progenitor cells and mature luminal cells were characterized as off, primed or active (Methods). The CCVs overlapping with enhancers are colored red.

(a known somatic breast cancer driver gene[15]), along with genes implicated by transcriptome-wide association studies (*LINC00886* (ref. [16]) and *YBEY*[17]).

We used linkage disequilibrium score regression to investigate genetic correlations[18,19] between subtypes and compare enrichment of genomic features[20] between luminal A-like and triple-negative subtypes (Methods). All subtypes were moderately to highly correlated, with luminal A-like and triple-negative having a correlation of 0.46 (s.e. = 0.05). The correlation between breast cancer in *BRCA1* carriers and triple-negative breast cancers in BCAC participants was 0.83 (s.e. = 0.08), suggesting a high degree of similarity in the genetic basis between these subtypes (Fig. 5 and Supplementary Table 16). To compare genomic enrichment, we first evaluated 53 annotations and found that 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 and Supplementary Fig. 9). On investigation of cell-specific enrichment of the histone markers H3K4me1, H3K3me3, H3K9ac and H3K27ac (Supplementary Note), we found both luminal A and triple-negative subtypes enriched for gastrointestinal cell types and suppression of central nervous system cell types (Supplementary Fig. 10).

The proportions of genome-wide chip heritability explained by the 32 novel variants plus 178 previously identified variants[1,2,21] were 54.2, 37.6 and 26.9% for luminal A-like, triple-negative and *BRCA1* carriers, respectively (Table 1 and Supplementary Table 18). These

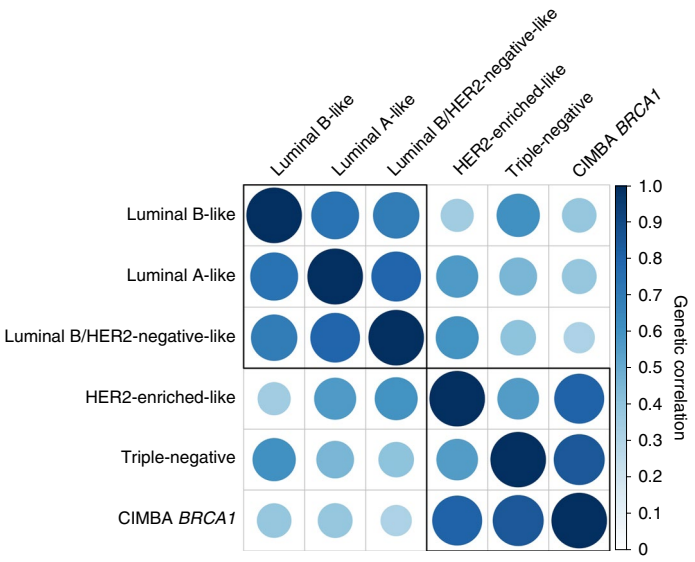


Fig. 5 | Genetic correlation between the five intrinsic-like breast cancer subtypes and breast cancer in *BRCA1* mutation carriers, estimated through linkage disequilibrium score regression. See Supplementary Table 16 for further details. Both the color and size of the circles reflect the strength of the genetic correlations.

210 variants explained approximately 18.3% of the twofold familial relative risk for invasive breast cancer, while all reliably imputable variants on the OncoArray explained 37.1% (Methods). The per-standard deviation ORs between polygenic risk scores (PRSs) for luminal A-like and triple-negative subtypes (Methods), which included 313 published variants[22] and 17 novel variants that were independent of the 313 variants (Supplementary Table 19), were 1.83 (95% CI = 1.78–1.88) and 1.65 (95% CI = 1.57–1.73), with corresponding areas under the receiver-operator curves of 66.09 and 63.58, respectively (Extended Data Fig. 2–6).

These analyses show the benefit of combining standard GWAS methods with methods accounting for underlying tumor heterogeneity. Moreover, these methods and results may help to clarify mechanisms predisposing to specific molecular subtypes, and provide precise risk estimates for subtypes to inform the development of subtype-specific PRSs[22]. However, to expand the generalizability of our findings, these analyses should be replicated and expanded in multi-ancestry populations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0609-2.

Received: 23 September 2019; Accepted: 5 March 2020;
Published online: 18 May 2020

References

1. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
2. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
3. Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat. Genet.* **45**, 392–398 (2013).
4. Zhang, H. et al. A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics. *Biostatistics* https://doi.org/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**, 353–361 (2013).
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**, 373–380 (2015).
7. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B* **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).
9. Spurdle, A. B. 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).
10. Phelan, C. M. et al. Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).
11. Stacey, S. N. et al. A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nat. Genet.* **43**, 1098–1103 (2011).
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).
13. Beesley, J. et al. Chromatin interactome mapping at 139 independent breast cancer risk signals. *Genome Biol.* **21**, 8 (2020).
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).
15. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
16. Ferreira, M. A. et al. Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. *Nat. Commun.* **10**, 1741 (2019).
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).
18. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
19. Bulik-Sullivan, B. K. et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
20. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
21. Ahearn, T. U. et al. Common breast cancer risk loci predispose to distinct tumor subtypes. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/733402v1 (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).

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

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2020

Haoyu Zhang[1,2,228], Thomas U. Ahearn[1,228], Julie Lecarpentier[3], Daniel Barnes[3], Jonathan Beesley[4], Guanghao Qi[2], Xia Jiang[5], Tracy A. O'Mara[4], Ni Zhao[2], Manjeet K. Bolla[6], Alison M. Dunning[3], Joe Dennis[6], Qin Wang[6], Zumuruda Abu Ful[7], Kristiina Aittomäki[8], Irene L. Andrulis[9], Hoda Anton-Culver[10], Volker Arndt[11], Kristan J. Aronson[12], Banu K. Arun[13], Paul L. Auer[14,15], Jacopo Azzollini[16], Daniel Barrowdale[17], Heiko Becher[18], Matthias W. Beckmann[19], Sabine Behrens[20], Javier Benitez[21], Marina Bermisheva[22], Katarzyna Bialkowska[23], Ana Blanco[24,25,26], Carl Blomqvist[27,28],

