Published in final edited form as:
Nat Genet. 2020 January ; 52(1): 56–73. doi:10.1038/s41588-019-0537-1.

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

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

[#] These authors contributed equally to this work.

Abstract

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 (CCVs) 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

[*]Correspondence: pkraft@hsph.harvard.edu (PK), amd24@medschl.cam.ac.uk (AMD).
[280]Senior author.

Data Availablity

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 Manjeet Bolla (bcac@medschl.cam.ac.uk)

Author Contributions

Conceptualization: L.Fa., H.As., J.Be., 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., A.M.D. Methodology: L.Fa., H.As., J.Be., D.R.B., J.Al., J.D.E., S.L.E., A.C.A., G.C.T., J.Si., D.F.E., P.K., A.M.D. Software: J.Be., J.P.T., M.L. Formal analysis: L.Fa., H.As., J.Be., D.R.B., J.Al., S.Ka., C.Tu., M.Mor., X.J. Resources: 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., W.Z. Data management and curation: K.Mi., J.D., M.K.B., Q.W., R.Ke., J.C.C. and M.K.S. Writing original draft: L.Fa., H.As., J.Be., G.C.T., D.F.E., P.K., A.M.D. Writing review and editing: D.R.B., J.Al., P.So., A.Le., V.N.K., J.D.E., S.L.E., A.C.A., J.Si. Visualization: L.Fa., H.As., J.Be., C.Tu. Supervision: A.C.A., G.C.T., J.Si., D.F.E., P.K., A.M.D. Funding acquisition: 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., A.M.D. All authors read and approved the final version of the manuscript.

Competing Interests Statement

The authors declare no competing interests.

Europe PMC Funders Author Manuscripts

pipeline for prioritizing genes as targets of those potentially causal variants, using gene expression (eQTL), 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.

Introduction

Genome-wide association studies (GWAS) 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 > 217K 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 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 breast cancer cases 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), deletion/insertions, and copy number variants (CNVs) with minor allele frequency (MAF) ≥ 0.1% within 152, previously defined, risk-associated regions (Supplementary Table 1; Figure 1). Multivariate logistic regression confirmed associations for 150/152 regions at a p-value < 10^{-4} significance threshold (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-values <10^{-6}) or *moderate* (10^{-6} < conditional p-values <10^{-4}).

From the 150 genomic regions we identified 352 independent risk signals containing 13,367 CCVs, 7,394 of these were within the 196 strong-evidence signals across 129 regions (Figures 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 is causal (Figures 2C-D, Table 1). Furthermore, within signals with few CCVs (<10), the mean

Europe PMC Funders Author Manuscripts

scaled CADD score was higher than in signals with more CCVs (13.1 Vs 6.7 for CCVs in exons; $P_{ttest} = 2.7x10^{-4}$) 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, Figure 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 (Figure 2F). More than one quarter of the CCVs (2,277) were accounted for by one signal, resulting from strong linkage disequilibrium with a copy number variant. The remaining analyses focused on the other 205 strong signals across 128 regions (Supplementary Table 2C).

The proportion of the familial relative risk of breast cancer (FRR) 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).

CCVs are over-represented in active gene-regulatory regions and transcription factor binding sites

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 a cell-specific TF binding or histone mark (determined through 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 Figure 1A, Supplementary Table 3). We found significant enrichment of CCVs (FDR 5%) in the following genomic-features:

(i) Open chromatin (determined by DNase-seq and FAIRE-seq) in ER-positive breast cancer cell-lines and normal breast (Figure 3A). 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]).

(ii) Actively transcribed genes in normal breast and ER-positive cell lines (defined by H3K36me3 or H3K79me2 histone marks, Figure 3A). Enrichment was larger

for ER-neutral CCVs than for those affecting either ER-positive or ER-negative tumors.

(iii) Gene regulatory regions. CCVs overlapped distal gene regulatory elements in ER-positive breast cancer cells lines (defined by H3K4me1 or H3K27ac marks, Figure 3B). This was confirmed using the 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 [28] and [27] (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 [27], showed that 73% of active enhancer regions overlapped by ER-positive CCVs in ER-positive cells (MCF-7), are inactive in the normal HMEC breast cell line; thus, these enhancers appear to be MCF-7-specific.

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, Figure 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 (Figure 3B). As a control, we performed similar analyses with autoimmune disease CCVs [29] (Methods) and relevant B and T cells (Figures 3B-E). The strongest evidence of enrichment of breast cancer CCVs was found at regulatory regions active in ER-positive cells (Figure 3B), whereas enrichment of autoimmune CCVs was in regulatory regions active in B and T cells (Figure 3E). We also compared the enrichment of our CCVs in enhancer-like and promoter-like regions (defined by ENCODE; Supplementary Figure 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 Figure 1B). These results highlight both the tissue- and disease-specificity of these histone marked gene regulatory regions.

(iv) We observed significant enrichment of CCVs in the binding sites for 40 transcription factor binding sites(TFBS) determined by ChIP-Seq (Figures 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 in ER-positive cell lines, and 16 in normal breast). These TFBSs overlap each other and histone marks of active regulatory regions (Supplementary Figure 2). Enrichment in five TFBSs (ESR1, FOXA1, GATA3, TCF7L2, E2F1) has been previously reported [2,30]. All 40 TFBSs were significantly enriched in ER-positive CCVs (Figure 3F), seven were also enriched in ER-negative CCVs and nine in ER-neutral CCVs (Figures 3G-H). ESR1, FOXA1, GATA3 and EP300 TFBSs were enriched in all CCV ER-subtypes. However, the enrichment for ESR1, FOXA1 or GATA3 was stronger for ER-positive CCVs than for ER-negative or ER-neutral.

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 fork head 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 (≥2-fold, Supplementary Table 4C). Forty-eight ER-positive signals (73%) had at least one CCV predicted to modify the binding affinity >10-fold. This analysis validates previous reports of breast cancer causal variants that alter DNA binding affinity for FOXA1 [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 (Figure 1). In brief, PAINTOR integrates genetic association results, linkage disequilibrium (LD) 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 artifacts due to differences in genotyping and imputation across platforms, we restricted PAINTOR analyses to cases and controls typed using the OncoArray (61% of the total). We identified seven variants with 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 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 (Figure 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 84/86 regions analyzed for overall breast cancer, with a maximum of 16.1 and a mean of 6.4. For ER-positive cancer, 46/47 regions had total posterior probability > 2.0 (maximum 18.3, mean 6.5) and for ER-negative, 17/23 regions had total posterior probability > 2.0 (maximum 9.1, mean 3.2).

Although for many regions we were not able to identify HPP variants, we were able to reduce the proportion of variants needed to account for 80% of the total posterior probability in a region to under 5% for 65 regions for overall, 43 for ER-positive, and 18 for ER-negative breast cancer (Supplementary Figure 3A-C). PAINTOR analyses were also able to reduce the set of likely causal variants in many cases. After summing the posterior probabilities for CCVs in each of the overall breast cancer signals, 39/100 strong-evidence

signals had a total posterior probability > 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 PP in each signal ranged 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 posterior probability > 1%, and only 48 (0.013%) of these had posterior probability > 30% (Supplementary Figure 3D). At ER-positive and ER-negative signals respectively, 3.1% and 1.6% of the non-CCVs at strong signals had posterior probability > 1%, and 40 (0.019%) and 3 (0.003%) of these had posterior probability > 30% (Figures S3E-F). For the non-CCVs at strong-evidence signals with posterior probability > 30%, the relatively high posterior probability may be driven by the addition of functional annotation. Indeed, the incorporation of functional annotations more than doubled the posterior probability for 64/88 variants when compared to 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) data sets to identify a credible set of variants associated with differences in gene expression (eVariants): 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 eVariants (for each gene eVariants 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 eVariants from both the NHS normal and breast cancer tissue studies (normal breast OR = 2.70, p-value = 1.7×10^{-5}; tumor tissue OR = 2.34, p-value = 2.6×10^{-4}; Supplementary Table 3). ER-neutral CCVs overlapped with eVariants in normal tissue more frequently than did ER-positive and ER-negative CCVs ($OR_{ER\text{-}neutral}$ = 3.51, p-value = 1.3×10^{-5}). Cancer risk CCVs overlapped credible eVariants in 128/205 (62%) signals in at least one of the datasets (Supplementary Table 5A-B). Sixteen additional variants with PP ≥ 30%, not included among the CCVs, also overlapped with a credible eVariant (Supplementary Table 5A-B).

Transcription factors and known somatic breast cancer drivers are overrepresented 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 67/5,375 CCVs, and 19/137 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). Four hundred and ninety-nine CCVs at 94 signals, and four additional HPPV (≥ 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 majority of common susceptibility variants are regulatory.

We applied an updated version of our pipeline INQUISIT - **in**tegrated expression **qu**antitative trait and ***in-silico*** prediction of GWAS **t**argets) [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 124/128 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; a 5-fold increase over expected from CCVs and 15-fold from HPPVs; p-value= 1×10^{-6}; Supplementary Figure 4A). Notably, thirteen 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. TF genes [40] were also enriched amongst high-confidence targets predicted from both CCVs (2-fold, p-value = 4.6×10^{-4}) and HPPVs (2.5-fold, p-value = 1.8×10^{-2}, Supplementary Figure 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-value = 4.3×10^{-8}, Supplementary Figure 4B, Figure 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 genes prioritized via multiple lines of evidence in Figure 4A.

Three examples of INQUISIT using genomic features to identify predict target genes. Based on capture Hi-C and ChIA-PET chromatin interaction data, *NRIP1* is a predicted target of intergenic CCVs and HPPVs at chr21q21 (Supplementary Figure 5A). Multiple target genes were predicted at chr22q12, including the driver genes *CHEK2* and *XBP1* (Supplementary Figure 5B). A third example at chr12q24.31 is a more complicated scenario with two Level 1 targets: *RPLP0* [41] and a modulator of mammary progenitor cell expansion, *MSI1* [42] (Supplementary Figure 5C).

Target gene pathways include DNA integrity-checkpoint, apoptosis, developmental processes and the immune system

We performed pathway analysis to identify common processes using INQUSIT high confidence target protein-coding genes (Figure 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 (Figure 5B). We found that 23% (14/60) of the ER-positive target genes were classified within developmental process pathways (including mammary development), 18% in immune system and a further 17% in nuclear receptors pathways. Of genes targeted by ER-neutral signals, 21% (18/87) were classified in developmental process pathways, 19% in in immune system pathways, and a further 18% in apoptotic process. The top themes of genes targeted

by ER-negative signals were DNA integrity checkpoint and immune system, each containing 19% (7/37) genes, and apoptotic processes (16%).

Novel pathways revealed by this study include TNF-related apoptosis-inducing ligand (TRAIL) signaling, the AP-2 transcription factors pathway, and regulation of IκB kinase/NF-κB signaling. Of note, the latter of these is specifically overrepresented among ER-negative target genes. We also found significant overrepresentation of additional carcinogenesis-linked pathways including cAMP, NOTCH, PI3K, RAS, WNT/Beta-catenin, and of receptor tyrosine kinases signaling, including FGFR, EGFR, or TGFBR [43–47]. Finally, our target genes are also significantly overrepresented in DNA damage checkpoint, DNA repair pathways, as well as programmed cell death pathways, such as apoptotic process, regulated necrosis, and death receptor signaling-related pathways.

Discussion

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

We observed most regions contain multiple independent signals, 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 have used two complementary approaches to identify likely causal variants within each region: a Bayesian approach, PAINTOR, which integrated genetic associations, LD and informative genomic features, providing complementary evidence supporting most associations found by the more traditional, multinomial regression approach, 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 have 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, including the stop-gain rs11571833 (*BRCA2* NP_000050.2:p.K3326*, Meeks et al., 2016)[48] and two *CHEK2* coding variants; the frameshift rs555607708 [49,50], and a missense variant, rs17879961 [51,52]. In addition, a splicing variant, rs10069690, in *TERT* 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 and as well as trying to predict their target gene(s). Looking 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 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 to 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 is little publicly available ChIP-Seq data on ER-negative breast cancer cell lines. The heterogeneity of ER-negative tumors may also 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* [54] and *MDM4* [55]. Most targets, however, 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 extensively studied [56], although not primarily in the context of susceptibility. Five ER-negative high confidence target genes (*ALK*, *CASP8*, *CFLAR*, *ESR1*, *TNFSF10*) lie in the I-κB kinase/NF-κB signaling pathway. Interestingly, ER-negative cells have high levels of NF-kB activity when compared to ER-positive [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.

