Breast cancer susceptibility: an integrative analysis of genomic data

Simone Mocellin, MD, 1-2; Sara Valpione, MD, 3; Carlo Riccardo Rossi, MD, 1-2, Karen Pooley, PhD, 4

1) Istituto Oncologico Veneto (IOV-IRCCS), Padova, Italy

2) Dept. Surgery Oncology and Gastroenterology, University of Padova, Italy

3) CRUK Manchester Institute, The University of Manchester, UK

4) Centre for Cancer Genetic Epidemiology, Dept. Public Health and Primary Care, University of Cambridge, UK

Corresponding author: Simone Mocellin

Istituto Oncologico Veneto (IOV-IRCCS), Padova, Italy

Dept. Surgery Oncology and Gastroenterology, University of Padova, Italy

Via Gattamelata 64

35128 Padova, Italy

E: simone.mocellin@unipd.it

P: 049 8215743

F: 049 8215575

Key words: breast cancer, susceptibility, risk, predisposition, germline DNA variant, single nucleotide polymorphism, SNP, knowledge-base, integrative analysis, network analysis, pathway analysis, genome-wide association study, GWAS, genomic data.

Abbreviations: GWAS: genome-wide association study; SNP: single nucleotide polymorphism; BCPG: breast cancer predisposition gene; LD: linkage disequilibrium

Abstract

Background

Genome wide association studies (GWAS) are greatly accelerating the pace of discovery of germline variants underlying the genetic architecture of sporadic breast cancer predisposition. We have built the first knowledge-base dedicated to this field and used it to generate hypotheses on the molecular pathways involved in disease susceptibility.

Methods

We gathered data on the common single nucleotide polymorphisms (SNPs) discovered by breast cancer risk GWAS. Information on SNP functional effect (including data on linkage disequilibrium, expression quantitative trait locus, and SNP relationship with regulatory motifs or promoter/enhancer histone marks) was utilized to select putative breast cancer predisposition genes (BCPGs). Ultimately, BCPGs were subject to pathway (gene set enrichment) analysis and network (protein-protein interaction) analysis.

Results

Data from 38 studies (28 original case-control GWAS enrolling 383,260 patients with breast cancer; and 10 GWAS meta-analyses) were retrieved. Overall, 281 SNPs were associated with the risk of breast cancer with a P-value <10E-06 and a minor allele frequency >1%. Based on functional information, we identified 296 putative BCPGs. Primary analysis showed that germline perturbation of classical cancer-related pathways (e.g., apoptosis, cell cycle, signal transduction including estrogen receptor signaling) play a significant role in breast carcinogenesis. Other less established pathways (such as ribosome and peroxisome machineries) were also highlighted. In the main subgroup analysis, we considered the BCPGs encoding transcription factors (n=36), which in turn target 252 genes. Interestingly, pathway and network analysis of these genes yielded results resembling those of primary analyses, suggesting that most of the effect of genetic variation on disease risk hinges upon transcriptional regulons.

Conclusions

This knowledge-base, which is freely available and will be annually updated, can inform future studies dedicated to breast cancer molecular epidemiology as well as genetic susceptibility and development.

Introduction

With a 10-12% lifetime risk, breast cancer is the most common cancer among women with about 1,700,000 new cases and more than 500,000 deaths each year worldwide (1). Breast cancer is a multifactorial disease stemming from a complex interplay between environmental, reproductive/endocrine and genetic risk factors. Dissecting the genetic architecture of breast cancer susceptibility is a pivotal step to understand the cascade of molecular events underlying breast carcinogenesis, which ultimately could lead to better preventive and therapeutic strategies according to the precision medicine principles (2).

Familial aggregation of breast cancer (which occurs in about 10% of cases) has led to family-based linkage analysis and positional cloning studies demonstrating that rare (<1%) germline DNA variation in high to moderate penetrance cancer predisposition genes - such as *BRCA1*, *BRCA2*, *PTEN*, *CHEK2*, *ATM*, *BRIP1* and *PALB2* - accounts for about 15-20% of the familial risk of this disease (3,4). The residual heritability for breast cancer is believed to be sustained by a polygenic model according to the common disease/common variant hypothesis. Subsequent case-control studies based on the candidate gene approach (also known as hypothesis testing approach) have identified some common germline variants (mainly single nucleotide polymorphisms, SNPs) linked to breast cancer risk, though the evidence quality is often low mainly because of small sample size and result heterogeneity (5,6). More recently, the completion of the Human Genome Project and the implementation of genome-wide association studies (GWAS) – based on a hypothesis generating (also known as data driven) approach and testing hundreds of thousands of SNPs at a time - has greatly accelerated the pace of discovery of low penetrance variants linked to the risk of many diseases, including several cancer types (7).