Natalia V. Bogdanova[29,30,31], Stig E. Bojesen[32,33,34,35], Bernardo Bonanni[36], Davide Bondavalli[36], Ake Borg[37], Hiltrud Brauch[38,39,40], Hermann Brenner[11,40,41], Ignacio Briceno[42], Annegien Broeks[43], Sara Y. Brucker[44], Thomas Brüning[45], Barbara Burwinkel[46,47], Saundra S. Buys[48], Helen Byers[49], Trinidad Caldés[50], Maria A. Caligo[51], Mariarosaria Calvello[36], Daniele Campa[20,52], Jose E. Castelao[53], Jenny Chang-Claude[20,54], Stephen J. Chanock[1], Melissa Christiaens[55], Hans Christiansen[31], Wendy K. Chung[56], Kathleen B. M. Claes[57], Christine L. Clarke[58], Sten Cornelissen[43], Fergus J. Couch[59], Angela Cox[60], Simon S. Cross[61], Kamila Czene[62], Mary B. Daly[63], Peter Devilee[64], Orland Diez[65], Susan M. Domchek[66], Thilo Dörk[30], Miriam Dwek[67], Diana M. Eccles[68], Arif B. Ekici[69], D. Gareth Evans[49,70], Peter A. Fasching[19,71], Jonine Figueroa[72], Lenka Foretova[73], Florentia Fostira[74], Eitan Friedman[75], Debra Frost[17], Manuela Gago-Dominguez[76,77], Susan M. Gapstur[78], Judy Garber[79], José A. García-Sáenz[50], Mia M. Gaudet[78], Simon A. Gayther[80], Graham G. Giles[81,82,83], Andrew K. Godwin[84], Mark S. Goldberg[85,86,87], David E. Goldgar[88], Anna González-Neira[35], Mark H. Greene[89], Jacek Gronwald[23], Pascal Guénel[90], Lothar Häberle[91], Eric Hahnen[92], Christopher A. Haiman[93], Christopher R. Hake[94], Per Hall[62,95], Ute Hamann[96], Elaine F. Harkness[97,98], Bernadette A. M. Heemskerk-Gerritsen[99], Peter Hillemanns[30], Frans B. L. Hogervorst[100], Bernd Holleczek[101], Antoinette Hollestelle[99], Maartje J. Hooning[99], Robert N. Hoover[1], John L. Hopper[82], Anthony Howell[102], Hanna Huebner[19], Peter J. Hulick[103], Evgeny N. Imyanitov[104], kConFab Investigators*, ABCTB Investigators*, Claudine Isaacs[105], Louise Izatt[106], Agnes Jager[99], Milena Jakimovska[107], Anna Jakubowska[23,108], Paul James[109], Ramunas Janavicius[110,111], Wolfgang Janni[112], Esther M. John[113], Michael E. Jones[114], Audrey Jung[20], Rudolf Kaaks[20], Pooja Middha Kapoor[20,115], Beth Y. Karlan[116], Renske Keeman[43], Sofia Khan[117], Elza Khusnutdinova[22,118], Cari M. Kitahara[119], Yon-Dschun Ko[120], Irene Konstantopoulou[74], Linetta B. Koppert[121], Stella Koutros[1], Vessela N. Kristensen[122,123], Anne-Vibeke Laenkholm[124], Diether Lambrechts[125,126], Susanna C. Larsson[127,128], Pierre Laurent-Puig[129], Conxi Lazaro[130], Emilija Lazarova[131], Flavio Lejbkowicz[7], Goska Leslie[6], Fabienne Lesueur[132], Annika Lindblom[133,134], Jolanta Lissowska[135], Wing-Yee Lo[38,136], Jennifer T. Loud[89], Jan Lubinski[23], Alicja Lukomska[23], Robert J. MacInnis[81,82], Arto Mannermaa[137,138,139], Mehdi Manoochehri[96], Siranoush Manoukian[16], Sara Margolin[95,140], Maria Elena Martinez[77,141], Laura Matricardi[142], Lesley McGuffog[6], Catriona McLean[143], Noura Mebirouk[144], Alfons Meindl[145], Usha Menon[146], Austin Miller[147], Elvira Mingazheva[118], Marco Montagna[142], Anna Marie Mulligan[148,149], Claire Mulot[129], Taru A. Muranen[117], Katherine L. Nathanson[66], Susan L. Neuhausen[150], Heli Nevanlinna[117], Patrick Neven[55], William G. Newman[49,70], Finn C. Nielsen[151], Liene Nikitina-Zake[152], Jesse Nodora[77,153], Kenneth Offit[154], Edith Olah[155], Olufunmilayo I. Olopade[156,157], Håkan Olsson[158,159], Nick Orr[160], Laura Papi[161], Janos Papp[155], Tjoung-Won Park-Simon[30], Michael T. Parsons[4], Bernard Peissel[16], Ana Peixoto[162], Beth Peshkin[163], Paolo Peterlongo[164], Julian Peto[6,165], Kelly-Anne Phillips[82,166,167], Marion Piedmonte[147], Dijana Plaseska-Karanfilska[107], Karolina Prajzendanc[23], Ross Prentice[14], Darya Prokofyeva[118], Brigitte Rack[112], Paolo Radice[168], Susan J. Ramus[169,170,171], Johanna Rantala[172], Muhammad U. Rashid[96,173], Gad Rennert[7], Hedy S. Rennert[7], Harvey A. Risch[174], Atocha Romero[175,176], Matti A. Rookus[177], Matthias Rübner[91], Thomas Rüdiger[178], Emmanouil Saloustros[179], Sarah Sampson[180], Dale P. Sandler[181], Elinor J. Sawyer[182], Maren T. Scheuner[183], Rita K. Schmutzler[92], Andreas Schneeweiss[47,184], Minouk J. Schoemaker[114], Ben Schöttker[11], Peter Schürmann[30], Leigha Senter[185], Priyanka Sharma[186], Mark E. Sherman[187], Xiao-Ou Shu[188], Christian F. Singer[189], Snezhana Smichkoska[131], Penny Soucy[190],

Melissa C. Southey[83], John J. Spinelli[191,192], Jennifer Stone[82,193], Dominique Stoppa-Lyonnet[194], EMBRACE Study*, GEMO Study Collaborators*, Anthony J. Swerdlow[114,195], Csilla I. Szabo[196], Rulla M. Tamimi[5,197,198], William J. Tapper[199], Jack A. Taylor[181,200], Manuel R. Teixeira[162,176], MaryBeth Terry[201], Mads Thomassen[202], Darcy L. Thull[203], Marc Tischkowitz[204,205], Amanda E. Toland[206], Rob A. E. M. Tollenaar[207], Ian Tomlinson[208,209], Diana Torres[96,210], Melissa A. Troester[211], Thérèse Truong[90], Nadine Tung[212], Michael Untch[213], Celine M. Vachon[214], Ans M. W. van den Ouweland[215], Lizet E. van der Kolk[100], Elke M. van Veen[49,70], Elizabeth J. vanRensburg[216], Ana Vega[24,25,26], Barbara Wappenschmidt[92], Clarice R. Weinberg[217], Jeffrey N. Weitzel[218], Hans Wildiers[55], Robert Winqvist[219,220,221,222], Alicja Wolk[108,127,128], Xiaohong R. Yang[1], Drakoulis Yannoukakos[74], Wei Zheng[188], Kristin K. Zorn[223], Roger L. Milne[81,82,83], Peter Kraft[5,198], Jacques Simard[190], Paul D. P. Pharoah[3,6], Kyriaki Michailidou[6,224,225], Antonis C. Antoniou[6], Marjanka K. Schmidt[43,226], Georgia Chenevix-Trench[4], Douglas F. Easton[3], Nilanjan Chatterjee[2,227,229] ✉ and Montserrat García-Closas[1,229]

[1]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health (NIH), Department of Health and Human Services, Bethesda, MD, USA. [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [3]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. [4]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [5]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [6]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [7]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel. [8]Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [9]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [10]Department of Epidemiology, Genetic Epidemiology Research Institute, University of California, Irvine, Irvine, CA, USA. [11]Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany. [12]Department of Public Health Sciences and Cancer Research Institute, Queen's University, Kingston, Ontario, Canada. [13]Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [14]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [15]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA. [16]Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy. [17]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, UK. [18]Institute of Medical Biometry and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [19]Department of Gynecology and Obstetrics, Comprehensive Cancer Center ER-EMN, University Hospital Erlangen, Friedrich–Alexander University of Erlangen–Nuremberg, Erlangen, Germany. [20]Division of Cancer Epidemiology, DKFZ, Heidelberg, Germany. [21]Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Valencia, Spain. [22]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia. [23]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland. [24]Molecular Medicine Unit, Fundación Pública Galega de Medicina Xenómica, Santiago de Compostela, Spain. [25]Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, Servicio Galego de Saude (SERGAS), Santiago de Compostela, Spain. [26]CIBERER, Santiago de Compostela, Spain. [27]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [28]Department of Oncology, Örebro University Hospital, Örebro, Sweden. [29]N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus. [30]Gynaecology Research Unit, Hannover Medical School, Hannover, Germany. [31]Department of Radiation Oncology, Hannover Medical School, Hannover, Germany. [32]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [33]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [34]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [35]Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [36]Division of Cancer Prevention and Genetics, European Institute of Oncology (IEO) IRCCS, Milan, Italy. [37]Department of Oncology, Lund University and Skåne University Hospital, Lund, Sweden. [38]Dr. Margarete Fischer-Bosch Institute of Clinical Pharmacology, Stuttgart, Germany. [39]iFIT Cluster of Excellence, University of Tübingen, Tübingen, Germany. [40]German Cancer Consortium (DKTK), DKFZ, Heidelberg, Germany. [41]Division of Preventive Oncology, DKFZ and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [42]Bioscience Department, Faculty of Medicine, Universidad de la Sabana, Chia, Colombia. [43]Division of Molecular Pathology, The Netherlands Cancer Institute—Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [44]Department of Women's Health, University of Tübingen, Tübingen, Germany. [45]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance (IPA), Ruhr University Bochum, Bochum, Germany. [46]Molecular Epidemiology Group (C080), DKFZ, Heidelberg, Germany. [47]Molecular Biology of Breast Cancer, University Women's Clinic Heidelberg, University of Heidelberg, Heidelberg, Germany. [48]Department of Medicine, Huntsman Cancer Institute, Salt Lake City, UT, USA. [49]Manchester 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. [50]Medical Oncology Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro de Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, Spain. [51]Section of Molecular Genetics, Department of Laboratory Medicine, University Hospital of Pisa, Pisa, Italy. [52]Department of Biology, University of Pisa, Pisa, Italy. [53]Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain. [54]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [55]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium. [56]Departments of Pediatrics and Medicine, Columbia University, New York, NY, USA. [57]Centre for Medical Genetics, Ghent University, Ghent, Belgium. [58]Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia. [59]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA. [60]Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of