Methods

Study samples

Epidemiological data for European women were obtained from 75 breast cancer case-control studies participating in the Breast Cancer Association Consortium (BCAC) (cases: 40,285 iCOGS, 69,615 OncoArray; cases with ER status available: 29,561 iCOGS, 55,081

OncoArray); controls: 38,058 iCOGS, 50,879 OncoArray). Details of the participating studies, genotyping calling and quality control are given in [2,22,23], respectively. Epidemiological data for *BRCA1* mutation carriers were obtained from 60 studies providing data to the Consortium of Investigators of Modifiers of *BRCA1* and *BRCA2* (CIMBA) (affected 1,591 iCOGS, 7,772 OncoArray; unaffected 1,665 iCOGS, 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 analyses.

Variant selection and genotyping

Similar approaches were used to select variants for inclusion on the iCOGS and OncoArray, which are described in detail elsewhere [2,21]. Both arrays including 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 using the Oncoarray: the regions were 1Mb intervals centred 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 MAF > 0.02 in Europeans were identified, and all those correlated with the published GWAS variants at r2 > 0.1 together with a set of variants designed to tag all remaining variants at r2 > 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 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 QC criteria.

We imputed genotypes for all remaining variants using IMPUTE2 [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 increasing both the buffer regions and the number of haplotypes to use as templates (more detailed description of the parameters used can be found in [21]). We thus genotyped or successfully imputed 639,118 variants (all with imputation info score ≥ 0.3 and minor allele frequency (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 (OR) and standard errors (SE) 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 false discovery rate (FDR) at 1% level for each threshold 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-values $<10^{-6}$ in the final model), and moderate-evidence signals (conditional p-values $<10^{-4}$ and $\geq10^{-6}$ 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): (a) common and low frequency variants (MAF ≥ 0.01) with logistic regression p-value inclusion threshold ≤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 (b) rarer variants (MAF ≥ 0.001 and < 0.01), with logistic regression inclusion p-value ≤ 0.0001. The same parameters used for adjustment in logistic regression were used in the multinomial regression analysis (R function *multinom*). 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 into 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-value $< 10^{-4}$) 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 demonstrated genome-wide significance, while 44 were marginally significant ($9.89\times10^{-5} \geq$ p-value $> 5\times10^{-8}$). For two regions there were no variants significant at $p<10^{-4}$ (chr14:104712261-105712261; rs10623258 multinomial regression p-value = 2.32×10^{-4}; chr19:10923703-11923703, rs322144, multinomial regression p-value = 3.90×10^{-3}). Four main differences in the datasets used here and in the previous paper may account for this: (i) our previous paper [2] included data from 11 additional GWAS (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 GWAS; (ii) 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 (iii) for the set of samples genotyped with both arrays, [2] used the iCOGS genotypes, while this study includes OncoArray genotypes to maximize the number of

samples genotyped with a larger coverage; and (iv) 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. We first we identified the index variant of the main signal in the region, and then ran multinomial logistic regression for all 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 credible causal variants (CCVs)—For each independently associated signal, we first defined credible candidate variants (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 Figure 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 Figure 6B). Where a model with a higher chi-square was obtained, the index variant was replaced by the CCV in the best model (Supplementary Figure 6C-D). This process was repeated until the model (i.e. the set of index variants) did not change further (Supplementary Figure 6G). This procedure was performed first for the set of strong signals (i.e. 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, the index variants for the weak signals (but not the strong signals) to vary. Using this procedure we could define the best model for 140/150 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, chr22:38068833-39859355). For these 10 regions, we defined the best model, from among all possible combinations of credible variants, as that with the largest chi-square value. Finally, 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 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 use 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-value < 5×10-8). Two hundred and five of these 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 info score in iCOGS. Out of these nine signals, two signals would be classified as main primary signals, rs114709821 at region chr1:145144984-146144984 (OncoArray imputation info score = 0.72), and rs540848673 at region chr1:149406413-150420734 (OncoArray imputation info 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 (i.e. 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 the 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-value < 1×10-4) between iCOGS and OncoArray, had MAF < 0.005, were poorly imputed (imputation info score < 0.3) or were imputed to iCOGS only (i.e. 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 prior QC 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 LD between a possible spurious association and the conditionally independent variants. High LD between a variant and a conditionally independent variant within its region causes model instability through collinearity and the convergence of the model likelihood maximization may not 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 fine-mapping signals, we used the formula:

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

here $\gamma' = \gamma\sqrt{p(1-p)}$, $\tau'^T = \tau\sqrt{p(1-p)}$, where p is a vector of allele frequencies, γ are the estimated per-allele odds ratios and τ 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 [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 maximise a conditional bivariate normal likelihood. Following [65] we then estimated mean square error estimates based on a weighted mean of the maximum likelihood estimates and the naïve estimates, which they show to be close to be unbiased in the 1df 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 familial breast cancer risk (FRR) to first degree relatives of affected women, using the formula $h^2 / (2log\lambda)$ where the FRR λ was assumed to be 2 [2].

eQTL analysis

Total RNA was extracted from normal breast tissue in formalin-fixed paraffin embedded breast cancer tissue blocks from 264 Nurses' Health Study (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 v1.18.012); quality control was performed using GlueQC and arrayQualityMetrics v3.24.014. Genome-wide data on variants were generated using the Illumina HumanHap 550 BeadChip as part of the Cancer Genetic Markers of Susceptibility initiative [66]. Imputation to the 1000KGP Phase 3 v5 ALL reference panel was performed using MACH to pre-phase measured genotypes and minimac to impute.

Expression analyses were performed using data from The Cancer Genome Atlas (TCGA) and Molecular Taxonomy of Breast Cancer International Consortium (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 [68,69]. Gene expression had been measured on the Illumina HiSeq 2000 RNA-Seq platform (gene-level RSEM normalized counts [70]), copy-number estimates were derived from the

Affymetrix SNP 6.0 (somatic copy-number alteration minus germline copy-number variation called using the GISTIC2 algorithm [71]), and methylation beta values 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 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 breast cancer patients 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 feature enrichment

We explored the overlap of CCVs and excluded variants with 90 transcription factors, 10 histone marks, and DNase hypersensitivity sites in in 15 breast cell lines, and eight normal human breast tissues. We analysed data from the Encyclopedia of DNA Elements (ENCODE) Project [74,75], Roadmap Epigenomics Projects [76], the International Human Epigenome Consortium [77, 27], Pellacani et al. [78], The Cancer Genome Atlas (TCGA) [33], the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) [34], ReMap database (We included 241 TF annotations from ReMap (of 2825 total) 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 (NCBI) Gene Expression Omnibus (GEO). Promoters were defined following the procedure defined in [78], that is +/- 2Kb from a gene transcription start site, using an updated version of the RefSeq genes (refGene, version updated 2017-04-11)[80]. Transcribed regions were defined using the same version of refSeq genes. lncRNA annotation was obtained from Gencode (v19)[81]

To include eQTL results in the enrichment analysis we (i) identified all the genes for which summary statistics were available; (ii) defined the most significant eQTL variant for each gene (index eQTL variant, p-value threshold ≤ 5×10-4); (iii) classified variants with p-values within two orders of magnitude of the index eVariant as the credible set of eQTL variants; ie. the best candidates to drive expression of the gene. Variants within at least one eQTL credible set were defined as eVariants. We evaluated the overlap between eQTL credible sets and CCVs (risk variants credible set). We evaluated the enrichment of CCVs for genomic feature using logistic regression, with CCV (vs 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 FDR 5% were included into 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 follow the same approach for the autoimmune related CCVs from [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 line, 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 the transcription factors significantly enriched in the genomic featured. For this we used two publicly available databases, Factorbook [82] and JASPAR 2016 [83]. For the search using Factorbook we included the motifs for the transcription factors discovered in the cell lines where a significant enrichment was found in our genomic features analysis. We also searched for all the available motifs for *Homo sapiens* at the JASPAR database (*JASPAR CORE 2016*, *TFBSTools* [84]) Using as reference the USCS sequence (*BSgenome.Hsapiens.USCS.hg19*) we created fasta sequences with the reference and alternative alleles for all the variants included in our analysis plus 20 bp flanking each variant. We used FIMO (version 4.11.2, Grant et al., 2011)[85] to scan all 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 TF binding motif when compared with the excluded variants. We ran these analyses for all 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 id: ENCFF674BKS) and evaluated whether the ER-positive CCVs overlap any of the motifs more frequently that the excluded variants. We also evaluated the change in total binding affinity caused by the ER-positive CCCR 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 id: ENCFF674BKS), and overlapped by ER-positive CCVs, were enriched in known TFBS motifs. We first 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. In order to create a control sequence set, we used the script included with the MEME Suite (*fasta-shuffle-letters*) to created 10 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 know TFBS consensus motifs when compared to the shuffled control sequences, or to 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 scoring method the total number of positions in the sequence whose motif score p-value is less than 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 artifacts due to mis-matches between the LD in study samples and the LD matrix supplied to PAINTOR, we used association logistic regression summary statistics from OncoArray data only and estimated the LD 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 stopping rule $p<0.001$); Model 2 (with those genomic features selected with stopping rule $p<0.01$); and Model 3 (with those genomic features selected with 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 $\Delta_{ij} = \left(\frac{100}{V+100}\right)\left(\acute{BIC}_l - \acute{BIC}_J\right)$. This assumes a N(0,100) prior on the difference, or roughly a 16% chance that model i would be decisively better than model j (i.e. $|BIC_i\text{-}BIC_j|>10$). 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 a model with fewer if the larger model was a considerably better fit—i.e. $\Delta_{ij} > 2$. Model 1 was the best fit according to this process for overall and ER-positive breast cancer; 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 LD 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; and 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 (Figure 1, Methods).