To date, tens of GWAS dedicated to breast cancer have been published, and many single SNPs have been associated with the risk of this malignancy (3,6). This has led to an overwhelming wealth of data which are often difficult to manage by the single reader, in part because most susceptibility loci are intergenic (and thus are linked neither to an obvious gene nor to an obvious functional effect), which hinders a straightforward biological interpretation typical of candidate gene studies.

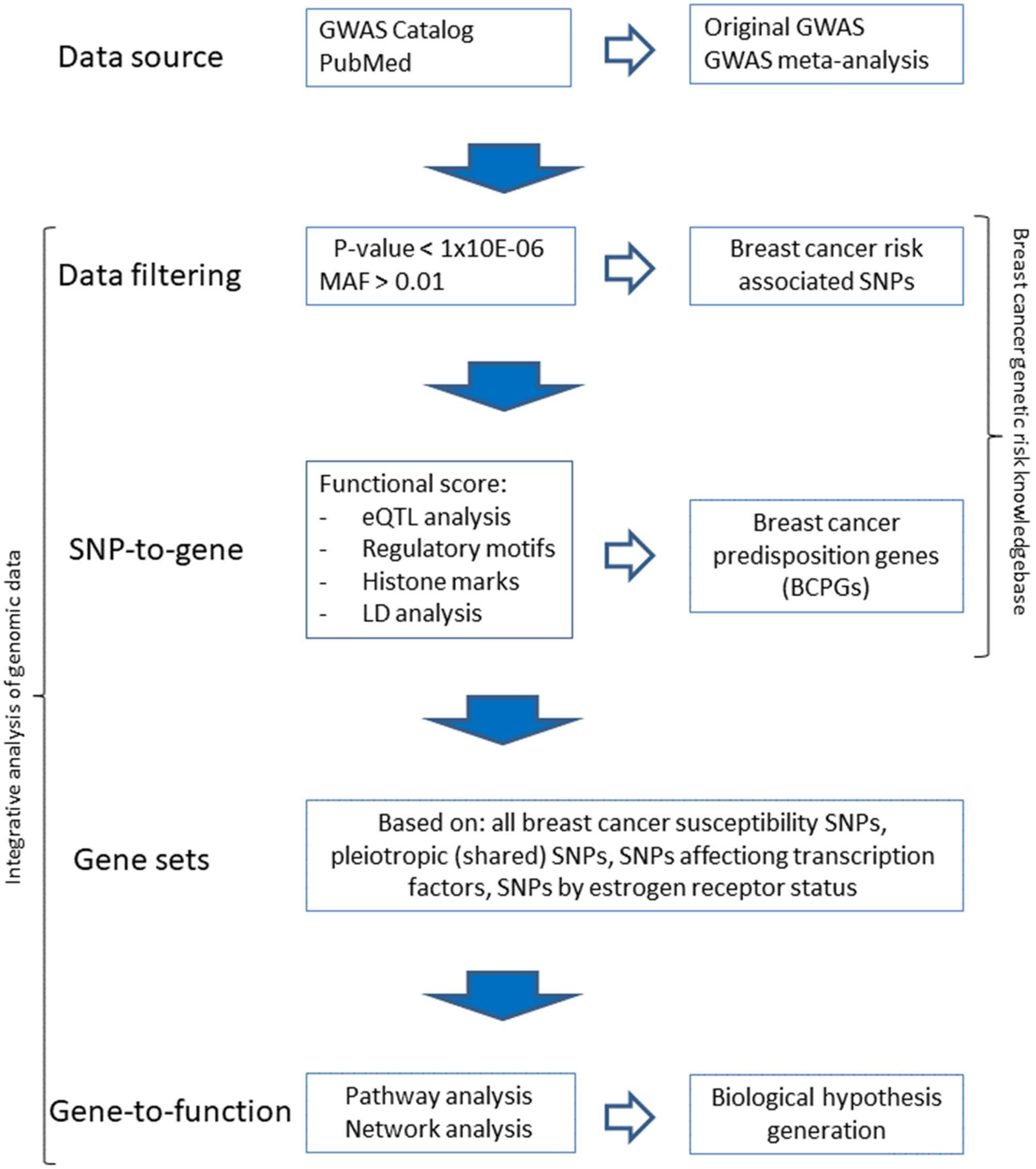
With the present work we intended to systematically review breast cancer GWAS findings in order to provide readers with the first publicly available knowledge-base dedicated to the relationship between germline genomic DNA variation and breast cancer risk. According to the above mentioned polygenic model of sporadic tumor inheritance and using modern SNP-to-gene and gene-to-function approaches such as integrative analysis of genomic data (8,9) as well as pathway and network analysis (10,11), we also aimed to suggest a biological interpretation of current findings. In particular, we tried to exploit the available GWAS evidence to comprehensively identify the cell pathways whose germline variation condition the predisposition to breast cancer, with an additional effort to prioritize genes/pathways/networks which could be of special relevance to inform future studies in the fields of both molecular epidemiology and biology of breast cancer.