Sheffield, Sheffield, UK. [61]Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK. [62]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [63]Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA. [64]Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands. [65]Oncogenetics Group, Vall d'Hebron Institute of Oncology (VHIO), Barcelona, Spain. [66]Department of Medicine, Abramson Cancer Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [67]Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK. [68]Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK. [69]Institute of Human Genetics, University Hospital Erlangen, Friedrich–Alexander University Erlangen–Nuremberg, Comprehensive Cancer Center ER-EMN, Erlangen, Germany. [70]Division 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. [71]David Geffen School of Medicine, Department of Medicine, Division of Hematology and Oncology, University of California, Los Angeles, Los Angeles, CA, USA. [72]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh Medical School, Edinburgh, UK. [73]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic. [74]Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research 'Demokritos', Athens, Greece. [75]The Susanne Levy Gertner Oncogenetics Unit, Chaim Sheba Medical Center, Ramat Gan, Israel. [76]Genomic Medicine Group, Galician Foundation of Genomic Medicine, IDIS, Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain. [77]Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. [78]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA. [79]Cancer Risk and Prevention Clinic, Dana-Farber Cancer Institute, Boston, MA, USA. [80]Center for Bioinformatics and Functional Genomics and the Cedars-Sinai Genomics Core, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [81]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia. [82]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Victoria, Australia. [83]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Melbourne, Victoria, Australia. [84]Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Kansas City, KS, USA. [85]Department of Medicine, McGill University, Montréal, Québec, Canada. [86]Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, Québec, Canada. [87]Breast Cancer Research Unit, Cancer Research Institute, University Malaya Medical Centre, Kuala Lumpur, Malaysia. [88]Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, USA. [89]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [90]Cancer and Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University of Paris-Saclay, Paris, France. [91]Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich–Alexander University Erlangen–Nuremberg, Comprehensive Cancer Center ER-EMN, Erlangen, Germany. [92]Center for Familial Breast and Ovarian Cancer, Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [93]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [94]Waukesha Memorial Hospital Pro Health Care, Waukesha, WI, USA. [95]Department of Oncology, Södersjukhuset, Stockholm, Sweden. [96]Molecular Genetics of Breast Cancer, DKFZ, Heidelberg, Germany. [97]Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. [98]Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester, UK. [99]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, the Netherlands. [100]Family Cancer Clinic, The Netherlands Cancer Institute—Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [101]Saarland Cancer Registry, Saarbrücken, Germany. [102]Division of Cancer Sciences, University of Manchester, Manchester, UK. [103]Center for Medical Genetics, NorthShore University HealthSystem, Evanston, IL, USA. [104]N.N. Petrov Institute of Oncology, St. Petersburg, Russia. [105]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA. [106]Clinical Genetics, Guy's and St. Thomas' NHS Foundation Trust, London, UK. [107]Research Centre for Genetic Engineering and Biotechnology 'Georgi D. Efremov', Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia. [108]Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland. [109]Parkville Familial Cancer Centre, Peter MacCallum Cancer Center, Melbourne, Victoria, Australia. [110]Hematology, Oncology and Transfusion Medicine Center, Department of Molecular and Regenerative Medicine, Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania. [111]State Research Institute Center for Innovative Medicine, Vilnius, Lithuania. [112]Department of Gynaecology and Obstetrics, University Hospital Ulm, Ulm, Germany. [113]Department of Medicine, Division of Oncology, Stanford Cancer Institute, School of Medicine, Stanford University, Stanford, CA, USA. [114]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. [115]Faculty of Medicine, University of Heidelberg, Heidelberg, Germany. [116]David Geffen School of Medicine, Department of Obstetrics and Gynecology, University of California, Los Angeles, Los Angeles, CA, USA. [117]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [118]Department of Genetics and Fundamental Medicine, Bashkir State Medical University, Ufa, Russia. [119]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [120]Department of Internal Medicine, Evangelische Kliniken Bonn, Johanniter Krankenhaus, Bonn, Germany. [121]Department of Surgical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, the Netherlands. [122]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway. [123]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway. [124]Department of Surgical Pathology, Zealand University Hospital, Slagelse, Denmark. [125]VIB Center for Cancer Biology, VIB, Leuven, Belgium. [126]Laboratory of Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium. [127]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [128]Department of Surgical Sciences, Uppsala University, Uppsala, Sweden. [129]Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris, France. [130]Molecular Diagnostic Unit, Hereditary Cancer Program, Bellvitge Biomedical Research Institute (IDIBELL), Catalan Institute of Oncology (ICO), CIBERONC, Barcelona, Spain. [131]Ss. Cyril and Methodius University in Skopje, Medical Faculty, University Clinic of Radiotherapy and Oncology, Skopje, Republic of North Macedonia. [132]Genetic Epidemiology of Cancer Team, INSERM U900, Paris, France. [133]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden. [134]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden. [135]Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Cancer Center, Oncology Institute, Warsaw, Poland. [136]University of Tübingen, Tübingen, Germany. [137]Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland. [138]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland. [139]Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland. [140]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden. [141]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, USA. [142]Immunology and Molecular Oncology Unit, Veneto Institute of Oncology (IOV)-IRCCS, Padua, Italy. [143]Department of Anatomical Pathology, The Alfred Hospital, Melbourne, Victoria, Australia. [144]Genetic Epidemiology of Cancer Team, INSERM U900, Institut Curie, PSL University, Mines ParisTech, Paris, France. [145]Department of Gynecology and Obstetrics, Ludwig Maximilian University of Munich, Munich, Germany. [146]MRC Clinical Trials Unit at UCL, Institute of Clinical Trials and Methodology, University College London, London, UK. [147]NRG Oncology, Statistics and Data Management Center, Roswell Park Cancer Institute, Buffalo, NY, USA. [148]Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada. [149]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [150]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA. [151]Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark. [152]Latvian Biomedical Research and Study Centre, Riga, Latvia. [153]Department of Family Medicine and Public Health, School of Medicine, University of California, San Diego, La Jolla, CA, USA. [154]Clinical Genetics Research Laboratory, Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [155]Department of Molecular Genetics, National Institute of Oncology, Budapest, Hungary.