Variant annotation

Variants 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 v1.6 batch was also used to annotate coding and splicing variants. PolyPhen-2 [91], SIFT [92], MAPP [93] were used to predict the consequence 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 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], chromatin interaction analysis by paired-end tag sequencing (ChIA-PET; [97]) and genome-wide chromosome conformation capture from HMECs (Hi-C, (Rao et al., 2014)). We used computational enhancer–promoter correlations (PreSTIGE [98], IM-PET (He et al., 2014), FANTOM5 [99] and super-enhancers [28]), results for breast tissue-specific expression variants (eVariants) from multiple independent studies (TCGA, METABRIC, NHS, Methods), allele-specific imbalance in gene expression [100], transcription factor and histone modification chromatin immunoprecipitation followed by sequencing (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 (TAD) boundaries from T47D cells (ENCODE, [101], Methods and Key Resources Table). 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 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/) [102]. 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 (eg CHi-C), computational prediction methods, allele-specific expression, or eVariant associations. All CCV 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 - +0.1 kb). Further points were awarded to such genes if there was evidence for eVariant 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 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 0-7 (distal); 0-3 (promoter) or 0-2 (coding). We converted these scores into 'confidence levels': Level 1 (highest confidence) when distal score > 4, promoter score >= 3 or coding score > 1; Level 2 when distal score <= 4 and >=1, promoter score = 1 or = 2, coding score = 1; and Level 3 when distal score < 1 and > 0, promoter score < 1 and > 0, and coding < 1 and > 0. For genes with multiple scores (for example, 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 prior to adding a point for driver gene status. Modifications to the pipeline since original publication [2] include:

- TAD boundary definitions from ENCODE T47D Hi-C analysis. Previously, we used regions from Rao, Cell 2013;
- eQTL: Addition of NHS normal and tumor samples
- allele-specific imbalance using TCGA and GTEx RNA-seq data [100]
- Capture Hi-C data from six breast cell lines [103]
- Additional biofeatures derived from global enrichment in this study
- Variants affecting sites of post-translational modification [102]

Multi-signal targets—To test if more genes were targeted by multiple signals than expected by chance, we modelled 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 [104] (http://download.baderlab.org/EM_Genesets/current_release/Human/symbol/). This database contains pathways from Reactome [105], NCI Pathway Interaction Database [106], GO (Gene Ontology) [107], HumanCyc [108], MSigdb [109], NetPath [110], and Panther [111]. 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 gene list. 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 Level1 genes that overlap with any of the genes in the pathway data set ($m_{\text{strong GO}} = 145$, $m_{\text{ER-positive GO}} = 50$, $m_{\text{ER-negative GO}} = 27$, $m_{\text{ER-neutral GO}} = 73$; $m_{\text{strong Pathways}} = 121$, $m_{\text{ER-positive Pathways}} = 38$, $m_{\text{ER-negative Pathways}} = 21$, $m_{\text{ER-neutral Pathways}} = 68$), and N is the number of genes in the pathway data set ($N_{\text{Genes GO}} = 14{,}252$, $N_{\text{Genes Pathways}} = 10{,}915$). We only included pathways that overlapped with at least two Level 1 genes. We used the Benjamini-Hochberg false discovery rate (FDR) [112] at 5% level.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Laura Fachal[1], Hugues Aschard[#2,3,4], Jonathan Beesley[#5], 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], 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[70,71,72], EMBRACE Collaborators[6], J. Margriet Collée[73], Sten Cornelissen[74], Fergus J. Couch[75], Angela Cox[50], Simon S. Cross[76], Cezary Cybulski[31], Kamila Czene[77], Mary B. Daly[78], Miguel de la Hoya[57], Peter Devilee[79,80], Orland Diez[81,82], Yuan Chun Ding[83], Gillian S. Dite[84], Susan M. Domchek[85], Thilo Dörk[37], Isabel dos-Santos-Silva[86], Arnaud Droit[8,87], Stéphane Dubois[8], Martine Dumont[8], Mercedes Duran[88], Lorraine

Durcan[89,90], Miriam Dwek[91], Diana M. Eccles[92], Christoph Engel[93], Mikael Eriksson[77], D. Gareth Evans[94,95], Peter A. Fasching[96,97], Olivia Fletcher[98], Giuseppe Floris[99], Henrik Flyger[100], Lenka Foretova[101], William D. Foulkes[102], Eitan Friedman[103,104], Lin Fritschi[105], Debra Frost[6], Marike Gabrielson[77], Manuela Gago-Dominguez[106,107], Gaetana Gambino[58], Patricia A. Ganz[108], Susan M. Gapstur[64], Judy Garber[109], José A. García-Sáenz[110], Mia M. Gaudet[64], Vassilios Georgoulias[111], Graham G. Giles[84,112,113], Gord Glendon[12], Andrew K. Godwin[114], Mark S. Goldberg[115,116], David E. Goldgar[117], Anna González-Neira[29], Maria Grazia Tibiletti[118], Mark H. Greene[119], Mervi Grip[120], Jacek Gronwald[31], Anne Grundy[121], Pascal Guénel[122], Eric Hahnen[123,124], Christopher A. Haiman[125], Niclas Håkansson[126], Per Hall[77,127], Ute Hamann[128], Patricia A. Harrington[1], Jaana M. Hartikainen[129,130,131], Mikael Hartman[132,133], Wei He[77], Catherine S. Healey[1], Bernadette A.M. Heemskerk-Gerritsen[134], Jane Heyworth[135], Peter Hillemanns[37], Frans B.L. Hogervorst[136], Antoinette Hollestelle[134], Maartje J. Hooning[134], John L. Hopper[84], Anthony Howell[137], Guanmengqian Huang[128], Peter J. Hulick[138,139], Evgeny N. Imyanitov[140], KConFab Investigators[60,61], HEBON Investigators[141], ABCTB Investigators[142], Claudine Isaacs[143], Motoki Iwasaki[144], Agnes Jager[134], Milena Jakimovska[145], Anna Jakubowska[31,146], Paul A. James[61,147], Ramunas Janavicius[148,149], Rachel C. Jankowitz[150], Esther M. John[151], Nichola Johnson[98], Michael E. Jones[152], Arja Jukkola-Vuorinen[153], Audrey Jung[154], Rudolf Kaaks[154], Daehee Kang[155,156,157], Pooja Middha Kapoor[154,158], Beth Y. Karlan[159,160], Renske Keeman[74], Michael J. Kerin[161], Elza Khusnutdinova[30,162], Johanna I. Kiiski[163], Judy Kirk[164], Cari M. Kitahara[165], Yon-Dschun Ko[166], Irene Konstantopoulou[167], Veli-Matti Kosma[129,130,131], Stella Koutros[168], Katerina Kubelka-Sabit[169], Ava Kwong[170,171,172], Kyriacos Kyriacou[7], Yael Laitman[103], Diether Lambrechts[173,174], Eunjung Lee[125], Goska Leslie[6], Jenny Lester[159,160], Fabienne Lesueur[71,72,175], Annika Lindblom[176,177], Wing-Yee Lo[44,45], Jirong Long[27], Artitaya Lophatananon[178,179], Jennifer T. Loud[119], Jan Lubiński[31], Robert J. MacInnis[84,112], Tom Maishman[89,90], Enes Makalic[84], Arto Mannermaa[129,130,131], Mehdi Manoochehri[128], Siranoush Manoukian[24], Sara Margolin[127,180], Maria Elena Martinez[107,181], Keitaro Matsuo[182,183], Tabea Maurer[184], Dimitrios Mavroudis[111], Rebecca Mayes[1], Lesley McGuffog[6], Catriona McLean[185], Noura Mebirouk[70,71,72], Alfons Meindl[186], Austin Miller[187], Nicola Miller[161], Marco Montagna[9], Fernando Moreno[110], Kenneth Muir[178,179], Anna Marie Mulligan[188,189], Victor M. Muñoz-Garzon[190], Taru A. Muranen[163], Steven A. Narod[191], Rami Nassir[192], Katherine L. Nathanson[85], Susan L. Neuhausen[83], Heli Nevanlinna[163], Patrick Neven[99], Finn C. Nielsen[193], Liene Nikitina-Zake[194], Aaron Norman[195], Kenneth Offit[196,197], Edith Olah[198], Olufunmilayo I. Olopade[199], Håkan Olsson[200], Nick Orr[201], Ana Osorio[28,29], V. Shane Pankratz[202], Janos Papp[198], Sue K. Park[155,156,157], Tjoung-Won Park-Simon[37], Michael T. Parsons[5], James Paul[203], Inge Sokilde Pedersen[204,205,206], Bernard Peissel[24], Beth Peshkin[143], Paolo Peterlongo[207], Julian Peto[86], Dijana Plaseska-Karanfilska[145], Karolina Prajzendanc[31], Ross Prentice[22], Nadege Presneau[91], Darya Prokofyeva[162], Miquel Angel Pujana[208], Katri Pylkäs[209,210], Paolo Radice[211], Susan J. Ramus[212,213], Johanna Rantala[214], Rohini Rau-Murthy[197], Gad Rennert[215], Harvey A. Risch[216], Mark Robson[197],

Atocha Romero[217], Caroline Maria Rossing[193], Emmanouil Saloustros[218], Estela Sánchez-Herrero[217], Dale P. Sandler[219], Marta Santamariña[28,220,221], Christobel Saunders[222], Elinor J. Sawyer[223], Maren T. Scheuner[32], Daniel F. Schmidt[84,224], Rita K. Schmutzler[123,124], Andreas Schneeweiss[55,225], Minouk J. Schoemaker[152], Ben Schöttker[18,226], Peter Schürmann[37], Christopher Scott[195], Rodney J. Scott[227,228,229], Leigha Senter[230], Caroline M Seynaeve[134], Mitul Shah[1], Priyanka Sharma[231], Chen-Yang Shen[232,233], Xiao-Ou Shu[27], Christian F. Singer[234], Thomas P. Slavin[235], Snezhana Smichkoska[236], Melissa C. Southey[113,237], John J. Spinelli[238,239], Amanda B. Spurdle[5], Jennifer Stone[84,240], Dominique Stoppa-Lyonnet[70,241,242], Christian Sutter[243], Anthony J. Swerdlow[152,244], Rulla M. Tamimi[3,4,245], Yen Yen Tan[246], William J. Tapper[92], Jack A. Taylor[219,247], Manuel R. Teixeira[248,249], Maria Tengström[129,250,251], Soo H. Teo[252,253], Mary Beth Terry[254], Alex Teulé[255], Mads Thomassen[256], Darcy L. Thull[257], Marc Tischkowitz[102,258], Amanda E. Toland[259], Rob A.E.M. Tollenaar[260], Ian Tomlinson[261,262], Diana Torres[48,128], Gabriela Torres-Mejía[263], Melissa A. Troester[264], Thérèse Truong[122], Nadine Tung[265], Maria Tzardi[266], Hans-Ulrich Ulmer[267], Celine M. Vachon[268], Christi J. van Asperen[269], Lizet E. van der Kolk[136], Elizabeth J. van Rensburg[270], Ana Vega[271], Alessandra Viel[272], Joseph Vijai[196,197], Maartje J. Vogel[136], Qin Wang[6], Barbara Wappenschmidt[123,124], Clarice R. Weinberg[273], Jeffrey N. Weitzel[235], Camilla Wendt[180], Hans Wildiers[99], Robert Winqvist[209,210], Alicja Wolk[126,274], Anna H. Wu[125], Drakoulis Yannoukakos[167], Yan Zhang[18,46], Wei Zheng[27], David Hunter[3,4], Paul D.P. Pharoah[1,6], Jenny Chang-Claude[154,184], Montserrat García-Closas[168,275], Marjanka K. Schmidt[74,276], Roger L. Milne[84,112,113], Vessela N. Kristensen[277,278], Juliet D. French[5], Stacey L. Edwards[5], Antonis C. Antoniou[6], Georgia Chenevix-Trench[5,280], Jacques Simard[8,280], Douglas F. Easton[1,6,280], Peter Kraft[3,4,280,*], Alison M. Dunning[1,280,*]

Affiliations

[1]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, CB1 8RN, 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, 02115, USA [4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, 02115, USA [5]Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, 4006, Australia [6]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK [7]Department of Electron Microscopy/Molecular Pathology and The Cyprus School of Molecular Medicine, The Cyprus Institute of Neurology & Genetics, Nicosia, 1683, Cyprus [8]Genomics Center, Centre Hospitalier Universitaire de Québec – Université Laval, Research Center, Québec City, QC, G1V 4G2, Canada [9]Immunology and Molecular Oncology Unit, Veneto Institute of Oncology IOV - IRCCS, Padua, 35128, Italy [10]Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki, 00290, Finland [11]Human Genotyping-CEGEN Unit, Human Cancer Genetic Program, Spanish National Cancer Research

