ARTICLES
https://doi.org/10.1038/s41588-019-0537-1

nature genetics

Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes

Genome-wide association studies have identified breast cancer risk variants in over 150 genomic regions, but the mechanisms underlying risk remain largely unknown. These regions were explored by combining association analysis with in silico genomic feature annotations. We defined 205 independent risk-associated signals with the set of credible causal variants in each one. In parallel, we used a Bayesian approach (PAINTOR) that combines genetic association, linkage disequilibrium and enriched genomic features to determine variants with high posterior probabilities of being causal. Potentially causal variants were significantly over-represented in active gene regulatory regions and transcription factor binding sites. We applied our INQUSIT pipeline for prioritizing genes as targets of those potentially causal variants, using gene expression (expression quantitative trait loci), chromatin interaction and functional annotations. Known cancer drivers, transcription factors and genes in the developmental, apoptosis, immune system and DNA integrity checkpoint gene ontology pathways were over-represented among the highest-confidence target genes.

Genome-wide association studies (GWASs) have identified genetic variants associated with breast cancer risk in more than 150 genomic regions[1,2]. However, the variants and genes driving these associations are mostly unknown, with fewer than 20 regions studied in detail[3–20]. Here, we aimed to fine-map all known breast cancer susceptibility regions using dense genotype data on >217,000 subjects participating in the Breast Cancer Association Consortium (BCAC) and the Consortium of Investigators of Modifiers of *BRCA1/2* (CIMBA). All samples were genotyped using the OncoArray[1,2,21] or the iCOGS chip[22,23]. Stepwise multinomial logistic regression was used to identify independent association signals in each region and to define credible causal variants (CCVs) within each signal. We found genomic features significantly overlapping the CCVs. We then used a Bayesian approach, integrating genomic features and genetic associations, to refine the set of likely causal variants and calculate their posterior probabilities. Finally, we integrated genetic and in silico epigenetic expression and chromatin conformation data to infer the likely target genes of each signal.

Results

Most breast cancer genomic regions contain multiple independent risk-associated signals. We included 109,900 cases of breast cancer and 88,937 controls, all of European ancestry, from 75 studies in the BCAC. Genotypes (directly observed or imputed) were available for 639,118 single nucleotide polymorphisms (SNPs), deletions/insertions and copy number variants (CNVs) with a minor allele frequency (MAF) $\geq 0.1\%$ within 152 previously defined, risk-associated regions (Supplementary Table 1 and Fig. 1). Multivariate logistic regression confirmed associations for 150 out of 152 regions at a significance threshold of $P < 10^{-4}$ (Supplementary Table 2a). To determine the number of independent risk signals within each region, we applied stepwise multinomial logistic regression, deriving the association of each variant, conditional on the more significant ones, in order of statistical significance. Finally, we defined CCVs in each signal as variants with conditional *P* values within two orders of magnitude of the index variant[24]. We classified the evidence for each independent signal, and its CCVs, as either strong (conditional $P < 10^{-6}$) or moderate ($10^{-6} <$ conditional $P < 10^{-4}$).

From the 150 genomic regions, we identified 352 independent risk signals containing 13,367 CCVs, 7,394 of which were within the 196 strong-evidence signals across 129 regions (Fig. 2a,b). The number of signals per region ranged from 1 to 11, with 79 (53%) containing multiple signals. We noted a wide range of CCVs per signal, but in 42 signals there was only a single CCV: for these signals, the simplest hypothesis is that the CCV was causal (Fig. 2c,d and Table 1). Furthermore, within signals with few CCVs (<10), the mean scaled combined annotation-dependent depletion score was higher than in signals with more CCVs (13.1 versus 6.7 for CCVs in exons; $P_{t\text{-test}} = 2.7 \times 10^{-4}$), suggesting that these are more likely to be functional.

The majority of breast tumors express the estrogen receptor (ER positive), but ~20% do not (ER negative); these two tumor types have distinct biological and clinical characteristics[25]. Using a case-only analysis for the 196 strong-evidence signals, we found 66 signals (34%; containing 1,238 CCVs) where the lead variant conferred a greater relative risk of developing ER-positive tumors (false discovery rate (FDR) = 5%), and 29 (15%; 646 CCVs) where the lead variant conferred a greater risk of ER-negative cancer tumors (FDR = 5%) (Supplementary Table 2b and Fig. 2e). The remaining 101 signals (51%; 5,510 CCVs) showed no difference by ER status (referred to as ER neutral).

Patients with *BRCA1* mutations are more likely to develop ER-negative tumors[26]. Hence, to increase our power to identify ER-negative signals, we performed a fixed-effects meta-analysis, combining association results from *BRCA1* mutation carriers in CIMBA with the BCAC ER-negative association results. This meta-analysis identified ten additional signals (seven ER-negative and three ER-neutral), making 206 strong-evidence signals (17% ER negative) containing 7,652 CCVs in total (Fig. 2f). More than one-quarter of the CCVs (2,277) were accounted for by one signal, resulting from strong linkage disequilibrium with a CNV. The remaining analyses focused on the other 205 strong signals across 128 regions (Supplementary Table 2c).

The proportion of the familial relative risk (FRR) of breast cancer explained by all 206 strong signals was 20.6%, compared with 17.6% when only the lead SNP for each region was considered. The proportion of the FRR explained increased by a further 3% (to 23.6%) when all 352 signals were considered (Supplementary Table 2d).

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

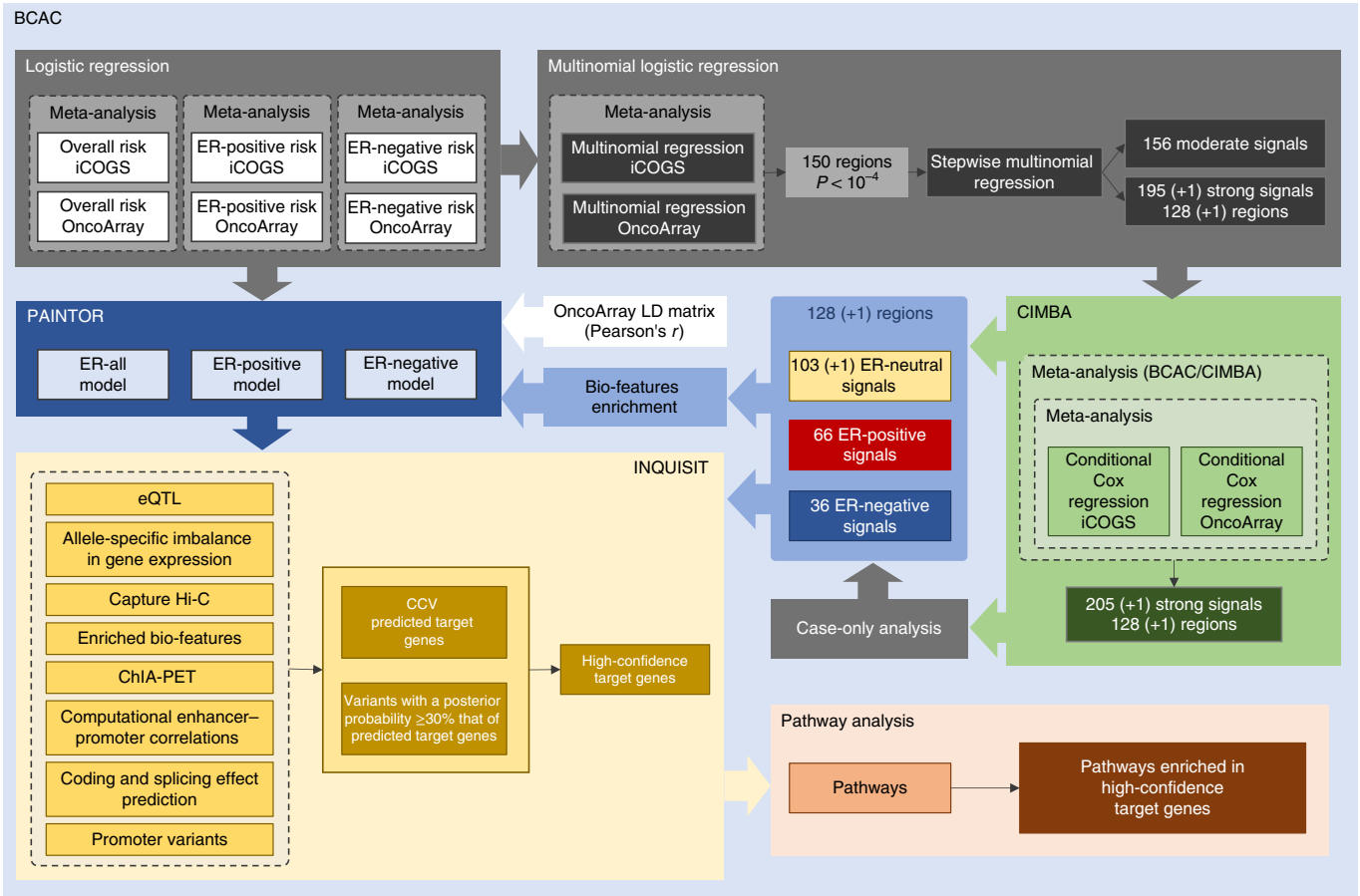

Fig. 1 | Flowchart summarizing the study design. Logistic regression summary statistics were used to select the final set of variants to run stepwise multinomial regression. These results were meta-analyzed with CIMBA to provide the final set of strong independent signals and their CCVs. Through case-only analysis, we identified significant differences in effect sizes between ER-positive and ER-negative breast cancer and used this to classify the phenotype for each independent signal. With these strong CCVs, we ran the bio-features enrichment analysis, which identified the features to be included in the PAINTOR models, together with the OncoArray logistic regression summary statistics and the OncoArray linkage disequilibrium. Both multinomial regression CCVs and PAINTOR high-posterior-probability (PP) variants were analyzed with INQUISIT to determine high-confidence target genes. Finally, we used the set of high-confidence target genes to identify enriched pathways. iCOGS and OncoArray Cox regression was conditional on the index variants from BCAC strong signals.

CCVs are over-represented in active gene regulatory regions and transcription factor binding sites (TFBSs). We constructed a database of mapped genomic features in seven primary cells derived from normal breast and 19 breast cell lines using publicly available data, resulting in 811 annotation tracks in total. These ranged from general features (such as whether a variant was in an exon or in open chromatin) to more specific features (such as cell-specific transcription factor binding or histone marks (determined through chromatin immunoprecipitation followed by sequencing (ChIP-Seq) experiments) in breast-derived cells or cell lines). Using logistic regression, we examined the overlap of these genomic features with the positions of 5,117 CCVs in the 195 strong-evidence BCAC signals versus the positions of 622,903 variants excluded as credible candidates in the same regions (Supplementary Fig. 1a and Supplementary Table 3). We found significant enrichment of CCVs (FDR = 5%) in four genomic features (open chromatin, actively transcribed genes, gene regulatory regions and binding sites), as described below.

Open chromatin. As shown in Fig. 3a, DNase I hypersensitive sites sequencing and formaldehyde-assisted isolation of regulatory elements sequencing showed significant enrichment of CCVs in open chromatin in ER-positive breast cancer cell lines and normal breast. Conversely, we found depletion of CCVs within heterochromatin (determined by the H3K9me3 mark in normal breast, and by chromatin state in ER-positive cells[27]).

Actively transcribed genes. Significant enrichment of CCVs was also found in actively transcribed genes in normal breast and ER-positive cell lines (as defined by H3K36me3 or H3K79me2 histone marks; Fig. 3a). Enrichment was larger for ER-neutral CCVs than for those affecting either ER-positive or ER-negative tumors.

Gene regulatory regions. CCVs overlapped distal gene regulatory elements in ER-positive breast cancer cells lines (defined by H3K4me1 or H3K27ac marks; Fig. 3b). This was confirmed using the Encyclopedia of DNA Elements (ENCODE) definition of active enhancers in MCF-7 cells (enhancer-like regions defined by combining DNase and H3K27ac marks), as well as the definition of refs. [27,28] (Supplementary Table 3). Under these more stringent definitions, enrichment among ER-positive CCVs was significantly larger than ER-negative or ER-neutral CCVs. Data from ref. [27] showed that 73% of active enhancer regions overlapped by ER-positive CCVs in ER-positive cells (MCF-7) are inactive in the normal human mammary epithelial (HMEC) breast cell line; thus, these enhancers appear to be MCF-7 specific.

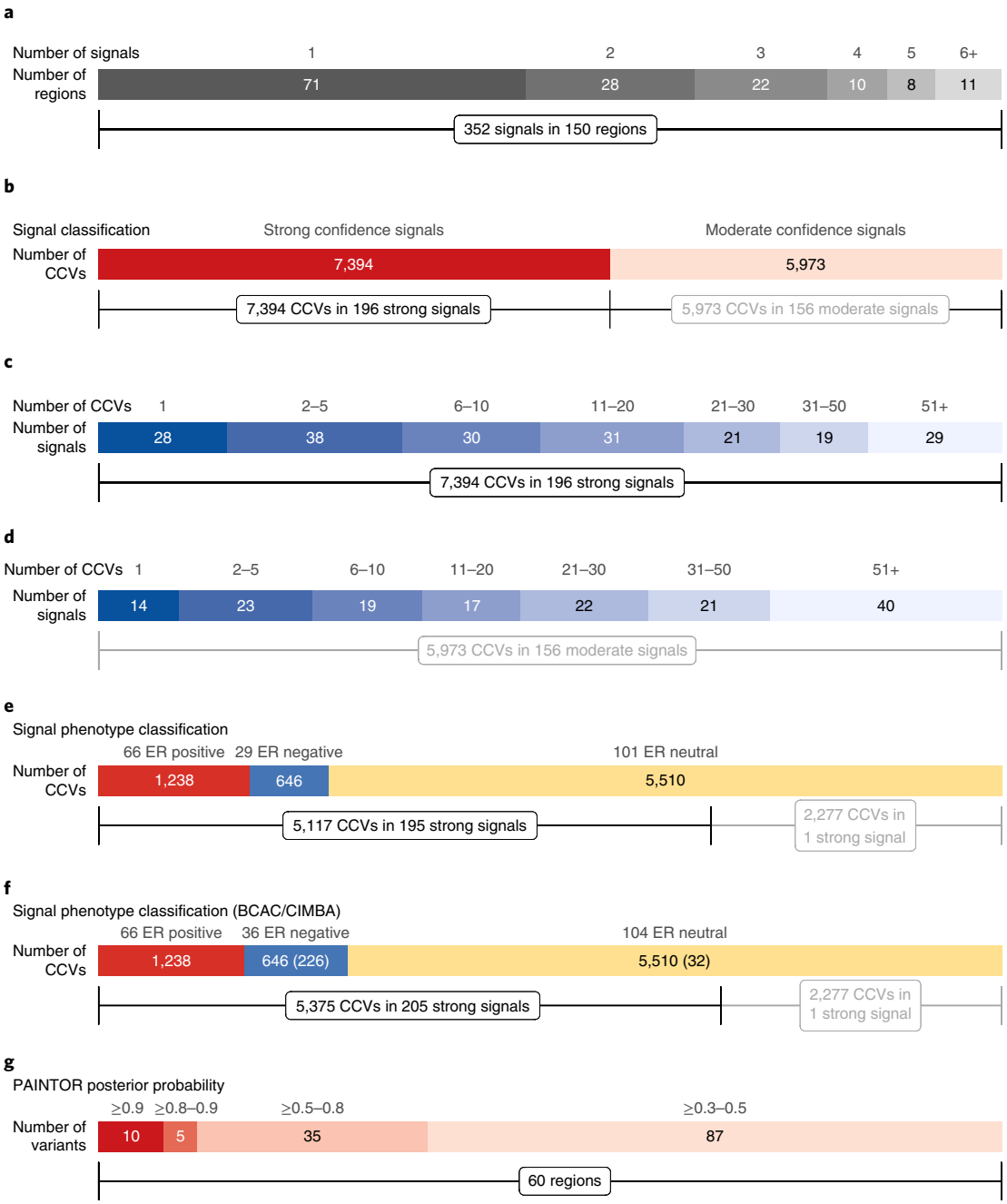

Fig. 2 | Determining independent risk signals and CCVs. a, Number of independent signals per region, identified through multinomial stepwise logistic regression. **b**, Signal classification as strong- or moderate-confidence signals. **c**,**d**, Number of CCVs per signal in strong- (**c**) and moderate-confidence signals (**d**), identified through multinomial stepwise logistic regression. **e**, Subtype classification of strong signals into ER positive, ER negative and signals equally associated with both phenotypes (ER neutral) from the BCAC analysis. **f**, Subtype classification from the meta-analysis of BCAC and CIMBA. Numbers in brackets show the numbers of CCVs from the meta-analysis of BCAC and CIMBA. **g**, Number of variants at different posterior probability thresholds. In total, 15 variants reached a posterior probability of ≥80% by at least one of the three models (ER all, ER positive or ER negative).

We also detected significant enrichment of CCVs in active promoters in ER-positive cells (defined by H3K4me3 marks in T-47D), although the evidence for this effect was weaker than for distal regulatory elements (defined by H3K27ac marks in MCF-7; Fig. 3b). Only ER-positive CCVs were significantly enriched in T-47D active promoters. Conversely, CCVs were depleted among repressed gene regulatory elements (defined by H3K27me3 marks) in normal breast (Fig. 3b). As a control, we performed similar analyses with autoimmune disease CCVs[29] (Methods) and relevant B and T cells (Fig. 3b–e). The strongest evidence of enrichment of breast cancer CCVs was found at regulatory regions active in ER-positive cells (Fig. 3b), whereas enrichment of autoimmune CCVs was in regulatory regions active in B and T cells (Fig. 3e). We also compared the enrichment of our CCVs in enhancer-like and promoter-like regions (defined by ENCODE; Supplementary Fig. 1b). The strongest evidence of enrichment of ER-positive CCVs in enhancer-like regions was found in MCF-7 cells—the only ER-positive cell line in ENCODE (Supplementary Fig. 1b). These results highlight both the tissue specificity and disease specificity of these histone-marked gene regulatory regions.