Materials and Methods

We collected GWAS findings on breast cancer risk (along with other genomic data, see below) to identify breast cancer risk associated SNPs, which were then linked to breast cancer predisposition genes (BCPGs): the data from this knowledge-base were used to perform pathway and network analysis. A flowchart of the study design is reported in **Figure 1**.

Figure 1

Study design: flow chart of the integrative analysis of genomic data on breast cancer susceptibility


Data source
GWAS Catalog
PubMed
Original GWAS
GWAS meta-analysis
Data filtering
P-value < 1x10E-06
MAF > 0.01
Breast cancer risk associated SNPs
SNP-to-gene
Functional score:
- eQTL analysis
- Regulatory motifs
- Histone marks
- LD analysis
Breast cancer predisposition genes (BCPGs)
Gene sets
Based on: all breast cancer susceptibility SNPs, pleiotropic (shared) SNPs, SNPs affectiong transcription factors, SNPs by estrogen receptor status
Gene-to-function
Pathway analysis
Network analysis
Biological hypothesis generation
Integrative analysis of genomic data
Breast cancer genetic risk knowledgebase

Breast cancer risk associated SNPs

GWAS addressing the role of germ-line single nucleotide polymorphisms (SNPs) in breast cancer susceptibility were retrieved in the GWAS Catalog repository (7) as well as by performing a systematic review in PubMed (search terms: "breast cancer", "GWAS"). GWAS meta-analyses were also included for data extraction. Searches were updated until the 11th of December 2017.

To be included in the knowledge-base, each SNP had to be associated with breast cancer risk with a nominal P-value lower than 1x10E-06 (genome-wide significance level) and have a minor allele frequency ≥ 1% in the general population of European ancestry.

SNP-to-gene analysis

Following the principles of integrative analysis of genomic data (8,9), the functional association between a breast cancer risk associated SNP and a gene (hereafter called BCPG) was scored according to three types of information:

A) SNP relationship with gene(s):

[Category 1 – score=2] This applies to within-gene non-synonymous variants (e.g., missense SNPs), variants associated with expression quantitative trait locus (eQTL) data (based on GTex portal database (12)), variants in high linkage disequilibrium (LD) (pairwise r-squared ≥ 0.8) with another SNP that is an eQTL hit, or variants in high LD with a within-gene non-synonymous variant;

[Category 2 – score=1] This applies to within-gene synonymous variants, or variants located in a non-coding gene region (e.g., intronic SNPs), or variants in high LD with another within gene SNP (synonymous variant, or variant located in a non-coding gene region);

[Category 3 – score=0] intergenic and non eQTL hit variants.

B) SNP relationship with regulatory motifs (i.e., variant potentially affecting the binding of transcription factors based on a library of position weight matrices scored on genomic sequences (13)):

[Category 1 – score=1] variant changing these motifs, or variant in high LD with another SNP changing these motifs;

[Category 2 - score = 0] no change of motif/LD with SNP changing motif.

C) SNP relationship with breast tissue specific promoter/enhancer histone marks (based on data from the Roadmap Epigenomics Project (14)):

[Category 1, score=1] variant co-localization with these marks, or variant in high LD with another SNP co-localising with these marks;

[Category 2, score=0] no co-localization with epigenetic marks/LD with SNP colocalising with epigenetic marks.

The principles underlying the above scoring system are analogous to those employed in well-established functional annotation databases (such as RegulomeDB (15) and HaploReg (16)). However, we added the information deriving from LD analysis (which was performed using the LDLink website (17)), which increases the likelihood of identifying additional functional variants relevant to disease susceptibility.

Gene-to-function: pathway and network analysis

Once BCPGs were identified, we used them to perform pathway and network analysis in order to identify biological functions whose genetic perturbations can predispose to breast cancer development (10,11).