[156]Center for Clinical Cancer Genetics, The University of Chicago, Chicago, IL, USA. [157]Department of Clinical Pathology, University of Melbourne, Melbourne, Victoria, Australia. [158]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden. [159]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [160]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, Northern Ireland, UK. [161]Unit of Medical Genetics, Department of Biomedical, Experimental and Clinical Sciences, University of Florence, Florence, Italy. [162]Department of Genetics, Portuguese Oncology Institute, Porto, Portugal. [163]Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA. [164]Genome Diagnostics Program, IFOM, FIRC Institute of Molecular Oncology, Milan, Italy. [165]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [166]Peter MacCallum Cancer Center, Melbourne, Victoria, Australia. [167]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. [168]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, INT, Milan, Italy. [169]Adult Cancer Program, Lowy Cancer Research Centre, University of NSW Sydney, Sydney, New South Wales, Australia. [170]School of Women's and Children's Health, Faculty of Medicine, University of NSW Sydney, Sydney, New South Wales, Australia. [171]The Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [172]Clinical Genetics, Karolinska Institutet, Stockholm, Sweden. [173]Department of Basic Sciences, Shaukat Khanum Memorial Cancer Hospital and Research Centre (SKMCH&RC), Lahore, Pakistan. [174]Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA. [175]Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain. [176]Biomedical Sciences Institute (ICBAS), University of Porto, Porto, Portugal. [177]Department of Epidemiology, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [178]Institute of Pathology, Staedtisches Klinikum Karlsruhe, Karlsruhe, Germany. [179]Department of Oncology, University Hospital of Larissa, Larissa, Greece. [180]Prevent Breast Cancer Centre and Nightingale Breast Screening Centre, Manchester University NHS Foundation Trust, Manchester, UK. [181]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Durham, NC, USA. [182]Research Oncology, Guy's Hospital, King's College London, London, UK. [183]Cancer Genetics and Prevention Program, University of California, San Francisco, San Francisco, CA, USA. [184]NCT, University Hospital and DKFZ, Heidelberg, Germany. [185]Clinical Cancer Genetics Program, Division of Human Genetics, Department of Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [186]Department of Internal Medicine, Division of Oncology, University of Kansas Medical Center, Westwood, Kansas City, KS, USA. [187]Department of Health Sciences Research, Mayo Clinic College of Medicine, Jacksonville, FL, USA. [188]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA. [189]Department of Obstetrics and Gynecology and Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria. [190]Genomics Center, Centre Hospitalier Universitaire de Québec—Université Laval Research Center, Québec City, Québec, Canada. [191]Population Oncology, BC Cancer, Vancouver, British Columbia, Canada. [192]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada. [193]Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Western Australia, Australia. [194]Department of Genetics, INSERM U830, Institut Curie, Paris Descartes Sorbonne-Paris Cité University, Paris, France. [195]Division of Breast Cancer Research, The Institute of Cancer Research, London, UK. [196]National Human Genome Research Institute, National Cancer Institute, Bethesda, MD, USA. [197]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [198]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [199]Faculty of Medicine, University of Southampton, Southampton, UK. [200]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Durham, NC, USA. [201]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA. [202]Department of Clinical Genetics, Odense University Hospital, Odense, Denmark. [203]Department of Medicine, Magee-Womens Hospital, School of Medicine, University of Pittsburgh , Pittsburgh, PA, USA. [204]Program in Cancer Genetics, Departments of Human Genetics and Oncology, McGill University, Montréal, Québec, Canada. [205]Department of Medical Genetics, National Institute for Health Research, Cambridge Biomedical Research Centre, University of Cambridge, Cambridge, UK. [206]Department of Cancer Biology and Genetics, The Ohio State University, Columbus, OH, USA. [207]Department of Surgery, Leiden University Medical Center, Leiden, the Netherlands. [208]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. [209]Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford, UK. [210]Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia. [211]Department 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. [212]Department of Medical Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [213]Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany. [214]Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA. [215]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, the Netherlands. [216]Department of Genetics, University of Pretoria, Pretoria, South Africa. [217]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Durham, NC, USA. [218]Clinical Cancer Genomics, City of Hope, Duarte, CA, USA. [219]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland. [220]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland. [221]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [222]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [223]Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [224]Biostatistics Unit, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [225]Cyprus School of Molecular Medicine, Nicosia, Cyprus. [226]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute—Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [227]Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD, USA. [228]These authors contributed equally: Haoyu Zhang, Thomas U. Ahearn. [229]These authors jointly supervised this work: Georgia Chenevix-Trench, Douglas F. Easton, Nilanjan Chatterjee, Montserrat García-Closas. *A list of members and affiliations appears in the Supplementary Note. ✉e-mail: nchatte2@jhu.edu

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, cases of invasive breast cancer, cases of carcinoma in situ and cases of unknown invasiveness. Information on clinicopathological characteristics was collected for the individual studies and combined in a central database after quality control checks. We used the BCAC database version 10 for these analyses. Among a subset of participants ($n=16,766$) who were genotyped on both the iCOGS and OncoArray arrays, we kept only the OncoArray data. One study contributing to the iCOGS dataset (Leuven Multidisciplinary Breast Centre) was excluded due to inflation of the test statistics that was not corrected by adjustment for the first ten principal components. 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 estrogen receptor, progesterone receptor, 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 characteristic correlations compared with invasive cases, which could potentially bias the estimates from the expectation–maximization based missing data handling 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 the footnote of Supplementary Table 2 for a further explanation of the excluded studies). We did not include the summary results from the 14,910 cases and 17,588 controls from the 11 other GWASs 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 (included in the expectation–maximization algorithm to better 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 were available from 58 studies from 24 countries. Following quality control and the removal of participants who 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 GWASs 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 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 of >0.005 in the overall breast cancer analysis and >0.01 in the subtype analysis.

Known breast cancer susceptibility variants. Previous studies identified susceptibility variants from genome-wide analyses at a significance level of $P<5.0\times10^{-8}$ for all breast cancer types, estrogen-receptor-negative or estrogen-receptor-positive breast cancer, in *BRCA1* or *BRCA2* mutation carriers, or in meta-analyses of these[1–3]. We defined known breast cancer susceptibility variants as those variants that were identified or replicated in previous BCAC analyses[1,2]. To help ensure that novel, independent susceptibility variants were identified, we excluded from these analyses variants within 500 kilobases (kb) of a previously published variant. These excluded regions were subject to separate, fine-mapping conditional analyses that were focused on identifying additional independent susceptibility variants in these regions[14].

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 ten principal components for ancestry informative variants. The methods for estimating principal components have been described elsewhere[1,2]. For the remaining GWASs, adjustment for inflation was done by adjusting for up to three principal components and using genomic control adjustment, as previously described[1]. We evaluated the associations between approximately 10.8 million variants with imputation quality scores (r^2) ≥ 0.3 and minor allele frequency (MAF) scores of >0.005. We excluded variants located within ±500 kb of, or in linkage disequilibrium ($r^2\geq0.1$) with, known susceptibility variants[21]. The association effect size estimates from these GWASs, as well as the previously derived estimates from the 11 other GWASs, were then combined using a fixed-effects meta-analysis. Since individual-level genotyping data were not available for some previous GWASs, 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 Sullivan[23].

Subtype analysis of BCAC data. We have 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 estrogen receptor, progesterone receptor and HER2-positive/negative status and grade of differentiation (low, intermediate or 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 OR parameters through a decomposition procedure resulting in a baseline parameter that represents a case–control OR of a baseline cancer subtype, and case–case OR 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 the underlying two-stage polytomous model. In the two-stage model, the frequency of different tumor subtypes corresponding to different combinations of the tumor characteristics is 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 themselves, the methodology accounts for the correlation among the tumor markers in a robust manner that does not require strong modeling assumptions.

To identify novel susceptibility loci, we used both a fixed-effects two-stage polytomous model and a mixed-effects two-stage polytomous model. The score test we developed based on the mixed-effects model allows coefficients associated with individual tumor characteristics to enter as either fixed- or random-effects terms. Our previous analyses have shown that incorporation of random-effects terms can improve the power of the score test by essentially reducing the effective degrees of freedom associated with fixed effects related to exploratory markers (that is, markers for which there is little previous evidence to suggest that they are a source of heterogeneity)[4]. In contrast, incorporation of fixed-effects terms can preserve distinct associations of known important tumor characteristics, such as estrogen receptor. In the mixed-effects two-stage polytomous model, we therefore kept estrogen receptor as a fixed effect, but modeled progesterone receptor, HER2 and grade as random effects. We evaluated variants with MAF > 0.01 (~10.0 million) and $r^2\geq0.3$, and excluded variants within ±500 kb of, or in linkage disequilibrium ($r^2\geq0.1$) 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 $P<5.0\times10^{-8}$ in either the fixed- or mixed-effects models.

