Computational and Systems Oncology

WILEY

COMPUTATIONAL AND SYSTEMS ONCOLOGY Open Access

ORIGINAL ARTICLE OPEN ACCESS

From Association to Mechanism: Regulatory Annotation and Pathway Mapping of Genes Surrounding Breast Cancer Risk Variants

Sultana Jannat[1] | Trina Mitra[2] | Nafisa Nawar Fariha[3] | Tasmia Tabassum[3] | Kulsuma Bahar Bethi[4] | Synchita Majumder Kaya[5] | Md. Arif Hossen[6] | Md. Saad Hossain[3] | Jobaier Ibne Deen[7] | Mohammed Emon[8] | Laila Khaleda[9] | Md. Hridoy Ahmed[9]

[1]Department of Pharmacy, International Islamic University Chittagong, Chittagong, Bangladesh | [2]Department of Pharmacy, Jashore University of Science and Technology, Jashore, Bangladesh | [3]Genetic Engineering and Biotechnology Department, East West University, Dhaka, Bangladesh | [4]Nutrition and Food Science, Institute of Nutrition and Food Science, University of Dhaka, Dhaka, Bangladesh | [5]Biotechnology and Genetic Engineering, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh | [6]Biochemistry and Molecular Biology, Mawlana Bhashani Science and Technology University, Tangail, Bangladesh | [7]Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet, Bangladesh | [8]Biotechnology and Genetic Engineering, Gopalganj Science and Technology University, Gopalganj, Bangladesh | [9]Genetic Engineering and Biotechnology Department, University of Chittagong, Chittagong, Bangladesh

Correspondence: Md. Hridoy Ahmed (hridoy.geb.cu@gmail.com)

Received: 31 March 2026 | **Revised:** 13 July 2026 | **Accepted:** 27 July 2026

Keywords: breast cancer | enhancer | genetic association | genetics | haplotype | hedgehog signaling pathway | linkage disequilibrium | wnt signaling pathway

ABSTRACT

Inherited factors account for a large share of breast cancer susceptibility, yet the biological consequences of most risk variants are still poorly understood. To address this gap, we studied 175 breast cancer risk variants confirmed by genome-wide association studies and gathered the genes that lie near them. Two complementary gene sets were assembled. The first collected every gene immediately flanking each risk variant, regardless of distance. The second kept only the closest gene together with genes tagged by variants in strong linkage disequilibrium within a 100 kb interval. Across both sets, the clearest biological signal pointed to androgen receptor-mediated signaling. Pathway mapping additionally implicated the Wnt/β-catenin and Hedgehog cascades, and this pattern was most evident in the flanking-gene set. With HaploReg, we annotated correlated variants at each locus and recovered features that mark active regulation. Enrichment testing returned a 4.8-fold excess of DNase I hypersensitivity in a breast cancer cell line and a 15.7-fold excess of enhancer motifs in a stem cell line. Together, these results indicate that breast cancer risk markers and their correlated variants influence the activity of neighboring genes and converge on a small number of signaling pathways central to mammary biology.

1 | Introduction

Genome-wide association studies have tied hundreds of common variants to cancer predisposition [1]. For breast cancer, however, each individual risk allele shifts the odds only modestly, rarely beyond an odds ratio of about 1.5. As sequencing throughput grows and study cohorts enlarge, the catalogue of susceptibility variants is expected to widen further. The harder task that follows is mechanistic, namely, pinpointing the causal variant at each locus and resolving how it acts.

Several large reference efforts make this task tractable. Projects, such as HapMap and 1000 Genomes, describe the structure of human variation, ENCODE catalogues the regulatory elements

that govern transcription, and The Cancer Genome Atlas profiles tumors at molecular resolution. Read together, these resources allow a single association signal to be placed in its regulatory and tumor-genomic context.

Breast cancer remains the leading cause of cancer death among women, and roughly one in seven women will receive the diagnosis during her lifetime. Its progression draws on many pathways, and the well-documented influence of family history points to a strong heritable component [2, 3].

Most of the 175 risk variants identified to date sit in noncoding intronic or intergenic DNA rather than in protein-coding sequence. This location suggests that their effect is regulatory. Annotating the elements around each variant, including neighboring genes, the networks those genes form, and the correlated variants in strong linkage disequilibrium, therefore, offers a route into the biology that shapes breast cancer risk. That annotation is the purpose of the present study.

2 | Methods

2.1 | Breast Cancer Risk Alleles

We assembled a panel of 175 breast cancer risk variants. These were drawn from the NHGRI GWAS catalogue as it stood in April 2013 and from the iCOGS effort of the International Collaborative Oncological Gene-Environment Study, keeping only variants that met the genome-wide significance threshold of $p < 5 \times 10^{-8}$ [1, 4]. The complete panel is given in Table S1.

2.2 | Gene List Construction

Building on the annotation strategy described by Loo and colleagues [5], we mapped each risk variant to nearby genes under two complementary rules. The first rule retained every gene immediately upstream and downstream of a variant, whatever its distance from the index marker. The second rule kept the closest gene together with any gene tagged by a correlated variant in strong linkage disequilibrium with the index marker, restricting the search to a 100 kb window on either side of the single nucleotide polymorphism. We chose the 100 kb limit because most enhancers lie within this distance of the genes they regulate.

2.3 | HaploReg Analysis

Correlated variants and their regulatory features were retrieved with HaploReg v2 (http://compbio.mit.edu/HaploReg) [6]. Linkage disequilibrium was defined in the European samples of the 1000 Genomes Project at a threshold of $r^2 \geq 0.8$. For each variant and its proxies, we extracted chromatin-state calls, DNase I hypersensitivity, evolutionary conservation from Site-specific Phylogenetic (SiPhy) analysis, and predicted effects on transcription-factor motifs, the last of these drawn from the TRANSFAC, JASPAR, and PBM collections [7, 8]. These per-variant annotations were then summarized across each gene set.

2.4 | Protein Motif Prediction

Coding variants were evaluated for their likely effect on protein function with PolyPhen and PROVEAN [9, 10]. For each amino-acid substitution we recorded the predicted category, treating scores toward the damaging end of the scale as evidence that the change could disturb protein function.

2.5 | Functional Network and Pathway Prediction

To find the networks and pathways linking our genes, we submitted each gene list to Ingenuity Pathway Analysis (IPA). A core analysis queried the IPA Knowledge Base to connect input genes through molecular interactions reported in the published literature, and the Upstream-Regulator module identified transcription factors, drugs, and other molecules whose known targets were over-represented in our lists. A regulator was judged significant when it passed a Fisher's exact test at $p < 1 \times 10^{-4}$.