Table 1 | Signals with single CCVs and variants with a posterior probability of >80%

Fine-mapping region[a]	Variant[b]	Ref/alt[c]	EAF[d]	PP[e]	Model[f]	Signal[g]	CCVs[h]	ER negative		ER positive		*P* value[i]	FP[j]	Predicted target gene(s)[k]	Confidence[l]
								OR[i]	95% CI	OR[i]	95% CI				
Chr1:120723447–121780613	rs11249433	A/G	0.42	0.57	ER all	Signal 1	1	1.02	0.99–1.04	1.13	1.11–1.15	8.11×10^{-60}	NA	NA	
Chr1:200937832–201937832	rs35383942	C/T	0.06	0.96	ER all	Signal 1	2	1.10	1.05–1.16	1.09	1.06–1.13	1.14×10^{-7}	D	*TNNI1*	Level 1
Chr2:201681247–202681247	rs3769821	C/T	0.66	0.40	ER all	Signal 1	1	0.94	0.92–0.97	0.95	0.93–0.96	1.46×10^{-12}	D	*ALS2CR12*	Level 1
Chr2:217405832–218796508	rs4442975[m]	G/T	0.48	0.84	ER all	Signal 1	1	0.94	0.92–0.97	0.86	0.85–0.87	2.50×10^{-90}	D	*IGFBP5*[n]	Level 2
Chr4:105569013–106856761	esv3601665	–/Alu	0.07	0.95	ER pos			1.01	0.95–1.08	1.10	1.06–1.14	3.27×10^{-6}	D	*ARHGEF38* and *AC004066.3*	Level 1
Chr5:779790–1797488	rs10069690	C/T	0.27	0.58	ER neg	Signal 1	1	1.18	1.15–1.21	1.03	1.01–1.05	1.20×10^{-34}	D	*SLC6A18* and*TERT*[n]	Level 2
Chr5:44013304–45206498	rs10941679	A/G	0.26	0.00	ER pos	Signal 1	1	1.04	1.02–1.07	1.17	1.15–1.19	1.50×10^{-77}	D	*MRPS30*	Level 2
	rs5867671	A/–	0.77	0.01	ER pos	Signal 2	1	0.91	0.89–0.94	0.99	0.97–1.01	2.25×10^{-9}	NA	NA	
Chr5:44013304–45206498	rs190443933	T/C	0.01	0.00	ER all	Signal 4	1	1.30	1.14–1.48	1.26	1.16–1.37	2.32×10^{-8}	NA	NA	
Chr5:55531884–56587883	rs984113	G/C	0.61	0.81	ER pos	Signal 2	1	0.96	0.93–0.98	0.96	0.94–0.97	3.51×10^{-8}	D	*MAP3K1*[n]	Level 2
	rs889310	C/T	0.56	0.84	ER pos	(Signal 6)	15	1.03	1.00–1.05	1.05	1.03–1.06	1.75×10^{-7}	D	*MAP3K1*[n]	Level 1
Chr6:15899557–16899557	rs3819405	C/T	0.32	0.96	ER all	Signal 1	1	0.97	0.95–1.00	0.95	0.94–0.97	1.14×10^{-7}	D	*ATXN1*, *RP1-151F17.1* and *RP1-151F17.2*	Level 2
Chr6:151418856–152937016	rs12173562	C/T	0.08	0.10	ER neg	Signal 1	1	1.30	1.25–1.36	1.14	1.11–1.18	3.98×10^{-40}	D	*ESR1*[n]	Level 1
	rs34133739	–/C	0.53	0.25	ER all	Signal 2	1	1.11	1.09–1.14	1.05	1.04–1.07	2.36×10^{-22}	D	*ESR1*[n]	Level 1
	rs851984	G/A	0.40	0.73	ER all	Signal 3	1	1.07	1.04–1.09	1.05	1.04–1.07	3.69×10^{-13}	D	*ESR1*[n]	Level 1
Chr7:130167121–131167121	rs68056147	G/A	0.30	0.84	ER all			1.04	1.01–1.07	1.05	1.03–1.06	3.07×10^{-7}	D	*MKLN1*	Level 2
Chr8:127424659–130041931	rs35961416	–/A	0.41	0.68	ER all	Signal 3	1	0.97	0.94–0.99	0.95	0.93–0.96	9.97×10^{-11}	D	*MYC*[n]	Level 1
Chr9:21247803–22624477	rs539723051	AAAA/–	0.33	0.43	ER all	Signal 1	1	1.08	1.05–1.11	1.06	1.04–1.08	1.81×10^{-15}	NA	NA	
Chr9:109803808–111395353	rs10816625	A/G	0.07	0.95	ER pos	Signal 3	1	1.06	1.01–1.11	1.13	1.10–1.16	3.62×10^{-15}	D	*KLF4*[n]	Level 2
	rs13294895	C/T	0.18	0.93	ER pos	Signal 4	1	1.01	0.98–1.05	1.09	1.07–1.11	4.00×10^{-17}	D	*KLF4*[n]	Level 1
Chr9:109803808–111395353	rs60037937	AA/–	0.22	0.68	ER pos	Signal 2	1	1.02	0.99–1.06	1.11	1.09–1.13	3.17×10^{-26}	D	*KLF4*[n] and *RAD23B*	Level 2
Chr10:63758684–65063702	rs10995201	A/G	0.15	0.31	ER all	Signal 1	1	0.91	0.88–0.94	0.87	0.85–0.89	1.40×10^{-37}	NA	NA	
Chr10:122593901–123849324	rs35054928	C/–	0.56	0.60	ER all	Signal 1	1	0.96	0.94–0.98	0.74	0.73–0.76	6.55×10^{-342}	D	*FGFR2*[n]	Level 1
	rs45631563[m]	A/T	0.04	0.93	ER pos	Signal 3	1	0.97	0.92–1.03	0.76	0.73–0.79	4.84×10^{-44}	C	*FGFR2*[n]	Level 2
	rs7899765	T/C	0.06	0.02	ER all	Signal 5	1	1.01	0.97–1.06	0.87	0.84–0.90	2.21×10^{-18}	D	*FGFR2*[n]	Level 1
Chr11:68831418–69879161	rs78540526	C/T	0.09	0.91	ER pos	Signal 1	1	1.01	0.97–1.06	1.40	1.36–1.44	2.77×10^{-145}	D	*CCND1*[n] and *MYEOV*	Level 1
Chr12:27639846–29034415	rs7297051	C/T	0.23	0.23	ER all	Signal 1	1	0.87	0.85–0.90	0.89	0.88–0.91	3.12×10^{-43}	D	*CCDC91*[n], *PTHLH*[n] and *RP11-967K21.1*	Level 2
Chr12:115336522–116336522	rs35422	G/A	0.57	0.58	ER pos	Signal 2	1	0.98	0.96–1.01	1.05	1.03–1.07	4.85×10^{-10}	D	*TBX3*	Level 1
Chr14:91341069–92368623	rs7153397	C/T	0.70	0.81	ER pos	Signal 1	3	1.01	0.99–1.04	1.06	1.04–1.08	3.25×10^{-11}	D and C	*CCDC88C*, *CTD-2547L24.4*, *C14orf159*, *GPR68*, *RPS6KA5*, *RP11-73M18.7* and *RP11-895M11.3*	Level 2
Chr16:52038825–53038825	rs4784227	C/T	0.27	0.95	ER pos	Signal 1	1	1.15	1.12–1.18	1.26	1.24–1.28	4.63×10^{-160}	D	*TOX3*[n]	Level 1
Chr18:23832476–25075396	rs180952292	T/C	0.01	0.01	ER neg	Signal 4	1	1.24	1.12–1.37	0.98	0.92–1.05	2.07×10^{-5}	NA	NA	
Chr18:41899590–42899590	rs9952980	T/C	0.34	0.95	ER all	Signal 2	3	0.97	0.94–0.99	0.95	0.93–0.96	7.43×10^{-12}	D	*SLC14A2*	Level 2
Chr20:5448227–6448227	rs16991615	G/A	0.07	0.97	ER all	Signal 1	1	1.09	1.04–1.15	1.07	1.04–1.11	7.89×10^{-7}	D and C	*GPCPD1* and *MCM8*	Level 2
Chr22:45783297–46783297	rs184070480	C/T	0.01	0	ER all	Signal 2	1	1.40	1.20–1.64	1.01	0.91–1.12	5.02×10^{-5}	D	*ATXN10* and *WNT7B*	Level 2

[a]Labels show the chromosome (chr) number and range of base pairs (from genome assembly GRCh37/hg19). [b]Current reference identification. [c]Reference (ref) versus alternative (alt) allele. [d]Effect (alt) allele frequency (EAF) in OncoArray. [e]Largest posterior probability (PP) in all of the evaluated models. [f]Model with which the variant reaches the largest PP (ER all, ER positive (ER pos) or ER negative (ER neg). [g]Signal where the variant is included. Numbers in brackets represent moderate-confidence signals. [h]Number of CCVs in the signal. [i]Multinomial logistic regression summary statistics (odds ratios (ORs), confidence intervals (CIs) and X^2 single-variant analysis *P* values), estimated using 67,136 ER-positive and 17,506 ER-negative cases, together with 88,937 controls. [j]Function prediction (FP) (distal regulation (D), coding (C) or not available (NA)). [k]Predicted target genes with the largest confidence level for each variant. [l]INQUISIT level of confidence. [m]Two variants reach a PP of >0.8 in both the ER-all and ER-pos models: rs4442975 (ER-pos PP = 0.83; ER-all PP = 0.84); and rs45631563 (ER-pos PP = 0.93; ER-all PP = 0.92). [n]Target genes with functional follow-up.

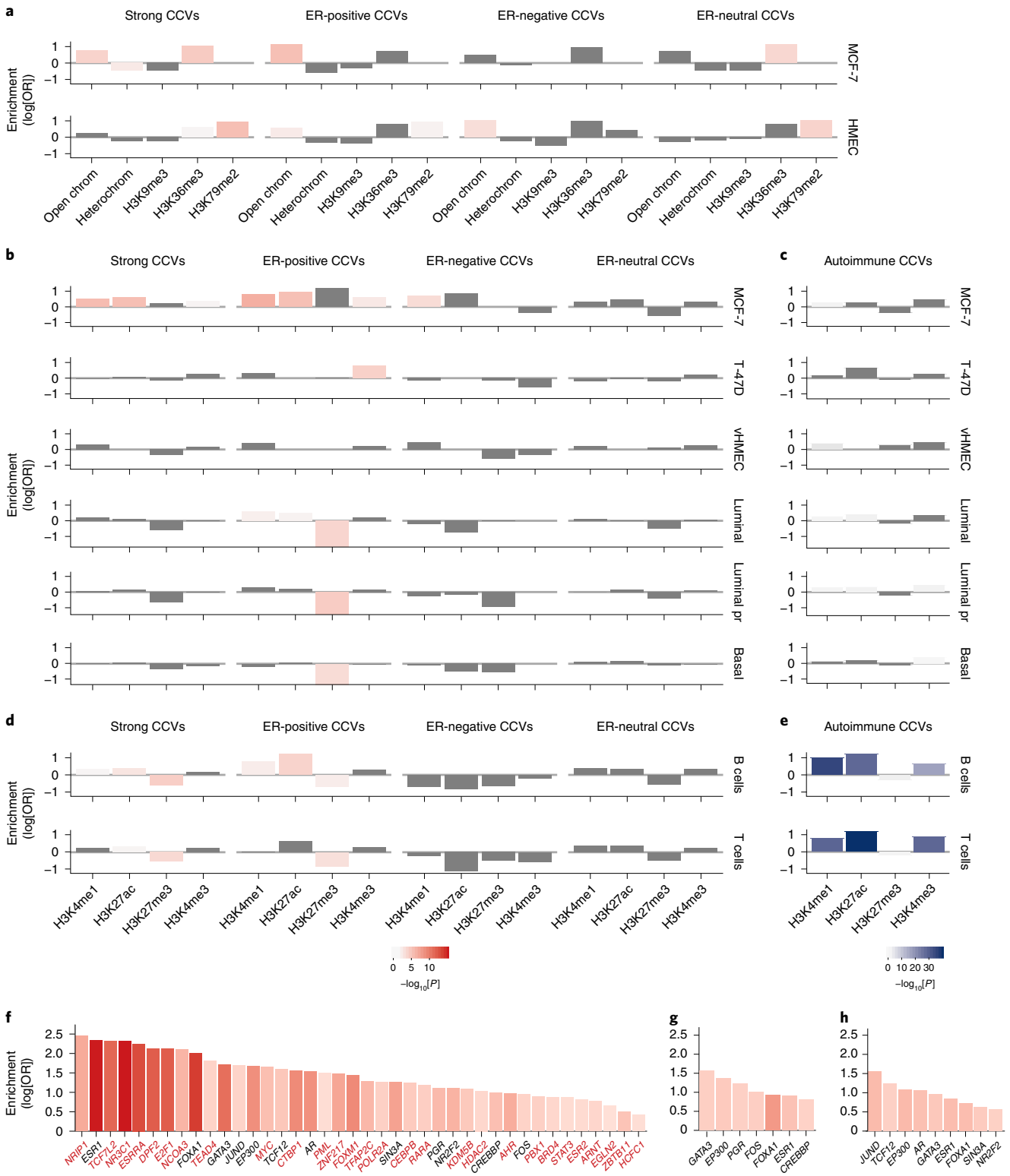

Fig. 3 | Overlap of CCVs with gene regulatory regions, gene bodies and TFBSs. a, Breast cancer CCVs overlap with chromatin states and broad breast cell epigenetic marks. HMEC, human mammary epithelial cells. Open chrom, open chromatin; heterochrom, heterochromatin **b**,**c**, Breast cancer CCVs (**b**) and autoimmune CCVs (**c**) overlap with breast cell epigenetic marks. vHMEC, variant HMEC. Luminal pr, luminal progenitor. **d**,**e**, Breast cancer CCVs (**d**) autoimmune CCVs (**e**) and overlap with autoimmune-related epigenetic marks. In **a**, **b** and **d**, the column 'strong CCVs' represents analysis with all CCVs at strong signals, while the remaining columns represent analysis of CCVs at strong signals stratified by phenotype. Logistic regression robust variance estimation for clustered observations was used, and Wald test X^2 *P* values were estimated using 67,136 ER-positive and 17,506 ER-negative cases, together with 88,937 controls. Non-significant *P* values are shown in dark gray. Significance was defined as an FDR of 5%, which corresponds to the following *P* value thresholds: $P = 1.66 \times 10^{-2}$ (strong signals); $P = 2.42 \times 10^{-2}$ (ER positive); $P = 3.02 \times 10^{-3}$ (ER negative); and $P = 1.76 \times 10^{-3}$ (ER neutral). **f**–**h**, Significant ER-positive (**f**), ER-negative (**g**) and ER-neutral CCVs (**h**) overlap with TFBSs. TFBSs found significant for ER-positive CCVs are highlighted in red (*x* axis labels).

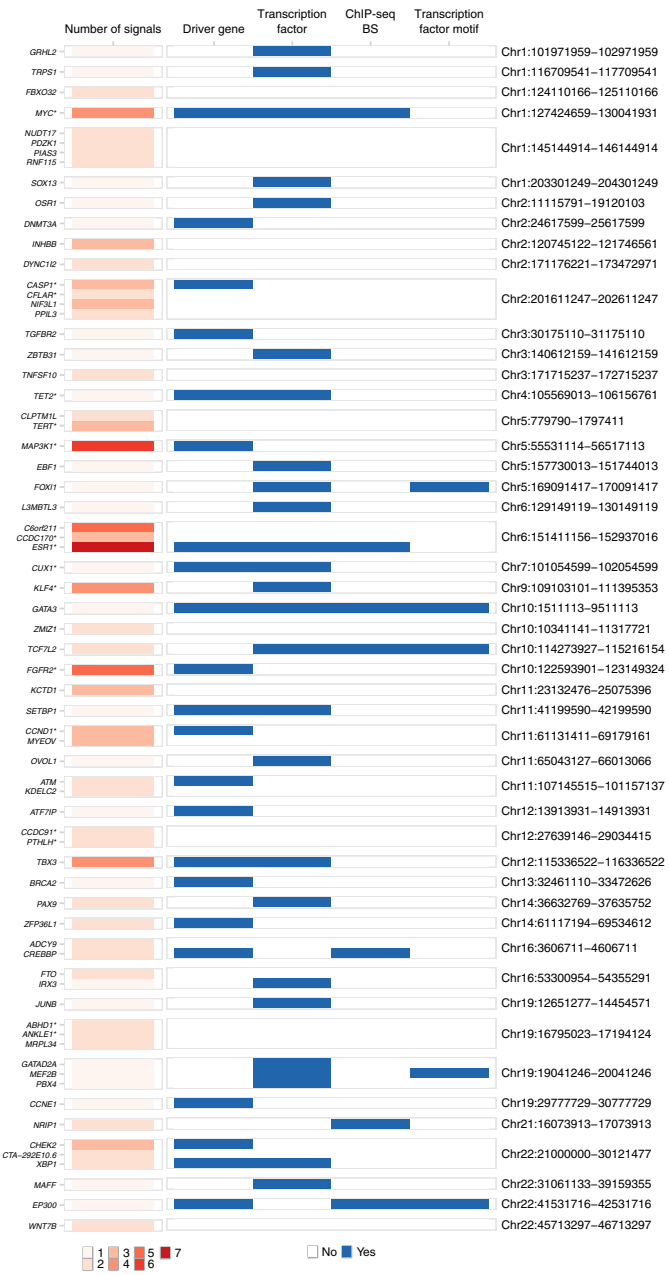

Fig. 4 | Predicted target genes are enriched in known breast cancer driver genes and transcription factors. Target genes ($n=69$) that fulfill at least one of the following criteria: (1) is targeted by more than one independent signal; (2) is a known driver gene; (3) is a known transcription factor gene; (4) its binding sites (as determined by ChIP-Seq (ChIP-Seq BS)) are significantly overlapped by CCVs; or (5) its consensus (transcription factor) motif is significantly overlapped by CCVs. Asterisks indicate genes with published functional follow-up.

Binding sites. We observed significant enrichment of CCVs in the binding sites for 40 TFBSs determined by ChIP-Seq (Fig. 3f–h). The majority of the experiments were performed in ER-positive cell lines (90 TFBSs; 20 with data in ER-negative cell lines, 76 with data in ER-positive cell lines and 16 with data in normal breast). These TFBSs overlap each other and histone marks of active regulatory regions (Supplementary Fig. 2). Enrichment in five TFBSs (*ESR1*, *FOXA1*, *GATA3*, *TCF7L2* and *E2F1*) has been reported previously[2,30]. All 40 TFBSs were significantly enriched in ER-positive CCVs (Fig. 3f), seven were also enriched in ER-negative CCVs and nine were enriched in ER-neutral CCVs (Fig. 3g–h). *ESR1*, *FOXA1*, *GATA3* and *EP300* TFBSs were enriched in all CCV ER subtypes. However, the enrichment for *ESR1*, *FOXA1* and *GATA3* was stronger for ER-positive CCVs than for ER-negative or ER-neutral CCVs.

CCVs significantly overlap consensus transcription factor binding motifs. We investigated whether CCVs were also enriched within consensus transcription factor binding motifs by conducting a motif search within active regulatory regions (ER-positive CCVs at H3K4me1 marks in MCF-7). We identified 30 motifs from eight transcription factor families, with enrichment in ER-positive CCVs (FDR = 10%; Supplementary Table 4a) and a further five motifs depleted among ER-positive CCVs. To assess whether the motifs appeared more frequently than by chance at active regulatory regions overlapped by our ER-positive CCVs, we compared motif presence in a set of randomized control sequences (Methods). Thirteen of 30 motifs were more frequent at active regulatory regions with ER-positive CCV enrichment; these included seven homeodomain motifs and two forkhead factors (Supplementary Table 4b).

When we looked at the change in predicted binding affinity, 57 ER-positive signals (86%) included at least one CCV predicted to modify the binding affinity of the enriched TFBSs (at least twofold; Supplementary Table 4c). Forty-eight ER-positive signals (73%) had at least one CCV predicted to modify the binding affinity greater than tenfold. This analysis validates previous reports of breast cancer causal variants that alter the DNA binding affinity for FOXA1 (refs. [3,30]).

Bayesian fine-mapping incorporating functional annotations and linkage disequilibrium. As an alternative statistical approach for inferring likely causal variants, we applied PAINTOR[31] to the same 128 regions (Fig. 1). In brief, PAINTOR integrates genetic association results, linkage disequilibrium structure and enriched genomic features in an empirical Bayes framework and derives the posterior probability of each variant being causal, conditional on available data. To eliminate artefacts due to differences in genotyping and imputation across platforms, we restricted PAINTOR analyses to cases and controls typed using OncoArray (61% of the total). We identified seven variants with a high posterior probability (HPP ≥ 80%) of being causal for overall breast cancer, and ten for the ER-positive subtype (Table 1); two of these had a HPP > 80% for both ER-positive and overall breast cancer. These 15 HPP variants (HPPVs; ≥80%) were distributed across 13 regions. We also identified an additional 35 variants in 25 regions with HPP (≥50 and <80%) for ER-positive, ER-negative or overall breast cancer (Fig. 2g).

Consistent with the CCV analysis, we found evidence that most regions contained multiple HPPVs; the sum of posterior probabilities across all variants in a region (an estimate of the number of distinct causal variants in the region) was >2.0 for 84 out of 86 regions analyzed for overall breast cancer, with a maximum of 16.1 and a mean of 6.4. For ER-positive cancer, 46 out of 47 regions had total posterior probability of >2.0 (maximum: 18.3; mean: 6.5). For ER-negative cancer, 17 out of 23 regions had a total posterior probability of >2.0 (maximum: 9.1; mean: 3.2).