Both fixed- and mixed-effects models adjusted for the top ten principal components and age. As age is correlated with tumor characteristics[25], we added age as a covariate to improve the statistical power of the expectation–maximization algorithm. Country was not adjusted for in the subtype analyses, since doing so required adequate sample sizes for 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 the 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 correlations[27]. 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-negative[28], 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 METAL[29]. 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 the 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 the association for all remaining variants within ±500 kb of the lead variants, irrespective of linkage disequilibrium. 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 were already genome-wide significant in the original GWAS scan and the conditional analyses restricted to local regions, we used a significance threshold of $P < 1 \times 10^{-6}$ to control for type I error[30].

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-effects model test to identify variants showing evidence of heterogeneity with respect to any of the underlying tumor markers, estrogen receptor, progesterone receptor, HER2 and/or grade. We accounted for multiple testing of the global heterogeneity test using an FDR < 0.05 under the Benjamini–Hochberg procedure[31]. Among the variants with observed heterogeneity, we then further used a fixed-effects two-stage model to evaluate the influence of specific tumor characteristic(s) driving observed heterogeneity, adjusted for the other markers in the model. We also fit separate fixed-effects two-stage models to estimate case–control ORs and 95% CIs for five surrogate intrinsic-like subtypes defined by combinations of estrogen receptor, progesterone receptor, HER2 and grade[8]: (1) luminal A-like (estrogen receptor$^+$ and/or progesterone receptor$^+$; HER2$^-$; grades 1 and 2); (2) luminal B/HER2-negative-like (estrogen receptor$^+$ and/or progesterone receptor$^+$; HER2$^-$; grade 3); (3) luminal B-like (estrogen receptor$^+$ and/or progesterone receptor$^+$; HER2$^+$); (4) HER2-enriched-like (estrogen receptor$^-$ and progesterone receptor$^-$; HER2$^+$); and (5) triple-negative (estrogen receptor$^-$; progesterone receptor$^-$; HER2$^-$). Furthermore, 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 the two-stage polytomous model accounting for missing tumor data.

CCVs. We defined credible sets of 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 subpopulations. We obtained enhancer maps for three enriched primary breast subpopulations (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 fragments per kilobase of transcript per million mapped reads thresholds, as previously described[12]. Briefly, genomic regions containing a high H3K4me1 signal observed in any cell type were used to define the superset of breast regulatory elements. A subpopulation 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 sites were characterized by H3K4me1-high/H3K27ac-low, and off sites were characterized 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 that exhibited differing states between at least one cell type as switch enhancers.

Genetic correlation analyses. We used linkage disequilibrium 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, iCOGS and CIMBA meta-analysis. The genetic correlation[18] analysis was restricted to the roughly 1 million variants included in HapMap 3 with a MAF of >1% and an imputation quality score r^2 of >0.3 in the OncoArray data. Since two-stage polytomous models integrated an imputation algorithm for missing tumor characteristic data, we modified the linkage disequilibrium 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:

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

where n is the total number of identified variants, $_{pi}$ is the MAF for the ith variant, $\hat{\beta}_i$ is the log[odds ratio] estimate for the ith 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 linkage disequilibrium score regression to estimate heritability (σ^2_{GWAS}) 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^2_{GWAS} is characterized by population variance of the underlying true polygenetic risk scores as $\sigma^2_{\mathrm{GWAS}} = \mathrm{Var}\left(\sum_{m=1}^{M} \beta_m G_m\right)$, where G_m is the standardized genotype for the mth variant, β_m is the true log[odds ratio] for the mth variant and M is 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:

$$\sum_{i=1}^{n} 2p_i(1-p_i)\left(\hat{\beta}_i^2 - \tau_i^2\right)/\sigma^2_{\mathrm{GWAS}}$$

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 of the familial aggregation of the disease[34]. Under the frailty scale, we define the broad sense heritability[35] as σ^2. The relationship between λ and σ^2 was shown to be $\sigma^2 = 2 \times \log[\lambda]$[34]. We assumed $\lambda = 2$ as the overall familial relative risk of invasive breast cancer[34]; thus, $\sigma^2 = 2 \times \log[2]$ and the proportion of the familial relative risk explained by identified susceptibility variants is $\sum_{i=1}^{n} p_i(1-p_i)\left(\hat{\beta}_i^2 - \tau_i^2\right)/\log[2]$, and the proportion of the familial relative risk explained by GWAS variants is $\sigma^2_{\mathrm{GWAS}}/[2\times \log[2]]$. Analyses 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 hereditability across five intrinsic subtypes using linkage disequilibrium score regression where the summary statistics were derived using either a 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 estrogen-receptor-specific breast cancer[22]. The 313 SNPs included SNPs that did not reach genome-wide significance. After excluding variants within 500 kb of the 313 SNPs or with a linkage disequilibrium $r^2 \geq 0.1$, 17 of the 32 novel variants were independent of the 313 SNPs. The BCAC data were split into a training dataset and a test dataset with proportions of 80 and 20%, respectively. Half of the test dataset were five studies nested within prospective cohorts, including Karolinska Mammography Project for Risk Prediction of Breast Cancer–Cohort Study, Mayo Mammography Health Study, The Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial, The Sister Study and UK Breakthrough Generations Study (Supplementary Table 2), and the other half were 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 estrogen-receptor-specific log[odds ratios] for 313 SNPs by respectively fitting standard and estrogen-receptor-specific logistic regression on the training dataset. We obtained the log[odds ratio] for 330 SNPs by fitting the fixed-effects 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

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 2SISTER, ABCS,

ABCS-F, ABCTB, BBCC, BCEES, BCFR-NY, BCFR-PA, BCINIS, BIGGS, BREOGAN, BSUCH, CECILE, CGPS, CNIO-BCS, CPSII, CTS, DIETCOMPLYF, ESTHER, FHRISK, FHRISK, GENICA, GEPARSIXTO, HABCS, HCSC, HEBCS, HMBCS, HUBCS, KARBAC, KARMA, KBCP, KCONFAB/AOCS, LMBC, MABCS, MBCSG, MCBCS, MISS, MMHS, MTLGEBCS, NBCS, NCBCS, OBCS, ORIGO, PKARMA, PREFACE, PROCAS, RBCS, SEARCH, SKKDKFZS, SUCCESSB, SUCCESSC, SZBCS, TNBCC, UCIBCS, UKBGS, UKOPS and USRT 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

The data analysis code relevant to 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 an 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.

References

23. Lin, D. Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* **85**, 862–872 (2009).
24. Chatterjee, N. A two-stage regression model for epidemiological studies with multivariate disease classification data. *J. Am. Stat. Assoc.* **99**, 127–138 (2004).
25. Anderson, W. F., Rosenberg, P. S., Prat, A., Perou, C. M. & Sherman, M. E. How many etiological subtypes of breast cancer: two, three, four, or more? *J. Natl. Cancer Inst.* **106**, dju165 (2014).
26. Barnes, D. R. et al. Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations. *Genet. Epidemiol.* **36**, 274–291 (2012).
27. Antoniou, A. C. et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. *Nat. Genet.* **42**, 885–892 (2010).
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**, 134–147 (2012).
29. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
30. Hendricks, A. E., Dupuis, J., Logue, M. W., Myers, R. H. & Lunetta, K. L. Correction for multiple testing in a gene region. *Eur. J. Hum. Genet.* **22**, 414–418 (2014).
31. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
32. Udler, M. S., Tyrer, J. & Easton, D. F. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. *Genet. Epidemiol.* **34**, 463–468 (2010).
33. Wellcome Trust Case Control Consortiumet al. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat. Genet.* **44**, 1294–1301 (2012).
34. Pharoah, P. D. et al. Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).
35. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

Acknowledgements

We thank all of the individuals who took part in these studies, and all of the researchers, clinicians, technicians and administrative staff who enabled this work to be carried out. This project has been funded in part with Federal funds from the National Cancer Institute Intramural Research Program, National Institutes of Health. 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 et 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) (1U19 CA148065 for the Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project and X01HG007492 to the Center for Inherited Disease Research 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 (1U19 CA148065) (DRIVE, part of the GAME-ON initiative). Linkage disequilibrium 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) and 634935 (BRIDGES)). CIMBA was funded by Cancer Research UK (C12292/A20861 and C12292/A11174). N.C. was funded by NHGRI (1R01 HG010480-01). For a full description of funding and acknowledgments, see the Supplementary Note.