2.6 | TCGA Analysis of Breast Tumor Tissue

Tumor-level alterations were examined through the cBioPortal interface to The Cancer Genome Atlas (TCGA) breast cancer cohort ($n = 236$) [11, 12]. For each gene we tabulated changes in expression, copy number, and mutation status, and we flagged a gene as altered when such changes were present in at least 5% of tumors.

2.7 | Statistical Analysis

Index variants were admitted at the genome-wide threshold of $p < 5 \times 10^{-8}$ and correlated variants at $r^2 \geq 0.8$ in the European 1000 Genomes panel. Enrichment of regulatory features was judged by comparing observed counts against those expected for matched reference variants, with significance from a binomial test and effect size expressed as the observed-to-expected fold change. Network and upstream-regulator overlaps were tested by Fisher's exact test, and the IPA network score equals the negative base-10 logarithm of that p value. Because the analysis works from a fixed list of known variants rather than a fresh sample, its weight rests on enrichment magnitude and significance rather than on a conventional power calculation.

3 | Results

3.1 | Functional Profiles of Genes Flanking Risk Variants

Mapping each of the 175 risk variants to its immediate neighbors produced a nonredundant set of 342 genes (Table S2). When this set was profiled in IPA, 61 genes carried a functional link to cancer, and 12 were specific to breast cancer, among them ARHGAP27, ARHGEF12, FGF10, BRCA2, HNF1B, KLK2, KLK3, LMTK2, NKX3-1, and SLC22A3. The analysis also resolved gene–gene interaction networks, which implies that subsets of these genes operate together. The leading network, organized around 30 focus genes, was tied to organismal, embryonic, and organ development (Figure 1).

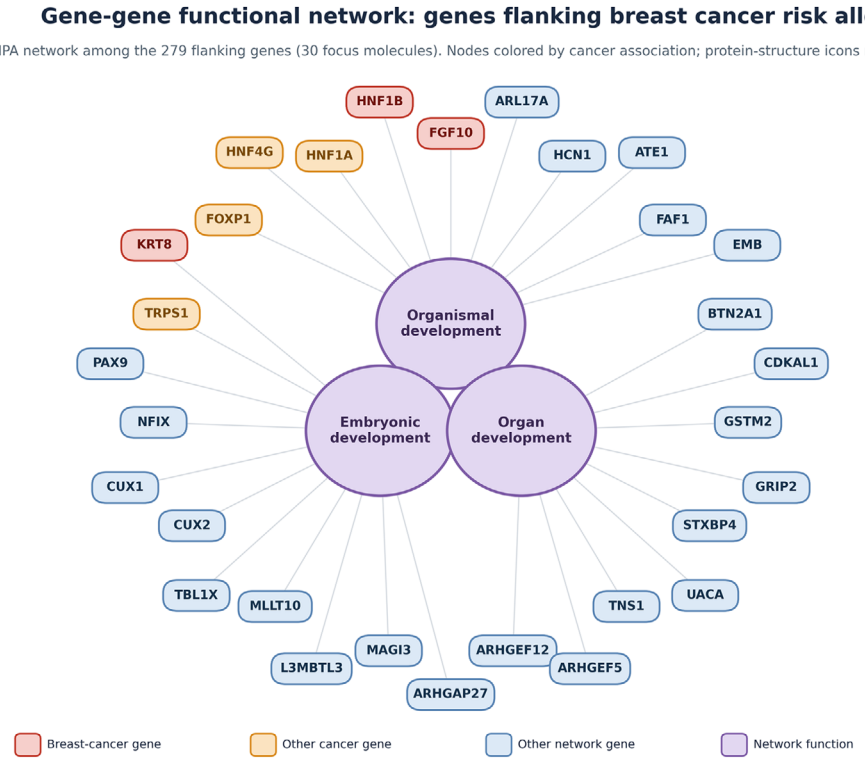

FIGURE 1 | Functional network formed by genes flanking breast cancer risk alleles. The 342 genes mapped around the 175 risk loci were analyzed in Ingenuity Pathway Analysis. The strongest network connects genes active in organismal, embryonic, and organ development. Nodes are colored by cancer association, with breast cancer genes in red, other cancer genes in amber, and the remaining genes in blue, whereas the central nodes mark the shared developmental roles. The protein-structure icons used in the original layout have been removed so that the gene labels read clearly.

Beyond this developmental network, the 342 genes populated further networks associated with connective tissue disorders, cancer, cell death and survival, cell-to-cell signaling, and organismal injury (Table 1). The breadth of these networks suggests that the genes surrounding breast cancer risk alleles are functionally interlinked and feed into processes that span cancer, signaling, development, and cell survival.

Interrogating the same 342 genes in the The Cancer Genome Atlas (TCGA) breast cancer series ($n = 236$) showed that 166 were altered in at least 8% of tumors, and eight were altered in more than 10% (Table S3).

To trace the signals that might tie these genes together, we applied the IPA Upstream-Regulator tool, which ranks candidate regulators by the overrepresentation of their targets in the input list. Of the 23 regulators returned, Hedgehog, cadmium chloride, and the Wnt effector β-catenin (CTNNB1) ranked highest, with Hedgehog showing the strongest enrichment (Fisher's exact test, $p = 1.3 \times 10^{-7}$; Table S4). These regulators point to receptor, Hedgehog, and Wnt/β-catenin signaling as recurring themes in breast cancer biology. Seventeen of the 342 genes lie downstream of these regulators, and because several of them answer to more than one regulator, the pathways themselves may intersect (Figure 2).

Concentrating on the 32 genes tied to the leading regulators, we asked whether they were altered in the 236 The Cancer Genome Atlas (TCGA) tumors. Most were, with 19 of 32 changed in at least 5% of cases. BRCA2 was the most frequently disrupted, carrying a mutation or deletion in 28 tumors. Other altered members of this regulatory network, each affected in more than 8% of tumors, included the cancer-susceptibility genes CASC15, CASC16, and CASC21, together with CASP8, Cyclin D1, CFLAR, ITGA6, CLDN11, FGFR2, FGF10, GATA5, and GREB1 (Table S3). These changes fit the idea that the risk variants modulate genes central to the onset and progression of breast cancer and that copy-number and expression shifts across this network contribute to tumor development. Figure 3 presents the top IPA functional networks overlapping the 279 genes flanking breast cancer risk alleles, ranked by the number of overlapping focus genes. Cancer-associated networks showed the strongest enrichment, with the highest overlap observed for the cancer/haematological and immunological disease network (30 genes), followed by cancer/hereditary disorder/organismal injury (28 genes). Additional enrichment was observed in DNA replication and repair, cellular assembly, lipid metabolism, and developmental pathways, indicating that breast cancer susceptibility genes are concentrated in multiple disease-related functional networks.