Centre, Madrid, 28029, Spain [12]Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, M5G 1X5, Canada [13]Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S 1A8, Canada [14]Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA, 92617, USA [15]N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, 223040, Belarus [16]Department of Pathology, Landspitali University Hospital, Reykjavik, 101, Iceland [17]BMC (Biomedical Centre), Faculty of Medicine, University of Iceland, Reykjavik, 101, Iceland [18]Division of Clinical Epidemiology and Aging Research, C070, German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany [19]Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON, K7L 3N6, Canada [20]Department of Breast Medical Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, 77030, USA [21]Institute of Human Genetics, Hannover Medical School, Hannover, 30625, Germany [22]Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, 98109, USA [23]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, 53205, USA [24]Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, 20133, Italy [25]High Risk and Cancer Prevention Group, Vall d'Hebron Institute of Oncology, Barcelona, 08035, Spain [26]Department of Medical Oncology, University Hospital of Vall d'Hebron, Barcelona, 08035, Spain [27]Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, 37232, USA [28]Centro de Investigación en Red de Enfermedades Raras (CIBERER), Madrid, 28029, Spain [29]Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid, 28029, Spain [30]Institute of Biochemistry and Genetics, Ufa Federal Research Centre of the Russian Academy of Sciences, Ufa, 450054, Russia [31]Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, 71-252, Poland [32]Cancer Genetics and Prevention Program, University of California San Francisco, San Francisco, CA, 94143-1714, USA [33]Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, 00290, Finland [34]Department of Oncology, Örebro University Hospital, Örebro, 70185, Sweden [35]International Epidemiology Institute, Rockville, MD, 20850, USA [36]Department of Radiation Oncology, Hannover Medical School, Hannover, 30625, Germany [37]Gynaecology Research Unit, Hannover Medical School, Hannover, 30625, Germany [38]Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark [39]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark [40]Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 2200, Denmark [41]Division of Cancer Prevention and Genetics, IEO, European Institute of Oncology IRCCS, Milan, 20141, Italy [42]Department of Oncology, Lund University and Skåne University Hospital, Lund, 222 41, Sweden [43]Institute of Medical Genetics and Applied Genomics, University of

Tübingen, Tübingen, 72074, Germany [44]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, 70376, Germany [45]iFIT-Cluster of Excellence, University of Tuebingen, Tuebingen, 72074, Germany [46]German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany [47]Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, 69120, 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, S10 2TN, UK [51]Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, V5Z 1L3, Canada [52]Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada [53]Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, 44789, Germany [54]Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany [55]Molecular Biology of Breast Cancer, University Womens Clinic Heidelberg, University of Heidelberg, Heidelberg, 69120, Germany [56]Department of Medicine, Huntsman Cancer Institute, Salt Lake City, UT, 84112, USA [57]Molecular Oncology Laboratory, CIBERONC, Hospital Clinico San Carlos, IdISSC (Instituto de Investigación Sanitaria del Hospital Clínico San Carlos), Madrid, 28040, Spain [58]SOD Genetica Molecolare, University Hospital, Pisa, Italy [59]Department of Internal Medicine, Huntsman Cancer Institute, Salt Lake City, UT, 84112, USA [60]Research Department, Peter MacCallum Cancer Center, Melbourne, Victoria, 3000, Australia [61]Sir Peter MacCallum Department of Oncology, The University of Melbourne, Melbourne, Victoria, 3000, Australia [62]Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany [63]Cancer Research UK Cambridge Research Institute, Li Ka Shing Centre, University of Cambridge, Cambridge, UK [64]Behavioral and Epidemiology Research Group, American Cancer Society, Atlanta, GA, 30303, USA [65]Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, 36312, Spain [66]CRCHU de Québec-Université Laval, axe oncologie, Québec, QC, G1S 4L8, Canada [67]Departments of Pediatrics and Medicine, Columbia University, New York, NY, 10032, USA [68]Centre for Medical Genetics, Ghent University, Gent, 9000, Belgium [69]Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, 2145, Australia [70]Department of Tumour Biology, INSERM U830, Paris, 75005, France [71]Institut Curie, Paris, 75005, France [72]Mines ParisTech, Fontainebleau, 77305, France [73]Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, 3015 CN, The Netherlands [74]Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, 1066 CX, The Netherlands [75]Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, 55905, USA [76]Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, S10 2TN, UK [77]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, 171 65,

Sweden [78]Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, 19111, USA [79]Department of Pathology, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands [80]Department of Human Genetics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands [81]Oncogenetics Group, Vall dHebron Institute of Oncology (VHIO), Barcelona, 8035, Spain [82]Clinical and Molecular Genetics Area, University Hospital Vall dHebron, Barcelona, 8035, Spain [83]Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, 91010, USA [84]Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria, 3010, Australia [85]Basser Center for BRCA, Abramson Cancer Center, University of Pennsylvania, Philadelphia, PA, 19066, USA [86]Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK [87]Département de Médecine Moléculaire, Faculté de Médecine, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, QC, G1V 0A6, Canada [88]Cáncer Hereditario, Instituto de Biología y Genética Molecular, IBGM, Universidad de Valladolid, Centro Superior de Investigaciones Científicas, UVA-CSIC, Valladolid, 47003, Spain [89]Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton, Southampton, SO17 1BJ, UK [90]Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, SO17 1BJ, UK [91]School of Life Sciences, University of Westminster, London, W1B 2HW, UK [92]Faculty of Medicine, University of Southampton, Southampton, SO17 1BJ, UK [93]Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, 04107, Germany [94]Genomic Medicine, Division of Evolution and Genomic Sciences, The University of Manchester, Manchester Academic Health Science Centre, Manchester Universities Foundation Trust, St. Mary's Hospital, Manchester, M13 9WL, UK [95]Genomic Medicine, North West Genomics hub, Manchester Academic Health Science Centre, Manchester Universities Foundation Trust, St. Mary's Hospital, Manchester, M13 9WL, UK [96]David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, 90095, USA [97]Department of Gynecology and Obstetrics, Comprehensive Cancer Center ER-EMN, University Hospital Erlangen, Friedrich-Alexander-University Erlangen-Nuremberg, Erlangen, 91054, Germany [98]The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, SW7 3RP, UK [99]Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, 3000, Belgium [100]Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, 2730, Denmark [101]Department of Cancer Epidemiology and Genetics, Masaryk Memorial Cancer Institute, Brno, 65653, Czech Republic [102]Program in Cancer Genetics, Departments of Human Genetics and Oncology, McGill University, Montréal, QC, H4A 3J1, Canada [103]The Susanne Levy Gertner Oncogenetics Unit, Chaim Sheba Medical Center, Ramat Gan, 52621, Israel [104]Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv, 69978, Israel [105]School of Public Health, Curtin University, Perth, Western Australia, 6102,

Australia [106]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, 15706, Spain [107]Moores Cancer Center, University of California San Diego, La Jolla, CA, 92037, USA [108]Schools of Medicine and Public Health, Division of Cancer Prevention & Control Research, Jonsson Comprehensive Cancer Centre, UCLA, Los Angeles, CA, 90096-6900, USA [109]Cancer Risk and Prevention Clinic, Dana-Farber Cancer Institute, Boston, MA, 02215, USA [110]Medical Oncology Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid, 28040, Spain [111]Department of Medical Oncology, University Hospital of Heraklion, Heraklion, 711 10, Greece [112]Cancer Epidemiology Division, Cancer Council Victoria, Melbourne, Victoria, 3004, Australia [113]Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, Victoria, 3168, Australia [114]Department of Pathology and Laboratory Medicine, Kansas University Medical Center, Kansas City, KS, 66160, USA [115]Department of Medicine, McGill University, Montréal, QC, H4A 3J1, Canada [116]Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, QC, H4A 3J1, Canada [117]Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, 84112, USA [118]UO Anatomia Patologica Ospedale di Circolo, ASST Settelaghi, Varese, Italy [119]Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, 20850-9772, USA [120]Department of Surgery, Oulu University Hospital, University of Oulu, Oulu, 90220, Finland [121]Centre de Recherche du Centre Hospitalier de Université de Montréal (CHUM), Université de Montréal, Montréal, QC, H2X 0A9, Canada [122]Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, 94805, France [123]Center for Hereditary Breast and Ovarian Cancer, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, 50937, Germany [124]Center for Integrated Oncology (CIO), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, 50937, Germany [125]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, 90033, USA [126]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, 171 77, Sweden [127]Department of Oncology, Södersjukhuset, Stockholm, 118 83, Sweden [128]Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany [129]Translational Cancer Research Area, University of Eastern Finland, Kuopio, 70210, Finland [130]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, 70210, Finland [131]Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, 70210, Finland [132]Saw Swee Hock School of Public Health, National University of Singapore, Singapore, 119077, Singapore [133]Department of Surgery, National University Health System, Singapore, 119228, Singapore [134]Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam,

3015 CN, The Netherlands [135]School of Population and Global Health, The University of Western Australia, Perth, Western Australia, 6009, Australia [136]Family Cancer Clinic, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, 1066 CX, The Netherlands [137]Division of Cancer Sciences, University of Manchester, Manchester, M13 9PL, UK [138]Center for Medical Genetics, NorthShore University HealthSystem, Evanston, IL, 60201, USA [139]The University of Chicago Pritzker School of Medicine, Chicago, IL, 60637, USA [140]N.N. Petrov Institute of Oncology, St. Petersburg, 197758, Russia [141]The Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON), Coordinating center: The Netherlands Cancer Institute, Amsterdam, 1066 CX, The Netherlands [142]Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research, University of Sydney, Sydney, New South Wales, 2145, Australia [143]Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, 20007, USA [144]Division of Epidemiology, Center for Public Health Sciences, National Cancer Center, Tokyo, 104-0045, Japan [145]Research Centre for Genetic Engineering and Biotechnology 'Georgi D. Efremov', Macedonian Academy of Sciences and Arts, Skopje, 1000, Republic of North Macedonia [146]Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin, 71-252, Poland [147]Parkville Familial Cancer Centre, Peter MacCallum Cancer Center, Melbourne, Victoria, 3000, Australia [148]Hematology, oncology and transfusion medicine center, Dept. of Molecular and Regenerative Medicine, Vilnius University Hospital Santariskiu Clinics, Vilnius, Lithuania [149]State Research Institute Centre for Innovative Medicine, Vilnius, Lithuania [150]Department of Medicine, Division of Hematology/Oncology, UPMC Hillman Cancer Center; University of Pittsburgh School of Medicine, Pittsburgh, PA 15232, USA [151]Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, 94304, USA [152]Division of Genetics and Epidemiology, The Institute of Cancer Research, London, SM2 5NG, UK [153]Department of Oncology, Tampere University Hospital, Tampere University and Tampere Cancer Center, Tampere, 33521, Finland [154]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, 69120, Germany [155]Department of Preventive Medicine, Seoul National University College of Medicine, Seoul, 03080, Korea [156]Department of Biomedical Sciences, Seoul National University Graduate School, Seoul, 03080, Korea [157]Cancer Research Institute, Seoul National University, Seoul, 03080, Korea [158]Faculty of Medicine, University of Heidelberg, Heidelberg, 69120, Germany [159]David Geffen School of Medicine, Department of Obstetrics and Gynecology, University of California at Los Angeles, Los Angeles, CA, 90095, USA [160]Women's Cancer Program at the Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, CA, 90048, USA [161]Surgery, School of Medicine, National University of Ireland, Galway, H91TK33, Ireland [162]Department of Genetics and Fundamental Medicine, Bashkir State Medical University, Ufa, 450000, Russia [163]Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, 00290, Finland [164]Familial Cancer Service, Weatmead Hospital,