Although for many regions we were not able to identify HPPVs, we were able to reduce the proportion of variants needed to account for 80% of the total posterior probability in a region to <5% for 65 regions for overall breast cancer, 43 regions for ER-positive breast cancer and 18 regions for ER-negative breast cancer (Supplementary Fig. 3a–c). PAINTOR analyses were also able to reduce the set of likely causal variants in many cases. After summation of the posterior probabilities for CCVs in each of the overall breast cancer signals, 39 out of 100 strong-evidence signals had a total posterior probability of >1.0. The number of CCVs in these signals ranged from 1 to 375 (median: 24), but the number of variants needed to capture 95% of the total posterior probability in each signal ranged

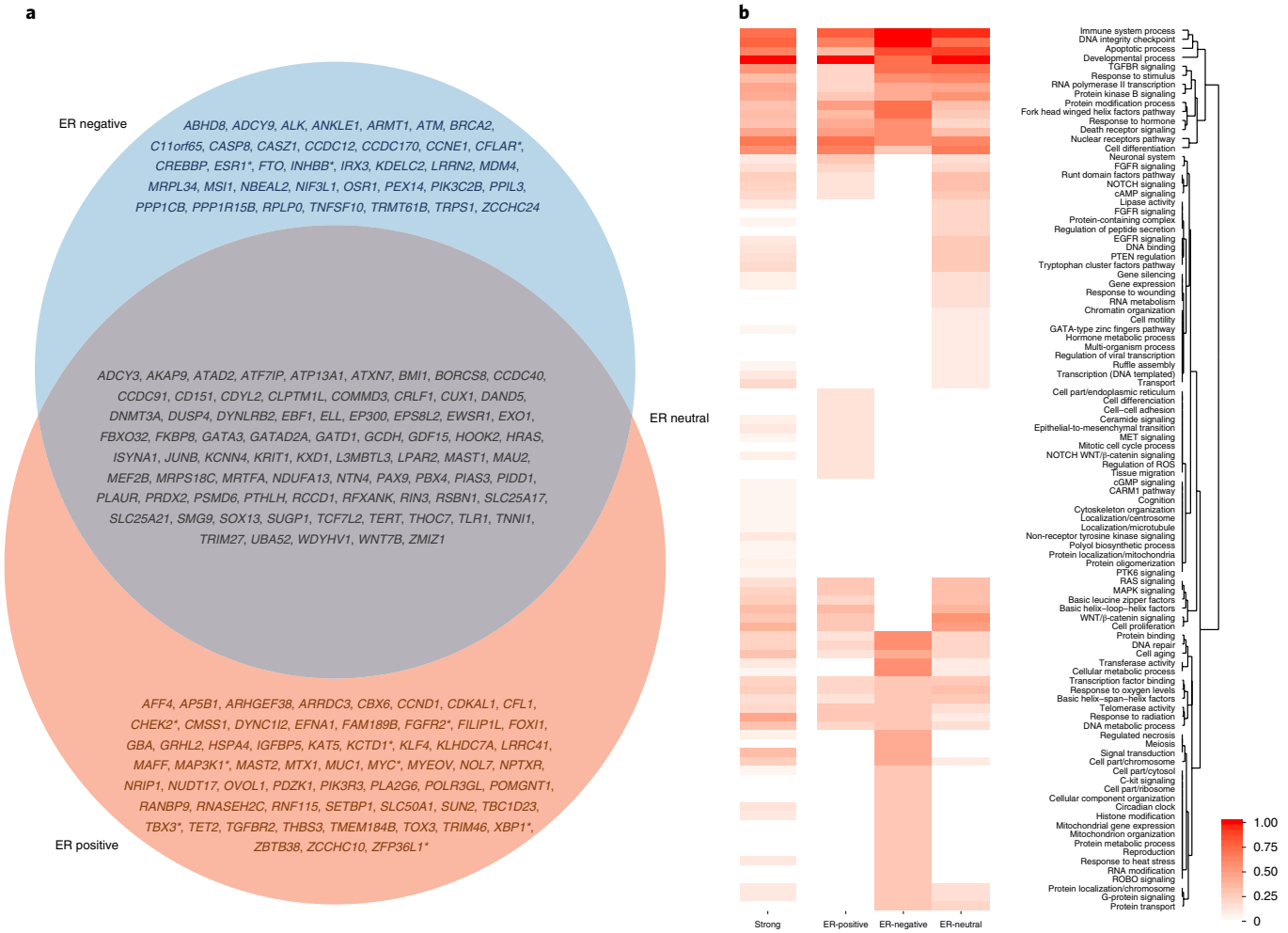

Fig. 5 | Predicted target genes by phenotype and significantly enriched pathways. a, Venn diagram showing the associated phenotype (ER positive, ER negative or ER neutral) for the level 1 target genes, predicted by the CCVs and HPPVs. Asterisks denote ER-positive or ER-negative target genes also targeted by ER-neutral signals. **b**, Heat map showing clustering of pathway themes over-represented by INQUISIT level 1 target genes. Colors represent the relative number of genes per phenotype within enriched pathways, grouped by common themes (ER positive, ER negative, ER neutral or all phenotypes together (strong)). cAMP, cyclic adenosine monophosphate; CARM1, coactivator associated arginine methyltransferase 1; cGMP, cyclic guanosine monophosphate; EGFR, epidermal growth factor receptor; FGFR, fibroblast growth factor receptor; GATA, GATA transcription factors; MAPK, mitogen activated protein kinase; MET, MET proto-oncogene receptor tyrosine kinase; NOTCH, notch protein; PTEN, phosphatase and tensin homolog; PTK6, protein tyrosine kinase 6; RAS, RAS protein; ROBO, roundabout receptors; ROS, reactive oxygen species; TGFBR, transforming growth factor beta receptor; WNT, WNT proteins.

from 1 to 115 (median: 12), representing an average reduction of 43% in the number of variants needed to capture the signal.

PAINTOR and CCV analyses were generally consistent, yet complementary. Only 3.3% of variants outside of the set of strong-signal CCVs for overall breast cancer had a posterior probability of >1%, and only 48 (0.013%) of these had a posterior probability of >30% (Supplementary Fig. 3d). At ER-positive and ER-negative signals, respectively, 3.1 and 1.6% of the non-CCVs at strong signals had a posterior probability of >1%, and 40 (0.019%) and 3 (0.003%) of these had a posterior probability of >30% (Supplementary Fig. 3e–f). For the non-CCVs at strong-evidence signals with a posterior probability of >30%, the relatively HPP may be driven by the addition of functional annotation. Indeed, the incorporation of functional annotations more than doubled the posterior probability for 64 out of 88 variants when compared with a PAINTOR model with no functional annotations.

CCVs co-localize with variants controlling local gene expression. We used four breast-specific expression quantitative trait loci (eQTL) datasets to identify a credible set of variants associated with differences in gene expression (expression variants): tumor tissue from the Nurses' Health Study (NHS)[32] and The Cancer Genome Atlas (TCGA)[33]; and normal breast tissue from the NHS and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[34]. We then examined the overlap of expression variants (for each gene, expression variants were defined as those variants that had a *P* value within two orders of magnitude of the variant most significantly associated with that gene's expression) with CCVs (Methods). There was significant overlap of CCVs with expression variants from both the NHS normal and breast cancer tissue studies (normal breast: odds ratio (OR) = 2.70; $P = 1.7 \times 10^{-5}$; tumor tissue: OR = 2.34; $P = 2.6 \times 10^{-4}$; Supplementary Table 3). ER-neutral CCVs overlapped with expression variants in normal tissue more frequently than ER-positive and ER-negative CCVs ($OR_{ER\ neutral} = 3.51$; $P = 1.3 \times 10^{-5}$). Cancer risk CCVs overlapped credible expression variants in 128 out of 205 signals (62%) in at least one of the datasets (Supplementary Table 5a,b). Sixteen additional variants with a posterior probability of ≥30%, not included among the

CCVs, also overlapped with a credible expression variant (Supplementary Table 5a,b).

Transcription factors and known somatic breast cancer drivers are over-represented among prioritized target genes. We assumed that causal variants function by affecting the behavior of a local target gene. However, it is challenging to define target genes or to determine how they may be affected by the causal variant. Few potentially causal variants directly affect protein coding: we observed 67 out of 5,375 CCVs and 19 out of 137 HPPVs (≥30%) in protein-coding regions. Of these, 33 (0.61%) were predicted to create a missense change, one a frameshift and another a stop gain, while 30 were synonymous (0.59%; Supplementary Table 5c). In total, 499 CCVs at 94 signals, and four additional HPPVs (≥30%), are predicted to create new splice sites or activate cryptic splice sites in 126 genes (Supplementary Table 5d). These results are consistent with previous observations that the majority of common susceptibility variants are regulatory.

We applied an updated version of our pipeline integrated expression quantitative trait and in silico prediction of GWAS targets (INQUISIT)[2] to prioritize potential target genes from 5,375 CCVs in strong signals and all 138 HPPVs (≥30%; Supplementary Table 2c). The pipeline predicted 1,204 target genes from 124 out of 128 genomic regions examined. As a validation, we examined the overlap between INQUISIT predictions and 278 established breast cancer driver genes[35–39]. Cancer driver genes were over-represented among high-confidence (level 1) targets, with a fivefold increase over expected levels from CCVs and a 15-fold increase from HPPVs ($P=1\times10^{-6}$; Supplementary Fig. 4a). Notably, 13 cancer driver genes (*ATAD2*, *CASP8*, *CCND1*, *CHEK2*, *ESR1*, *FGFR2*, *GATA3*, *MAP3K1*, *MYC*, *SETBP1*, *TBX3*, *XBP1* and *ZFP36L1*) were predicted from the HPPVs derived from PAINTOR. Cancer driver gene status was consequently included as an additional weighting factor in the INQUISIT pipeline. Transcription factor genes[40] were also enriched among high-confidence targets predicted from both CCVs (twofold; $P=4.6\times10^{-4}$) and HPPVs (2.5-fold; $P=1.8\times10^{-2}$; Supplementary Fig. 4a).

In total, INQUISIT identified 191 target genes supported by strong evidence (Supplementary Table 6). Significantly more genes were targeted by multiple independent signals ($n=165$) than expected by chance ($P=4.3\times10^{-8}$; Supplementary Fig. 4b and Fig. 4). Six high-confidence predictions came only from HPPVs, although three of these (*IGFBP5*, *POMGNT1* and *WDYHV1*) had been predicted at lower confidence from CCVs. Target genes included 20 that were prioritized via potential coding/splicing changes (Supplementary Table 7), ten via promoter variants (Supplementary Table 8) and 180 via distal regulatory variants (Supplementary Table 9). We illustrate the genes prioritized via multiple lines of evidence in Fig. 4.

Three examples of INQUISIT using genomic features to predict target genes. On the basis of capture Hi-C and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data, *NRIP1* is a predicted target of intergenic CCVs and HPPVs at chr21q21 (Supplementary Fig. 5a). Multiple target genes were predicted at chr22q12, including the driver genes *CHEK2* and *XBP1* (Supplementary Fig. 5b). A third example at chr12q24.31 is a more complicated scenario with two level 1 targets: *RPLP0* (ref. [41]) and a modulator of mammary progenitor cell expansion, *MSI1* (ref. [42]) (Supplementary Fig. 5c).

Target gene pathways include DNA integrity checkpoint, apoptosis and developmental processes and the immune system. We performed pathway analysis to identify common processes using INQUSIT high-confidence target protein-coding genes (Fig. 5a) and identified 488 Gene Ontology terms and 307 pathways at an FDR of 5% (Supplementary Table 10). These were grouped into 98 themes by common ancestor Gene Ontology terms, pathways or transcription factor classes (Fig. 5b). We found that 23% (14/60) of the ER-positive target genes were classified within developmental process pathways (including mammary development), 18% were classified in immune system pathways and a further 17% were classified in nuclear receptor pathways. Of the genes targeted by ER-neutral signals, 21% (18/87) were classified in developmental process pathways, 19% were classified in immune system pathways and a further 18% were classified in apoptotic process pathways. The top themes of genes targeted by ER-negative signals were DNA integrity checkpoint processes and the immune system, each of which contained 19% of genes (7/37), and apoptotic processes (16%).

Novel pathways revealed by this study include tumor necrosis factor-related apoptosis-inducing ligand (TRAIL) signaling, the AP-2 transcription factors pathway, and regulation of IκB kinase/nuclear factor-κB (NF-κB) signaling. Of note, the latter of these is specifically over-represented among ER-negative target genes. We also found significant over-representation of additional carcinogenesis-linked pathways, including cyclic adenosine monophosphate, NOTCH, phosphoinositide 3-kinase, RAS and WNT/β-catenin, and of receptor tyrosine kinase signaling, including fibroblast growth factor receptor, epidermal growth factor receptor and transforming growth factor-β receptor[43–47]. Finally, our target genes are also significantly over-represented in DNA damage checkpoint and DNA repair pathways, as well as programmed cell death pathways, such as apoptotic processes, regulated necrosis and death receptor signaling-related pathways.

Discussion

We have performed multiple complementary analyses on 150 breast cancer-associated regions originally found by GWASs, and identified 362 independent risk signals, 205 of these with high confidence ($P<10^{-6}$). The inclusion of these new variants increases the explained proportion of familial risk by 6% compared with that explained by the lead signals alone.

We observed that most regions contain multiple independent signals, with the greatest number (nine) in the region surrounding *ESR1* and its co-regulated genes, and on 2q35, where *IGFBP5* appears to be a key target. We used two complementary approaches to identify likely causal variants within each region: a Bayesian approach, PAINTOR (which integrated genetic associations, linkage disequilibrium and informative genomic features, providing complementary evidence) and a more traditional, multinomial regression approach. PAINTOR supported most associations found by multinomial regression and also identified additional variants. Specifically, the Bayesian method highlighted 15 variants that are highly likely to be causal (HPP ≥ 80%). From these approaches, we identified a single variant, likely to be causal, at each of 34 signals (Table 1). Of these, only rs16991615 (*MCM8*; NP_115874.3:p.E341K) and rs7153397 (*CCDC88C*; NM_001080414.2:c.5058+1342G>A; a cryptic splice-donor site) were predicted to affect protein-coding sequences. However, in other signals, we also identified four coding changes previously recognized as deleterious: the stop gain rs11571833 (*BRCA2*; NP_000050.2:p.K3326*)[48]; two *CHEK2* coding variants (the frameshift rs555607708 (refs. [49,50]) and a missense variant, rs17879961 (refs. [51,52])); and a splicing variant (rs10069690 in *TERT*, which results in the truncated protein INS1b[19], decreased telomerase activity, telomere shortening and increased DNA damage response[53]).

Having identified potential causal variants within each signal, we aimed to uncover their functions at the DNA level, as well as trying to predict their target gene(s). Across all 150 regions, a notable feature is that many likely causal variants implicated in ER-positive cancer risk lie in gene regulatory regions marked as open and active in ER-positive breast cells, but not in other cell types. Moreover,

a significant proportion of potential causal variants overlap the binding sites for transcription factor proteins ($n=40$ from ChIP-Seq) and co-regulators ($n=64$ with the addition of computationally derived motifs). Furthermore, nine proteins also appear in the list of high-confidence target genes; hence, the following genes and their products have been implicated by two different approaches: *CREBBP*, *EP300*, *ESR1*, *FOXI1*, *GATA3*, *MEF2B*, *MYC*, *NRIP1* and *TCF7L2*. Most proteins encoded by these genes already have established roles in estrogen signaling. *CREBBP*, *EP300*, *ESR1*, *GATA3* and *MYC* are also known cancer driver genes that are frequently somatically mutated in breast tumors.

In contrast with ER-positive signals, we identified fewer genomic features enriched in ER-negative signals. This may reflect the common molecular mechanisms underlying their development, but the power of this study was limited, despite including as many patients with ER-negative tumors as possible from the BCAC and CIMBA consortia. Less than 20% of genomic signals confer a greater risk of ER-negative cancer and there are few publicly available ChIP-Seq data on ER-negative breast cancer cell lines. The heterogeneity of ER-negative tumors also may have limited our power. Nevertheless, we have identified 35 target genes for ER-negative likely causal variants. Some of these already had functional evidence supporting their role: including *CASP8* (ref. [54]) and *MDM4* (ref. [55]). However, most targets currently have no reported function in ER-negative breast cancer development.

Finally, we examined the Gene Ontology pathways in which target genes most often lie. Of note, 14% (25/180) of all high-confidence target genes and 19% of ER-negative target predictions are in immune system pathways. Among the significantly enriched pathways were T cell activation, interleukin signaling, Toll-like receptor cascades and I-κB kinase/NF-κB signaling, as well as processes leading to activation and perpetuation of the innate immune system. The link between immunity, inflammation and tumorigenesis has been studied extensively[56], although not primarily in the context of susceptibility. Five ER-negative high-confidence target genes (*ALK*, *CASP8*, *CFLAR*, *ESR1* and *TNFSF10*) lie in the IκB kinase/NF-κB signaling pathway. Interestingly, ER-negative cells have high levels of NF-kB activity compared with ER-positive cells[57]. A recent expression–methylation analysis on breast cancer tumor tissue also identified clusters of genes correlated with DNA methylation levels: one enriched in ER signaling genes and a second in immune pathway genes[58].

These analyses provide strong evidence for more than 200 independent breast cancer risk signals, identify the plausible cancer variants and define likely target genes for the majority of these. However, notwithstanding the enrichment of certain pathways and transcription factors, the biological basis underlying most of these signals remains poorly understood. Our analyses provide a rational basis for such future studies into the biology underlying breast cancer susceptibility.

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/s41588-019-0537-1.