3.2 | Functional Profiles of Genes Near Variants in High LD With Risk Alleles

We next widened the analysis to correlated variants. In total, 3021 unique variants, including the 175 index markers among them, were in strong linkage disequilibrium with the risk alleles. Within 100 kb of these variants lay 191 annotated genes and three miRNAs (Table S5). Network analysis placed 32 of the 191 genes in the leading network, which is linked to cancer, hereditary disorder,

TABLE 1 | Interaction networks among genes flanking breast cancer risk variants, together with the diseases and biological functions associated with each network.

ID	Molecules in network	Score	Focus molecules	Top diseases and functions
1	ADRB, Alp, ANXA13, AP4B1, Beta-catenin/TCF, BMP, CaMKII, COL8A1, collagen, Cop9 signalosome, CREB5, ESRRG, Gli, Hedgehog, HIVEP3, Hsp27, LMX1B, MN1, MYL3, NFkB (complex), osteocalcin, PDGF (complex), PDGF BB, PIDD1, PLA2G6, PP1 protein complex group, proinsulin, PTHLH, ROCK, RTKN2, SETBP1, Smad, SSBP4, TCF/LEF, TENT5A	26	16	Cancer, gastrointestinal disease, organismal injury, and abnormalities
2	ADAP2, AKT1, APP, ATAD5, C11orf65, C2CD6, CALML3, CTNNB1, DGCR5, ERMN, FAM227B, GRB2, KCNH7, KCTD1, KREMEN1, LETM2, MCM8, mir-1180, mir-147, mir-637, MTCP1, NAA25, NKX2-8, PAK1/2/3, PPP1R17, RB1, RNF115, RSPH10B/RSPH10B2, SLC26A7, TNRC6B, VCL, WFDC21P, ZC3H12D, ZMIZ1, ZNF322	22	14	Cancer, organismal injury and abnormalities, respiratory disease
3	20:1 cholesteryl ester, 4732491K20Rik, ARAP2, BLVRA, c-Myc/N-Myc, cholesteryl 9-hexadecenoate, cholesteryl arachidonate, cholesteryl docosatetraenoate, cholesteryl eicosatrienoate, cholesteryl myristate, cholesteryl stearate, DEFB114, DLEU1, EFR3B, EHBP1L1, GIPR, HYCC2, KRAS, LINC-PINT, MAPK1, Max-Myc, MYC, *N*-arachidonoyl *L*-serine, NSF, OR5V1, PCAT1, Peg3, PPA2, SKIDA1, SLC22A4, SLC25A21, Sprr2g, TNF, vaccenic acid, ZNF184	22	14	Lipid metabolism, molecular transport, small molecule biochemistry
4	ARHGAP27, ARHGEF12, ARHGEF5, ARL17A/ARL17B, ATE1, BTN2A1, CDKAL1, CG, CUX1, CUX2, EMB, FAF1, FGF10, FOXP1, glutathione peroxidase, GRIP2, GSTM2, GTPase, HCN1, HNF1A, HNF1B, HNF4G, Jnk, KRT8, L3MBTL3, MAGI3, MLLT10, NFIX, PAX9, Sod, STXBP4, TBL1X, TNS1, TRPS1, UACA	61	30	Cancer, hematological disease, immunological disease
5	ASTN2, ATG10, BRCA2, CD160, CHEK2, CNTD1, cytochrome *C*, DNA damage response, EXO1, GREB1, homologous repair, homologous repair syndrome, INCENP, JMJD1C, K channel, KCNU1, LATS, MDM4, MTMR11, NEK10, PALB2, PHLDA3, POGLUT3, RAD51B, SCAMP2, SMC2, TEAD, TP53, TRIM46, TTC28, VGLL3, XPNPEP3, ZBED9, ZKSCAN3, ZNF365	55	28	Cancer, hereditary disorder, organismal injury and abnormalities
6	ABCA8, actin, alpha catenin, ANKIB1, ASH1L, Ck2, CNTNAP1, DCLRE1B, EPB41L3, F actin, H4C13, Hif1, histone h2a, histone h3, histone h4, LGR6, MAPT, ngf, PIP4K2A, RALY, RIN3, RNF146, Rnr, RPS18, Spectrin, TCF7L2, TERT, TET2, TTBK1, tubulin, Vegf, WDR43, XRCC6, ZBTB38, ZBTB7B	39	22	Dermatological diseases and conditions, DNA replication, recombination, and repair, organismal injury and abnormalities
7	20s proteasome, 26s proteasome, AMFR, ATXN1, CBX8, CDKN2B-AS1, CMSS1, collagenase, DAP3, DNAJC1, ESR1, H2BC6, HSP, Hsp70, Hsp90, HSPA4, MAFB, MFN1, MIER3, NBPF10 (includes others), Nos, PCNT, PEX14, PP2A, PRKAA, protein phosphatase, RNA polymerase II, SUB1, TEFM, Tlr, TOX3, TRAK2, transcription factor, Ubiquitin, ZC3H7B	37	21	Cardiovascular disease, cellular assembly and organization, organismal injury and abnormalities
8	ADAM29, alpha tubulin, calcineurin A, calpain, CDYL2, CK1, collagen Type I (complex), collagen Type IV, DNER, ERBB4, ERK, Fcer1, Fgf, FGFR, FGFR2, focal adhesion kinase, G protein alpha I, H/K/NRAS, Integrin, KRIT1, Mapk, metalloprotease, Mmp, MRTFA, NREP, Rab5, Rac	18	12	Embryonic development, organ development, organ morphology
9	14-3-3, AMPK, BCR (complex), CASP8, caspase, Caspase 3/7, CD3, Cyclin A, Cyclin B, EGR2, EIF2AK3, FOS (family), Ifn gamma, Ige, IkB, IKK (complex), Importin alpha, JINK1/2, KCNK2, MAP1LC3, MAP2K1/2, MAP3K1, Mek, NELFA, Nfat (family), PARP, PARP8, PITPNB, Pkc(s), SIAH2, SLC14A2, TAB2, TCR, Tnf (family), TSPAN16	18	12	Cellular function and maintenance, embryonic development, organismal development

(Continues)

TABLE 1 | (Continued)