For pathway analysis purposes we utilized gene set enrichment analysis (GSEA) as performed by the EnrichR web server (18). Hypergeometric distribution was used to calculate the statistical significance of gene overlapping (19), followed by correction for multiple hypotheses testing (using the false discovery rate [FDR] method (20)). Only pathways with a FDR <0.05 were considered of interest.

Also protein-protein interaction (PPI) networks can be employed to select gene sets. In contrast to pathways, networks are not based on specific biological functions but are built on the basis of both direct (physical) and indirect (genetic) interactions between gene products (proteins). For network analysis, we utilized the STRING 10.5 web server (21). In order to consider only highly reliable information on protein-protein interactions (PPI), we set the interaction score to ≥ 0.7 (high confidence). The resulting network provides information of the degree of overall connectivity across imputed gene products (as quantified by the ratio between observed and expected interactions [a.k.a. "edges"] between proteins [a.k.a. "nodes"], and formally tested by means of a PPI enrichment test). Then, molecular clusters (subnetworks) can be identified that can be utilized for gene set enrichment analysis (only subnetworks with at least three BCPGs were considered). When the network connectivity is low, the PPI database can be exploited to add first-shell interactors (we chose to add no more than 10 such interactors to avoid data over-interpretation) and then re-run pathway analysis. Ultimately, this data augmentation process increases the likelihood of identifying relevant biological pathways which would be otherwise overlooked when starting with only few BCPGs belonging to a given pathway.

Other analyses

Within the frame of network analysis, we searched for so called "hub proteins": these are molecules with a high degree of connectivity with the other network components and thus are likely to play a dominant role in the activity of the network itself (they are also known as "influencers") (22). To this aim, we used the EsyN webtool (23) to calculate the collective influence score, which defined as the product of a node-reduced degree (number of edges minus one) times the sum of the reduced degree of the nodes that are two steps away (a.k.a. radius) (24).

Finally, in order to provide further information beyond the *cis* effects of included variants (as done in the above analyses), we explored the potential effect *in trans* of breast cancer associated SNPs. To this aim, we first identified the transcription factors among the putative BCPGs: then, the genes whose expression were regulated by these transcription factors (identified by using the Uniprot (25) and TRRUST (26) repositories) were input in both pathway and network analysis to assess the cellular functions potentially affected by germline variation linked to cancer risk.

Results

Breast cancer risk associated SNPs

We found 38 studies (published between 2007 and 2017) which matched our inclusion criteria (27-64). Of these, 28 were original case-control GWAS (overall enrolling 383,260 patients with breast cancer) and 10 were meta-analyses of previously published case-control GWAS (overall enrolling 239,271 patients with breast cancer) (**Supplementary Table 1**).

In most studies, patients and controls were of European ancestry (71% among original GWAS; 97% among GWAS meta-analyses); in the remaining studies, individuals were mainly of Asian ancestry among original GWAS (96%) and African-American among GWAS meta-analyses (100%). Only one original study was dedicated to male breast cancer. As regards tumor subtype by estrogen receptor expression, two original GWAS were dedicated to receptor negative and two to receptor positive breast cancer, whereas four GWAS meta-analyses were dedicated to ER negative cases. In the original articles associations were reported (and are reported hereafter in the text) as per-allele odds ratios (ORs).

Overall, 281 SNPs were associated with the risk of breast cancer with a P-value <10E-06 and a MAF >0.01 (**Supplementary Table 2**); the median minor allele frequency was 0.28 (interquartile range [IQR]: 0.16-0.39); the median OR was 0.93 (IQR: 0.91-0.95) and 1.11 (IQR: 1.06-1.19) for protective and risk alleles, respectively.

Chromosome distribution showed an over-representation of chromosome 5 (signals observed: 32; expected: 15; FDR: 0.0003) and chromosome 19 (signals observed: 12; expected: 5; FDR: 0.014).

Linkage disequilibrium (LD) analysis of the 281 SNPs showed that 48 polymorphisms were tagged by one or more other variants (LD r-squared >0.8), leading to the identification of 233 breast cancer predisposition loci (**Supplementary Table 3**).

Out of 281 reported SNPs associated with breast cancer risk at a genome-wide significance level, only 34 (12.1%) were reported by two or more data sources.