Received: 13 December 2018; Accepted: 24 October 2019;
Published online: 7 January 2020

References

1. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
2. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
3. Ghoussaini, M. et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nat. Commun.* **4**, 4999 (2014).
4. Wyszynski, A. et al. An intergenic risk locus containing an enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression. *Hum. Mol. Genet.* **25**, 3863–3876 (2016).
5. Guo, X. et al. Fine-scale mapping of the 4q24 locus identifies two independent loci associated with breast cancer risk. *Cancer Epidemiol. Biomark. Prev.* **24**, 1680–1691 (2015).
6. Glubb, D. M. et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am. J. Hum. Genet.* **96**, 5–20 (2015).
7. Dunning, A. M. et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate *ESR1*, *RMND1* and *CCDC170*. *Nat. Genet.* **48**, 374–386 (2016).
8. Shi, J. et al. Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. *Int. J. Cancer* **139**, 1303–1317 (2016).
9. Orr, N. et al. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. *Hum. Mol. Genet.* **24**, 2966–2984 (2015).
10. Darabi, H. et al. Polymorphisms in a putative enhancer at the 10q21.2 breast cancer risk locus regulate *NRBF2* expression. *Am. J. Hum. Genet.* **97**, 22–34 (2015).
11. Darabi, H. et al. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). *Sci. Rep.* **6**, 32512 (2016).
12. Meyer, K. B. et al. Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am. J. Hum. Genet.* **93**, 1046–1060 (2013).
13. Betts, J. A. et al. Long noncoding RNAs *CUPID1* and *CUPID2* mediate breast cancer risk at 11q13 by modulating the response to DNA damage. *Am. J. Hum. Genet.* **101**, 255–266 (2017).
14. French, J. D. et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.* **92**, 489–503 (2013).
15. Ghoussaini, M. et al. Evidence that the 5p12 variant rs10941679 confers susceptibility to estrogen-receptor-positive breast cancer through *FGF10* and *MRPS30* regulation. *Am. J. Hum. Genet.* **99**, 903–911 (2016).
16. Horne, H. N. et al. Fine-mapping of the 1p11.2 breast cancer susceptibility locus. *PLoS ONE* **11**, e0160316 (2016).
17. Zeng, C. et al. Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. *Breast Cancer Res.* **18**, 64 (2016).
18. Lin, W. Y. et al. Identification and characterization of novel associations in the *CASP8/ALS2CR12* region on chromosome 2 with breast cancer risk. *Hum. Mol. Genet.* **24**, 285–298 (2015).
19. Bojesen, S. E. et al. Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer. *Nat. Genet.* **45**, 371–384.e2 (2013).
20. Lawrenson, K. et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. *Nat. Commun.* **7**, 12675 (2016).
21. Amos, C. I. et al. The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomark. Prev.* **26**, 126–135 (2017).
22. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.* **45**, 353–361.e2 (2013).
23. 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).
24. 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).
25. Mavaddat, N., Antoniou, A. C., Easton, D. F. & Garcia-Closas, M. Genetic susceptibility to breast cancer. *Mol. Oncol.* **4**, 174–191 (2010).
26. Lakhani, S. R. et al. Prediction of *BRCA1* status in patients with breast cancer using estrogen receptor and basal phenotype. *Clin. Cancer Res.* **11**, 5175–5180 (2005).
27. Taberlay, P. C., Statham, A. L., Kelly, T. K., Clark, S. J. & Jones, P. A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.* **24**, 1421–1432 (2014).
28. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
29. Farh, K. K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
30. Cowper-Sal lari, R. et al. Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
31. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

32. Quiroz-Zarate, A. et al. Expression quantitative trait loci (QTL) in tumor adjacent normal breast tissue and breast tumor tissue. *PLoS ONE* **12**, e0170181 (2017).
33. Cancer Genome Atlas Research Networket al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
34. Curtis, C. et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
35. Ciriello, G. et al. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163**, 506–519 (2015).
36. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
37. Pereira, B. et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (2016).
38. Cancer Genome Atlas Network Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
39. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
40. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
41. Artero-Castro, A. et al. Disruption of the ribosomal P complex leads to stress-induced autophagy. *Autophagy* **11**, 1499–1519 (2015).
42. Wang, X. Y. et al. Musashi1 modulates mammary progenitor cell expansion through proliferin-mediated activation of the Wnt and Notch pathways. *Mol. Cell Biol.* **28**, 3589–3599 (2008).
43. Vijayan, D., Young, A., Teng, M. W. L. & Smyth, M. J. Targeting immunosuppressive adenosine in cancer. *Nat. Rev. Cancer* **17**, 709–724 (2017).
44. Takebe, N. et al. Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. *Nat. Rev. Clin. Oncol.* **12**, 445–464 (2015).
45. Thorpe, L. M., Yuzugullu, H. & Zhao, J. J. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. *Nat. Rev. Cancer* **15**, 7–24 (2015).
46. Nusse, R. & Clevers, H. Wnt/β-catenin signaling, disease, and emerging therapeutic modalities. *Cell* **169**, 985–999 (2017).
47. Massague, J. TGFβ signalling in context. *Nat. Rev. Mol. Cell Biol.* **13**, 616–630 (2012).
48. Meeks, H. D. et al. BRCA2 polymorphic stop codon K3326X and the risk of breast, prostate, and ovarian cancers. *J. Natl Cancer Inst.* **108**, djv315 (2016).
49. CHEK2 Breast Cancer Case-Control Consortium CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am. J. Hum. Genet.* **74**, 1175–1182 (2004).
50. Schmidt, M. K. et al. Age- and tumor subtype-specific breast cancer risk estimates for CHEK2*1100delC carriers. *J. Clin. Oncol.* **34**, 2750–2760 (2016).
51. Kilpivaara, O. et al. CHEK2 variant I157T may be associated with increased breast cancer risk. *Int. J. Cancer* **111**, 543–547 (2004).
52. Muranen, T. A. et al. Patient survival and tumor characteristics associated with CHEK2:p.I157T—findings from the Breast Cancer Association Consortium. *Breast Cancer Res.* **18**, 98 (2016).
53. Killedar, A. et al. A common cancer risk-associated allele in the *hTERT* locus encodes a dominant negative inhibitor of telomerase. *PLoS Genet.* **11**, e1005286 (2015).
54. De Basio, A. et al. Unusual roles of caspase-8 in triple-negative breast cancer cell line MDA-MB-231. *Int. J. Oncol.* **48**, 2339–2348 (2016).
55. Haupt, S. et al. Targeting Mdmx to treat breast cancers with wild-type p53. *Cell Death Dis.* **6**, e1821 (2015).
56. Pandya, P. H., Murray, M. E., Pollok, K. E. & Renbarger, J. L. The immune system in cancer pathogenesis: potential therapeutic approaches. *J. Immunol. Res.* **2016**, 4273943 (2016).
57. Gionet, N., Jansson, D., Mader, S. & Pratt, M. A. NF-κB and estrogen receptor α interactions: differential function in estrogen receptor-negative and -positive hormone-independent breast cancer cells. *J. Cell Biochem.* **107**, 448–459 (2009).
58. Fleischer, T. et al. DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* **8**, 1379 (2017).

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

Laura Fachal[1], Hugues Aschard[2,3,4,282], Jonathan Beesley[5,282], Daniel R. Barnes[6], Jamie Allen[6], Siddhartha Kar[1], Karen A. Pooley[6], Joe Dennis[6], Kyriaki Michailidou[6,7], Constance Turman[4], Penny Soucy[8], Audrey Lemaçon[8], Michael Lush[6], Jonathan P. Tyrer[1], Maya Ghoussaini[1], Mahdi Moradi Marjaneh[5,280], Xia Jiang[3], Simona Agata[9], Kristiina Aittomäki[10], M. Rosario Alonso[11], Irene L. Andrulis[12,13], Hoda Anton-Culver[14], Natalia N. Antonenkova[15], Adalgeir Arason[16,17], Volker Arndt[18], Kristan J. Aronson[19], Banu K. Arun[20], Bernd Auber[21], Paul L. Auer[22,23], Jacopo Azzollini[24], Judith Balmaña[25,26], Rosa B. Barkardottir[16,17], Daniel Barrowdale[6], Alicia Beeghly-Fadiel[27], Javier Benitez[28,29], Marina Bermisheva[30], Katarzyna Białkowska[31], Amie M. Blanco[32], Carl Blomqvist[33,34], William Blot[27,35], Natalia V. Bogdanova[15,36,37], Stig E. Bojesen[38,39,40], Manjeet K. Bolla[6], Bernardo Bonanni[41], Ake Borg[42], Kristin Bosse[43], Hiltrud Brauch[44,45,46], Hermann Brenner[18,46,47], Ignacio Briceno[48,49], Ian W. Brock[50], Angela Brooks-Wilson[51,52], Thomas Brüning[53], Barbara Burwinkel[54,55], Saundra S. Buys[56], Qiuyin Cai[27], Trinidad Caldés[57], Maria A. Caligo[58], Nicola J. Camp[59], Ian Campbell[60,61], Federico Canzian[62], Jason S. Carroll[63], Brian D. Carter[64], Jose E. Castelao[65], Jocelyne Chiquette[66], Hans Christiansen[36], Wendy K. Chung[67], Kathleen B. M. Claes[68], Christine L. Clarke[69], GEMO Study Collaborators[70], EMBRACE Collaborators[70], J. Margriet Collée[71], Sten Cornelissen[72], Fergus J. Couch[73], Angela Cox[50], Simon S. Cross[74], Cezary Cybulski[31], Kamila Czene[75], Mary B. Daly[76], Miguel de la Hoya[57], Peter Devilee[77,78], Orland Diez[79,80], Yuan Chun Ding[81], Gillian S. Dite[82], Susan M. Domchek[83], Thilo Dörk[37], Isabel dos-Santos-Silva[84], Arnaud Droit[8,85], Stéphane Dubois[8], Martine Dumont[8], Mercedes Duran[86], Lorraine Durcan[87,88], Miriam Dwek[89], Diana M. Eccles[90], Christoph Engel[91], Mikael Eriksson[75], D. Gareth Evans[92,93], Peter A. Fasching[94,95], Olivia Fletcher[96], Giuseppe Floris[97], Henrik Flyger[98], Lenka Foretova[99], William D. Foulkes[100], Eitan Friedman[101,102],

Lin Fritschi[103], Debra Frost[6], Marike Gabrielson[75], Manuela Gago-Dominguez[104,105], Gaetana Gambino[58], Patricia A. Ganz[106], Susan M. Gapstur[64], Judy Garber[107], José A. García-Sáenz[108], Mia M. Gaudet[64], Vassilios Georgoulias[109], Graham G. Giles[82,110,111], Gord Glendon[12], Andrew K. Godwin[112], Mark S. Goldberg[113,114], David E. Goldgar[115], Anna González-Neira[29], Maria Grazia Tibiletti[116], Mark H. Greene[117], Mervi Grip[118], Jacek Gronwald[31], Anne Grundy[119], Pascal Guénel[120], Eric Hahnen[121,122], Christopher A. Haiman[123], Niclas Håkansson[124], Per Hall[75,125], Ute Hamann[126], Patricia A. Harrington[1], Jaana M. Hartikainen[127,128,129], Mikael Hartman[130,131], Wei He[75], Catherine S. Healey[1], Bernadette A. M. Heemskerk-Gerritsen[132], Jane Heyworth[133], Peter Hillemanns[37], Frans B. L. Hogervorst[134], Antoinette Hollestelle[132], Maartje J. Hooning[132], John L. Hopper[82], Anthony Howell[135], Guanmengqian Huang[126], Peter J. Hulick[136,137], Evgeny N. Imyanitov[138], KConFab Investigators[70], HEBON Investigators[70], ABCTB Investigators[70], Claudine Isaacs[139], Motoki Iwasaki[140], Agnes Jager[132], Milena Jakimovska[141], Anna Jakubowska[31,142], Paul A. James[61,143], Ramunas Janavicius[144,145], Rachel C. Jankowitz[146], Esther M. John[147], Nichola Johnson[96], Michael E. Jones[148], Arja Jukkola-Vuorinen[149], Audrey Jung[150], Rudolf Kaaks[150], Daehee Kang[151,152,153], Pooja Middha Kapoor[150,154], Beth Y. Karlan[155,156], Renske Keeman[72], Michael J. Kerin[157], Elza Khusnutdinova[30,158], Johanna I. Kiiski[159], Judy Kirk[160], Cari M. Kitahara[161], Yon-Dschun Ko[162], Irene Konstantopoulou[163], Veli-Matti Kosma[127,128,129], Stella Koutros[164], Katerina Kubelka-Sabit[165], Ava Kwong[166,167,168], Kyriacos Kyriacou[7], Yael Laitman[101], Diether Lambrechts[169,170], Eunjung Lee[123], Goska Leslie[6], Jenny Lester[155,156], Fabienne Lesueur[171,172,173], Annika Lindblom[174,175], Wing-Yee Lo[44], Jirong Long[27], Artitaya Lophatananon[176], Jennifer T. Loud[117], Jan Lubiński[31], Robert J. MacInnis[82,110], Tom Maishman[87,88], Enes Makalic[82], Arto Mannermaa[127,128,129], Mehdi Manoochehri[126], Siranoush Manoukian[24], Sara Margolin[125,177], Maria Elena Martinez[105,178], Keitaro Matsuo[179,180], Tabea Maurer[181], Dimitrios Mavroudis[109], Rebecca Mayes[1], Lesley McGuffog[6], Catriona McLean[182], Noura Mebirouk[171,172,183], Alfons Meindl[184], Austin Miller[185], Nicola Miller[157], Marco Montagna[9], Fernando Moreno[108], Kenneth Muir[176], Anna Marie Mulligan[186,187], Victor M. Muñoz-Garzon[188], Taru A. Muranen[159], Steven A. Narod[189], Rami Nassir[190], Katherine L. Nathanson[83], Susan L. Neuhausen[81], Heli Nevanlinna[159], Patrick Neven[97], Finn C. Nielsen[191], Liene Nikitina-Zake[192], Aaron Norman[193], Kenneth Offit[194,195], Edith Olah[196], Olufunmilayo I. Olopade[197], Håkan Olsson[198], Nick Orr[199], Ana Osorio[28,29], V. Shane Pankratz[200], Janos Papp[196], Sue K. Park[151,152,153], Tjoung-Won Park-Simon[37], Michael T. Parsons[5], James Paul[201], Inge Sokilde Pedersen[202,203,204], Bernard Peissel[24], Beth Peshkin[139], Paolo Peterlongo[205], Julian Peto[84], Dijana Plaseska-Karanfilska[141], Karolina Prajzendanc[31], Ross Prentice[22], Nadege Presneau[89], Darya Prokofyeva[158], Miquel Angel Pujana[206], Katri Pylkäs[207,208], Paolo Radice[209], Susan J. Ramus[210,211], Johanna Rantala[212], Rohini Rau-Murthy[195], Gad Rennert[213], Harvey A. Risch[214], Mark Robson[195], Atocha Romero[215], Maria Rossing[191], Emmanouil Saloustros[216], Estela Sánchez-Herrero[215], Dale P. Sandler[217], Marta Santamariña[28,218,219], Christobel Saunders[220], Elinor J. Sawyer[221], Maren T. Scheuner[32], Daniel F. Schmidt[82,222], Rita K. Schmutzler[121,122], Andreas Schneeweiss[55,223], Minouk J. Schoemaker[148], Ben Schöttker[18,224], Peter Schürmann[37], Christopher Scott[193], Rodney J. Scott[225,226,227], Leigha Senter[228], Caroline M. Seynaeve[132], Mitul Shah[1], Priyanka Sharma[229], Chen-Yang Shen[230,231], Xiao-Ou Shu[27], Christian F. Singer[232], Thomas P. Slavin[233], Snezhana Smichkoska[234], Melissa C. Southey[111,235], John J. Spinelli[236,237], Amanda B. Spurdle[5], Jennifer Stone[82,238], Dominique Stoppa-Lyonnet[183,239,240],

Christian Sutter[241], Anthony J. Swerdlow[148,242], Rulla M. Tamimi[3,4,243], Yen Yen Tan[244], William J. Tapper[90], Jack A. Taylor[217,245], Manuel R. Teixeira[246,247], Maria Tengström[127,248,249], Soo Hwang Teo[250,251], Mary Beth Terry[252], Alex Teulé[253], Mads Thomassen[254], Darcy L. Thull[255], Marc Tischkowitz[100,256], Amanda E. Toland[257], Rob A. E. M. Tollenaar[258], Ian Tomlinson[259,260], Diana Torres[48,126], Gabriela Torres-Mejía[261], Melissa A. Troester[262], Thérèse Truong[120], Nadine Tung[263], Maria Tzardi[264], Hans-Ulrich Ulmer[265], Celine M. Vachon[266], Christi J. van Asperen[267], Lizet E. van der Kolk[134], Elizabeth J. van Rensburg[268], Ana Vega[269], Alessandra Viel[270], Joseph Vijai[194,195], Maartje J. Vogel[134], Qin Wang[6], Barbara Wappenschmidt[121,122], Clarice R. Weinberg[271], Jeffrey N. Weitzel[233], Camilla Wendt[177], Hans Wildiers[97], Robert Winqvist[207,208], Alicja Wolk[124,272], Anna H. Wu[123], Drakoulis Yannoukakos[163], Yan Zhang[18,46], Wei Zheng[27], David Hunter[273], Paul D. P. Pharoah[1,6], Jenny Chang-Claude[150,181], Montserrat García-Closas[164,274], Marjanka K. Schmidt[72,275], Roger L. Milne[82,110,111], Vessela N. Kristensen[276,277,278,279], Juliet D. French[5], Stacey L. Edwards[5], Antonis C. Antoniou[6], Georgia Chenevix-Trench[5,281], Jacques Simard[8,281], Douglas F. Easton[1,6,281], Peter Kraft[3,4,281*] and Alison M. Dunning[1,281*]

[1]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK. [2]Centre de Bioinformatique Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France. [3]Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [5]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [6]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK. [7]Department of Electron Microscopy/Molecular Pathology and The Cyprus School of Molecular Medicine, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus. [8]Genomics Center, Centre Hospitalier Universitaire de Québec, Université Laval Research Center, Québec City, Québec, Canada. [9]Immunology and Molecular Oncology Unit, Veneto Institute of Oncology (IOV), IRCCS, Padua, Italy. [10]Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [11]Human Genotyping-CEGEN Unit, Human Cancer Genetic Program, Spanish National Cancer Research Centre, Madrid, Spain. [12]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [13]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [14]Department of Epidemiology, Genetic Epidemiology Research Institute, University of California, Irvine, Irvine, CA, USA. [15]N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus. [16]Department of Pathology, Landspitali University Hospital, Reykjavik, Iceland. [17]BMC (Biomedical Centre), Faculty of Medicine, University of Iceland, Reykjavik, Iceland. [18]Division of Clinical Epidemiology and Aging Research (C070), German Cancer Research Center (DKFZ), Heidelberg, Germany. [19]Department of Public Health Sciences and Cancer Research Institute, Queen's University, Kingston, Ontario, Canada. [20]Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [21]Institute of Human Genetics, Hannover Medical School, Hannover, Germany. [22]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [23]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA. [24]Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy. [25]High Risk and Cancer Prevention Group, Vall Hebron Institute of Oncology, Barcelona, Spain. [26]Department of Medical Oncology, Vall Hebron University Hospital, Barcelona, Spain. [27]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA. [28]Centro de Investigación en Biomédica en Red de Enfermedades Raras (CIBERER), Madrid, Spain. [29]Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [30]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, Russia. [31]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland. [32]Cancer Genetics and Prevention Program, University of California, San Francisco, San Francisco, CA, USA. [33]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [34]Department of Oncology, Örebro University Hospital, Örebro, Sweden. [35]International Epidemiology Institute, Rockville, MD, USA. [36]Department of Radiation Oncology, Hannover Medical School, Hannover, Germany. [37]Gynaecology Research Unit, Hannover Medical School, Hannover, Germany. [38]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [39]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [40]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [41]Division of Cancer Prevention and Genetics, European Institute of Oncology (IEO), IRCCS, Milan, Italy. [42]Department of Oncology, Lund University and Skåne University Hospital, Lund, Sweden. [43]Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany. [44]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany. [45]iFIT Cluster of Excellence, University of Tuebingen, Tuebingen, Germany. [46]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany. [47]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany. [48]Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia. [49]Medical Faculty, Universidad de La Sabana, Bogota, Colombia. [50]Sheffield Institute for Nucleic Acids (SInFoNiA), Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK. [51]Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia, Canada. [52]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, British Columbia, Canada. [53]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany. [54]Molecular Epidemiology Group (C080), German Cancer Research Center (DKFZ), Heidelberg, Germany. [55]Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg, University of Heidelberg, Heidelberg, Germany. [56]Department of Medicine, Huntsman Cancer Institute, Salt Lake City, UT, USA. [57]Molecular Oncology Laboratory, CIBERONC, Hospital Clinico San Carlos, Instituto de Investigación Sanitaria del Hospital Clínico San Carlos (IdISSC), Madrid, Spain. [58]SOD Genetica Molecolare, University Hospital, Pisa, Italy. [59]Department of Internal Medicine, Huntsman Cancer Institute, Salt Lake City, UT, USA. [60]Research Department, Peter MacCallum Cancer Center, Melbourne, Victoria, Australia. [61]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, Australia.