ID	Molecules in network	Score	Focus molecules	Top diseases and functions
10	ATXN7, CARD6, Cdk, Ctbp, Cyclin D, Cyclin E, E2f, ERK1/2, ETS, GADD45, GAREM1, GATAD2A, Gcn5l, H19, Hdac, HISTONE, histone H2b, LDL-cholesterol, LINE1, Mta, NF-Y, Pias, PRC2, Rb, RBBP8, SGCE, SIN3A, Smad2/3-Smad4, Srebp, SUMO, TBX3, TCF, Top2, WNT3, ZBTB46	16	11	Embryonic development, organismal development, tissue development
11	ADCY, AKAP9, ANTXR1, calmodulin, Cdc2, CDKN1A, COL1A1, Collagen alpha-1, Collagen Type i (family), collagen(s), EMID1, Fc gamma receptor, FSH, GNAO1, GNRH, IgG, Immunoglobulin, Insulin, ITPR1, Laminin (complex), Lh, NMDA Receptor, P38 MAPK, p70 S6k, Pde4, PDE4D, PHGDH, Pka, Pka catalytic subunit, PLC, Ppp2c, SPAG6, Tgf beta, TGFBR2, TSH	16	11	Embryonic development, organismal development, tissue development
12	ADCY3, ADSS1, beta-estradiol, Brd4, CAK, CCDC170, CCNH, CDK6, CUZD1, DBH-AS1, Ewsr1, FAM3D, FAM83F, GDF6, GPR26, GPR6, GRHL1, HORMAD2-AS1, HSPA8, JAZF1, Klre1, LINC00511, MAPK14, OSBPL7, P-TEFb, PRR15L, SHOC1, SLC35D3, SORCS3, STAT5A, TCEAL3, TENM2, YJEFN3, ZNF804A, ZSCAN16-AS1	16	11	Nervous system development and function, nucleic acid metabolism, small molecule biochemistry
13	Ap1, APOBEC3A, calcineurin protein(s), CFLAR, Creb, EBF1, EBF2, ELL, FRY, growth hormone, Gsk3, HDL, Ifn, IFN alpha/beta, IFN Beta, Igm, IL1, IL12 (complex), IL12 (family), IRF1, KIR, LDL, LSP1, NFAT (complex), NFkB (family), Notch, Nr1h, OTUD7B, PEPCK, PI3K (complex), PI3K (family), pro-inflammatory cytokine, TLR1, Tnf receptor, Trk receptor	14	10	Cellular development, embryonic development, nervous system development and function
14	Ahr-aryl hydrocarbon-Arnt, Akt, AMPA Receptor, BAIAP2L1, Bcl9-Cbp/p300-Ctnnb1-Lef/Tcf, Cbp/p300, CCDC88C, Ciap, Collagen Type VI, DNA-PK, Dynein, Esr1-Esr1-estrogen-estrogen, estrogen receptor, Foxo, Hat, histone deacetylase, MAC, N-cor, NCOA1, NRIP1, Nuclear factor 1, PKNOX2, PVT1, Rar, RPA, Rxr, SMAD1/5/9, Stat5 dimer, TACC2, TFIIH, thymidine kinase, thyroid hormone receptor, TIP60, ZFPM2, ZNF322	12	9	Cellular development, cellular growth and proliferation, developmental disorder
15	CCND1, CILP, CRLR-RAMP2 (complex), CRLR-RAMP3 (complex), cytokine, FCGR1BP, FIP1L1-PDGFRA, FLACC1, FRS2-GRB2-SHP2, FTO, Hsd3b1, IFN lambda 2/3, IgG1, inositol 4,5-bisphosphate, Interferon alpha, KANSL1, KCNN4, Klra17, MHC Class II (complex), MHC II, MTORC1, NOTCH2, Ntrk1 dimer, PI3K p85, PTPN11, RAS, Rsk, Sh2b3, SHP2-PI3K-GAB2, Smad2/3, Sos, SRC (family), STAT5a/b, Tcrb-V3, THSD4	10	8	Hematological system development and function, inflammatory response, tissue morphology
16	AQP4-AS1-211, METTL3	2	1	Cell cycle, cell death and survival, cell morphology
17	ANKLE1, BANF1	2	1	Cell cycle, connective tissue disorders, DNA replication, recombination, and repair
18	miR-4693-3p (miRNAs w/seed GAGAGUG), miR-4693-5p (miRNAs w/seed UACUGUG), MIR4693	2	1	

Note: The Score is the negative base-10 logarithm of the *p* value from a right-tailed Fisher's exact test, so a higher score reflects a lower probability that the focus genes were assembled into the network by chance and therefore marks a more significant network.

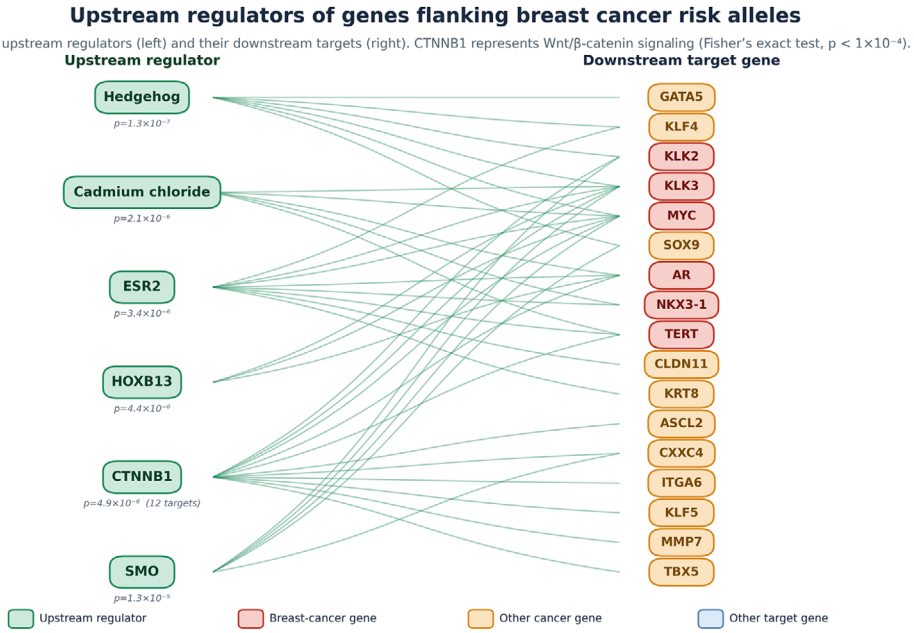

FIGURE 2 | Upstream regulators of the flanking gene set. Candidate regulators identified with the IPA Upstream-Regulator tool appear on the left and their downstream targets on the right. Each edge marks a regulator-target relationship that passed a Fisher's exact test at $p < 1 \times 10^{-4}$, and target genes are shaded by cancer association. IPA, Ingenuity Pathway Analysis.