Author Contributions

Writing group: H.Z., T.U.A., D. Barnes, J. Beesley, G.Q., X.J., T.A.O., R.L.M., P.K., J. Simard, P.D.P.P., K.M., A.C.A., M.K.S., G.C.-T., D.F.E., N.C., M.G.-C. Statistical analysis: H.Z., T.U.A., J. Lecarpentier, D. Barnes, J. Beesley, G.Q., X.J., T.A.O., N.Z. Provision of DNA samples and/or phenotypic data: M.K.B., A.M.D., J.D., Q.W., Z.A.F., K.A., I.L.A., H.A.-C., V.A., K.J.A., B.K.A., P.L.A., J.A., D. Barrowdale, H. Becher, M.W.B., S.B., J. Benitez, M.B., K.B., A. Blanco, C.B., N.V.B., S.E.B., B. Bonanni, D. Bondavalli, A. Borg, H. Brauch, H. Brenner, I.B., A. Broeks, S.Y.B., T.B., B. Burwinkel, S.S.B., H. Byers, T.C., M.A.C., M.C., D.C., J.E.C., J.C.-C., S.J.C., M.C., H.C., W.K.C., K.B.M.C., C.L.C., S.C., F.J.C., A.C., S.S.C., K.C., M.B.D., P.D., O.D., S.M.D., T.D., M.D., D.M.E., A.B.E., D.G.E., P.A.F., J.F., L.F., F.F., E.F., D.F., M.G.-D., S.M.G., J. Garber, J.A.G.-S., M.M.G., S.A.G., G.G.G., A.K.G., M.S.G., D.E.G., A.G.-N., M.H.G., J. Gronwald, P.G., L.H., E.H., C.A.H., C.R.H., P. Hall, U.H., E.F.H., B.A.M.H.-G., P. Hillemanns, F.B.L.H., B.H., A. Hollestelle, M.J.H., R.N.H., J.L.H., A. Howell, H.H., P.J.H., E.N.I., C.I., L.I., A. Jager, M.J., A. Jakubowska, P.J., R.J., W.J., E.M.J., M.E.J., A. Jung, R. Kaaks, P.M.K., B.Y.K., R. Keeman, S. Khan, E.K., C.M.K., Y.D.K., I.K., L.B.K., S. Koutros, V.N.K., A.-V.L., D.L., S.C.L., P.L.-P., C.L., E.L., F. Lejbkowicz, G.L., F. Lesueur, A. Lindblom, J. Lissowska, W.-Y.L., J.T.L., J. Lubinski, A. Lukomska, R.J.M., A. Mannermaa, M. Manoochehri, S. Manoukian, S. Margolin, M.E.M., L. Matricardi, L. McGuffog, C. McLean, N.M., A. Meindl, U.M., A. Miller, E.M., M. Montagna, A.M.M., C. Mulot, T.A.M., K.L.N., S.L.N., H.N., P.N., W.G.N., F.C.N., L.N.-Z., J.N., K.O., E.O., O.I.O., H.O., N.O., L.P., J. Papp, T.-W.P.-S., M.T.P., B. Peissel, A.P., B. Peshkin, P.P., J. Peto, K.-A.P., M.P., D.P.-K., K.P., R.P., D.P., B.R., P.R., S.J.R., J.R., M.U.R., G.R., H.S.R., H.A.R., A.R., M.A.R., R.R., T.R., E.S., S. Sampson, D.P.S., E.J.S., M.T.S., R.K.S., A.S., M.J.S., B.S., P. Schürmann, L.S., P. Sharma, M.E.S., X.-O.S., C.F.S., S. Smichkoska, P. Soucy, M.C.S., J.J.S., J. Stone, D.S.-L., A.J.S., C.I.S., R.M.T., W.J.T., J.A.T., M.R.T., M. Terry, M. Thomassen, D.L.T., M. Tischkowitz, A.E.T., R.A.E.M.T., I.T., D.T., M.A.T., T.T., N.T., M.U., C.M.V., A.M.W.V.D.O., L.E.V.D.K., E.M.V.V., E.J.V.R., A.V., B.W., C.R.W., J.N.W., H.W., R.W., A.W., X.R.Y., D.Y., W.Z., K.K.Z., R.L.M., P.K., J. Simard, P.D.P.P., K.M., A.C.A., M.K.S., G.C.-T., D.F.E., M.G.-C. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

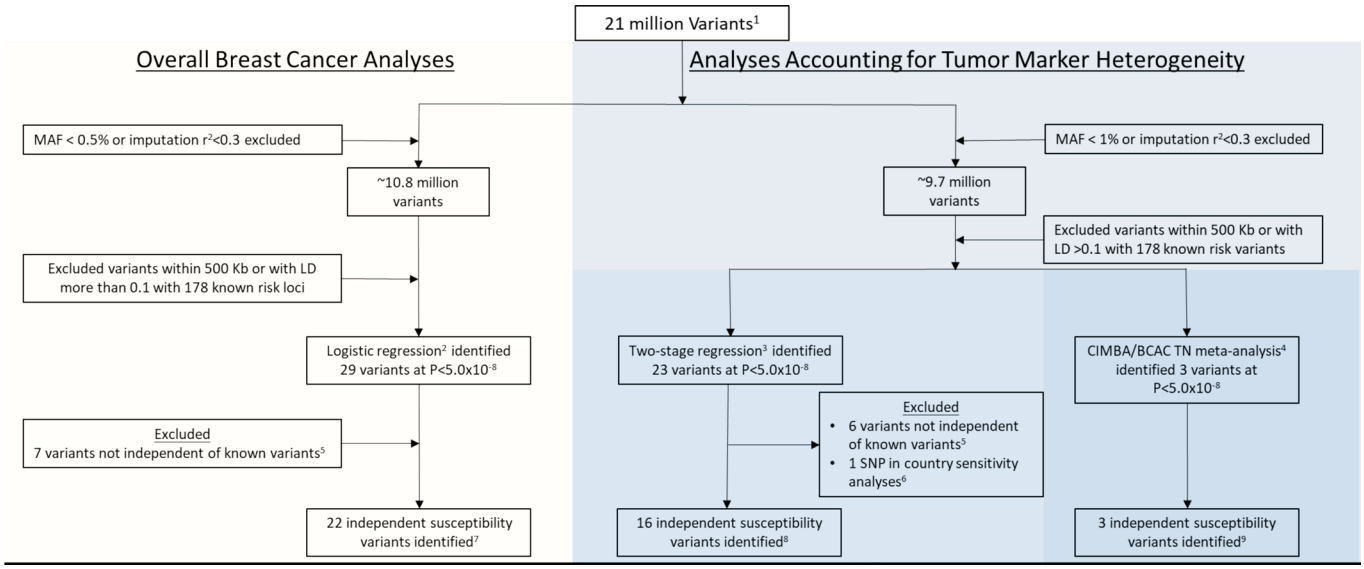
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-020-0609-2.