[62]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany. [63]Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK. [64]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, USA. [65]Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain. [66]Axe Oncologie, Centre de Recherche, Centre Hospitalier Universitaire de Québec, Université Laval, Québec, Québec, Canada. [67]Departments of Pediatrics and Medicine, Columbia University, New York, NY, USA. [68]Centre for Medical Genetics, Ghent University, Ghent, Belgium. [69]Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia. [70]A full list of members and affiliations appears at the end of the paper. [71]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, the Netherlands. [72]Division of Molecular Pathology, Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [73]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA. [74]Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK. [75]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden. [76]Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA. [77]Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands. [78]Department of Human Genetics, Leiden University Medical Center, Leiden, the Netherlands. [79]Oncogenetics Group, Vall d'Hebron Institute of Oncology, Barcelona, Spain. [80]Clinical and Molecular Genetics Area, Vall Hebron University Hospital, Barcelona, Spain. [81]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA. [82]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, Australia. [83]Basser Center for BRCA, Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, USA. [84]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [85]Département de Médecine Moléculaire, Faculté de Médecine, Centre de Recherche, Centre Hospitalier Universitaire de Québec, Laval University, Québec City, Québec, Canada. [86]Cáncer Hereditario, Instituto de Biología y Genética Molecular (IBGM), Universidad de Valladolid Centro Superior de Investigaciones Científicas (UVA-CSIC), Valladolid, Spain. [87]Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton, Southampton, UK. [88]Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK. [89]School of Life Sciences, University of Westminster, London, UK. [90]Faculty of Medicine, University of Southampton, Southampton, UK. [91]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany. [92]Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester Academic Health Science Centre, Manchester, UK. [93]North West Genomics Laboratory Hub, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK. [94]David Geffen School of Medicine, Department of Medicine, Division of Hematology and Oncology, University of California, Los Angeles, Los Angeles, CA, USA. [95]Department of Gynecology and Obstetrics, Comprehensive Cancer Center ER-EMN, University Hospital Erlangen, Friedrich–Alexander University Erlangen–Nuremberg, Erlangen, Germany. [96]Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK. [97]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium. [98]Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark. [99]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, Czech Republic. [100]Program in Cancer Genetics, Departments of Human Genetics and Oncology, McGill University, Montréal, Québec, Canada. [101]The Suzanne Levy-Gertner Oncogenetics Unit, Chaim Sheba Medical Center, Ramat Gan, Israel. [102]Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, Israel. [103]School of Public Health, Curtin University, Perth, Western Australia, Australia. [104]Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela, Spain. [105]Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA. [106]Schools of Medicine and Public Health, Division of Cancer Prevention and Control Research, Jonsson Comprehensive Cancer Centre, University of California, Los Angeles, Los Angeles, CA, USA. [107]Cancer Risk and Prevention Clinic, Dana-Farber Cancer Institute, Boston, MA, USA. [108]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. [109]Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece. [110]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, Australia. [111]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Melbourne, Victoria, Australia. [112]Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Kansas City, KS, USA. [113]Department of Medicine, McGill University, Montréal, Québec, Canada. [114]Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, Québec, Canada. [115]Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, USA. [116]U.O. Anatomia Patologica Ospedale di Circolo, ASST Settelaghi, Varese, Italy. [117]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [118]Department of Surgery, Oulu University Hospital, University of Oulu, Oulu, Finland. [119]Centre de Recherche du Centre Hospitalier de l'Université de Montréal (CHUM), Université de Montréal, Montréal, Québec, Canada. [120]Cancer and Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Paris, France. [121]Center for Hereditary Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [122]Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany. [123]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. [124]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [125]Department of Oncology, Södersjukhuset, Stockholm, Sweden. [126]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany. [127]Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland. [128]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland. [129]Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland. [130]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, Singapore. [131]Department of Surgery, National University Health System, Singapore, Singapore. [132]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, the Netherlands. [133]School of Population and Global Health, The University of Western Australia, Perth, Western Australia, Australia. [134]Family Cancer Clinic, The Netherlands Cancer Institute – Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [135]Division of Cancer Sciences, University of Manchester, Manchester, UK. [136]Center for Medical Genetics, NorthShore University HealthSystem, Evanston, IL, USA. [137]The University of Chicago Pritzker School of Medicine, Chicago, IL, USA. [138]N.N. Petrov Institute of Oncology, St. Petersburg, Russia. [139]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA. [140]Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo, Japan. [141]Research Centre for Genetic Engineering and Biotechnology 'Georgi D. Efremov', Macedonian Academy of Sciences and Arts, Skopje, Republic of North Macedonia. [142]Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, Poland. [143]Parkville Familial Cancer Centre, Peter MacCallum Cancer Center, Melbourne, Victoria, Australia. [144]Hematology, Oncology and Transfusion Medicine Center, Department of Molecular and Regenerative Medicine, Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania. [145]State Research Institute Centre for Innovative Medicine, Vilnius, Lithuania. [146]Department of Medicine, Division of Hematology/Oncology, UPMC Hillman Cancer Center, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [147]Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA. [148]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK. [149]Department of Oncology, Tampere University Hospital, Tampere University and Tampere Cancer Center, Tampere, Finland. [150]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [151]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, Korea. [152]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, Korea. [153]Cancer Research Institute, Seoul National University, Seoul, Korea. [154]Faculty of Medicine, University of Heidelberg, Heidelberg, Germany.

[155]David Geffen School of Medicine, Department of Obstetrics and Gynecology, University of California, Los Angeles, Los Angeles, CA, USA. [156]Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [157]Surgery, School of Medicine, National University of Ireland, Galway, Ireland. [158]Department of Genetics and Fundamental Medicine, Bashkir State Medical University, Ufa, Russia. [159]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland. [160]Familial Cancer Service, Weatmead Hospital, Sydney, New South Wales, Australia. [161]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA. [162]Department of Internal Medicine, Evangelische Kliniken Bonn, Johanniter Krankenhaus, Bonn, Germany. [163]Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research 'Demokritos', Athens, Greece. [164]Division of Cancer Epidemiology and Genetics, Department of Health and Human Services, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA. [165]Department of Histopathology and Cytology, Clinical Hospital 'Acibadem Sistina', Skopje, Republic of North Macedonia. [166]Hong Kong Hereditary Breast Cancer Family Registry, Cancer Genetics Centre, Happy Valley, Hong Kong. [167]Department of Surgery, The University of Hong Kong, Pok Fu Lam, Hong Kong. [168]Department of Surgery, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong. [169]VIB Center for Cancer Biology, VIB, Leuven, Belgium. [170]Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, Belgium. [171]Institut Curie, Paris, France. [172]Mines ParisTech, Paris, France. [173]Genetic Epidemiology of Cancer Team, INSERM U900, Paris, France. [174]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden. [175]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, Sweden. [176]Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, Faculty of Biology, Medicine and Health, The University of Manchester, Manchester, UK. [177]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, Sweden. [178]Department of Family Medicine and Public Health, University of California, San Diego, La Jolla, CA, USA. [179]Division of Cancer Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Japan. [180]Division of Cancer Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, Japan. [181]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [182]Anatomical Pathology, The Alfred Hospital, Melbourne, Victoria, Australia. [183]Department of Tumour Biology, INSERM U830, Paris, France. [184]Department of Gynecology and Obstetrics, University of Munich, Munich, Germany. [185]NRG Oncology, Statistics and Data Management Center, Roswell Park Cancer Institute, Buffalo, NY, USA. [186]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [187]Laboratory Medicine Program, University Health Network, Toronto, Ontario, Canada. [188]Radiation Oncology, Hospital Meixoeiro-XXI de Vigo, Vigo, Spain. [189]Women's College Research Institute, University of Toronto, Toronto, Ontario, Canada. [190]Department of Pathology, School of Medicine, Umm Al-Qura University, Holy Makkah, Saudi Arabia. [191]Center for Genomic Medicine at Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark. [192]Latvian Biomedical Research and Study Centre, Riga, Latvia. [193]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA. [194]Clinical Genetics Research Laboratory, Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [195]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [196]Department of Molecular Genetics, National Institute of Oncology, Budapest, Hungary. [197]Center for Clinical Cancer Genetics, The University of Chicago, Chicago, IL, USA. [198]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden. [199]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, UK. [200]University of New Mexico Health Sciences Center, University of New Mexico, Albuquerque, NM, USA. [201]Cancer Research UK Clinical Trials Unit, Institute of Cancer Sciences, University of Glasgow, Glasgow, UK. [202]Molecular Diagnostics, Aalborg University Hospital, Aalborg, Denmark. [203]Clinical Cancer Research Center, Aalborg University Hospital, Aalborg, Denmark. [204]Department of Clinical Medicine, Aalborg University, Aalborg, Denmark. [205]Genome Diagnostics Program, IFOM—the FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan, Italy. [206]ProCURE, Catalan Institute of Oncology, Bellvitge Biomedical Research Institute (IDIBELL), Barcelona, Spain. [207]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland. [208]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland. [209]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, Italy. [210]School of Women's and Children's Health, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia. [211]The Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [212]Clinical Genetics, Karolinska Institutet, Stockholm, Sweden. [213]Clalit National Israeli Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, Israel. [214]Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, USA. [215]Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain. [216]Department of Oncology, University Hospital of Larissa, Larissa, Greece. [217]Epidemiology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA. [218]Fundación Pública Galega de Medicina Xenómica, Santiago de Compostela, Spain. [219]Instituto de Investigación Sanitaria de Santiago de Compostela, Santiago de Compostela, Spain. [220]School of Medicine, University of Western Australia, Perth, Western Australia, Australia. [221]Research Oncology, Guy's Hospital, King's College London, London, UK. [222]Faculty of Information Technology, Monash University, Melbourne, Victoria, Australia. [223]National Center for Tumor Diseases, University Hospital and German Cancer Research Center, Heidelberg, Germany. [224]Network Aging Research, University of Heidelberg, Heidelberg, Germany. [225]Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, New South Wales, Australia. [226]Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Newcastle, New South Wales, Australia. [227]Hunter Medical Research Institute, John Hunter Hospital, Newcastle, New South Wales, Australia. [228]Clinical Cancer Genetics Program, Division of Human Genetics, Department of Internal Medicine, The Comprehensive Cancer Center, The Ohio State University, Columbus, OH, USA. [229]Department of Internal Medicine, Division of Medical Oncology, University of Kansas Medical Center, Westwood, KS, USA. [230]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan. [231]School of Public Health, China Medical University, Taichung, Taiwan. [232]Department of Obstetrics and Gynecology and Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria. [233]Clinical Cancer Genomics, City of Hope, Duarte, CA, USA. [234]University Clinic of Radiotherapy and Oncology, Medical Faculty, Ss. Cyril and Methodius University in Skopje, Skopje, Republic of North Macedonia. [235]Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, Australia. [236]Population Oncology, BC Cancer, Vancouver, British Columbia, Canada. [237]School of Population and Public Health, University of British Columbia, Vancouver, British Columbia, Canada. [238]The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Western Australia, Australia. [239]Service de Génétique, Institut Curie, Paris, France. [240]Université Paris Descartes, Paris, France. [241]Institute of Human Genetics, University Hospital Heidelberg, Heidelberg, Germany. [242]Division of Breast Cancer Research, The Institute of Cancer Research, London, UK. [243]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [244]Department of Obstetrics and Gynecology, Medical University of Vienna, Vienna, Austria. [245]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, NC, USA. [246]Department of Genetics, Portuguese Oncology Institute, Porto, Portugal. [247]Biomedical Sciences Institute (ICBAS), University of Porto, Porto, Portugal. [248]Cancer Center, Kuopio University Hospital, Kuopio, Finland. [249]Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio, Finland. [250]Breast Cancer Research Programme, Cancer Research Malaysia, Kuala Lumpur, Malaysia. [251]Department of Surgery, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia. [252]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA. [253]Hereditary Cancer Program, ONCOBELL-IDIBELL-IDIBGI-IGTP, Catalan Institute of Oncology, CIBERONC, Barcelona, Spain. [254]Department of Clinical Genetics, Odense University Hospital, Odence, Denmark. [255]Department of Medicine, Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA. [256]Department of Medical Genetics, University of Cambridge,

Cambridge, UK. [257]Department of Cancer Biology and Genetics, The Ohio State University, Columbus, OH, USA. [258]Department of Surgery, Leiden University Medical Center, Leiden, the Netherlands. [259]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, UK. [260]Wellcome Trust Centre for Human Genetics and NIHR Oxford Biomedical Research Centre, University of Oxford, Oxford, UK. [261]Center for Population Health Research, National Institute of Public Health, Cuernavaca, Mexico. [262]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. [263]Department of Medical Oncology, Beth Israel Deaconess Medical Center, Boston, MA, USA. [264]Department of Pathology, University Hospital of Heraklion, Heraklion, Greece. [265]Frauenklinik der Stadtklinik Baden-Baden, Baden-Baden, Germany. [266]Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, USA. [267]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. [268]Department of Genetics, University of Pretoria, Pretoria, South Africa. [269]Fundación Pública Galega de Medicina Xenómica-SERGAS, Grupo de Medicina Xenómica-USC, CIBERER, IDIS, Santiago de Compostela, Spain. [270]Division of Functional Onco-genomics and Genetics, Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, Aviano, Italy. [271]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, USA. [272]Department of Surgical Sciences, Uppsala University, Uppsala, Sweden. [273]Nuffield Department of Population Health, University of Oxford, Oxford, UK. [274]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK. [275]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute–Antoni van Leeuwenhoek Hospital, Amsterdam, the Netherlands. [276]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, Norway. [277]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway. [278]The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON) Coordinating Center, The Netherlands Cancer Institute, Amsterdam, the Netherlands. [279]Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, Australia. [280]Present address: UK Dementia Research Institute, Imperial College London, London, UK. [281]These authors jointly supervised this work: Georgia Chenevix-Trench, Jacques Simard, Douglas F. Easton, Peter Kraft, Alison M. Dunning. [282]These authors contributed equally: Hugues Aschard, Jonathan Beesley. *e-mail: pkraft@hsph.harvard.edu; amd24@medschl.cam.ac.uk

GEMO Study Collaborators

Véronique Mari[283], Pascaline Berthet[284], Laurent Castera[284], Dominique Vaur[284], Hakima Lallaoui[285], Yves-Jean Bignon[286], Nancy Uhrhammer[286], Valérie Bonadona[287], Christine Lasset[287], Françoise Révillion[288], Paul Vennin[288], Daniele Muller[289], Denise Molina Gomes[290], Olivier Ingster[291], Isabelle Coupier[292], Pascal Pujol[292], Marie-Agnès Collonge-Rame[293], Isabelle Mortemousque[294], Odile Bera[295], Mickaelle Rose[295], Amandine Baurand[296], Geoffrey Bertolone[296], Laurence Faivre[296], Hélène Dreyfus[297], Dominique Leroux[297], Laurence Venat-Bouvet[298], Stéphane Bézieau[299], Capucine Delnatte[299], Jean Chiesa[300], Brigitte Gilbert-Dussardier[301], Paul Gesta[302], Fabienne Prieur Prieur[303], Myriam Bronner[304], Johanna Sokolowska[304], Florence Coulet[305], Nadia Boutry-Kryza[306], Alain Calender[306], Sophie Giraud[306], Mélanie Leone[306], Sandra Fert-Ferrer[307], Dominique Stoppa-Lyonnet[308], Yue Jiao[309], Fabienne Lesueur Lesueur[309], Noura Mebirouk[309], Emmanuelle Barouk-Simonet[310], Virginie Bubien[310], Michel Longy[310], Nicolas Sevenet[310], Laurence Gladieff[311], Christine Toulas[311], Audrey Reimineras[312], Hagay Sobol[312], Brigitte Bressac-de Paillerets[313], Odile Cabaret[313], Olivier Caron[313], Marine Guillaud-Bataille[313], Etienne Rouleau[313], Muriel Belotti[241], Bruno Buecher[241], Sandrine Caputo[241], Chrystelle Colas[314], Antoine De Pauw[241], Emmanuelle Fourme[315], Marion Gauthier-Villars[241], Lisa Golmard[241], Virginie Moncoutier[241] and Claire Saule[241]

[283]Centre Antoine Lacassagne, Nice, France. [284]Centre François Baclesse, Caen, France. [285]Centre Hospitalier de La Rochelle, La Rochelle, France. [286]Centre Jean Perrin, Clermont-Ferrand, France. [287]Centre Léon Bérard, Lyon, France. [288]Centre Oscar Lambret, Lille, France. [289]Centre Paul Strauss, Strasbourg, France. [290]CHI Poissy, Poissy, France. [291]CHU Angers, Angers, France. [292]CHU Arnaud-de-Villeneuve, Montpellier, France. [293]CHU Besançon, Besançon, France. [294]CHU Bretonneau, Tours and Centre Hospitalier de Bourges, Orléans, France. [295]CHU de Martinique, Fort de France, France. [296]CHU Dijon, Dijon, France. [297]CHU Grenoble, Grenoble, France. [298]CHU Limoges, Limoges, France. [299]CHU Nantes, Nantes, France. [300]CHU Nîmes Carémeau, Nîmes, France. [301]CHU Poitiers, Poitiers, France. [302]CHU Poitiers, Centre Hospitalier d'Angoulême and Centre Hospitalier de Niort, Poitiers, France. [303]CHU Saint Etienne, Saint Etienne, France. [304]CHU Vandoeuvre-les-Nancy, Vandoeuvre-les-Nancy, France. [305]Groupe Hospitalier Pitié-Salpétrière, Paris, France. [306]Centre Léon Bérard, Hospices Civils de Lyon, Lyon, France. [307]Hôtel Dieu Centre Hospitalier, Chambéry, France. [308]Inserm, U830, Service de Génétique, Institut Curie, Université Paris Descartes, Paris, France. [309]Inserm, U900, Institut Curie, PSL University, Mines ParisTech, Paris, France. [310]Institut Bergonié, Bordeaux, France. [311]Institut Claudius Regaud, Toulouse, France. [312]Institut Paoli Calmettes, Marseille, France. [313]Service de Génétique, Gustave Roussy, VIllejuif, France. [314]Service de Génétique, Institut Curie, Paris, France. [315]Service de Génétique, Institut Curie, Paris, France.

EMBRACE Collaborators

Alan Donaldson[316], Alex Murray[317], Angela Brady[318], Carole Brewer[319], Caroline Pottinger[320], Clare Miller[321], David Gallagher[322], Helen Gregory[323], Jackie Cook[324], Jacqueline Eason[325], Julian Adlard[326], Julian Barwell[327], Kai-Ren Ong[328], Katie Snape[329], Lisa Walker[330], Louise Izatt[331], Lucy Side[332], Marc Tischkowitz[333], Mark T. Rogers[334], Mary E. Porteous[335], Munaza Ahmed[336], Patrick J. Morrison[337], Paul Brennan[338], Ros Eeles[339] and Rosemarie Davidson[340]

[316]Clinical Genetics Department, St Michael's Hospital, Bristol, UK. [317]All Wales Medical Genetics Service, Singleton Hospital, Swansea, UK. [318]North West Thames Regional Genetics Service, Kennedy Galton Centre, The North West London Hospitals NHS Trust, Harrow, UK. [319]Department of Clinical Genetics, Royal Devon and Exeter Hospital, Exeter, UK. [320]All Wales Medical Genetics Service, Glan Clwyd Hospital, Rhyl, UK. [321]Department of Clinical Genetics, Alder Hey Hospital, Liverpool, UK. [322]Mater Private Hospital, Dublin, UK. [323]North of Scotland Regional Genetics Service, NHS Grampian and University of Aberdeen, Aberdeen, UK. [324]Sheffield Clinical Genetics Service, Sheffield Children's Hospital, Sheffield, UK. [325]Nottingham Clinical Genetics Service, Nottingham University Hospitals NHS Trust, Notingham, UK. [326]Yorkshire Regional Genetics Service, Chapel Allerton Hospital, Leeds, UK. [327]Leicestershire Clinical Genetics Service, University Hospitals of Leicester NHS Trust, Leicester, UK. [328]West Midlands Regional Genetics Service, Birmingham Women's Hospital Healthcare NHS Trust, Birmingham, UK. [329]Medical Genetics Unit, St George's, University of London, London, UK. [330]Oxford Regional Genetics Service, Churchill Hospital, Oxford, UK. [331]Clinical Genetics, Guy's and St. Thomas' NHS Foundation Trust, London, UK. [332]Wessex Clinical Genetics Service, The Princess Anne Hospital, Southampton, UK. [333]Department of Medical Genetics, Addenbrooke's Treatment Centre, Addenbrooke's Hospital, Cambridge, UK. [334]All Wales Medical Genetics Service, University Hospital of Wales, Cardiff, UK. [335]South East of Scotland Regional Genetics Service, Western General Hospital, Edinburgh, UK. [336]North East Thames Regional Genetics Service, Great Ormond Street Hospital for Children NHS Trust, London, UK. [337]Northern Ireland Regional Genetics Centre, Belfast City Hospital, Belfast, UK. [338]Institute of General Medicine, International Centre for Life, Newcastle, UK. [339]Oncogenetics Team, The Institute of Cancer Research and Royal Marsden NHS Foundation Trust, Sutton, UK. [340]Department of Clinical Genetics, South Glasgow University Hospitals, Glasgow, UK.