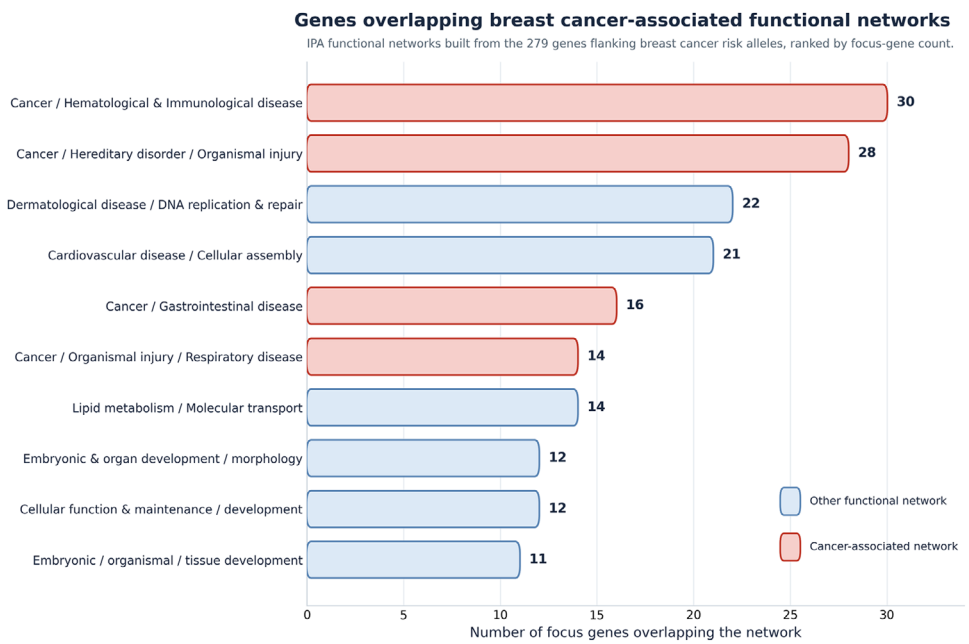

FIGURE 3 | Genes shared with breast cancer-associated functional networks. Each bar gives the number of focus genes assigned to an IPA network built from the 342 flanking genes. Networks with a direct cancer link are drawn in red. IPA, Ingenuity Pathway Analysis.

and organismal injury (Figure 4). The remaining associations spanned hereditary disorders, organismal injury, DNA replication and repair, cell morphology, and cellular maintenance (Table 2). Of the 102 focal molecules carried forward to The Cancer Genome Atlas (TCGA), 69 were altered in at least 8% of tumors and five in more than 10% (Table S5). Upstream-regulator analysis of these 102 molecules returned 20 significant regulators (Fisher's exact test, $p < 1 \times 10^{-4}$; Table S6), of which tamoxifen showed the strongest enrichment ($p = 5.6 \times 10^{-8}$). The five leading regulators, namely, tamoxifen, ARID1A, β-estradiol, TP53, and cisplatin, shared many downstream targets (Figure 5).

Tamoxifen, an antiestrogen, blocks the action of estrogen on breast tissue and can restrain the growth of estrogen-dependent tumors [13]. The shared target genes included CASP8, Cyclin D1, and CFLAR, the cancer-susceptibility genes CASC15, CASC16, and CASC21, and ITGA6, CLDN11, FGFR2, and FGF10, which take part in cell adhesion and the control of apoptosis [14].

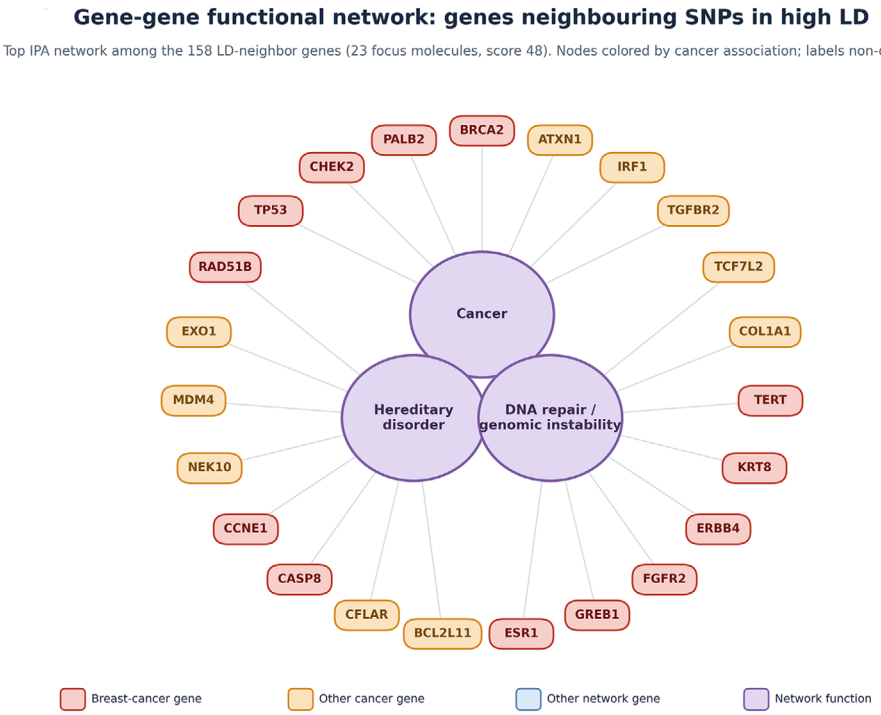

FIGURE 4 | Functional network formed by genes near variants in high linkage disequilibrium with breast cancer risk alleles. The leading network, with a score of 48 and 32 focus genes, ties to cancer, hereditary disorder, and genomic instability. Node colors match the scheme in Figure 1, and labels are spaced to prevent overlap. IPA, Ingenuity Pathway Analysis.

Taken together, these observations imply that variants close to the index markers, or in strong agreement with them, exert real control over genes in the breast cancer signaling program. Within the same The Cancer Genome Atlas (TCGA) cohort, six of the nine genes tied to the top five regulators were altered in at least 5% of tumors. The most common change was deletion of CASC15 (12.7%), whereas CASP8, CFLAR, ITGA6, FGFR2, and FGF10 were each altered in more than 8% of cases (Table S5).

3.3 | Functional Annotation of Risk Alleles and Variants in High LD

We then asked where the variants sit relative to gene structure and whether they carry regulatory marks. Each gene at the target loci was inspected to place its variants in exons, promoters, or enhancers, and HaploReg supplied ENCODE-based annotations that flag noncoding variants of likely function [6].

Among the 175 index variants (Table S7), two were missense changes in exons, one fell in a 3′ UTR, 29 lay in introns, and seven occupied evolutionarily conserved positions identified by SiPhy analysis [8]. Publicly available ENCODE ChIP-seq data, read through HaploReg, recorded transcription-factor binding at 22 of the variants [7]. In addition, 47 variants fell in DNase-hypersensitive regions, 33 carried enhancer histone marks, and five carried promoter marks. At some loci, these features coincided. The variant rs11568818 on 11q22.2, located 182 bp upstream of MMP7, combined sequence conservation, enhancer marks, and DNase hypersensitivity across several cell types, and ENCODE ChIP-seq confirmed transcription-factor binding there. Expression quantitative trait locus work has linked this locus to MMP7 transcription in liver tissue [15]. No new ChIP-seq was generated for this study, and the underlying datasets can be obtained from the ENCODE data portal (https://www.encodeproject.org).