Whereas most studies (n=21, 55.3%) enrolled women with unspecified sporadic breast cancer, subgroups were specifically investigated by others: estrogen receptor negative breast cancer (n=7); estrogen receptor positive tumor (n=2); triple negative tumor (n=1); breast cancer in BRCA1/2 mutation carriers (studies, n=3); early onset breast cancer (n=1); breast cancer in post-menopausal women (n=1); lobular carcinoma (n=1); and breast cancer in males (n=1).

Breast cancer predisposition genes

The majority of SNPs were located within coding genes (n=160, 56.9%). More specifically, SNPs were located in gene 3'-UTR (n=7, 2.5%), intron (n=140, 49.8%), exon (n=13, 4.6%; of these: missense, n=8, synonymous, n=4 and non-sense [stop gain], n=1). The remaining SNPs were intergenic region (n=95, 33.8%) and within non-coding genes (n=17, 9.2%). Of note, 6 intergenic SNPs (2.1%) were in high LD with non-tested SNPs located within a gene and other 8 SNPs (2.8%) were in high LD with non-tested missense variants.

As regards eQTL analysis, 107 SNPs (38.1%) were directly associated with a significant effect on the expression of one or more genes, and 3 SNPs (1.1%) were in high LD with SNPs with an eQTL effect. A large majority of 229 variants (81.5%) were associated with changing regulatory motifs, with 43 SNPs (15.3%) in high LD with those 229 variants, whereas only 9 SNPs had no impact on regulatory motifs. In addition, 79 SNPs (28.1%) co-localized with promoter/enhancer histone marks, with 107 variants (38.1%) in high LD with those 79 SNPs, and 95 SNPs (33.8%) having no such property.

Based on the above information, we associated the 281 SNPs linked with breast cancer risk to 334 genes with a score ranging from 0 to 4 (**Supplementary Table 4**): SNPs with low (0-1), intermediate (2) and high (3-4) functional score were 30 (10.7%), 68 (24.2%), and 183 (65.1%), respectively. In order to exclude genes with low level of evidence of association with breast cancer risk SNPs, we further considered only SNPs with a score equal or greater than 2 (n=251). With this cut off, we identified 296 putative BCPGs, which were the genes utilized in the following primary analysis. These genes code for known proteins in most cases (n=255, 86.1%).

Pathway and network analysis

Primary gene set enrichment analysis demonstrated that the 296 BCPGs are enriched in genes involved in apoptotic pathway and peroxisome machinery, as illustrated in **Table 1**.

Table 1: Pathway analysis of breast cancer predisposition genes (BCPGs). Overlap: number of BCPGs over number of pathway genes. FDR: false discovery rate.

PRIMARY ANALYSIS (296 BCPGs)			
Pathway	**Overlap**	**FDR**	**Genes**
Apoptosis	9/140	0.04330	PIDD1;DFFA;BCL2L11;CASP8;TNFSF10;ITPR1;PIK3R3;CTSW;CFLAR
Peroxisomal importomer	18/82	<0.00001	DCLRE1B;TCF7L2;DFFA;CDKN2B;STXBP4;PDE4D;MKL1;EBF1;LSP1;PEX14;RAD51B; BABAM1;TOX3;ADAM29; RALY;MDM4;LGR6;CASC16
FIRST-SHELL AUGMENTATION (296 BCPGs + 10 interactors)			
Pathway	**Overlap**	**FDR**	**Genes**
Prostate cancer	6/89	0.04198	TCF7L2;CCND1;CCNE1;PIK3R3;FGFR2;CREB5
Apoptosis	8/140	0.02506	DFFA;BCL2L11;CASP8;TNFSF10;ITPR1;PIK3R3;CTSW;CFLAR
FoxO signaling	8/133	0.02373	CDKN2B;BCL2L11;CCND1;IRS1;TNFSF10;PIK3R3;FBXO32;TGFBR2
Ribosome	10/137	0.00261	RPS16;RPS15A;RPS29;RPLP1;RPS3;RPLP2;RPS2;MRPL34;RPS13;RPS23
Non-alcoholic fatty liver disease	9/151	0.01670	ITCH;NDUFA13;BCL2L11;CASP8;IRS1;NDUFB3;NDUFA2;PIK3R3;COX6A1
SMALL SUBNETWORK (5 BCPGs)			
Pathway	**Overlap**	**FDR**	**Genes**
Translation Initiation	3/114	0.00001	RPLP2;EIF2S2;RPS23
Ribosome formation	3/180	0.00003	RPLP2;WDR43;RPS23
LARGE SUBNETWORK (59 BCPGs)			
Pathway	**Overlap**	**FDR**	**Genes**
Pathways in cancer	11/397	<0.00001	TCF7L2;CDKN2B;CASP8;CCND1;CCNE1;ADCY3;LPAR2;PIK3R3;BRCA2;FGFR2;TGFBR2
Pancreatic cancer	4/66	0.00057	CCND1;PIK3R3;BRCA2;TGFBR2
Prostate cancer	6/89	<0.00001	TCF7L2;CCND1;CCNE1;PIK3R3;FGFR2;CREB5
Colorectal cancer	4/62	0.00049	TCF7L2;CCND1;PIK3R3;TGFBR2
Small cell lung cancer	4/86	0.00118	CDKN2B;CCND1;CCNE1;PIK3R3
Endometrial cancer	3/52	0.00258	TCF7L2;CCND1;PIK3R3
Acute myeloid leukemia	3/57	0.00312	TCF7L2;CCND1;PIK3R3
Chronic myeloid leukemia	3/73	0.00520	CCND1;PIK3R3;TGFBR2
Proteoglycans in cancer	5/203	0.00219	PPP1CB;CCND1;ERBB4;PIK3R3;ESR1