KConFab Investigators

Adrienne Sexton[341], Alice Christian[342], Alison Trainer[343], Allan Spigelman[344], Andrew Fellows[345], Andrew Shelling[346], Anna De Fazio[347], Anneke Blackburn[348], Ashley Crook[349], Bettina Meiser[350], Briony Patterson[351], Christine Clarke[352], Christobel Saunders[353], Clare Hunt[354], Clare Scott[355], David Amor[356], Deb Marsh[357], Edward Edkins[358], Elizabeth Salisbury[359], Eric Haan[360], Eveline Neidermayr[343], Finlay Macrea[361], Gelareh Farshid[362], Geoff Lindeman[363], Georgia Trench[364], Graham Mann[365], Graham Giles[366], Grantley Gill[367], Heather Thorne[343], Ian Campbell[368], Ian Hickie[369], Ingrid Winship[370], James Flanagan[371], James Kollias[372], Jane Visvader[373], Jennifer Stone[374], Jessica Taylor[375], Jo Burke[376], Jodi Saunus[377], John Forbes[378], John Hopper[379], Jonathan Beesley[380], Judy Kirk[381], Juliet French[382], Kathy Tucker[383], Kathy Wu[384], Kelly Phillips[385], Lara Lipton[386], Leslie Andrews[387], Lizz Lobb[388], Logan Walker[389], Maira Kentwell[390], Mandy Spurdle[391], Margaret Cummings[392], Margaret Gleeson[393], Marion Harris[394], Mark Jenkins[395], Mary Anne Young[396], Martin Delatycki[356], Mathew Wallis[397], Matthew Burgess[398], Melanie Price[399], Melissa Brown[400], Melissa Southey[401], Michael Bogwitz[402], Michael Field[403], Michael Friedlander[404], Michael Gattas[405], Mona Saleh[406], Nick Hayward[407], Nick Pachter[408], Paul Cohen[409], Pascal Duijf[410], Paul James[411], Pete Simpson[412], Peter Fong[413], Phyllis Butow[414], Rachael Williams[344], Rick Kefford[415], Rodney Scott[416], Roger Milne[366], Rosemary Balleine[417], Sarah-Jane Dawson[345], Sheau Lok[418], Shona O'Connell[419], Sian Greening[420], Sophie Nightingale[421], Stacey Edwards[380], Stephen Fox[422], Sue-Anne McLachlan[423], Sunil Lakhani[424] and Yoland Antill[425]

[341]Familial Cancer Centre, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [342]Genetics Department, Central Region Genetics Service, Wellington Hospital, Wellington, New Zealand. [343]The Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [344]Family Cancer Clinic, St Vincent's Hospital, Sydney, New South Wales, Australia. [345]The Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [346]Obstetrics and Gynaecology, University of Auckland, Auckland, New Zealand. [347]Department of Gynaecological Oncology, Westmead Institute for Cancer Research, Westmead Hospital, Sydney, New South Wales, Australia. [348]Australian National University, Canberra, Australian Capital Territory, Australia. [349]Department of Clinical Genetics, Royal North Shore Hospital, Sydney, New South Wales, Australia. [350]Prince of Wales Hospital, The Uiversity of New South Wales, Sydney, New South Wales,

Australia. [351]Clinical Genetics Service, Royal Hobart Hospital, Hobart Tasmania, Australia. [352]Westmead Institute for Cancer Research, University of Sydney, Westmead Hospital, Sydney, New South Wales, Australia. [353]School of Surgery and Pathology, Queen Elizabeth II Medical Centre, Perth, Western Australia, Australia. [354]Southern Health Familial Cancer Centre, Monash Medical Centre, Melbourne, Victoria, Australia. [355]The Walter and Eliza Hall Institute of Medical Research, c/o Royal Melbourne Hospital, Melbourne, Victoria, Australia. [356]Genetic Health Services Victoria, Royal Children's Hospital, Melbourne, Victoria, Australia. [357]Kolling Institute of Medical Research, Royal North Shore Hospital, Sydney, New South Wales, Australia. [358]Clinical Chemistry, Princess Margret Hospital for Children, Perth, Western Australia, Australia. [359]Anatomical Pathology, Prince of Wales Hospital, Sydney, New South Wales, Australia. [360]Department of Medical Genetics, Women's and Children's Hospital, Adelaide, South Australia, Australia. [361]Family Cancer Clinic, The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [362]SA Tissue Pathology, IMVS, Adelaide, South Australia, Australia. [363]Breast Cancer Laboratory, The Walter and Eliza Hall Institute of Medical Research, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [364]Queensland Institute of Medical Research, Royal Brisbane Hospital, Brisbane, Queensland, Australia. [365]Westmead Institute for Cancer Research, Westmead Millennium Institute, Sydney, New South Wales, Australia. [366]Anti-Cancer Council of Victoria, Melbourne, Victoria, Australia. [367]Department of Surgery, Royal Adelaide Hospital, Adelaide, South Australia, Australia. [368]Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [369]Brain and Mind Centre, Sydney, New South Wales, Australia. [370]Department of Medicine, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [371]Epigenetics Unit, Department of Surgery and Oncology, Imperial College, London, England. [372]Breast Endocrine and Surgical Unit, Royal Adelaide Hospital, Adelaide, South Australia, Australia. [373]The Walter and Eliza Hall Institute of Medical Research, c/o Royal Melbourne Hospital, Melbourne, Victoria, Australia. [374]Centre for Genetic Origins of Health and Disease, University of Western Australia, Perth, Western Australia, Australia. [375]Familial Cancer and Genetics Medicine, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [376]Royal Hobart Hospital, Hobart, Tasmania, Australia. [377]Breast Pathology, University of Queensland Centre for Clinical Research, Royal Brisbane and Women's Hospital, Brisbane, Queensland, Australia. [378]Surgical Oncology, University of Newcastle, Newcastle Mater Hospital, Newcastle, New South Wales, Australia. [379]Centre for MEGA Epidemiology, Melbourne, Victoria, Australia. [380]Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [381]Familial Cancer Service, Department of Medicine, Westmead Hospital, Sydney, New South Wales, Australia. [382]School of Molecular and Microbial Sciences, University of Queensland, Brisbane, Queensland, Australia. [383]Heredity Cancer Clinic, Prince of Wales Hospital, Sydney, New South Wales, Australia. [384]Family Cancer Clinic, St Vincent's Hospital, Sydney, Australia. [385]Department of Medical Oncology, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [386]Medical Oncology and Clinical Haematology Unit, Western Hospital, Melbourne, Victoria, Australia. [387]Hereditary Cancer Clinic, Prince of Wales Hospital, Sydney, New South Wales, Australia. [388]School of Medicine, University of Notre Dame, Sydney, New South Wales, Australia. [389]Department of Pathology, University of Otago, Christchurch, New Zealand. [390]The Family Cancer Clinic, Royal Melbourne Hospital, Melbourne, Victoria, Australia. [391]Queensland Institute of Medical Research, Brisbane, Queensland, Australia. [392]Department of Pathology, University of Queensland Medical School, Brisbane, Queensland, Australia. [393]Hunter Family Cancer Service, Sydney, New South Wales, Australia. [394]Family Cancer Clinic, Monash Medical Centre, Melbourne, Victoria, Australia. [395]Centre for MEGA Epidemiology, The University of Melbourne, Melbourne, Victoria, Australia. [396]The Garvan Institute of Medical Research, The Kinghorn Cancer Centre, Sydney, New South Wales, Australia. [397]The Family Cancer Clinic, Austin Health, Melbourne, Victoria, Australia. [398]Clinical Genetics Service, Austin Health, Melbourne, Victoria, Australia. [399]Medical Psychology, University of Sydney, Sydney, New South Wales, Australia. [400]University of Queensland, Brisbane, Queensland, Australia. [401]Genetic Epidemiology Laboratory, Department of Pathology, The University of Melbourne, Melbourne, Victoria, Australia. [402]Familial Cancer Centre, The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [403]Clinical Genetics, Royal North Shore Hospital, Sydney, New South Wales, Australia. [404]Department of Medical Oncology, Prince of Wales Hospital, Sydney, New South Wales, Australia. [405]Queensland Clinical Genetic Service, Royal Children's Hospital, Brisbane, Queensland, Australia. [406]Centre for Genetic Education, Prince of Wales Hospital, Sydney, New South Wales, Australia. [407]Queensland Institute for Medical Research, Royal Brisbane Hospital, Brisbane, Queensland, Australia. [408]Genetic Services of WA, King Edward Memorial Hospital, Perth, Western Australia, Australia. [409]Gynaecological Cancer Research, St John of God Subiaco Hospital, Perth, Western Australia, Australia. [410]The University of Queensland Diamantina Institute, Brisbane, Queensland, Australia. [411]Family Cancer Clinic, The Peter MacCallum Cancer Centre, Melbourne, Australia. [412]The University of Queensland, Brisbane, Queensland, Australia. [413]Regional Cancer and Blood Services Auckland City Hospital, Auckland, New Zealand. [414]Medical Psychology Unit, Royal Prince Alfred Hospital, Sydney, New South Wales, Australia. [415]Medical Oncology Department, Westmead Hospital, Sydney, New South Wales, Australia. [416]Hunter Area Pathology Service, John Hunter Hospital, Newcastle, New South Wales, Australia. [417]Department of Medical Oncology, Westmead Hospital, Sydney, New South Wales, Australia. [418]Department of Medical Oncology, The Royal Melbourne Hospital, Melbourne, Victoria, Australia. [419]Southern Health Familial Cancer Centre, Melbourne, Victoria, Australia. [420]Illawarra Cancer Centre, Wollongong Hospital, Wollongong, New South Wales, Australia. [421]Western Health and Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [422]Pathology Department, Peter MacCallum Cancer Centre, VCCC Building, Melbourne, Victoria, Australia. [423]Department of Oncology, St Vincent's Hospital, Melbourne, Victoria, Australia. [424]UQ Centre for Clinical Research, University of Queensland, The Royal Brisbane and Women's Hospital Herston, Brisbane, Queensland, Australia. [425]The Family Cancer Clinic, Cabrini Hospital, Melbourne, Victoria, Australia.

HEBON Investigators

Cora Aalfs[426], Hanne Meijers-Heijboer[427], Klaartje van Engelen[426], Hans Gille[427], Ingrid Boere[428], Margriet Collée[429], Carolien van Deurzen[430], Maartje Hooning[428], Inge-Marie Obdeijn[431], Ans van den Ouweland[429], Caroline Seynaeve[428], Sabine Siesling[432], Janneke Verloop[432], Christi van Asperen[433], Peter Devilee[434], Twiggy van Cronenburg[433], Rien Blok[435], Maaike de Boer[435], Encarna Gómez Garcia[435], Muriel Adank[436], Frans Hogervorst[436], Denise Jenner[437], Flora van Leeuwen[437], Matti Rookus[437], Nicola Russell[438], Marjanka Schmidt[439], Sandra van den Belt-Dusebout[440], Carolien Kets[441], Arjen Mensenkamp[441], Truuske de Bock[442], Annemieke van der Hout[443], Marian Mourits[444], Jan Oosterwijk[443], Margreet Ausems[445] and Marco Koudijs[445]

[426]Department of Clinical Genetics, Academic Medical Center, Amsterdam, the Netherlands. [427]Department of Clinical Genetics, VU University Medical Centre, Amsterdam, the Netherlands. [428]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, the Netherlands. [429]Department of Clinical Genetics, Family Cancer Clinic, Erasmus University Medical Center, Rotterdam, the Netherlands.

[430]Department of Pathology, Family Cancer Clinic, Erasmus University Medical Center, Rotterdam, the Netherlands. [431]Department of Radiology, Family Cancer Clinic, Erasmus University Medical Center, Rotterdam, the Netherlands. [432]The Netherlands Comprehensive Cancer Organisation (IKNL), Utrecht, the Netherlands. [433]Department of Clinical Genetics, Leiden University Medical Center, Leiden, the Netherlands. [434]Department of Human Genetics and Department of Pathology, Leiden University Medical Center, Leiden, the Netherlands. [435]Department of Clinical Genetics, Maastricht University Medical Center, Maastricht, the Netherlands. [436]Family Cancer Clinic, Netherlands Cancer Institute, Amsterdam, the Netherlands. [437]Department of Epidemiology, Netherlands Cancer Institute, Amsterdam, the Netherlands. [438]Department of Radiotherapy, Netherlands Cancer Institute, Amsterdam, the Netherlands. [439]Division of Psychosocial Research and Epidemiology, Netherlands Cancer Institute, Amsterdam, the Netherlands. [440]The Nationwide Network and Registry of Histo- and Cytopathology (PALGA), Houten, the Netherlands. [441]Department of Human Genetics, Radboud University Medical Center, Nijmegen, the Netherlands. [442]Department of Oncological Epidemiology, University Medical Center, Groningen University, Groningen, the Netherlands. [443]Department of Genetics, University Medical Center Groningen, Groningen, the Netherlands. [444]Department of Gynaecological Oncology, University Medical Center, Groningen University, Groningen, the Netherlands. [445]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, the Netherlands.

ABCTB Investigators

Christine Clarke[446], Deborah Marsh[447], Rodney Scott[448,449], Robert Baxter[450], Desmond Yip[451,452], Jane Carpenter[453], Alison Davis[454,455], Nirmala Pathmanathan[456,457], Peter Simpson[458], Dinny Graham[446] and Mythily Sachchithananthan[446]

[446]Centre for Cancer Research, The Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. [447]University of Technology Sydney, Translational Oncology Group, School of Life Sciences, Faculty of Science, Sydney, New South Wales, Australia. [448]School of Biomedical Sciences, University of Newcastle, Newcastle, UK. [449]Hunter Medical Research Institute and NSW Health Pathology North, Newcastle, Australia. [450]Kolling Institute of Medical Research, University of Sydney, Sydney, New South Wales, Australia. [451]Epigenetics and Transcription Laboratory, Melanie Swan Memorial Translational Centre, Sci-Tech, University of Canberra, Canberra, Australian Capital Territory, Australia. [452]Department of Medical Oncology, The Canberra Hospital, Canberra, Australian Capital Territory, Australia. [453]Scientific Platforms, The Westmead Institute for Medical Research, The University of Sydney, Sydney, New South Wales, Australia. [454]The Canberra Hospital, Canberra, Australian Capital Territory, Australia. [455]The Australian National University, Canberra, Australian Capital Territory, Australia. [456]Westmead Breast Cancer Institute, Western Sydney Local Health District, Sydney, New South Wales, Australia. [457]University of Sydney, Western Clinical School, Sydney, New South Wales, Australia. [458]UQ Centre for Clinical Research, Faculty of Medicine, The University of Queensland, Brisbane, Queensland, Australia.

Methods

Study samples. Epidemiological data for European women were obtained from 75 breast cancer case–control studies participating in the BCAC (cases: 40,285 iCOGS and 69,615 OncoArray; cases with ER status available: 29,561 iCOGS and 55,081 OncoArray; controls: 38,058 iCOGS and 50,879 OncoArray). Details of the participating studies, genotype calling and quality control are given in refs. [2,22,23], respectively. Epidemiological data for *BRCA1* mutation carriers were obtained from 60 studies providing data to the CIMBA (affected: 1,591 iCOGS and 7,772 OncoArray; unaffected: 1,665 iCOGS and 7,780 OncoArray). This dataset has been described in detail previously[1,59,60]. All studies provided samples of European ancestry. Any non-European samples were excluded from the analyses.

Variant selection and genotyping. Similar approaches were used to select variants for inclusion on the iCOGS and OncoArray, and these are described in detail elsewhere[2,21]. Both arrays included a dense coverage of variants across known susceptibility regions (at the time of their design), with sparser coverage of the rest of the genome. Twenty-one known susceptibility regions were selected for dense genotyping using iCOGS and 73 regions were selected for OncoArray. These regions were 1-megabase (Mb) intervals centered on the published lead GWAS hit (combined into larger intervals where these overlapped). For iCOGS, all known variants from the March 2010 release of the 1000 Genomes Project with a MAF > 0.02 in Europeans were identified, and all those correlated with the published GWAS variants at $r^2 > 0.1$ (r^2, Pearson's squared correlation coefficient), together with a set of variants designed to tag all remaining variants at $r^2 > 0.9$, were selected to be included in the array (http://ccge.medschl.cam.ac.uk/files/2014/03/iCOGS_detailed_lists_ALL1.pdf). For OncoArray, all designable variants correlated with the known hits at $r^2 > 0.6$, plus all variants from lists of potentially functional variants on RegulomeDB and a set of variants designed to tag all of the remaining variants at $r^2 > 0.9$, were selected. In total, across the 152 regions considered here, 26,978 iCOGS- and 58,339 OncoArray-genotyped variants passed the quality control criteria.

We imputed genotypes for all of the remaining variants by using IMPUTE2 (ref. [61]) and the October 2014 release of the 1000 Genomes Project as a reference. Imputation was conducted independently in the iCOGS and OncoArray subsets. To improve accuracy at low-frequency variants, we used the standard IMPUTE2 MCMC algorithm for follow-up imputation, which includes no pre-phasing of the genotypes and increased both the buffer regions and the number of haplotypes to use as templates (a more detailed description of the parameters used can be found in ref. [21]). We thus genotyped or successfully imputed 639,118 variants (all with an imputation info score ≥ 0.3 and a MAF ≥ 0.001 in both iCOGS and OncoArray datasets). Imputation summaries and coverage for each of the analyzed regions stratified by allele frequency can be found in Supplementary Table 1b.

BCAC statistical analyses. Per-allele odds ratios and standard errors were estimated for each variant using logistic regression. We ran this analysis separately for iCOGS and OncoArray, and for overall, ER-positive and ER-negative breast cancer. The association between each variant and breast cancer risk was adjusted by study (iCOGS) or country (OncoArray), and eight (iCOGS) or ten (OncoArray) ancestry-informative principal components. The statistical significance for each variant was derived using a Wald test.

Defining appropriate significance thresholds for association signals. To establish an appropriate significance threshold for independent signals, all variants evaluated in the meta-analysis were included in logistic forward selection regression analyses for overall breast cancer risk in iCOGS, run independently for each region. We evaluated five *P* value thresholds for inclusion: $<1\times10^{-4}$, $<1\times10^{-5}$, $<1\times10^{-6}$, $<1\times10^{-7}$ and $<1\times10^{-8}$. The most parsimonious iCOGS models were tested in OncoArray, and the FDR at the 1% level for each threshold was estimated using the Benjamini–Hochberg procedure. At a 1% FDR threshold, 72% of associations, significant at $P<10^{-4}$, were replicated on iCOGS, and 94% of associations, significant at $P<10^{-6}$, were replicated on OncoArray. Based on these results, two categories were defined: strong-evidence signals (conditional $P<10^{-6}$ in the final model) and moderate-evidence signals (conditional $P<10^{-4}$ and $P\geq10^{-6}$ in the final model).