Extending the analysis to 551 variants, comprising the 175 index markers and their strong proxies, HaploReg placed 14 in exons, 6 in 5′ UTRs, 31 in 3′ UTRs, 109 in putative promoters within 1.5 kb of a transcription start site, 72 in introns, and 319 in intergenic DNA (Table S8).

Fifteen variants fell within candidate promoters, including regions that control CCDC170 and the noncoding gene cancer susceptibility candidate 8, and further variants appeared in the 5′ and 3′ UTRs of TP53, ANKLE1, ABHD8, MDM4, SSBP4, and SLC4A7. Because untranslated regions help set translational efficiency, mRNA stability, and expression level, variants there may carry regulatory weight [16]. Testing the combined set for regulatory enrichment, enhancer marks were overrepresented 15.7-fold relative to expectation ($p = 2.9 \times 10^{-4}$), with the strongest signal in the H1 embryonic stem cell line. DNase I hypersensitivity was likewise enriched in cells of breast origin, reaching 4.8-fold ($p = 1.8 \times 10^{-3}$) above expectation in a breast cancer cell line. The convergence of enhancer and DNase marks in both stem and breast cells is strong evidence that the risk variants and their proxies act through regulatory DNA.

Of the 14 exonic variants in this set, 8 were missense and 6 synonymous. One missense change, rs6929137 at chr6:151615542, lies in strong agreement with its index marker ($r^2 = 0.99$ in Europeans) and swaps valine for isoleucine in CCDC170. PolyPhen scored this substitution at 0.761, which points to a probable loss of CCDC170

TABLE 2 | Interaction networks among genes near variants in high linkage disequilibrium with breast cancer risk variants, together with their associated diseases and biological functions.

ID	Molecules in network	Score	Focus molecules	Top diseases and functions
1	ADP Ribosyl Cyclase, ADSS1, AQP4-AS1-211, beta-estradiol, C1QTNF4, CCDC170, DICER1-AS1, FIRRE, GRHL1, IFITM1	32	13	Cell death and survival, organismal development
2	Cbp/p300, Cop9 signalosome, Creb, CSDC2, Ctbp, ELL, ERK1/2, estrogen receptor, Gcn5I, GREB1	21	12	Embryonic development, organismal development
3	ADAM29, Ap1, Collagen Type I (complex), collagen, Cytokin, ERBB4, Fc gamma receptor, GIPR, Hif1, Ifn	18	11	Cell-to-cell signaling and interaction, hematological
4	ABHD8, ASB16, BTBD6, CHMP4C, EME2, FBXW8, H1-2, H2AX, KLHL5, KREMEN1	16	10	Cancer, dermatological diseases, and conditions
5	3′ overhanging ssDNA-DSBs:MRN:S1981, Ac-K3016-ATM:KAT5:BRCA-c complex:ExO1, DNA2:BLM, WRN:S990, Ac-K1249-BRIP1:RAD5	48	32	Cancer hereditary disorder, organismal injury
6	26s proteasome, ATXN1, CASP8, caspase, CBX8, CCNE1, Cdc2, CDKN2B-AS1, Ck2, CMSS1	40	20	Connective tissue development and function
7	ABCA8, actin, ADCY3, alpha catenin, ANKIB1, BTN2A1, CDKAL1, COL1A1, Collagen Alpha 1, Collagen Type I (family)	37	19	Cellular development, embryonic development
8	14-3-3, Activin (family), Akt, Alp, ARRDC3, BAIAP2L1, BMP, CCDC88c, COL8A1, collagen	25	14	Cardiovascular disease, cardiovascular system development
9	ADCY, ADRB, ARHGEF5, calmodulin, CFLAR, CG, Fcer1, FOXP1, FSH, glutathione peroxidase	25	14	Development disorder, neurological disease
10	16:0/20:4 phosphatidylcholine,1D-myo-inositol 1,3,4,5,6-pentakisphosphate, 3830403N18Rik/Xlr, AKT1, ARL17B, ATAD5	14	9	Cancer, cell death and survival, organismal injury
11	AMPD2, AMPK, BCL2L11, BCR (complex), calcineurin proteins, calpain, CD3, ERK, FGFR2, focal adhesion kinase	12	8	Organ morphology, organismal development
12	ANKLE1, BANF1	2	1	Cell cycle, connective tissue disorders, DNA replication

Note: The Score is the negative base-10 logarithm of the *p* value from a right-tailed Fisher's exact test, so a higher score reflects a lower probability that the focus genes were assembled into the network by chance and therefore marks a more significant network.

function. A second variant, rs16885577 at chr8:36930961, is itself an index risk marker, and rs4149909 at chr1:241860596 ($r^2 = 0.82$ with its index marker) produces a further amino-acid change.

The functional consequences of these substitutions were weighed with PolyPhen and PROVEAN, which judged most of them unlikely to disturb protein function. One case of note is rs11571833, an index missense variant at chr13:32398489 in the tumor suppressor BRCA2. PolyPhen returned a low score of 0.02, placing it at the benign end of the scale.

4 | Discussion

This study set out to read 175 known breast cancer risk alleles in functional terms, asking which genes they sit near and which biological processes they touch. We approached the question with two gene-mapping rules. One listed annotated genes flanking each index variant without regard to distance. The other followed linkage disequilibrium, gathering genes near variants in strong agreement ($r^2 \geq 0.8$) within 100 kb of the index marker. Comparing the two lists, 23 genes were unique to the LD route, 42 to the flanking route, and 64 were shared. Most of these genes bear on tumor development and breast cancer in particular.