Wentworthville, New South Wales, 2145, Australia [165]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, 20892, USA [166]Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, 53177, Germany [167]Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research 'Demokritos', Athens, 15310, Greece [168]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD, 20850, USA [169]Department of Histopathology and Cytology, Clinical Hospital Acibadem Sistina, Skopje, 1000, Republic of North Macedonia [170]Hong Kong Hereditary Breast Cancer Family Registry, Cancer Genetics Centre, Happy Valley, Hong Kong [171]Department of Surgery, The University of Hong Kong, Pok Fu Lam, Hong Kong [172]Department of Surgery, Hong Kong Sanatorium and Hospital, Happy Valley, Hong Kong [173]VIB Center for Cancer Biology, VIB, Leuven, 3001, Belgium [174]Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven, 3000, Belgium [175]Genetic Epidemiology of Cancer team, Inserm U900, Paris, 75005, France [176]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, 171 76, Sweden [177]Department of Clinical Genetics, Karolinska University Hospital, Stockholm, 171 76, Sweden [178]Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, CV4 7AL, UK [179]Institute of Population Health, University of Manchester, Manchester, M13 9PL, UK [180]Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm, 118 83, Sweden [181]Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA, 92093, USA [182]Division of Cancer Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, 464-8681, Japan [183]Division of Cancer Epidemiology, Nagoya University Graduate School of Medicine, Nagoya, 466-8550, Japan [184]Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, 20246, Germany [185]Anatomical Pathology, The Alfred Hospital, Melbourne, Victoria, 3004, Australia [186]Department of Gynecology and Obstetrics, University of Munich, Campus Großhadern, Munich, 81377, Germany [187]NRG Oncology, Statistics and Data Management Center, Roswell Park Cancer Institute, Buffalo, NY, 14263, USA [188]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, M5S 1A8, Canada [189]Laboratory Medicine Program, University Health Network, Toronto, ON, M5G 2C4, Canada [190]Radiation Oncology, Hospital Meixoeiro-XXI de Vigo, Vigo, 36214, Spain [191]Women's College Research Institute, University of Toronto, Toronto, ON, M5S 1A8, Canada [192]Department of Biochemistry and Molecular Medicine, University of California Davis, Davis, CA, 95817, USA [193]Center for Genomic Medicine, Rigshospitalet, Copenhagen University Hospital, Copenhagen, DK-2100, Denmark [194]Latvian Biomedical Research and Study Centre, Riga, Latvia [195]Department of Health Sciences Research, Mayo Clinic, Rochester, MN, 55905, USA [196]Clinical Genetics Research Lab, Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA

[197]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA [198]Department of Molecular Genetics, National Institute of Oncology, Budapest, 1122, Hungary [199]Center for Clinical Cancer Genetics, The University of Chicago, Chicago, IL, 60637, USA [200]Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, 222 42, Sweden [201]Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast, Ireland, BT7 1NN, UK [202]University of New Mexico Health Sciences Center, University of New Mexico, Albuquerque, NM, 87131, USA [203]Cancer Research UK Clinical Trials Unit, Institute of Cancer Sciences, University of Glasgow, Glasgow, G12 0YN, UK [204]Molecular Diagnostics, Aalborg University Hospital, Aalborg, 9000, Denmark [205]Clinical Cancer Research Center, Aalborg University Hospital, Aalborg, 9000, Denmark [206]Department of Clinical Medicine, Aalborg University, Aalborg, 9000, Denmark [207]Genome Diagnostics Program, IFOM - the FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan, 20139, Italy [208]Translational Research Laboratory, IDIBELL (Bellvitge Biomedical Research Institute),Catalan Institute of Oncology, CIBERONC, Barcelona, 08908, Spain [209]Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, 90570, Finland [210]Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, 90570, Finland [211]Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan, 20133, Italy [212]School of Women's and Children's Health, Faculty of Medicine, University of NSW Sydney, Sydney, New South Wales, 2052, Australia [213]The Kinghorn Cancer Centre, Garvan Institute of Medical Research, Sydney, New South Wales, 2010, Australia [214]Clinical Genetics, Karolinska Institutet, Stockholm, 171 76, Sweden [215]Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa, 35254, Israel [216]Chronic Disease Epidemiology, Yale School of Public Health, New Haven, CT, 06510, USA [217]Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, 28222, Spain [218]Department of Oncology, University Hospital of Larissa, Larissa, 411 10, Greece [219]Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, 27709, USA [220]Fundación Pública Galega Medicina Xenómica, Santiago De Compostela, 15706, Spain [221]Instituto de Investigación Sanitaria de Santiago de Compostela, Santiago De Compostela, 15706, Spain [222]School of Medicine, University of Western Australia, Perth, Western Australia, Australia [223]Research Oncology, Guy’s Hospital, King's College London, London, SE1 9RT, UK [224]Faculty of Information Technology, Monash University, Melbourne, Victoria, 3800, Australia [225]National Center for Tumor Diseases, University Hospital and German Cancer Research Center, Heidelberg, 69120, Germany [226]Network Aging Research, University of Heidelberg, Heidelberg, 69115, Germany [227]Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, New South Wales, 2305, Australia [228]Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Callaghan, New South Wales, 2308, Australia [229]Hunter

Medical Research Institute, John Hunter Hospital, Newcastle, New South Wales, 2305, Australia [230]Clinical Cancer Genetics Program, Division of Human Genetics, Department of Internal Medicine, The Comprehensive Cancer Center, The Ohio State University, Columbus, OH, 43210, USA [231]Department of Internal Medicine, Division of Medical Oncology, University of Kansas Medical Center, Westwood, KS, 66205, USA [232]Institute of Biomedical Sciences, Academia Sinica, Taipei, 115, Taiwan [233]School of Public Health, China Medical University, Taichung, Taiwan [234]Dept of OB/GYN and Comprehensive Cancer Center, Medical University of Vienna, Vienna, 1090, Austria [235]Clinical Cancer Genomics, City of Hope, Duarte, CA, 91010, USA [236]Ss. Cyril and Methodius University in Skopje, Medical Faculty, University Clinic of Radiotherapy and Oncology, Skopje, 1000, Republic of North Macedonia [237]Department of Clinical Pathology, The University of Melbourne, Melbourne, Victoria, 3010, Australia [238]Population Oncology, BC Cancer, Vancouver, BC, V5Z 1G1, Canada [239]School of Population and Public Health, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada [240]The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Western Australia, 6000, Australia [241]Service de Génétique, Institut Curie, Paris, 75005, France [242]Université Paris Descartes, Paris, 75006, France [243]Institute of Human Genetics, University Hospital Heidelberg, Heidelberg, 69120, Germany [244]Division of Breast Cancer Research, The Institute of Cancer Research, London, SW7 3RP, UK [245]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, 02115, USA [246]Dept of OB/GYN, Medical University of Vienna, Vienna, 1090, Austria [247]Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, 27709, USA [248]Department of Genetics, Portuguese Oncology Institute, Porto, 4220-072, Portugal [249]Biomedical Sciences Institute (ICBAS), University of Porto, Porto, 4050-013, Portugal [250]Cancer Center, Kuopio University Hospital, Kuopio, 70210, Finland [251]Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio, 70210, Finland [252]Breast Cancer Research Programme, Cancer Research Malaysia, Subang Jaya, Selangor, 47500, Malaysia [253]Department of Surgery, Faculty of Medicine, University Malaya, Kuala Lumpur, 50603, Malaysia [254]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, 10032, USA [255]Hereditary Cancer Program, ONCOBELL-IDIBELL-IDIBGI-IGTP, Catalan Institute of Oncology, CIBERONC, Barcelona, Spain [256]Department of Clinical Genetics, Odense University Hospital, Odence C, 5000, Denmark [257]Department of Medicine, Magee-Womens Hospital, University of Pittsburgh School of Medicine, Pittsburgh, PA, 15213, USA [258]Department of Medical Genetics, University of Cambridge, Cambridge, CB2 0QQ, UK [259]Department of Cancer Biology and Genetics, The Ohio State University, Columbus, OH, 43210, USA [260]Department of Surgery, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands [261]Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT, UK [262]Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre,

University of Oxford, Oxford, OX3 7BN, UK [263]Center for Population Health Research, National Institute of Public Health, Cuernavaca, Morelos, 62100, Mexico [264]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 [265]Department of Medical Oncology, Beth Israel Deaconess Medical Center, Boston, MA, 02215, USA [266]Department of Pathology, University Hospital of Heraklion, Heraklion, 711 10, Greece [267]Frauenklinik der Stadtklinik Baden-Baden, Baden-Baden, 76532, Germany [268]Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, MN, 55905, USA [269]Department of Clinical Genetics, Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands [270]Department of Genetics, University of Pretoria, Arcadia, 0007, South Africa [271]Fundación Pública galega Medicina Xenómica-SERGAS, Grupo de Medicina Xenómica-USC, CIBERER, IDIS, Santiago de Compostela, Spain [272]Division of Functional onco-genomics and genetics, Centro di Riferimento Oncologico di Aviano (CRO), IRCCS, Aviano, 33081, Italy [273]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, 27709, USA [274]Department of Surgical Sciences, Uppsala University, Uppsala, 751 05, Sweden [275]Division of Genetics and Epidemiology, Institute of Cancer Research, London, SM2 5NG, UK [276]Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, 1066 CX, The Netherlands [277]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo, 0379, Norway [278]Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, 0450, Norway

Acknowledgments

We thank all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. This work was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 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 Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project (NIH Grants U19 CA148065 and X01HG007492); and Cancer Research UK (C1287/A10118 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 The National Institute of Health (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 Supplementary Note.

References

1. Milne RL, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet. 2017; 49:1767–1778. [PubMed: 29058716]

2. Michailidou K, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017; 551:92–+. [PubMed: 29059683]
3. Ghoussaini M, et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nat Commun. 2014; 4:4999. [PubMed: 25248036]
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. 2016; 25:3863–3876. [PubMed: 27402876]
5. Guo X, et al. Fine-scale mapping of the 4q24 locus identifies two independent loci associated with breast cancer risk. Cancer Epidemiol Biomarkers Prev. 2015; 24:1680–91. [PubMed: 26354892]
6. Glubb DM, 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. 2015; 96:5–20. [PubMed: 25529635]
7. Dunning AM, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet. 2016; 48:374–86. [PubMed: 26928228]
8. Shi J, et al. Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. Int J Cancer. 2016; 139:1303–1317. [PubMed: 27087578]
9. Orr N, et al. Fine-mapping identifies two additional breast cancer susceptibility loci at 9q31.2. Hum Mol Genet. 2015; 24:2966–84. [PubMed: 25652398]
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. 2015; 97:22–34. [PubMed: 26073781]
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. 2016; 6:32512. [PubMed: 27600471]
12. Meyer KB, et al. Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am J Hum Genet. 2013; 93:1046–60. [PubMed: 24290378]
13. Betts JA, 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. 2017; 101:255–266. [PubMed: 28777932]
14. French JD, et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am J Hum Genet. 2013; 92:489–503. [PubMed: 23540573]
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. 2016; 99:903–911. [PubMed: 27640304]
16. Horne HN, et al. Fine-Mapping of the 1p11.2 Breast Cancer Susceptibility Locus. PLoS One. 2016; 11:e0160316. [PubMed: 27556229]
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. 2016; 18:64. [PubMed: 27459855]
18. Lin WY, et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. Hum Mol Genet. 2015; 24:285–98. [PubMed: 25168388]
19. Bojesen SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet. 2013; 45:371–84. [PubMed: 23535731]
20. Lawrenson K, et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. Nat Commun. 2016; 7:12675. [PubMed: 27601076]
21. Amos CI, et al. The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev. 2017; 26:126–135. [PubMed: 27697780]
22. Michailidou K, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet. 2013; 45:353–61. [PubMed: 23535729]