Identification of independent signals. To identify independent signals, we ran multinomial stepwise regression analyses, separately in iCOGS and OncoArray, for all variants displaying evidence of association ($n_{variants}=202,749$). We selected two sets of well-imputed variants (imputation info score ≥ 0.3 in both iCOGS and OncoArray): (1) common and low-frequency variants (MAF ≥ 0.01) with a logistic regression *P* value inclusion threshold of ≤0.05 in either the iCOGS or OncoArray datasets for at least one of the three phenotypes (overall, ER positive and ER negative breast cancer); and (2) rarer variants (MAF ≥ 0.001 and <0.01), with a logistic regression inclusion *P* value of ≤ 0.0001. The same parameters used for adjustment in logistic regression were used in the multinomial regression analysis (R function multinom), which simultaneously estimated per-allele odds ratios for ER-positive and ER-negative breast cancer. The multinomial regression estimates were combined using a fixed-effects meta-analysis weighted by the inverse variance. Variants with the lowest conditional *P* value from the meta-analysis of both European cohorts at each step were included in the multinomial regression model. However, if the new variant to be included in the model caused collinearity problems due to high correlation with an already selected variant, or showed high heterogeneity ($P<10^{-4}$) between iCOGS and OncoArray after being conditioned by the variant(s) in the model, we dropped the new variant and repeated this process.

At 105 of 152 evaluated regions, the main signal showed genome-wide significance, while 44 were marginally significant ($9.89\times10^{-5}\geq P>5\times10^{-8}$). For two regions, there were no variants significant at $P<10^{-4}$ (chr14:104712261–105712261; rs10623258; multinomial regression $P=2.32\times10^{-4}$; chr19:10923703–11923703; rs322144; multinomial regression $P=3.90\times10^{-3}$). Four main differences in the datasets used here and in the previous paper may account for this: (1) our previous paper[2] included data from 11 additional GWASs (14,910 cases and 17,588 controls) that have not been included in the present analysis in order to minimize differences in array coverage, and because ER status data were substantially incomplete and individual-level data were not available for all GWASs; (2) the present analysis was based on estimating separate risks for ER-positive and ER-negative disease, whereas in our previous paper the outcome was overall breast cancer risk; ER status was available for only 73% of the iCOGS and 79% of the OncoArray breast cancer cases; (3) for the set of samples genotyped with both arrays, ref. [2] used the iCOGS genotypes, while the present study included OncoArray genotypes to maximize the number of samples genotyped with a larger coverage; and (4) the imputation procedure was modified (in particular using one-step imputation without pre-phasing) to improve the imputation accuracy of less frequent variants.

We used a forward stepwise approach to define the number of independent signals within each associated genomic region. First, we identified the index variant of the main signal in the region, and then ran multinomial logistic regression for all of the other variants, adjusted by the index variant, to identify additional variants that remained independently significant within the model. We repeated this process, adjusting for identified index variants, until no more additional variants could be added. In this way, we found from 1–11 independent signals within the 150 regions that containing a genome-wide significant main signal.

Selection of a set of CCVs. For each independently associated signal, we first defined CCVs likely to drive its association as those variants with *P* values within two orders of magnitude of the most significant variant for that signal, after adjusting for the index variant of other signals within that region (as identified in the forward stepwise regression above; Supplementary Fig. 6a)[24]. For each region, we then attempted to obtain the best-fitting model by successively fitting models in which the index variant for each signal was replaced by other CCVs for that signal, adjusting for the index variants for the other signals (Supplementary Fig. 6b). Where a model with a higher chi-squared value was obtained, the index variant was replaced by the CCV in the best model (Supplementary Fig. 6c,d). This process was repeated until the model (that is, the set of index variants) did not change further (Supplementary Fig. 6g). This procedure was performed first for the set of strong signals (that is, considering models including only the strong signals). Once a final model had been obtained for the strong signals, the index variants for the strong signals were considered fixed and the process was repeated for all signals, this time allowing the index variants for the weak signals (but not the strong signals) to vary. Using this procedure, we could define the best model for 140 out of 150 regions, but for ten regions this approach did not converge (chr4:175328036–176346426, chr5:55531884–56587883, chr6:151418856–152937016, chr8:75730301–76917937, chr10:80341148–81387721, chr10:122593901–123849324, chr12:115336522–116336522, chr14:36632769–37635752, chr16:3606788–4606788 and chr22:38068833–39859355). For these ten regions, we defined the best model, from among all possible combinations of credible variants, as that with the largest chi-squared value. Finally, we redefined the set of CCVs for each signal using the conditional *P* values, after adjusting for the revised set of index variants. Again, for the strong signals, we conditioned on the index variants for the other strong signals, while for the weak signals we conditioned on the index variants for all of the other signals.

Case-only analysis. Differences in the effect size between ER-positive and ER-negative disease for each index-independent variant were assessed using a case-only analysis. We performed logistic regression with ER status as the dependent variable and the lead variant at each strong signal in the fine-mapping region as the independent variables. We used FDR (5%) to adjust for multiple testing.

OncoArray-only stepwise analysis. To evaluate whether the lower coverage in iCOGS could affect the identification of independent signals, we ran stepwise multinomial regression using only the OncoArray dataset. We identified 249 independent signals. Ninety-two signals, in 67 fine-mapping regions, achieved a genome-wide significance level (conditional $P<5\times10^{-8}$). Of these, 205 signals were also identified in the meta-analysis with iCOGS. Nine independent variants across ten regions were not evaluated in the combined analysis due to their low imputation information score in iCOGS. Of these nine signals, two signals would be classified as main primary signals: rs114709821 at region chr1:145144984–146144984 (OncoArray imputation information score = 0.72);

and rs540848673 at region chr1:149406413–150420734 (OncoArray imputation information score = 0.33). Given the low number of additional signals identified in the OncoArray dataset alone, all analyses were based on the combined iCOGS/OncoArray dataset.

CIMBA statistical analysis. CIMBA provided data from 60 retrospective cohort studies consisting of 9,445 unaffected and 9,363 affected female *BRCA1* mutation carriers of European ancestry. Unconditional (that is, single-variant) analyses were performed using a score test based on the retrospective likelihood of observing the genotype conditional on the disease phenotype[62,63]. Conditional analyses, where more than one variant is analyzed simultaneously, cannot be performed in this score test framework. Therefore, conditional analyses were performed by Cox regression, allowing for adjustment of the conditionally independent variants identified by the BCAC/DRIVE analyses. All models were stratified by country and birth cohort, and adjusted for relatedness (unconditional models used kinship-adjusted standard errors based on the estimated kinship matrix; conditional models used cluster robust standard errors based on phenotypic family data).

Data from the iCOGS array and OncoArray were analyzed separately and combined to give an overall *BRCA1* association by fixed-effects meta-analysis. Variants were excluded from further analyses if they exhibited evidence of heterogeneity (heterogeneity $P < 1 \times 10^{-4}$) between iCOGS and OncoArray, had a MAF < 0.005, were poorly imputed (imputation information score < 0.3) or were imputed to iCOGS only (that is, they must have been imputed to OncoArray or iCOGS and OncoArray).

Meta-analysis of ER-negative cases in BCAC with *BRCA1* mutation carriers from CIMBA. *BRCA1* mutation carrier association results were combined with the BCAC multinomial regression ER-negative association results in a fixed-effects meta-analysis. Variants considered for analysis must have passed all previous quality control steps and have had MAF ≥ 0.005. All meta-analyses were performed using the METAL software[64]. Instances where spurious associations might occur were investigated by assessing the linkage disequilibrium between a possible spurious association and the conditionally independent variants. High linkage disequilibrium between a variant and a conditionally independent variant within its region causes model instability through collinearity, and convergence of the model likelihood maximization may not be reliable. Where the association appeared to be driven by collinearity, the signals were excluded.

Heritability estimation. To estimate the frailty-scale heritability due to all of the fine-mapping signals, we used the formula:

$$h^2 = 2\left(\gamma'^T R \gamma' - \tau'^T I \tau'\right)$$

Here, $\gamma' = \gamma\sqrt{\mathbf{p}(1-\mathbf{p})}$ and $\tau' = \tau\sqrt{\mathbf{p}(1-\mathbf{p})}$, where $\mathbf{p}$ is a vector of allele frequencies, γ are the estimated per-allele odds ratios, τ are the corresponding standard errors and R is the correlation matrix of genotype frequencies.

To adjust for the overestimation resulting from only including signals passing a given significance threshold, we adapted the approach of ref. [65], based on maximizing the likelihood, conditional on the test statistic passing the relevant threshold. Since our analyses were based on estimating ER-negative and ER-positive odds ratios simultaneously, the method needed to be adapted to maximize a conditional bivariate normal likelihood. Following ref. [65], we then estimated mean square error estimates based on a weighted mean of the maximum likelihood estimates and the naïve estimates, which were shown to be unbiased in the 1-degree of freedom case. The estimated effect sizes for overall breast cancer were computed as a weighted mean of the ER-negative and ER-positive estimates, based on the proportions of each subtype in the whole study (weights: 0.21 and 0.79). The results were then expressed in terms of the proportion of the FRR to first-degree relatives of affected women, using the formula $h^2/(2\log[\lambda])$, where the FRR λ was assumed to be 2 (ref. [2]).

eQTL analysis. Total RNA was extracted from normal breast tissue in formalin-fixed paraffin-embedded breast cancer tissue blocks from 264 NHS participants[32]. Transcript expression levels were measured using the Glue Grant Human Transcriptome Array version 3.0 at the Molecular Biology Core Facilities, Dana-Farber Cancer Institute. Gene expression was normalized and summarized into $\log_2$ values using RMA (Affymetrix Power Tools version 1.18.012). Quality control was performed using GlueQC and arrayQualityMetrics version 3.24.014. Genome-wide data on variants were generated using the Illumina HumanHap550 BeadChip as part of the Cancer Genetic Markers of Susceptibility initiative[66]. Imputation to the 1000KGP Phase 3 version 5 ALL reference panel was performed using MACH to pre-phase measured genotypes, and minimac to impute.

Expression analyses were performed using data from the TCGA and METABRIC projects[34,38]. The TCGA eQTL analysis was based on 458 breast tumors that had matched gene expression, copy number and methylation profiles, together with the corresponding germline genotypes available. All 458 individuals were of European ancestry, as ascertained using the genotype data and the Local Ancestry in Admixed Populations (LAMP) software package (LAMP estimate cut-off > 95% European)[67]. Germline genotypes were imputed into the 1000 Genomes Project reference panel (October 2014 release) using IMPUTE version 2 (refs. [68,69]). Gene expression had been measured on the Illumina HiSeq 2000 RNA sequencing (RNA-Seq) platform (gene-level RSEM normalized counts[70]), copy number estimates were derived from Affymetrix SNP 6.0 (somatic copy number alteration minus germline copy number variation called using the GISTIC2 algorithm[71]), and methylation beta values were measured on the Illumina Infinium HumanMethylation450. Expression QTL analysis focused on all variants within each of the 152 genomic intervals that had been subjected to fine-mapping for their association with breast cancer susceptibility. Each of these variants was evaluated for its association with the expression of every gene within 2 Mb that had been profiled for each of the three data types. The effects of tumor copy number and methylation on gene expression were first regressed out using a method described previously[72]. eQTL analysis was performed by linear regression, with residual gene expression as the outcome, germline SNP genotype dosage as the covariate of interest, and ESR1 expression and age as additional covariates, using the R package Matrix eQTL[73].

The METABRIC eQTL analysis was based on 138 normal breast tissue samples resected from patients with breast cancer of European ancestry. Germline genotyping for the METABRIC study was also done on the Affymetrix SNP 6.0 array, and gene expression in the METABRIC study was measured using the Illumina HT12 microarray platform (probe-level estimates). No adjustment was implemented for somatic copy number and methylation status since we were evaluating eQTLs in normal breast tissue. All other steps were identical to the TCGA eQTL analysis described above.

Genomic features enrichment. We explored the overlap of CCVs and excluded variants with 90 transcription factors, ten histone marks and DNase hypersensitivity sites in 15 breast cell lines and eight normal human breast tissues. We analyzed data from the Encyclopedia of DNA Elements (ENCODE) Project[74,75], Roadmap Epigenomics Projects[76], the International Human Epigenome Consortium[27,77], Pellacani et al.[78], TCGA[33], METABRIC[34], the ReMap database (we included 241 transcription factor annotations from ReMap (from a total of 2,825), which showed at least 2% overlap for any of the phenotype SNP sets)[79] and other data obtained through the National Center for Biotechnology Information Gene Expression Omnibus. Promoters were defined following the procedure defined in ref. [78] (that is, ±2 kilobases (kb) from a gene transcription start site) using an updated version of the RefSeq genes (refGene version updated 11 April 2017)[80]. Transcribed regions were defined using the same version of RefSeq genes. lncRNA annotation was obtained from GENCODE (version 19)[81]

To include eQTL results in the enrichment analysis we: (1) identified all of the genes for which summary statistics were available; (2) defined the most significant eQTL variant for each gene (index eQTL variant; P value threshold $\leq 5 \times 10^{-4}$); and (3) classified variants with P values within two orders of magnitude of the index expression variant as the credible set of eQTL variants (that is, the best candidates to drive expression of the gene). Variants within at least one eQTL credible set were defined as expression variants. We evaluated the overlap between eQTL credible sets and CCVs (risk variants credible set). We evaluated the enrichment of CCVs for genomic features using logistic regression, with CCV (versus non-CCV variants) being the outcome. To adjust for the correlation among variants in the same fine-mapping region, we used robust variance estimation for clustered observations (R function multiwaycov). The associated variants at an FDR of 5% were included in a stepwise forward logistic regression procedure to select the most parsimonious model. A likelihood ratio test was used to compare multinomial logistic regression models with and without equality effect constraints to evaluate whether there was heterogeneity among the effect sizes for ER positive, ER negative or signals equally associated with both phenotypes (ER neutral).

To validate the disease specificity of the regulatory regions identified through this analysis, we followed the same approach for the autoimmune-related CCVs from ref. [29] ($n = 4{,}192$). Variants excluded as candidate causal variants, and within 500 kb upstream and downstream of the index variant for each signal, were classified as excluded variants ($n = 1{,}686{,}484$). We then tested the enrichment for both the breast cancer and autoimmune CCVs with breast and T and B cell enhancers. We also evaluated the overlap of our CCVs with ENCODE enhancer-like and promoter-like regions for 111 tissues, primary cells, immortalized cell lines and in vitro-differentiated cells. Of these, 73 had available data for both enhancer- and promoter-like regions.

Transcription binding site motif analysis. We conducted a search to find motif occurrences for transcription factors at active regulatory regions significantly enriched in CCVs. For this, we used two publicly available databases: Factorbook[82] and JASPAR 2016 (ref. [83]). For the search using Factorbook, we included the motifs for the transcription factors discovered in the cell lines where significant enrichment was found in our genomic features analysis. We also searched for all of the available motifs for *Homo sapiens* in the JASPAR database (JASPAR CORE 2016; TFBSTools[84]). Using the USCS sequence (BSgenome.Hsapiens.USCS.hg19) as a reference, we created fasta sequences with the reference and alternative alleles for all of the variants included in our analysis plus 20 base pairs flanking each variant. We used FIMO (version 4.11.2; Grant et al.[85]) to scan all of the fasta sequences, searching for the JASPAR and Factorbook motifs to identify any overlap of any

of the alleles for each of the variants (setting the *P* value threshold to 10^{-3}). We subsequently determined whether our CCVs were more frequency overlapping a particular transcription factor binding motif when compared with the excluded variants. We ran these analyses for all of the strong signals, but also strong signals stratified by ER status. Also, we subset this analysis to the variants located at regulatory regions in an ER-positive cell line (MCF-7 marked by H3K4me1; ENCODE identification: ENCFF674BKS) and evaluated whether the ER-positive CCVs overlapped any of the motifs more frequently than the excluded variants. We also evaluated the change in total binding affinity caused by the ER-positive CCV alternative allele for all but one (2:217955891:T:<CN0>:0) of the ER-positive CCVs (MatrixRider[86]).

Subsequently, we evaluated whether the MCF-7 regions demarked by H3K4me1 (ENCODE identification: ENCFF674BKS) and overlapped by ER-positive CCVs were enriched in known TFBS motifs. First, we subset the ENCODE bed file ENCFF674BKS to identify MCF-7 H3K4me1 peaks overlapped by the ER-positive CCVs ($n=107$), as well as peaks only overlapped by excluded variants ($n=11{,}099$), using BEDTools[87]. We created fasta format sequences using genomic coordinate data from the intersected bed files. To create a control sequence set, we used the script included with the MEME Suite (fasta-shuffle-letters) to create ten shuffled copies of each sequence overlapped by ER-positive CCVs ($n=1{,}070$). We then used AME[88] to interrogate whether the 107 MCF-7 H3K4me1 genomic regions overlapped by ER-positive CCVs were enriched in known TFBS consensus motifs when compared with the shuffled control sequences, or with the MCF-7 H3K4me1 genomic regions overlapped only by excluded variants. We used the command line version of AME (version 4.12.0), selecting as a scoring method the total number of positions in the sequence whose motif score *P* value was $<10^{-3}$, and using a one-tailed Fisher's exact test as the association test.

PAINTOR analysis. To further refine the set of CCVs, we performed empirical Bayes fine-mapping using PAINTOR to integrate marginal genetic association summary statistics, linkage disequilibrium patterns and biological features[31,89]. PAINTOR derives jointly the posterior probability for causality of all variants along the respective contribution of genomic features, in order to maximize the log-likelihood of the data across all regions. PAINTOR does not assume a fixed number of causal variants in each region, although it implicitly penalizes non-parsimonious causal models. We applied PAINTOR separately to association results for overall breast cancer (in 85 regions determined to have at least one ER-neutral association or ER-positive and ER-negative association), ER-positive breast cancer (in 48 regions determined to have at least one ER-positive-specific association) and ER-negative breast cancer (in 22 regions determined to have at least one ER-negative-specific association). To avoid artefacts due to mismatches between the linkage disequilibrium in study samples and the linkage disequilibrium matrix supplied to PAINTOR, we used association logistic regression summary statistics from OncoArray data only, and estimated the linkage disequilibrium structure in the OncoArray sample. For each endpoint, we fit four models with increasing numbers of genomic features selected from the stepwise enrichment analyses described above: model 0 (with no genomic features; assumes each variant is equally likely to be causal a priori); model 1 (with those genomic features selected with the stopping rule $P<0.001$); model 2 (with those genomic features selected with the stopping rule $P<0.01$); and model 3 (with those genomic features selected with the stopping rule $P<0.05$).

We used the Bayesian information criterion (BIC) to choose the best-fitting model for each outcome. As PAINTOR estimates the marginal log-likelihood of the observed *Z* scores using Gibbs sampling, we used a shrunk mean BIC across multiple Gibbs chains to account for the stochasticity in the log-likelihood estimates. We ran PAINTOR four times to generate four independent Gibbs chains, and estimated the BIC difference between model *i* and model *j* as $\Delta_{ij} = \left(\frac{100}{V+100}\right)(\mathrm{BIC}_i - \mathrm{BIC}_j)$. This assumes an $n(0,100)$ prior on the difference, or roughly a 16% chance that model *i* would be decisively better than model *j* (that is $|\mathrm{BIC}_i - \mathrm{BIC}_j| > 10$). We then proceeded to choose the best-fitting model in a stepwise fashion: starting with a model with no annotations, we selected a model with more annotations in favor of a model with fewer if the larger model was a considerably better fit (that is, $\Delta_{ij} > 2$). Model 1 was the best fit according to this process for overall and ER-positive breast cancer, while model 0 was the best fit for ER-negative breast cancer.

Differences between the PAINTOR and CCV outputs may be due to several factors. By considering functional enrichment and joint linkage disequilibrium among all SNPs, PAINTOR may refine the set of likely causal variants. Rather than imposing a hard threshold, PAINTOR allows for a gradient of evidence supporting causality; in addition, the two sets of calculations are based on different summary statistics: CCV analyses used both iCOGS and OncoArray genotypes, while PAINTOR used only OncoArray data (Fig. 1 and Methods).