The flanking rule produced 342 genes, of which 61 were cancer-related and 12 specific to breast cancer. The LD rule produced 191 genes, about half cancer-related and 19 specific to breast cancer. The leading networks from the two routes carried different emphases (Tables 1 and 2), yet both foregrounded receptor signaling, in line with its established role in the disease. The flanking list went further. Its leading regulators, headed by Hedgehog and including cadmium chloride and the Wnt effector β-catenin (CTNNB1), implicated the Hedgehog and Wnt/β-catenin cascades. The leading LD regulators, by contrast, namely, tamoxifen, ARID1A, β-estradiol, TP53, and cisplatin, centered on

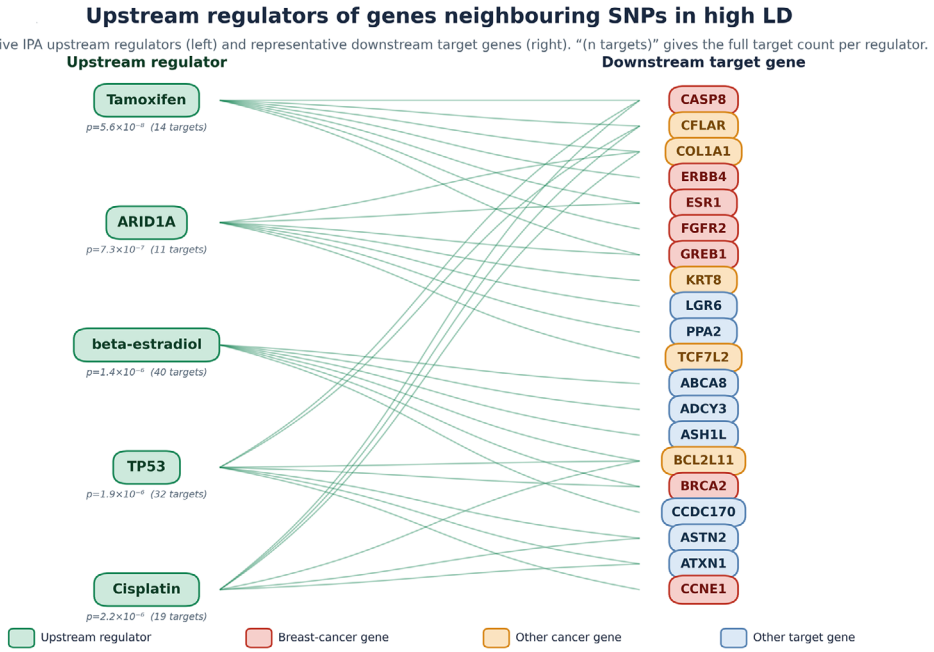

FIGURE 5 | Upstream regulators of the LD-based gene set. The five strongest regulators, namely, tamoxifen, ARID1A, β-estradiol, TP53, and cisplatin, are shown with representative targets. The number in parentheses after each regulator gives its full target count. IPA, Ingenuity Pathway Analysis.

estrogen and receptor signaling and did not highlight Wnt or Hedgehog. This pattern places index variants beside genes that serve particular signaling networks, which hints at functional ties worth pursuing [5].

Recent experiments show that blocking Hedgehog and receptor signaling together can suppress castration-resistant breast cancer, which argues that the two cooperate rather than act alone [17]. Because risk alleles appear to govern their proximal genes, it is worth considering that several variants jointly shape a higher order network controlling specific pathways. Direct biological tests will be needed to confirm these roles and any synergy between them.

Using HaploReg, we broadened the search to more than 2600 additional variants in strong linkage disequilibrium with the index markers, which gives a fuller view of the genetic landscape behind breast cancer risk [5]. This wider lens exposed regulatory roles that a focus on the 175 index variants alone would miss, and it enlarged the set of genes tied to these signals. Most of the added variants fell in intergenic DNA, with smaller fractions in promoters, introns, and untranslated regions, and fourteen in exons.

Among the index variants, rs11571833 carried clear marks of tumor-suppressor involvement through its likely effect on BRCA2 on chromosome 13. BRCA2 guards genomic stability and repairs damaged DNA, and its mutation raises the lifetime risk of breast cancer to between 45% and 85%, often with earlier onset than in the wider population. A second variant, rs515726080, disrupts PALB2, the partner that brings BRCA1 and BRCA2 together in homologous-recombination repair. Carriers of damaging PALB2 mutations face an estimated 33%–58% risk of breast cancer by the age of 70, along with raised risks of pancreatic and ovarian disease, which makes allele-specific control of PALB2 a worthwhile target for further study. A third variant, rs2046210, in strong agreement with an index marker ($r^2 = 0.99$), sits within ESR1. ESR1 encodes estrogen receptor alpha, the transcriptional relay through which estrogen drives growth, proliferation, and differentiation in the breast. About three-quarters of breast cancers are hormone-receptor-positive, and in these tumors estrogen binding to the receptor sustains cancer-cell survival, which is why the estrogen axis has proved such a productive treatment target.

These results belong to the wider effort to turn statistical association into mechanism. The bulk of breast cancer risk variants fall outside coding sequence, and accumulating evidence indicates that they act by tuning regulatory elements rather than by altering proteins. Our finding, that the variants and their proxies concentrate enhancer marks and DNase hypersensitivity in mammary epithelial and embryonic stem cells, fits that picture, and it carries the earlier annotation of 86 established risk variants by Loo and colleagues onto a larger and more current catalogue [5].

The recurrence of androgen and estrogen receptor signaling merits emphasis. The androgen receptor is present in most estrogen-receptor-positive tumors and in a fair share of triple-negative disease, where it shapes proliferation and the response to endocrine therapy. The clustering of KLK2, KLK3, NKX3-1, and AR among the targets of our leading regulators suggests a hormonal axis that the index variants may help to set, consistent with the long-standing link between cumulative hormone exposure and breast cancer risk [2, 3].

Several limits frame these conclusions. The work is computational and depends on curated interaction databases. The Cancer Genome Atlas (TCGA) series examined here holds 236

tumors, and the regulatory predictions still await testing in allele-specific reporter, chromatin, and expression-quantitative-trait experiments. Even so, the way independent risk loci converge on a handful of developmental and hormonal pathways yields a set of hypotheses that can be put to the test.

A further consideration concerns the three-dimensional organization of the genome. Both mapping rules used here rest on linear proximity, whether an immediate neighbor or a variant within 100 kb, yet chromatin folds into loops and topologically associating domains that can bring a regulatory variant into contact with a gene lying far away on the linear sequence. Chromosome-conformation methods such as Hi-C have shown that this architecture differs between normal mammary epithelial cells and breast cancer cells, with reorganized compartments and domain boundaries that accompany changes in gene expression [18, 19]. Applied directly to risk loci, Capture Hi-C across mammary and breast cancer cell lines has tied credible causal variants to distal target genes at scores of independent risk signals, several of which lie beyond the windows examined here [20]. Because higher-order folding may route the influence of a risk variant to genes outside its immediate neighborhood, some regulatory targets will fall beyond the reach of the two proximity-based approaches used in this study. Integrating breast-cancer-specific Hi-C maps with the annotations reported here is therefore a natural next step, one that could sharpen the assignment of variants to the genes they actually control.

5 | Conclusions

In sum, this analysis shows that breast cancer risk markers and the variants correlated with them lie within reach of genes that carry regulatory marks and assemble into coherent networks. Both gene-mapping strategies point to receptor signaling, and the flanking-gene set adds Hedgehog and Wnt/β-catenin involvement. By drawing on analytical tools and the large catalogue of noncoding regulatory regions, the study highlights how GWAS markers may steer the behavior of neighboring genes and how those genes may act together within the pathways that underpin inherited breast cancer susceptibility [5].