23. Michailidou K, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nature Genetics. 2015; 47:373–U127. [PubMed: 25751625]
24. Udler MS, Tyrer J, Easton DF. Evaluating the power to discriminate between highly correlated SNPs in genetic association studies. Genet Epidemiol. 2010; 34:463–8. [PubMed: 20583289]
25. Mavaddat N, Antoniou AC, Easton DF, Garcia-Closas M. Genetic susceptibility to breast cancer. Mol Oncol. 2010; 4:174–91. [PubMed: 20542480]
26. Lakhani SR, et al. Prediction of BRCA1 status in patients with breast cancer using estrogen receptor and basal phenotype. Clin Cancer Res. 2005; 11:5175–80. [PubMed: 16033833]
27. Taberlay PC, Statham AL, Kelly TK, Clark SJ, Jones PA. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 2014; 24:1421–32. [PubMed: 24916973]
28. Hnisz D, et al. Super-enhancers in the control of cell identity and disease. Cell. 2013; 155:934–47. [PubMed: 24119843]
29. Farh KK, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature. 2015; 518:337–43. [PubMed: 25363779]
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. 2012; 44:1191–8. [PubMed: 23001124]
31. Kichaev G, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 2014; 10:e1004722. [PubMed: 25357204]
32. Quiroz-Zarate A, et al. Expression Quantitative Trait loci (QTL) in tumor adjacent normal breast tissue and breast tumor tissue. PLoS One. 2017; 12:e0170181. [PubMed: 28152060]
33. Cancer Genome Atlas Research, N. et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013; 45:1113–20. [PubMed: 24071849]
34. Curtis C, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012; 486:346–52. [PubMed: 22522925]
35. Ciriello G, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell. 2015; 163:506–19. [PubMed: 26451490]
36. Nik-Zainal S, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. Nature. 2016; 534:47–54. [PubMed: 27135926]
37. Pereira B, et al. The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. Nat Commun. 2016; 7:11479. [PubMed: 27161491]
38. Cancer Genome Atlas, N. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]
39. Bailey MH, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018; 173:371–385 e18. [PubMed: 29625053]
40. Lambert SA, et al. The Human Transcription Factors. Cell. 2018; 172:650–665. [PubMed: 29425488]
41. Artero-Castro A, et al. Disruption of the ribosomal P complex leads to stress-induced autophagy. Autophagy. 2015; 11:1499–519. [PubMed: 26176264]
42. Wang XY, et al. Musashi1 modulates mammary progenitor cell expansion through proliferin-mediated activation of the Wnt and Notch pathways. Mol Cell Biol. 2008; 28:3589–99. [PubMed: 18362162]
43. Vijayan D, Young A, Teng MWL, Smyth MJ. Targeting immunosuppressive adenosine in cancer. Nat Rev Cancer. 2017; 17:709–724. [PubMed: 29059149]
44. Takebe N, et al. Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. Nat Rev Clin Oncol. 2015; 12:445–64. [PubMed: 25850553]
45. Thorpe LM, Yuzugullu H, Zhao JJ. PI3K in cancer: divergent roles of isoforms, modes of activation and therapeutic targeting. Nat Rev Cancer. 2015; 15:7–24. [PubMed: 25533673]
46. Nusse R, Clevers H. Wnt/beta-Catenin Signaling, Disease, and Emerging Therapeutic Modalities. Cell. 2017; 169:985–999. [PubMed: 28575679]
47. Massague J. TGFbeta signalling in context. Nat Rev Mol Cell Biol. 2012; 13:616–30. [PubMed: 22992590]

48. Meeks HD, et al. BRCA2 Polymorphic Stop Codon K3326X and the Risk of Breast, Prostate, and Ovarian Cancers. J Natl Cancer Inst. 2016; 108
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. 2004; 74:1175–82. [PubMed: 15122511]
50. Schmidt MK, et al. Age- and Tumor Subtype-Specific Breast Cancer Risk Estimates for CHEK2*1100delC Carriers. J Clin Oncol. 2016; 34:2750–60. [PubMed: 27269948]
51. Kilpivaara O, et al. CHEK2 variant I157T may be associated with increased breast cancer risk. Int J Cancer. 2004; 111:543–7. [PubMed: 15239132]
52. Muranen TA, et al. Patient survival and tumor characteristics associated with CHEK2:p.I157T - findings from the Breast Cancer Association Consortium. Breast Cancer Res. 2016; 18:98. [PubMed: 27716369]
53. Killedar A, et al. A Common Cancer Risk-Associated Allele in the hTERT Locus Encodes a Dominant Negative Inhibitor of Telomerase. PLoS Genet. 2015; 11:e1005286. [PubMed: 26053551]
54. De Blasio A, et al. Unusual roles of caspase-8 in triple-negative breast cancer cell line MDA-MB-231. Int J Oncol. 2016; 48:2339–48. [PubMed: 27082853]
55. Haupt S, et al. Targeting Mdmx to treat breast cancers with wild-type p53. Cell Death Dis. 2015; 6:e1821. [PubMed: 26181202]
56. Pandya PH, Murray ME, Pollok KE, Renbarger JL. The Immune System in Cancer Pathogenesis: Potential Therapeutic Approaches. J Immunol Res. 2016; 2016
57. Gionet N, Jansson D, Mader S, Pratt MA. NF-kappaB and estrogen receptor alpha interactions: Differential function in estrogen receptor-negative and -positive hormone-independent breast cancer cells. J Cell Biochem. 2009; 107:448–59. [PubMed: 19350539]
58. Fleischer T, et al. DNA methylation at enhancers identifies distinct breast cancer lineages. Nat Commun. 2017; 8:1379. [PubMed: 29123100]
59. Couch FJ, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet. 2013; 9:e1003212. [PubMed: 23544013]
60. Gaudet MM, et al. Identification of a BRCA2-specific modifier locus at 6p24 related to breast cancer risk. PLoS Genet. 2013; 9:e1003173. [PubMed: 23544012]
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. 2007; 39:906–13. [PubMed: 17572673]
62. Antoniou AC, 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. 2007; 81:1186–200. [PubMed: 17999359]
63. Barnes DR, et al. Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations. Genet Epidemiol. 2012; 36:274–91. [PubMed: 22714938]
64. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010; 26:2190–1. [PubMed: 20616382]
65. Zhong H, Prentice RL. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. Biostatistics. 2008; 9:621–34. [PubMed: 18310059]
66. Hunter DJ, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet. 2007; 39:870–4. [PubMed: 17529973]
67. Baran Y, et al. Fast and accurate inference of local ancestry in Latino populations. Bioinformatics. 2012; 28:1359–67. [PubMed: 22495753]
68. Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet. 2012; 44:955–9. [PubMed: 22820512]
69. Genomes Project C. et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012; 491:56–65. [PubMed: 23128226]

70. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]
71. Mermel CH, et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. 2011; 12:R41. [PubMed: 21527027]
72. Li Q, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell. 2013; 152:633–41. [PubMed: 23374354]
73. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012; 28:1353–8. [PubMed: 22492648]
74. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
75. Sloan CA, et al. ENCODE data at the ENCODE portal. Nucleic Acids Res. 2016; 44:D726–32. [PubMed: 26527727]
76. Roadmap Epigenomics C. et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015; 518:317–30. [PubMed: 25693563]
77. Stunnenberg HG, International Human Epigenome C. Hirst M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. Cell. 2016; 167:1897.
78. Pellacani D, et al. Analysis of Normal Human Mammary Epigenomes Reveals Cell-Specific Active Enhancer States and Associated Transcription Factor Networks. Cell Rep. 2016; 17:2060–2074. [PubMed: 27851968]
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. 2018; 46:D267–D275. [PubMed: 29126285]
80. Pruitt KD, et al. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 2014; 42:D756–63. [PubMed: 24259432]
81. Harrow J, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012; 22:1760–74. [PubMed: 22955987]
82. Wang J, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012; 22:1798–812. [PubMed: 22955990]
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. 2016; 44:D110–5. [PubMed: 26531826]
84. Tan G, Lenhard B. TFBSTools: an R/bioconductor package for transcription factor binding site analysis. Bioinformatics. 2016; 32:1555–6. [PubMed: 26794315]
85. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011; 27:1017–8. [PubMed: 21330290]
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. 2015; 10:e0143627. [PubMed: 26599758]
87. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–2. [PubMed: 20110278]
88. McLeay RC, Bailey TL. Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data. BMC Bioinformatics. 2010; 11:165. [PubMed: 20356413]
89. Kichaev G, et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics. 2017; 33:248–255. [PubMed: 27663501]
90. McLaren W, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016; 17:122. [PubMed: 27268795]
91. Adzhubei IA, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–9. [PubMed: 20354512]
92. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4:1073–81. [PubMed: 19561590]

Europe PMC Funders Author Manuscripts

93. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005; 15:978–86. [PubMed: 15965030]
94. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J Comput Biol. 2004; 11:377–94. [PubMed: 15285897]
95. Desmet FO, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Res. 2009; 37:e67. [PubMed: 19339519]
96. Beesley J, et al. Chromatin interactome mapping at 141 independent breast cancer risk signals.
97. Fullwood MJ, et al. An oestrogen-receptor-alpha-bound human chromatin interactome. Nature. 2009; 462:58–64. [PubMed: 19890323]
98. 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. 2014; 24:1–13. [PubMed: 24196873]
99. Andersson R, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507:455–461. [PubMed: 24670763]
100. Moradi Marjaneh M, et al. High-throughput allelic expression imbalance analyses identify 14 candidate breast cancer risk genes.
101. Dixon JR, et al. Integrative detection and analysis of structural variation in cancer genomes. Nat Genet. 2018; 50:1388–1398. [PubMed: 30202056]
102. Yang Y, et al. AWESOME: a database of SNPs that affect protein post-translational modifications. Nucleic Acids Res. 2019; 47:D874–D880. [PubMed: 30215764]
103. Beesley J, et al. Chromatin interactome mapping at 139 independent breast cancer risk signals. 2019
104. Merico D, Isserlin R, Bader GD. Visualizing gene-set enrichment results using the Cytoscape plug-in enrichment map. Methods Mol Biol. 2011; 781:257–77. [PubMed: 21877285]
105. Vastrik I, et al. Reactome: a knowledge base of biologic pathways and processes. Genome Biol. 2007; 8:R39. [PubMed: 17367534]
106. Schaefer CF, et al. PID: the Pathway Interaction Database. Nucleic Acids Res. 2009; 37:D674–9. [PubMed: 18832364]
107. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25:25–9. [PubMed: 10802651]
108. Romero P, et al. Computational prediction of human metabolic pathways from the complete human genome. Genome Biol. 2005; 6:R2. [PubMed: 15642094]
109. Subramanian A, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005; 102:15545–50. [PubMed: 16199517]
110. Kandasamy K, et al. NetPath: a public resource of curated signal transduction pathways. Genome Biol. 2010; 11:R3. [PubMed: 20067622]
111. Thomas PD, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13:2129–41. [PubMed: 12952881]
112. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B-Methodological. 1995; 57:289–300.

Europe PMC Funders Author Manuscripts

Editorial summary

Fine-mapping of causal variants and integration of epigenetic and chromatin conformation data identify likely target genes for 150 breast cancer risk regions.

Europe PMC Funders Author Manuscripts

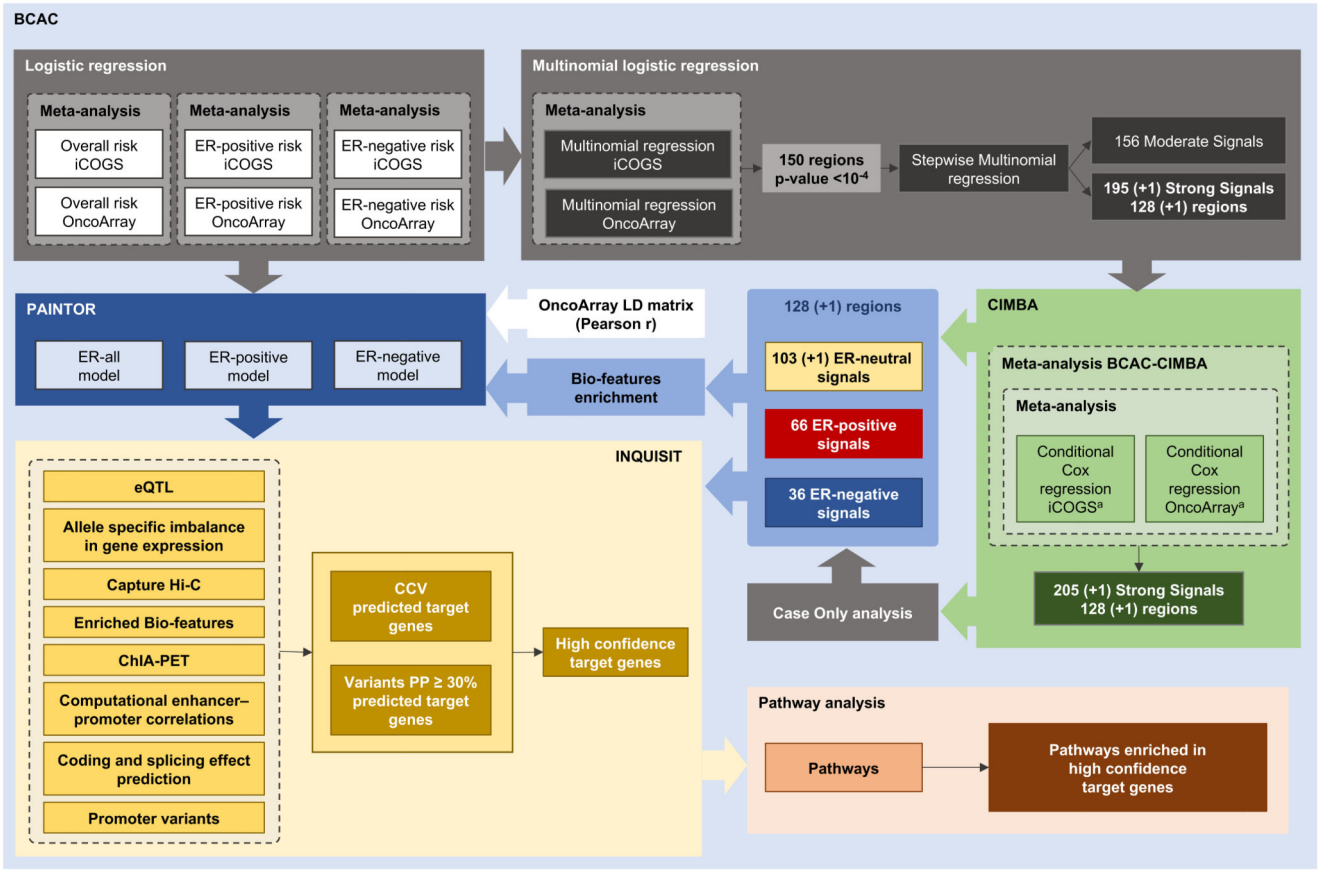

Figure 1. Flowchart summarizing the study design.