Variant annotation. Variant genome coordinates were converted to assembly GRCh38 with liftOver and uploaded to Variant Effect Predictor[90] to determine their effect on genes, transcripts and protein sequence. The commercial software Alamut Batch version 1.6 was also used to annotate coding and splicing variants. PolyPhen-2 (ref. [91]), SIFT[92] and MAPP[93] were used to predict the consequences of missense coding variants. MaxEntScan[94], Splice-Site Finder and Human Splicing Finder[95] were used to predict splicing effects.

INQUISIT analysis. *Logic underlying INQUISIT predictions.* Briefly, genes were considered to be potential targets of candidate causal variants through effects on: (1) distal gene regulation; (2) proximal regulation; or (3) a gene's coding sequence. We intersected CCV positions with multiple sources of genomic information, including chromatin interactions from capture Hi-C experiments performed in a panel of six breast cell lines[96], ChIA-PET[97] and Hi-C[98]. We used computational enhancer–promoter correlations (PreSTIGE[99], IM-PET[100], FANTOM5 (ref. [101]) and super-enhancers[28]), results for breast tissue-specific expression variants from multiple independent studies (TCGA, METABRIC and NHS; Methods), allele-specific imbalance in gene expression[102], transcription factor and histone modification ChIP-Seq from the ENCODE and Roadmap Epigenomics Projects, together with the genomic features found to be significantly enriched as described above, gene expression RNA-Seq from several breast cancer lines and normal samples, and topologically associated domain boundaries from T-47D cells (ENCODE[103]; Methods). To assess the impact of intragenic variants, we evaluated their potential to alter splicing using Alamut Batch to identify new and cryptic donors and acceptors, and several tools to predict the effects of coding sequence changes (see 'Variant annotation' section). Variants potentially affecting post-translational modifications were downloaded from the 'A Website Exhibits SNP On Modification Event' database (http://www.awesome-hust.com/)[104]. The output from each tool was converted to a binary measure to indicate deleterious or tolerated predictions.

Scoring hierarchy. Each target gene prediction category (distal, promoter or coding) was scored according to different criteria. Genes predicted to be distally regulated targets of CCVs were awarded points based on physical links (for example, CHi-C), computational prediction methods, allele-specific expression or expression variant associations. All CCVs and HPPVs were considered as potentially involved in distal regulation. Intersection of a putative distal enhancer with genomic features found to be significantly enriched (see 'Genomic features enrichment' for details) were further upweighted. Multiple independent interactions were awarded an additional point. CCVs and HPPVs in gene proximal regulatory regions were intersected with histone ChIP-Seq peaks characteristic of promoters and assigned to the overlapping transcription start sites (defined as −1.0 kb to +0.1 kb). Further points were awarded to such genes if there was evidence of expression variant association or allele-specific expression, while a lack of expression resulted in down-weighting as potential targets. Potential coding changes, including missense, nonsense and predicted splicing alterations, resulted in the addition of one point to the encoded gene for each type of change, while lack of expression reduced the score. We added an additional point for predicted target genes that were also breast cancer drivers. For each category, scores ranged from 0 to 7 (distal), 0–3 (promoter) or 0 to 2 (coding). We converted these scores into 'confidence levels': level 1 (highest confidence; distal score > 4, promoter score ≥ 3 and coding score > 1); level 2 (1 ≤ distal score ≤ 4, promoter score = 1 or 2 and coding score = 1); and level 3 (0 < distal score < 1, 0 < promoter score < 1 and 0 < coding < 1). For genes with multiple scores (for example, those predicted as targets from multiple independent risk signals or predicted to be impacted in several categories), we recorded the highest score. Driver and transcription factor gene enrichment analysis was carried out using INQUISIT scores before adding a point for driver gene status. Modifications to the pipeline since original publication[2] included:

- Topologically associated domain boundary definitions from ENCODE T-47D Hi-C analysis. Previously, we used regions from ref. [98].
- eQTL (addition of NHS normal and tumor samples).
- Allele-specific imbalance using TCGA and Genotype-Tissue Expression RNA-Seq data[102].
- Capture Hi-C data from six breast cell lines[96].
- Additional bio-features derived from global enrichment in this study.
- Variants affecting sites of post-translational modification[104].

Multi-signal targets. To test whether more genes were targeted by multiple signals than would be expected by chance, we modeled the number of signals per gene by negative binomial regression (R function glm.nb; package MASS) and Poisson regression (R function glm; package stats) with ChIA-PET interactions as a covariate, and adjusted by fine-mapping region. Likelihood ratio tests were used to compare goodness of fit. Rootograms were created using the R function rootogram (package vcd).

Pathway analysis. The pathway gene set database dated 1 September 2018 was used[105] (http://download.baderlab.org/EM_Genesets/current_release/Human/symbol/). This database contains pathways from Reactome[106], the NCI Pathway Interaction Database[107], Gene Ontology[108], HumanCyc[109], MSigdb[110], NetPath[111] and Panther[112]. All duplicated pathways, defined in two or more databases, were included. To provide more biologically meaningful results, only pathways that contained ≤200 genes were used.

We interrogated the pathway annotation sets with the list of high-confidence (level 1) INQUISIT genes. The significance of over-representation of the INQUISIT genes within each pathway was assessed with a hypergeometric test using the R function phyper as follows:

$$P(x|n,m,N) = 1 - \sum_{i=0}^{x-1} \frac{\binom{m}{i}\binom{N-m}{n-i}}{\binom{N}{n}}$$

where x is the number of level 1 genes that overlap with any of the genes in the pathway, n is the number of genes in the pathway, m is the number of level 1 genes that overlap with any of the genes in the pathway dataset ($m_{\text{strong GO}}=145$; $m_{\text{ER-positive GO}}=50$; $m_{\text{ER-negative GO}}=27$; $m_{\text{ER-neutral GO}}=73$; $m_{\text{strong pathways}}=121$; $m_{\text{ER-positive pathways}}=38$; $m_{\text{ER-negative pathways}}=21$; $m_{\text{ER-neutral pathways}}=68$) and N is the number of genes in the pathway dataset ($N_{\text{genes GO}}=14{,}252$; $N_{\text{genes pathways}}=10{,}915$). We only included pathways that overlapped with at least two level 1 genes. We used the Benjamini–Hochberg FDR[113] at the 5% level.

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

Data availability

The credible set of causal variants (determined by either multinomial stepwise regression or PAINTOR) is provided in Supplementary Table 2c. Further information and requests for resources should be directed to M.K.B. (bcac@medschl.cam.ac.uk).

References

59. Couch, F. J. et al. Genome-wide association study in *BRCA1* mutation carriers identifies novel loci associated with breast and ovarian cancer risk. *PLoS Genet.* **9**, e1003212 (2013).
60. Gaudet, M. M. et al. Identification of a *BRCA2*-specific modifier locus at 6p24 related to breast cancer risk. *PLoS Genet.* **9**, e1003173 (2013).
61. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
62. Antoniou, A. C. et al. *RAD51* 135G→C modifies breast cancer risk among *BRCA2* mutation carriers: results from a combined analysis of 19 studies. *Am. J. Hum. Genet.* **81**, 1186–1200 (2007).
63. 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).
64. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
65. Zhong, H. & Prentice, R. L. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9**, 621–634 (2008).
66. Hunter, D. J. et al. A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39**, 870–874 (2007).
67. Baran, Y. et al. Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics* **28**, 1359–1367 (2012).
68. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
69. Genomes Project, C. et al. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
70. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
71. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
72. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
73. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).
74. The ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
75. Sloan, C. A. et al. ENCODE data at the ENCODE portal. *Nucleic Acids Res.* **44**, D726–D732 (2016).
76. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
77. Stunnenberg, H. G. International Human Epigenome Consortium & Hirst, M. The International Human Epigenome Consortium: a blueprint for scientific collaboration and discovery. *Cell* **167**, 1145–1149 (2016).
78. 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).
79. Cheneby, J., Gheorghe, M., Artufel, M., Mathelier, A. & Ballester, B. ReMap 2018: an updated atlas of regulatory regions from an integrative analysis of DNA-binding ChIP-Seq experiments. *Nucleic Acids Res.* **46**, D267–D275 (2018).
80. Pruitt, K. D. et al. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.* **42**, D756–D763 (2014).
81. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
82. Wang, J. et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
83. Mathelier, A. et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **44**, D110–D115 (2016).
84. Tan, G. & Lenhard, B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* **32**, 1555–1556 (2016).
85. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
86. Grassi, E., Zapparoli, E., Molineris, I. & Provero, P. Total binding affinity profiles of regulatory regions predict transcription factor binding and gene expression in human cells. *PLoS ONE* **10**, e0143627 (2015).
87. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
88. McLeay, R. C. & Bailey, T. L. Motif enrichment analysis: a unified framework and an evaluation on ChIP data. *BMC Bioinformatics* **11**, 165 (2010).
89. Kichaev, G. et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).
90. McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
91. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
92. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
93. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15**, 978–986 (2005).
94. Yeo, G. & Burge, C. B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **11**, 377–394 (2004).
95. Desmet, F. O. et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* **37**, e67 (2009).
96. Beesley, J. et al. Chromatin interactome mapping at 139 independent breast cancer risk signals. Preprint at *bioRxiv* https://doi.org/10.1101/520916 (2019).
97. Fullwood, M. J. et al. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
98. Rao, S. S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
99. Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Res.* **24**, 1–13 (2014).
100. He, B. et al. Global view of enhancer-promoter interactome in human cells. *Proc. Natl Acad. Sci. USA* **111**, e2191–e21999 (2014).
101. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
102. Moradi Marjaneh, M. et al. High-throughput allelic expression imbalance analyses identify 14 candidate breast cancer risk genes. Preprint at *bioRxiv* https://doi.org/10.1101/521013 (2019).
103. Dixon, J. R. et al. Integrative detection and analysis of structural variation in cancer genomes. *Nat. Genet.* **50**, 1388–1398 (2018).
104. Yang, Y. et al. AWESOME: a database of SNPs that affect protein post-translational modifications. *Nucleic Acids Res.* **47**, D874–D880 (2019).
105. Merico, D., Isserlin, R. & Bader, G. D. Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. *Methods Mol. Biol.* **781**, 257–277 (2011).
106. Vastrik, I. et al. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* **8**, R39 (2007).
107. Schaefer, C. F. et al. PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
108. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
109. Romero, P. et al. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* **6**, R2 (2005).

110. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
111. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol.* **11**, R3 (2010).
112. Thomas, P. D. et al. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* **13**, 2129–2141 (2003).
113. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).

Acknowledgements

We thank all of the individuals who took part in these studies, as well as all of the researchers, clinicians, technicians and administrative staff who enabled this work to be carried out. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Marie Sklodowska-Curie grant agreement number 656144. Genotyping of the OncoArray was principally funded from three sources: the PERSPECTIVE project (funded by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the 'Ministère de l'Économie de la Science et de l'Innovation du Québec' (through Genome Québec) and the Quebec Breast Cancer Foundation); the NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and the Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project (NIH grants U19 CA148065 and X01HG007492); and Cancer Research UK (C1287/A10118, C8197/A16565 and C1287/A16563). BCAC is funded by Cancer Research UK (C1287/A16563), by the European Community's Seventh Framework Programme under grant agreement 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the 'CIHR Team in Familial Risks of Breast Cancer' program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec (grant PSR-SIIRI-701). Combining of the GWAS data was supported in part by NIH Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE; part of the GAME-ON initiative). For a full description of funding and acknowledgments, see the Supplementary Note.

Author contributions

L.Fa., H.A., J.Bee., D.R.B., J.Al., S.Ka., K.A.P., K.Mi., P.So., A.Le., M.Gh., P.D.P.P., J.C.C., M.G.C., M.K.S., R.L.M., V.N.K., J.D.E., S.L.E., A.C.A., G.C.T., J.Si., D.F.E., P.K. and A.M.D. conceived of the study idea. L.Fa., H.A., J.Bee., D.R.B., J.Al., J.D.E., S.L.E., A.C.A., G.C.T., J.Si., D.F.E., P.K. and A.M.D. developed the methodology. J.Bee., J.P.T. and M.L. provided software. L.Fa., H.A., J.Bee., D.R.B., J.Al.., S.Ka., C.Tu., M.Mor. and X.J. performed a formal analysis. S.A., K.A., M.R.A., I.L.A., H.A.C., N.N.A., A.A., V.A., K.J.A., B.K.A., B.A., P.L.A., J.Az., J.Ba., R.B.B., D.B., A.B.F., J.Ben., M.B., K.B., A.M.B., C.B., W.B., N.V.B., S.E.B., B.Bo., A.B., H.Bra., H.Bre., I.B., I.W.B., A.B.W., T.B., B.Bu., S.S.B., Q.C., T.C., M.A.C., N.J.C., I.C., F.C., J.S.C., B.D.C., J.E.C., J.C., H.C., W.K.C., K.B.M., C.L.C., J.M.C., S.C., F.J.C., A.C., S.S.C., C.C., K.C., M.B.D., M.D.H., P.D., O.D., Y.C.D., G.S.D., S.M.D., T.D., I.D.S., A.D., S.D., M.Dum., M.Dur., L.D., M.Dw., D.M.E., C.E., M.E., D.G.E., P.A.F., U.F., O.F., G.F., H.F., L.Fo., W.D.F., E.F., L.Fr., D.F., M.Ga., M.G.D., G.Ga., P.A.G., S.M.G., J.Ga., J.A.G., M.M.G., V.G., G.G.G., G.Gl., A.K.G., M.S.G., D.E.G., A.G.N., M.H.G., M.Gr., J.Gr., A.G., P.G., E.H., C.A.H., N.H., P.Ha., U.H., P.A.H., J.M.H., M.H., W.H., C.S.H., B.A.M., J.H., P.Hi., F.B.L., A.H., M.J.H., J.L.H., A.Ho., G.H., P.J.H., E.N.I., C.I., M.I., A.Jag., M.J., A.Jak., P.J., R.J., R.C.J., E.M.J., N.J., M.E.J., A.Juk., A.Jun., R.Ka., D.K., B.Pes., R.Ke., M.J.K., E.K., J.I.K., J.K., C.M.K., Y.K., I.K., V.K., S.Ko., K.K.S., T.K., A.K., K.K., Y.L., D.L., E.L., G.L., J.Le., F.L., A.Li., W.L., J.Lo., A.Lo., J.T.L., J.Lu., R.J.M., T.M., E.M., A.Ma., M.Ma., S.Man., S.Mag., M.E.M., K.Ma., D.M., R.M., L.M., C.M., N.Me., A.Me., P.M., A.Mi., N.Mi., M.Mo., F.M., A.M.M., V.M.M., T.A., S.A.N., R.N., K.L.N., N.Z.N., H.N., P.N., F.C.N., L.N.Z., A.N., K.O., E.O., O.I.O., H.O., N.O., A.O., V.S.P., J.Pa., S.K.P., T.W.P.S., M.T.P., J.Pau., I.S.P., B.Pei., B.Y.K., P.P., J.Pe., D.P.K., K.Pr., R.P., N.P., D.P., M.A.P., K.Py., P.R., S.J.R., J.R., R.R.M., G.R., H.A.R., M.R., A.R., C.M.R., E.S., E.S.H., D.P.S., M.Sa., C.Sa., E.J.S., M.T.S., D.F.S., R.K.S., A.S., M.J.S., B.S., P.Sc., C.Sc., R.J.S., L.S., C.M.D., M.Sh., P.Sh., C.Y.S., X.S., C.F.S., T.P.S., S.S., M.C.S., J.J.S., A.B.S., J.St., D.S.L., C.Su., A.J.S., R.M.T., Y.Y.T., W.J.T., J.A.T., M.R.T., M.Te., S.H., M.B.T., A.T., M.Th., D.L.T., M.G.T., M.Ti., A.E.T., R.A.E., I.T., D.T., G.T.M., M.A.T., N.T., M.Tz., H.U.U., C.M.V., C.J.A., L.E.K., E.J.R., A.Ve., A.Vi., J.V., M.J.V., Q.W., B.W., C.R.W., J.N.W., C.W., H.W., R.W., A.W., A.H.W., D.Y., Y.Z. and W.Z. provided resources. K.Mi., J.D., M.K.B., Q.W., R.Ke., J.C.C. and M.K.S. curated and managed the data. L.Fa., H.A., J.Bee., G.C.T., D.F.E., P.K. and A.M.D. wrote the original draft of the manuscript. D.R.B., J.Al., P.So., A.Le., V.N.K., J.D.E., S.L.E., A.C.A. and J.Si wrote and edited the manuscript. L.Fa., H.A., J.Bee. and C.Tu visualized the results. A.C.A., G.C.T., J.Si., D.F.E., P.K. and A.M.D. supervised the project. L.Fa., P.D.P.P., J.C.C., M.G.C., M.K.S., R.L.M., V.N.K., J.D.E., S.L.E., A.C.A., G.C.T., J.Si., D.F.E., P.K. and A.M.D. acquired funding. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-019-0537-1.

Correspondence and requests for materials should be addressed to P.K. or A.M.D.

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

nature research

Corresponding author(s): Alison Dunning, Peter Kraft

Last updated by author(s): Aug 10, 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	No software was used for the data collection.
Data analysis	The software used have been described in details in Online Methods section. Softwares included: IMPUTE2, MEME Suite (FIMO, AME, fasta-shuffle-letters), Meta, R (R libraries: stats, nnet, MASS, vcd, TFBSTools, MatrixRider, multinomRob, multiwaycov), Bedtools, MACs, Variant Effect Predictor, Alamut® Batch v1.6 (tools PolyPhen-2 , SIFT, MAPP, MaxEntScan, Splice-Site Finder, Human Splicing Finder), PAINTOR, liftover, RMA, GlueQC,, arrayQualityMetrics v3.24.014, MACH, Local Ancestry in admixed Populations, GISTIC2. The custom scripts used during the study are available from the corresponding author on reasonable request.

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

The credible set of causal variants (determined by either multinomial stepwise regression and PAINTOR) is provided in Supplementary Table S2C. Further information and requests for resources should be directed to and will be fulfilled by Manjeet Bolla (bcac@medschl.cam.ac.uk)

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	No sample size calculation was made. We aimed to bring together the largest possible sample size (109,900 breast cancer cases and 88,937 controls of European ancestry) with GWAS imputed up to 1000 Genomes Project Panel to study the role of genetic variants in breast cancer. The sample size included in this study (N ~ 199,000) is 2 times larger than previous breast cancer fine-mapping studies (N ~ 100,000 samples).
Data exclusions	Established protocols were used to conduct rigorous data quality control for each GWAS at the study level (more details can be found in Michailidou et al. Nature 2017 and Amos et al. Cancer Epidemiol Biomarkers Prev 2017). Imputed variants were excluded for the following reasons: (i) info score < 0.3 and (ii) minor allele frequencies (MAF) < 0.001 in both the iCOGS and OncoArray datasets. Only samples with ER status were included in the multinomial regression analyses (cases with ER status available: 29,561 iCOGS, 55,081 OncoArray). More details about the data exclusions can be found in the Online Methods section.
Replication	This was an observational study - analyses were based on all available data. All signals show no statistically significant heterogeneity between effect sizes estimated in iCOGS and OncoArray cohorts.
Randomization	Not relevant because the study is not experimental.
Blinding	Not relevant because the study is not experimental.

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	Analyses were conducted on breast cancer cases and controls of European ancestry. The association between each variant and breast cancer risk was adjusted by study (iCOGS) or country (OncoArray), and eight (iCOGS) or ten (OncoArray) ancestry-informative principal components.
Recruitment	Epidemiological data for European women were obtained from 75 breast cancer case-control studies participating in the Breast Cancer Association Consortium (BCAC). The majority of studies are population-based case–control studies, or case–control studies nested within population-based cohorts, but a subset of studies oversampled cases with a family history of the disease. Subjects included from CIMBA are women of European ancestry aged 18 years or older with a pathogenic variant for BRCA1. The majority of the participants were sampled through cancer genetics clinics. Multiple members of the same family were included in some instances.
Ethics oversight	All participating studies were approved by their appropriate ethics review board and all subjects provided informed consent.

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