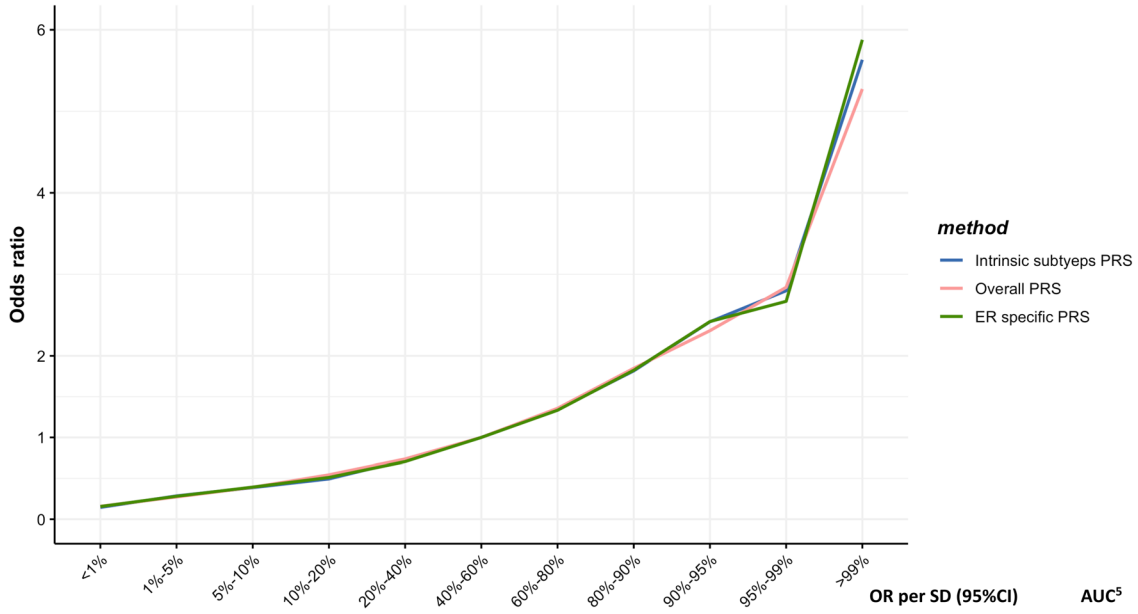
Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0609-2.

Correspondence and requests for materials should be addressed to N.C.

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



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>1x10-6) of 178 known susceptibility variants (see Methods). (6) See Supplementary Fig. 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.



												OR per SD (95%CI)	AUC[5]
Intrinsic subtypes PRS ORs[1]	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 ORs[2]	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 ORs[3]	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**)

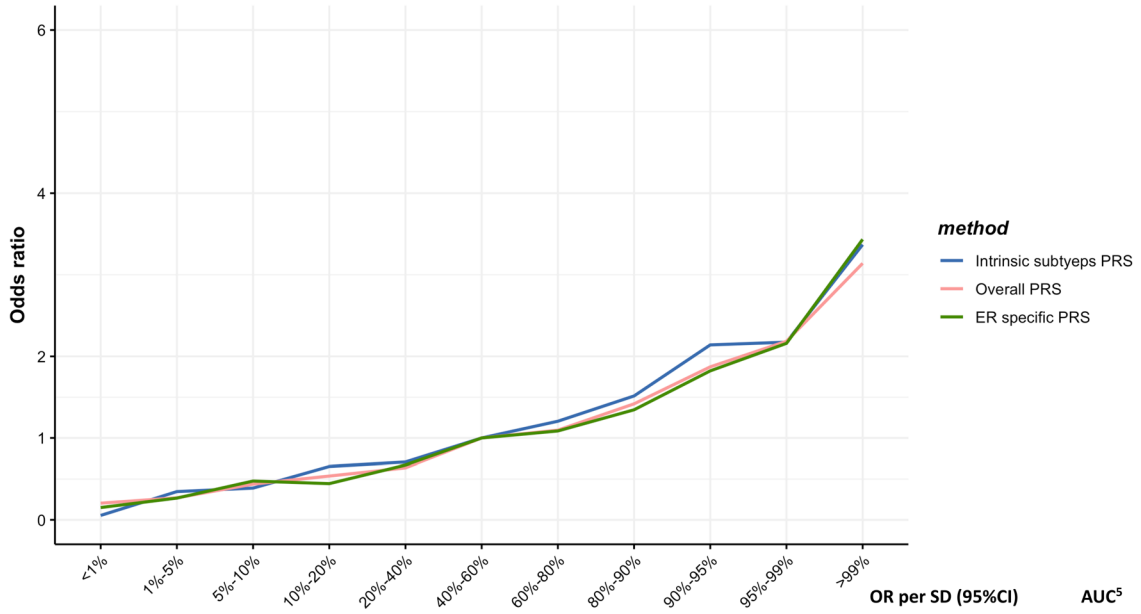
[2] Overall breast cancer PRS with 313 SNPs previously reported[22]

[3] ER-specific PRS with 313 SNPs previously reported[22]

[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-like[4] risk in the test dataset. Odds ratios for different quantiles of the PRS against the middle quantile (40%–60%) 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 ORs[1]	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 ORs[2]	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 ORs[3]	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**)

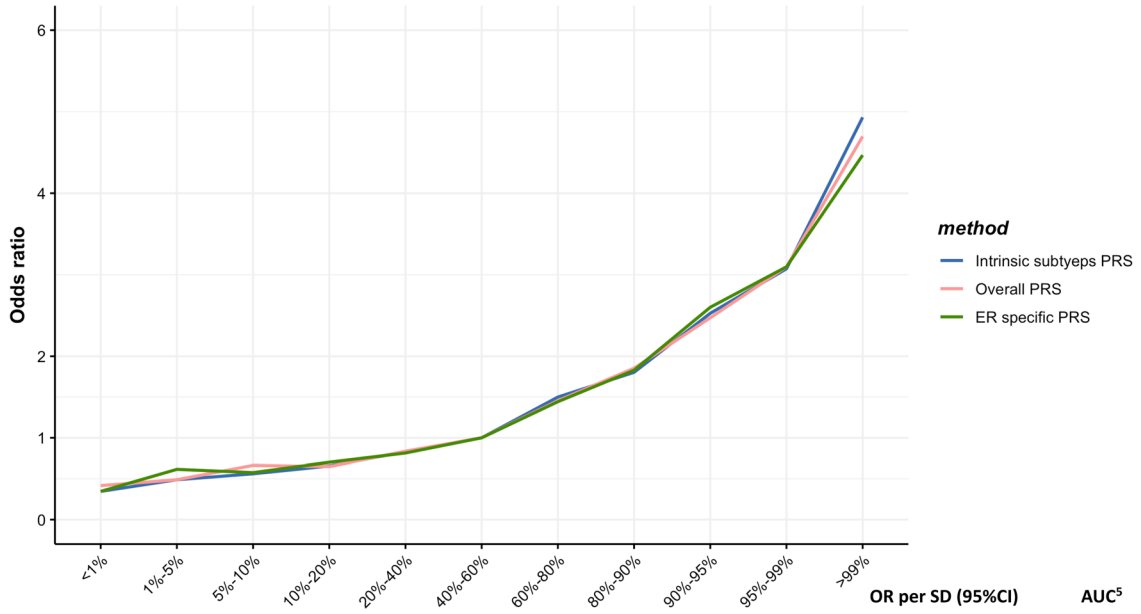
[2] Overall breast cancer PRS with 313 SNPs previously reported[22]

[3] ER-specific PRS with 313 SNPs previously reported[22]

[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-like[4] risk in the test dataset. Odds ratios for different quantiles of the PRS against the middle quantile (40%–60%) 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).