Logistic regression summary statistics were used to select the final set of variants to run stepwise multinomial regression. These results were meta-analysed with CIMBA to provide the final set of strong independent signals and their CCVs. Through a 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 LD. Both multinomial regression CCVs and PAINTOR high Posterior Probability 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.

[a] conditional on the index variants from BCAC strong signals.

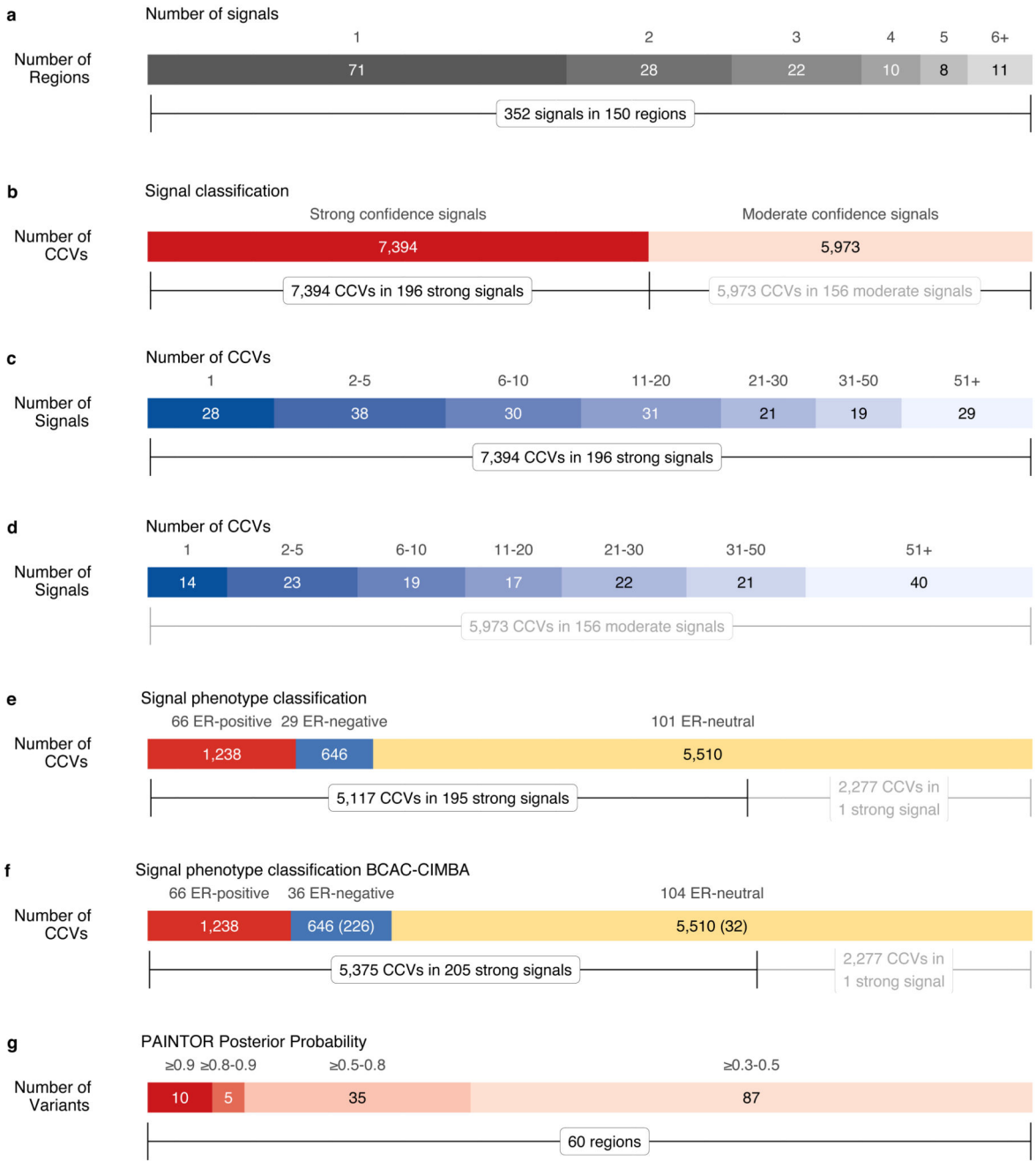

Figure 2. Determining independent risk signals and credible candidate variants (CCVs).

(a) Number of independent signals per region identified through multinomial stepwise logistic regression. (b) Signal classification according to their confidence into strong and moderate confidence signals. (c) Number of CCVs per signal at strong confidence signals identified through multinomial stepwise logistic regression. (d) Number of CCVs per signal at moderate confidence signals identified through multinomial stepwise regression. (e) Subtype classification of strong signals into ER-positive, ER-negative and signals equally associated with both phenotypes (ER-neutral) from BCAC analysis. (f) Subtype

classification from the meta-analysis of BCAC and CIMBA. Between brackets, number of CCVs from the meta-analysis of BCAC and CIMBA. (g) Number of variants at different posterior probability thresholds. 15 variants reach a PP ≥ 80% by at least one of the three models (ER-all, ER-positive, ER-negative).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

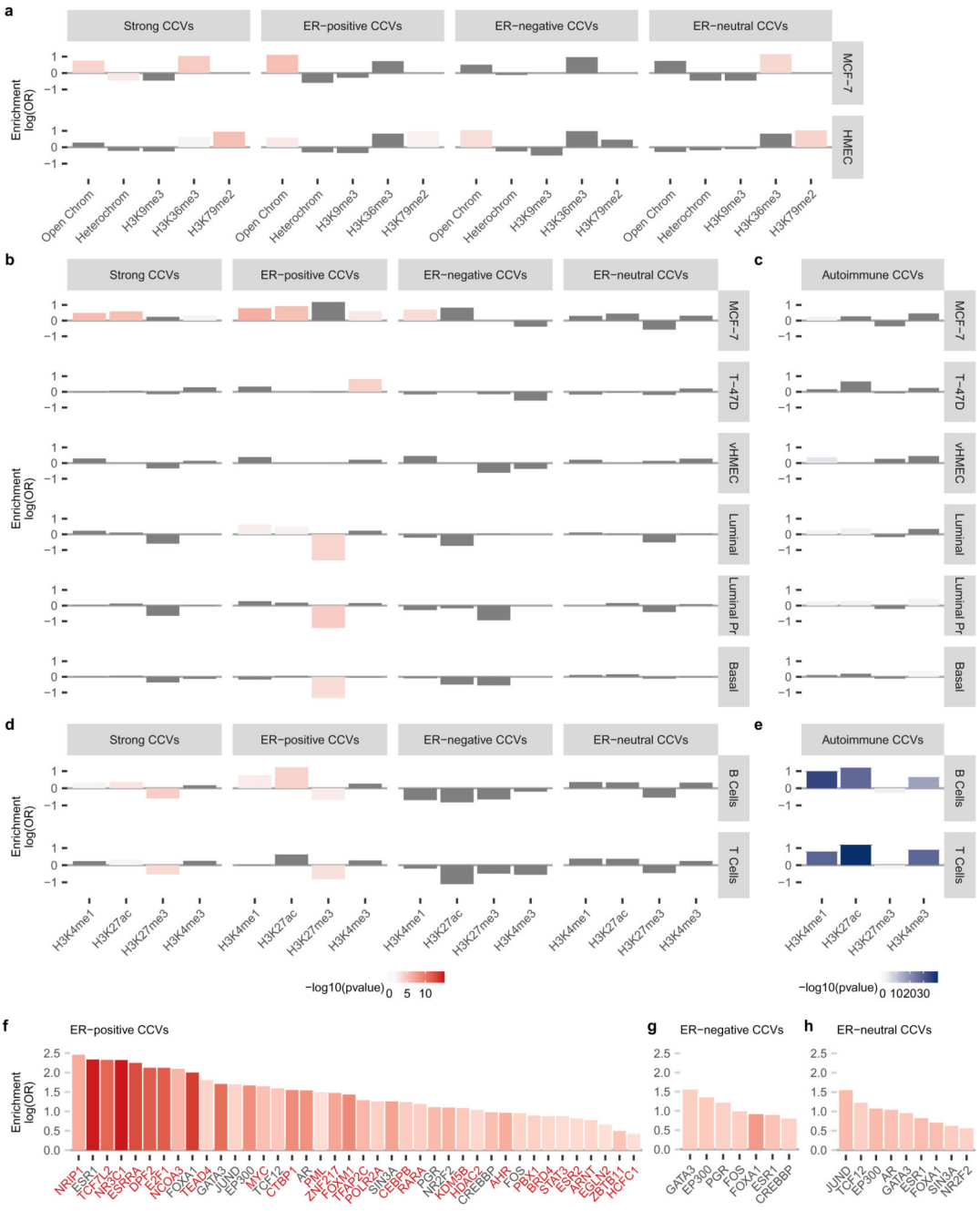

Figure 3. Overlap of CCVs with gene regulatory regions gene bodies and transcription factor binding sites.

(a) Breast cancer CCVs overlap with chromatin states and broad breast cells epigenetic marks. (b) Breast cancer CCVs overlap with breast cells epigenetic marks. (c) Autoimmune CCVs overlap with breast cells epigenetic marks. (d) Breast cancer CCVs overlap with autoimmune-related epigenetic marks. (e) Autoimmune CCVs overlap with autoimmune-related epigenetic marks. (f) Significant ER-positive CCVs overlap with transcription factors binding sites. TFBSs found significant for ER-positive CCVs are highlighted in red (x axis labels). (g) Significant ER-negative CCVs overlap with transcription factors binding sites.

(h) Significant ER-neutral CCVs overlap with transcription factors binding sites. Strong column: analysis with all CCVs at strong signals. ER-positive, ER-negative, ER-neutral: analysis of CCVs at strong signals stratified by phenotype. Logistic regression robust variance estimation for clustered observations, Wald test X^2 p-values estimated using 67,136 ER-positive and 17,506 ER-negative cases, together with 88,937 controls.
Non-significant p-values are noted as dark grey. Significance defined as FDR 5%, which corresponds to the following P-value thresholds: Strong signals P-value = $1.66x10^{-2}$, ER-positive P-value = $2.42x10^{-2}$; ER-negative P-value $3.02x10^{-3}$; ER-neutral P-value = $1.76x10^{-3}$.