Acknowledgments

We are grateful to BioPC-A Bioinformatics Laboratory for Research and Training, for its central role in shaping this study.

Funding

The authors have nothing to report.

Consent

All authors consented to publication of this work in this journal.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The datasets generated or analyzed during this study are available from the corresponding author on reasonable request.

Peer Review

For transparency, the peer review documents associated with this article are available at https://doi.org/10.1002/cso2.70024.

References

1. D. Welter, J. MacArthur, J. Morales, et al., "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations," *Nucleic Acids Research* 42, no. D1 (2014): D1001–D1006, https://doi.org/10.1093/nar/gkt1229.

2. T. J. Key, P. K. Verkasalo, and E. Banks, "Epidemiology of Breast Cancer," *Lancet Oncology* 2, no. 3 (2001): 133–140, https://doi.org/10.1016/S1470-2045(00)00254-0.

3. J. D. Yager and N. E. Davidson, "Estrogen Carcinogenesis in Breast Cancer," *New England Journal of Medicine* 354, no. 3 (2006): 270–282, https://doi.org/10.1056/NEJMra050776.

4. K. Michailidou, P. Hall, A. Gonzalez-Neira, et al., "Large-Scale Genotyping Identifies 41 New Loci Associated With Breast Cancer Risk," *Nature Genetics* 45, no. 4 (2013): 353–361, https://doi.org/10.1038/ng.2563.

5. L. W. M. Loo, A. Y. W. Fong, I. Cheng, and L. Le Marchand, "In Silico Functional Pathway Annotation of 86 Established Prostate Cancer Risk Variants," *PLoS ONE* 10, no. 2 (2015): e0117873, https://doi.org/10.1371/journal.pone.0117873.

6. L. D. Ward and M. Kellis, "HaploReg: A Resource for Exploring Chromatin States, Conservation, and Regulatory Motif Alterations Within Sets of Genetically Linked Variants," *Nucleic Acids Research* 40, no. D1 (2012): D930–D934, https://doi.org/10.1093/NAR/GKR917.

7. I. Dunham, A. Kundaje, S. F. Aldred, et al., "An Integrated Encyclopedia of DNA Elements in the Human Genome," *Nature* 489, no. 7414 (2012): 57–74, https://doi.org/10.1038/nature11247.

8. M. Garber, M. Guttman, M. Clamp, M. C. Zody, N. Friedman, and X. Xie, "Identifying Novel Constrained Elements by Exploiting Biased Substitution Patterns," *Bioinformatics* 25, no. 12 (2009): i54–i62, https://doi.org/10.1093/BIOINFORMATICS/BTP190.

9. I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, "Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2," *Current Protocols in Human Genetics* 7, no. S76 (2013): 7.20.1–7.20.41, https://doi.org/10.1002/0471142905.HG0720S76.

10. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, and A. P. Chan, "Predicting the Functional Effect of Amino Acid Substitutions and Indels," *PLoS ONE* 7, no. 10 (2012): e46688, https://doi.org/10.1371/JOURNAL.PONE.0046688.

11. E. Cerami, J. Gao, U. Dogrusoz, et al., "The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data," *Cancer Discovery* 2, no. 5 (2012): 401–404, https://doi.org/10.1158/2159-8290.CD-12-0095.

12. J. Gao, B. A. Aksoy, U. Dogrusoz, et al., "Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal," *Science Signaling* 6, no. 269 (2013): pl1, https://doi.org/10.1126/scisignal.2004088.

13. J. Cuzick, I. Sestak, S. Cawthorn, et al., "Tamoxifen for Prevention of Breast Cancer: Extended Long-Term Follow-Up of the IBIS-I Breast Cancer Prevention Trial," *Lancet Oncology* 16, no. 1 (2015): 67–75, https://doi.org/10.1016/S1470-2045(14)71171-4.

14. P. Mita, J. N. Savas, N. Djouder, et al., "Regulation of Androgen Receptor-Mediated Transcription by RPB5 Binding Protein URI/RMP," *Molecular and Cellular Biology* 31, no. 17 (2011): 3639–3652, https://doi.org/10.1128/mcb.05429-11.

15. E. E. Schadt, C. Molony, E. Chudin, et al., "Mapping the Genetic Architecture of Gene Expression in Human Liver," *PLoS Biology* 6, no. 5 (2008): e107, https://doi.org/10.1371/JOURNAL.PBIO.0060107.

16. L. W. Barrett, S. Fletcher, and S. D. Wilton, "Regulation of Eukaryotic Gene Expression by the Untranslated Gene Regions and Other Non-

Coding Elements," *Cellular and Molecular Life Sciences* 69, no. 21 (2012): 3613–3634, https://doi.org/10.1007/s00018-012-0990-9.

17. P. S. Gowda, J. D. Deng, S. Mishra, et al., "Inhibition of Hedgehog and Androgen Receptor Signaling Pathways Produced Synergistic Suppression of Castration-Resistant Prostate Cancer Progression," *Molecular Cancer Research* 11, no. 11 (2013): 1448–1461, https://doi.org/10.1158/1541-7786.MCR-13-0278.

18. A. R. Barutcu, B. R. Lajoie, R. P. McCord, et al., "Chromatin Interaction Analysis Reveals Changes in Small Chromosome and Telomere Clustering Between Epithelial and Breast Cancer Cells," *Genome Biology* 16, no. 1 (2015): 214, https://doi.org/10.1186/s13059-015-0768-0.

19. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, et al., "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science* 326, no. 5950 (2009): 289–293, https://doi.org/10.1126/science.1181369.

20. J. Beesley, H. Sivakumaran, M. Moradi Marjaneh, et al., "Chromatin Interactome Mapping at 139 Independent Breast Cancer Risk Signals," *Genome Biology* 21, no. 1 (2020): 8, https://doi.org/10.1186/s13059-019-1877-y.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Supporting File 1: cso270024-sup-0001-SuppMat1.xlsx **Supporting File 2**: cso270024-sup-0002-SuppMat2.xlsx **Supporting File 3**: cso270024-sup-0003-SuppMat3.xlsx **Supporting File 4**: cso270024-sup-0004-SuppMat4.xlsx **Supporting File 5**: cso270024-sup-0005-SuppMat5.xlsx **Supporting File 6**: cso270024-sup-0006-SuppMat6.xlsx **Supporting File 7**: cso270024-sup-0007-SuppMat7.xlsx **Supporting File 8**: cso270024-sup-0008-SuppMat8.xlsx