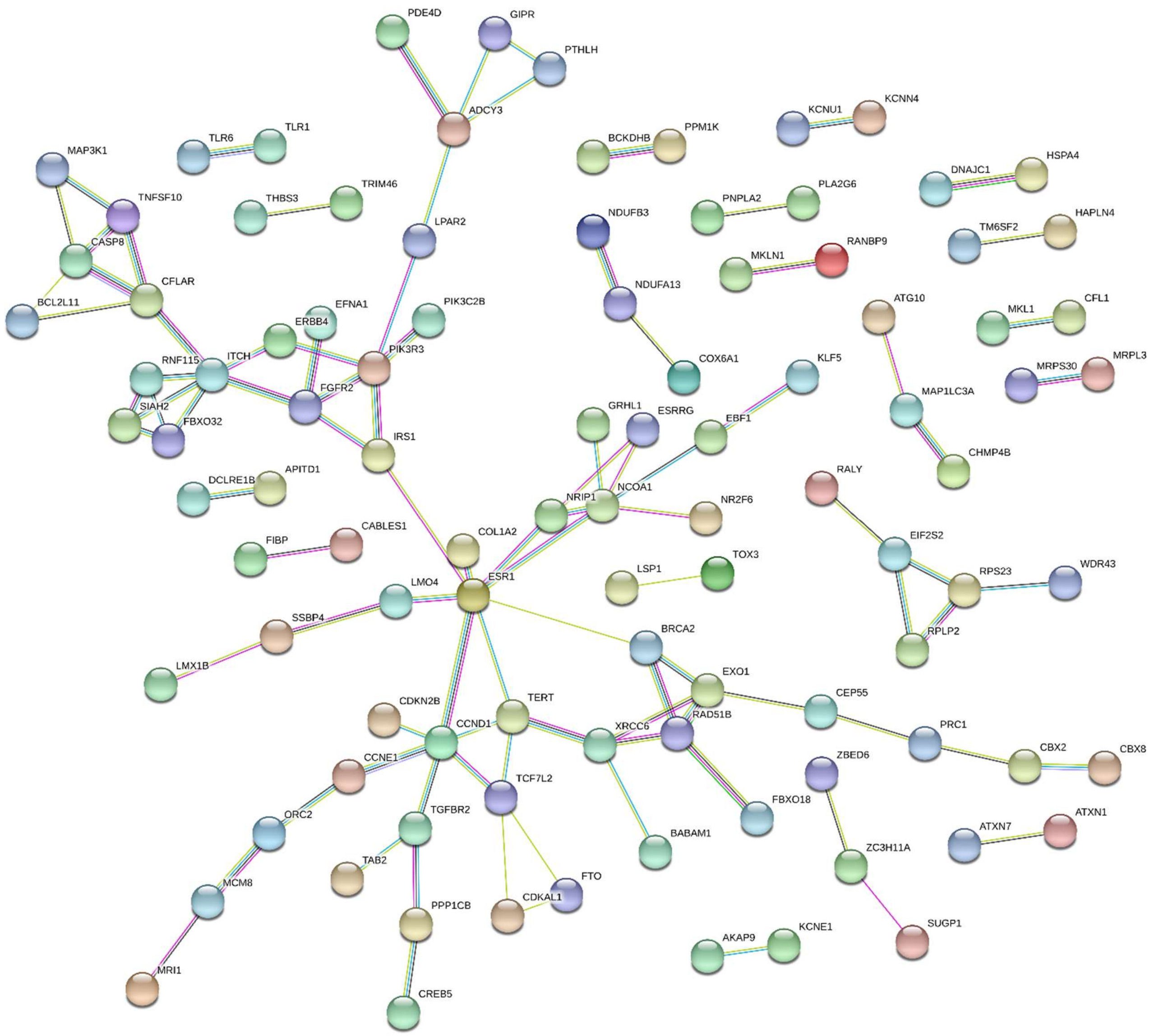
MicroRNAs in cancer	4/297	0.03057	BCL2L11;CCND1;CCNE1;IRS1
FoxO signaling	8/133	<0.00001	CDKN2B;BCL2L11;CCND1;IRS1;TNFSF10;PIK3R3;FBXO32;TGFBR2
PI3K-Akt signaling	10/341	<0.00001	EFNA1;COL1A2;BCL2L11;CCND1;CCNE1;IRS1;LPAR2;PIK3R3;FGFR2;CREB5
p53 signaling	3/69	0.00481	CASP8;CCND1;CCNE1
cAMP signaling	6/199	0.00048	PPP1CB;GIPR;PDE4D;ADCY3;PIK3R3;CREB5
Rap1 signaling	5/211	0.00249	EFNA1;ADCY3;LPAR2;PIK3R3;FGFR2
AMPK signaling	4/124	0.00258	CCND1;IRS1;PIK3R3;CREB5
Hippo signaling	4/153	0.00481	PPP1CB;TCF7L2;CCND1;TGFBR2
MAPK signaling	4/255	0.02078	MAP3K1;TAB2;FGFR2;TGFBR2
Phospholipase D signaling	3/144	0.02410	ADCY3;LPAR2;PIK3R3
TNF signaling	6/110	0.00002	ITCH;CASP8;PIK3R3;TAB2;CFLAR;CREB5
Longevity regulation	4/94	0.00147	IRS1;ADCY3;PIK3R3;CREB5