												OR per SD (95%CI)	AUC[5]
Intrinsic subtypes PRS ORs[1]	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 ORs[2]	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 ORs[3]	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**)

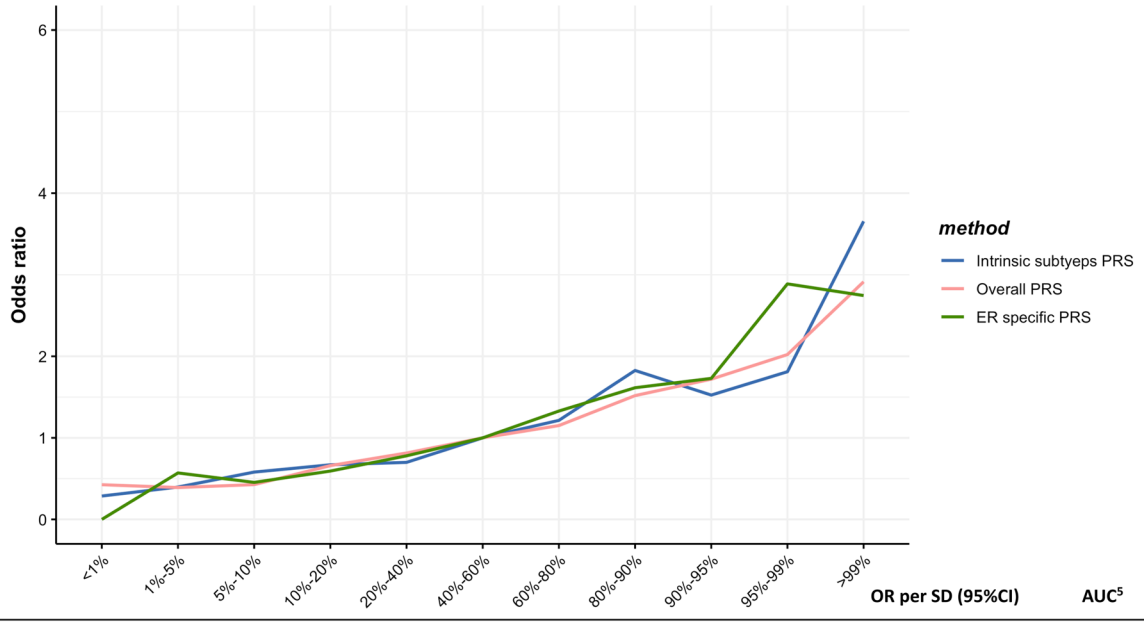
[2] Overall breast cancer PRS with 313 SNPs previously reported[22]

[3] ER-specific PRS with 313 SNPs previously reported[22]

[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-like[4] risk in the test dataset. Odds ratios for different quantiles of the PRS against the middle quantile (40%–60%) of the PRS. The odds ratios were estimated using the test dataset like (n = 1,682 Luminal B-like cases, n = 20,815 controls).



												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**)

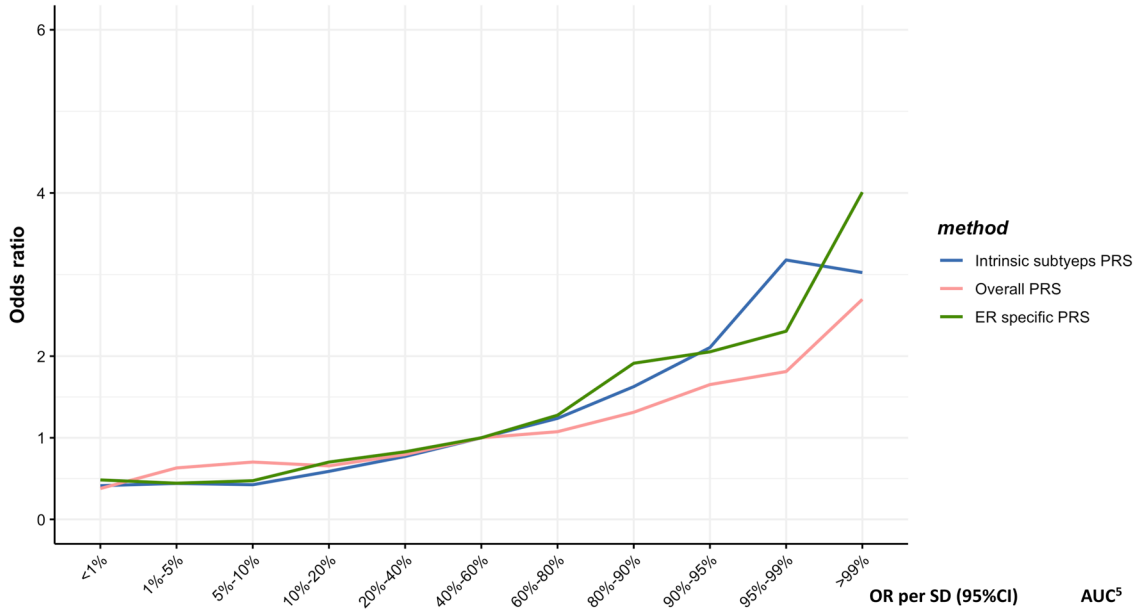
[2] Overall breast cancer PRS with 313 SNPs previously reported[22]

[3] ER-specific PRS with 313 SNPs previously reported[22]

[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-like[4] risk in the test dataset. Odds ratios for different quantiles of the PRS against the middle quantile (40%–60%) of the PRS. The odds ratios were estimated using the test dataset like (n = 718 HER2-enriched-like, n = 20,815 controls).



												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 reported[22]

[3] ER-specific PRS with 313 SNPs previously reported[22]

[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-negative[4] risk in the test dataset. Odds ratios for different quantiles of the PRS against the middle quantile (40%–60%) of the PRS. The odds ratios were estimated using the test dataset like (n = 2,006 triple-negative cases, n = 20,815 controls).

nature research

Corresponding author(s): Nilanjan Chatterjee

Last updated by author(s): Oct 1, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used SAS and R to access and manage the data.
Data analysis	We used these softwares to finish all the analysis: 1) R version 3.6.0 2) LDSC version 1.0.1 3) METAL 2011/03/25 version 4) MatrixEQTL v2.2. The analysis code could be found in GitHub repository (https://github.com/andrewhaoyu/breast_cancer_data_analysis). The analysis completed in the submitted paper could be reproduced through the code in the GitHub repository screened on Sep 13, 2019 with git commit id as 872fc6e.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

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

through the BCAC DACC (http://bcac.ccge.medschl.cam.ac.uk/).
A subset of the CIMBA data that support the findings of this study is publically available via dbGaP (accession number phs001321.v1.p1). The complete dataset will not be made publically available due to restraints imposed by the ethics committees of individual studies; requests for data can be made to the corresponding author or the Data Access Coordination Committee (DACC) of CIMBA (http://cimba.ccge.medschl.cam.ac.uk). CIMBA DACC approval is required to access data from studies 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. The CIMBA complete summary results are available through: http://cimba.ccge.medschl.cam.ac.uk/oncoarray-complete-summary-results/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We analyzed all the studies available for breast cancer genome-wide association studies (GWAS) and pathology information from the largest consortium.
Data exclusions	We excluded OncoArray data from Norway (the Norwegian Breast Cancer Study) because there are no controls available from Norway with OncoArray data. We also excluded one Study (Leuven Multidisciplinary Breast Centre) contributing to the iCOGs dataset due to inflation of the test statistics that was not corrected by adjustment for the first ten PCs.
Replication	We analyzed the all studies available instead of dividing the dataset into discovery and replication sets because of statistical power. We checked consistency of the results across studies/countries for replicability.
Randomization	This was an observational genetic association study, hence randomization was not relevant. The analyses were adjusted for country and ancestry informative principal components, as described in the methods
Blinding	The laboratories conducting the genotyping did not have access to the phenotypic data (i.e. were blinded). Moreover, genotype calling was automated. The phenotype and genotype data were only combined during the statistical analysis

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The analysis of data from Breast Cancer Association Consortium (BCAC) included women aged 16 years or older 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. 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 analysis. All participating studies were approved by their appropriate ethnics or institutional review board and all participants provided informed consent. The total sample size from BCAC including iCOGs, OncoArray, and other GWAS data, comprised 133,384 cases and 113,789 controls. Participants included from Consortium of Investigators of Modifiers of BRCA1/2 (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 member of the same family were enrolled. OncoArray genotype data was avaiable from 58 studies from 24 countries. Following quality control and removal of participants that overlapped with BCAC OncoArray study, data were available on 15,566 BRCA1 mutation carriers, of whom 7,784 were affected with breast cancer. 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.
Recruitment	Participants were recruited from epidemiology studies and selection factors are very unlikely to be related to genotypes.
Ethics oversight	All the studies were approved by local IRBs.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.