Europe PMC Funders Author Manuscripts

Figure 4. Predicted target genes are enriched in known breast cancer driver genes and transcription factors.

79 target genes that fulfil at least one of the following criteria: are targeted by more than one independent signal, are known driver genes, transcription factor genes, or their binding sites (ChIP-Seq BS) or consensus motif (TF Motif) are significantly overlapped by CCVs.

*Genes with published functional follow up.

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

Figure 5. Predicted target genes by phenotype and significantly enriched pathways.

(a) Venn diagram showing the associated phenotype (ER-positive, ER-negative, ER-neutral) for the Level 1 target genes, predicted by the CCVs and HPPVs. * ER-positive or ER-negative target genes also targeted by ER-neutral signals. (b) Heatmap showing clustering of pathway themes over-represented by INQUISIT Level 1 target genes. Color represents the relative number of genes per phenotype within enriched pathways, grouped by common themes. ER-positive, ER-negative, ER-neutral, and all phenotypes together (strong).

Europe PMC Funders Author Manuscripts

Table 1

Signals with single CCVs and variants with PP > 80%

								ER-negative		ER-positive					
Fine-mapping region[a]	**Variant**[b]	**Ref/Alt**[c]	**EAF**[d]	**PP**[e]	**Model**[f]	**Signal**[g]	**N CCV**[h]	**OR**[i]	**(95%CI)**	**OR**[i]	**(95%CI)**	**P-value**[i]	**FP**[j]	**Predicted target gene(s)**[k]	**Confidence**[l]
chr1:120723447 -121780613	rs11249433	A/G	0.42	0.57	ERALL	Signal 1	1	1.02	(0.99-1.04)	1.13	(1.11-1.15)	$8.11x10^{-60}$	na	na	
chr1:200937832 -201937832	rs35383942	C/T	0.06	0.96	ERALL	Signal 1	2	1.10	(1.05-1.16)	1.09	(1.06-1.13)	$1.14x10^{-7}$	D	*TNNI1*	Level 1
chr2:201681247 -202681247	rs3769821	C/T	0.66	0.40	ERALL	Signal 1	1	0.94	(0.92-0.97)	0.95	(0.93-0.96)	$1.46x10^{-12}$	D	*ALS2CR12*	Level 1
chr2:217405832 -218796508	rs4442975 [n]	G/T	0.48	0.84	ERALL	Signal 1	1	0.94	(0.92-0.97)	0.86	(0.85-0.87)	$2.50x10^{-90}$	D	*IGFBP5*[m]	Level 2
chr4:105569013 -106856761	esv3601665	-/Alu	0.07	0.95	ERPOS			1.01	(0.95-1.08)	1.10	(1.06-1.14)	$3.27x10^{-6}$	D	*ARHGEF38, AC004066.3*	Level 1
chr5:779790 -1797488	rs10069690	C/T	0.27	0.58	ERNEG	Signal 1	1	1.18	(1.15-1.21)	1.03	(1.01-1.05)	$1.20x10^{-34}$	D	*SLC6A18, TERT*[m]	Level 2
chr5:44013304 -45206498	rs10941679	A/G	0.26	0.00	ERPOS	Signal 1	1	1.04	(1.02-1.07)	1.17	(1.15-1.19)	$1.50x10^{-77}$	D	*MRPS30*	Level 2
	rs5867671	A/-	0.77	0.01	ERPOS	Signal 2	1	0.91	(0.89-0.94)	0.99	(0.97-1.01)	$2.25x10^{-9}$	na	na	
chr5:44013304 -45206498	rs190443933	T/C	0.01	0.00	ERALL	Signal 4	1	1.30	(1.14-1.48)	1.26	(1.16-1.37)	$2.32x10^{-8}$	na	na	
chr5:55531884 -56587883	rs984113	G/C	0.61	0.81	ERPOS	Signal 2	1	0.96	(0.93-0.98)	0.96	(0.94-0.97)	$3.51x10^{-8}$	D	*MAP3K1*[m]	Level 2
	rs889310	C/T	0.56	0.84	ERPOS	(Signal 6)	15	1.03	(1.00-1.05)	1.05	(1.03-1.06)	$1.75x10^{-7}$	D	*MAP3K1*[m]	Level 1
chr6:15899557 -16899557	rs3819405	C/T	0.32	0.96	ERALL	Signal 1	1	0.97	(0.95-1.00)	0.95	(0.94-0.97)	$1.14x10^{-7}$	D	*ATXN1, RP1-151F17.1, RP1-151F17.2*	Level 2
chr6:151418856 -152937016	rs12173562	C/T	0.08	0.10	ERNEG	Signal 1	1	1.30	(1.25-1.36)	1.14	(1.11-1.18)	$3.98x10^{-40}$	D	*ESR1*[m]	Level 1
	rs34133739	-/C	0.53	0.25	ERALL	Signal 2	1	1.11	(1.09-1.14)	1.05	(1.04-1.07)	$2.36x10^{-22}$	D	*ESR1*[m]	Level 1
	rs851984	G/A	0.40	0.73	ERALL	Signal 3	1	1.07	(1.04-1.09)	1.05	(1.04-1.07)	$3.69x10^{-13}$	D	*ESR1*[m]	Level 1

Europe PMC Funders Author Manuscripts

								ER-negative		ER-positive					
Fine-mapping region[a]	**Variant**[b]	**Ref/Alt**[c]	**EAF**[d]	**PP**[e]	**Model**[f]	**Signal**[g]	**N CCV**[h]	**OR**[i]	**(95%CI)**	**OR**[i]	**(95%CI)**	**P-value**[i]	**FP**[j]	**Predicted target gene(s)**[k]	**Confidence**[l]
chr7:130167121-131167121	rs68056147	G/A	0.30	0.84	ERALL			1.04	(1.01-1.07)	1.05	(1.03-1.06)	$3.07x10^{-7}$	D	*MKLN1*	Level 2
chr8:127424659-130041931	rs35961416	-/A	0.41	0.68	ERALL	Signal 3	1	0.97	(0.94-0.99)	0.95	(0.93-0.96)	$9.97x10^{-11}$	D	*MYC*[m]	Level 1
chr9:21247803-22624477	rs539723051	AAAA/-	0.33	0.43	ERALL	Signal 1	1	1.08	(1.05-1.11)	1.06	(1.04-1.08)	$1.81x10^{-15}$	na	na	
chr9:109803808-111395353	rs10816625	A/G	0.07	0.95	ERPOS	Signal 3	1	1.06	(1.01-1.11)	1.13	(1.10-1.16)	$3.62x10^{-15}$	D	*KLF4*[m]	Level 2
	rs13294895	C/T	0.18	0.93	ERPOS	Signal 4	1	1.01	(0.98-1.05)	1.09	(1.07-1.11)	$4.00x10^{-17}$	D	*KLF4*[m]	Level 1
chr9:109803808-111395353	rs60037937	AA/-	0.22	0.68	ERPOS	Signal 2	1	1.02	(0.99-1.06)	1.11	(1.09-1.13)	$3.17x10^{-26}$	D	*KLF4*[m], *RAD23B*	Level 2
chr10:63758684-65063702	rs10995201	A/G	0.15	0.31	ERALL	Signal 1	1	0.91	(0.88-0.94)	0.87	(0.85-0.89)	$1.40x10^{-37}$	na	na	
chr10:122593901-123849324	rs35054928	C/-	0.56	0.60	ERALL	Signal 1	1	0.96	(0.94-0.98)	0.74	(0.73-0.76)	$6.55x10^{-342}$	D	*FGFR2*[m]	Level 1
	rs45631563[n]	A/T	0.04	0.93	ERPOS	Signal 3	1	0.97	(0.92-1.03)	0.76	(0.73-0.79)	$4.84x10^{-44}$	C	*FGFR2*[m]	Level 2
	rs7899765	T/C	0.06	0.02	ERALL	Signal 5	1	1.01	(0.97-1.06)	0.87	(0.84-0.90)	$2.21x10^{-18}$	D	*FGFR2*[m]	Level 1
chr11:68831418-69879161	rs78540526	C/T	0.09	0.91	ERPOS	Signal 1	1	1.01	(0.97-1.06)	1.40	(1.36-1.44)	$2.77x10^{-145}$	D	*CCND1*[m], *MYEOV*	Level 1
chr12:27639846-29034415	rs7297051	C/T	0.23	0.23	ERALL	Signal 1	1	0.87	(0.85-0.90)	0.89	(0.88-0.91)	$3.12x10^{-43}$	D	*CCDC91*[m], *PTHLH*[m], *RP11-967K21.1*	Level 2
chr12:115336522-116336522	rs35422	G/A	0.57	0.58	ERPOS	Signal 2	1	0.98	(0.96-1.01)	1.05	(1.03-1.07)	$4.85x10^{-10}$	D	*TBX3*	Level 1
chr14:91341069-92368623	rs7153397	C/T	0.70	0.81	ERPOS	Signal 1	3	1.01	(0.99-1.04)	1.06	(1.04-1.08)	$3.25x10^{-11}$	D,C	*CCDC88C, CTD-2547L24.4, C14orf159, GPR68, RPS6KA5, RP11-73M18.7, RP11-895M11.3*	Level 2
chr16:52038825-53038825	rs4784227	C/T	0.27	0.95	ERPOS	Signal 1	1	1.15	(1.12-1.18)	1.26	(1.24-1.28)	$4.63x10^{-160}$	D	*TOX3*[m]	Level 1

Europe PMC Funders Author Manuscripts

								ER-negative		ER-positive					
Fine-mapping region[a]	**Variant**[b]	**Ref/Alt**[c]	**EAF**[d]	**PP**[e]	**Model**[f]	**Signal**[g]	**N CCV**[h]	**OR**[i]	**(95%CI)**	**OR**[i]	**(95%CI)**	**P-value**[i]	**FP**[j]	**Predicted target gene(s)**[k]	**Confidence**[l]
chr18:23832476-25075396	rs180952292	T/C	0.01	0.01	ERNEG	Signal 4	1	1.24	(1.12-1.37)	0.98	(0.92-1.05)	$2.07x10^{-5}$	na	na	
chr18:41899590-42899590	rs9952980	T/C	0.34	0.95	ERALL	Signal 2	3	0.97	(0.94-0.99)	0.95	(0.93-0.96)	$7.43x10^{-12}$	D	*SLC14A2*	Level 2
chr20:5448227-6448227	rs16991615	G/A	0.07	0.97	ERALL	Signal 1	1	1.09	(1.04-1.15)	1.07	(1.04-1.11)	$7.89x10^{-7}$	D, C	*GPCPD1, MCM8*	Level 2
chr22:45783297-46783297	rs184070480	C/T	0.01	0.00	ERALL	Signal 2	1	1.40	(1.20-1.64)	1.01	(0.91-1.12)	$5.02x10^{-5}$	D	*ATXN10, WNT7B*	Level 2

[a] GRCh37/hg19, bp

[b] Current reference ID

[c] Reference (Ref) versus Alternative (Alt) Allele

[d] Effect allele (Alt allele) frequency in OncoArray

[e] PP: Posterior probability. Largest posterior probability in all evaluated models

[f] Model where the variant reaches the largest posterior probability

[g] Signal where the variant is included. Between brackets moderate confidence signals.

[h] Number of CCVs in the signal

[i] Multinomial logistic regression summary statistics, X^2 single variant analysis p-value, estimated using 67,136 ER-positive and 17,506 ER-negative cases, together with 88,937 controls.

[j] D: Distal regulation, P: proximal regulation, C: coding; na: prediction non available

[k] Predicted target genes with the largest confidence level for each variant. Between brackets, largest confidence level. na: prediction non available

[l] INQUISIT level of confidence

[m] Target genes with functional follow up

[n] Two variants reach PP> 0.8 in both the ERall and ERpos models; rs4442975: ERpos PP = 0.83, ERall PP = 0.84; rs45631563: ERpos PP = 0.93, ERall PP = 0.92