Apoptosis	5/140	0.00065	BCL2L11;CASP8;TNFSF10;PIK3R3;CFLAR
Cell cycle	4/124	0.00258	CDKN2B;CCND1;CCNE1;ORC2
Viral carcinogenesis	6/205	0.00049	CDKN2B;CASP8;CCND1;CCNE1;PIK3R3;CREB5
Hepatitis B	6/146	0.00010	MAP3K1;CASP8;CCND1;CCNE1;PIK3R3;CREB5
Measles	5/136	0.00061	CCND1;CCNE1;TNFSF10;PIK3R3;TAB2
HTLV-I infection	7/258	0.00022	CDKN2B;MAP3K1;TERT;CCND1;ADCY3;PIK3R3;TGFBR2
Herpes simplex infection	3/185	0.04043	PPP1CB;CASP8;TAB2
Estrogen signaling	4/99	0.00170	ADCY3;PIK3R3;ESR1;CREB5
Thyroid hormone signaling	4/118	0.00249	NCOA1;CCND1;PIK3R3;ESR1
Oxytocin signaling	4/158	0.00511	PPP1CB;CCND1;ADCY3;PIK3R3
Prolactin signaling	3/72	0.00513	CCND1;PIK3R3;ESR1
Focal adhesion	4/202	0.01145	PPP1CB;COL1A2;CCND1;PIK3R3
Toll-like receptor signaling	3/106	0.01327	CASP8;PIK3R3;TAB2
Endocytosis	4/259	0.02105	ITCH;ERBB4;FGFR2;TGFBR2

Network analysis suggested that BCPGs protein products did not have more interactions among themselves than expected for a random protein set of equal size drawn from the proteome (observed edges: 98; expected edges: 83; PPI enrichment test P-value: 0.0527), indicating that these proteins are not remarkably biologically connected as a group. When 10 first-shell interactors were added to the network, ribosome proteins were then included in the enrichment list (**Table 1**).

Network analysis also enabled us to identify one large (n=59) and one small (n=5) subnetwork (**Figure 2**): the former was enriched in several cancer-related pathways, including apoptosis and estrogen receptor signaling, whereas the latter was enriched in ribosome machinery components (see **Table 1**). Finally, influence analysis of the large subnetwork identified estrogen receptor 1 (ESR1) as the most influential protein (**Suppementary Table 5**).

Figure 2

Network analysis: network plot of protein-protein interactions regarding the products of the putative breast cancer predisposition genes identified through the integrative analysis of GWAS data

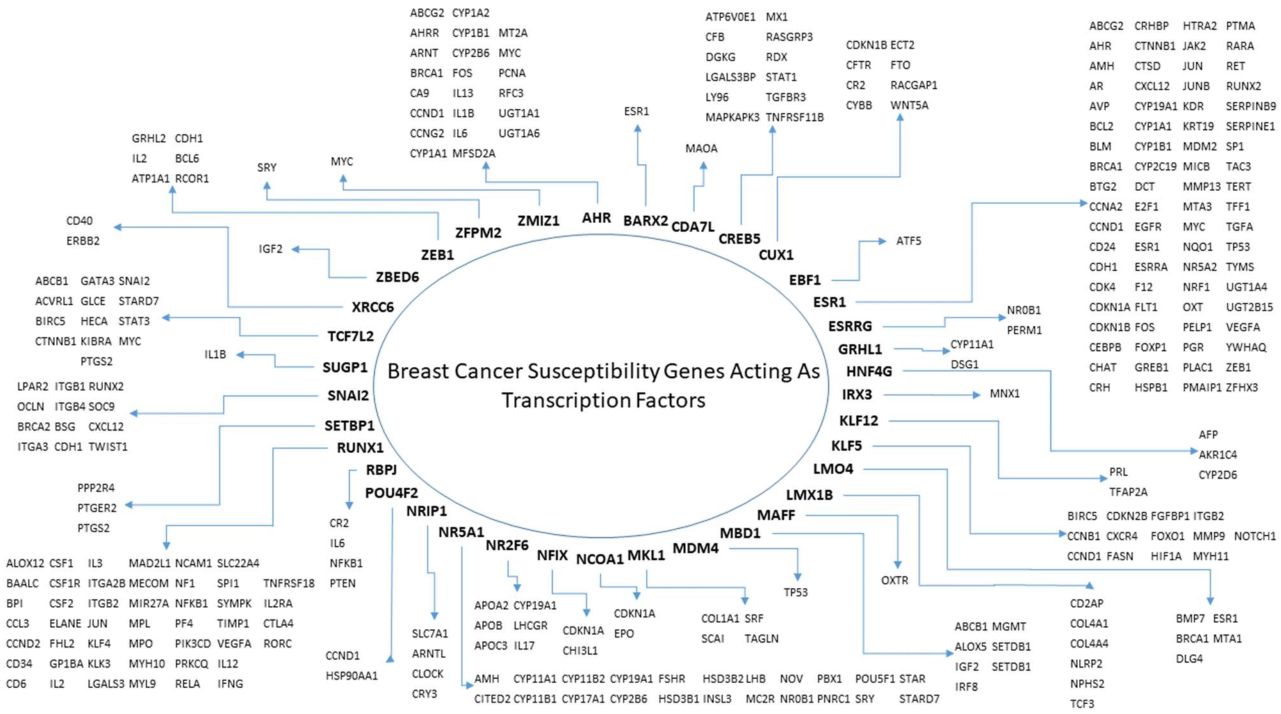

PDE4D
GIPR
PTHLH
ADCY3
KCNU1
KCNN4
TLR6
TLR1
BCKDHB
PPM1K
MAP3K1
HSPA4
DNAJC1
TNFSF10
THBS3
TRIM46
PNPLA2
PLA2G6
NDUFB3
HAPLN4
TM6SF2
LPAR2
CASP8
MKLN1
RANBP9
CFLAR
NDUFA13
ATG10
BCL2L11
EFNA1
PIK3C2B
MKL1
CFL1
ERBB4
PIK3R3
MRPL3
MRPS30
RNF115
ITCH
COX6A1
KLF5
FGFR2
MAP1LC3A
SIAH2
GRHL1
ESRRG
FBXO32
EBF1
IRS1
CHMP4B
DCLRE1B
APITD1
RALY
NRIP1
NCOA1
NR2F6
CABLES1
FIBP
COL1A2
EIF2S2
TOX3
LSP1
WDR43
RPS23
LMO4
ESR1
SSBP4
BRCA2
RPLP2
LMX1B
EXO1
CEP55
TERT
CDKN2B
RAD51B
PRC1
CCND1
XRCC6
CBX2
CBX8
CCNE1
ZBED6
TCF7L2
FBXO18
ORC2
TGFBR2
ATXN7
ATXN1
BABAM1
ZC3H11A
TAB2
FTO
MCM8
CDKAL1
PPP1CB
SUGP1
AKAP9
KCNE1
MRI1
CREB5

Subgroup analysis

In a first subgroup analysis we considered the BCPGs encoding transcription factors: there were 36 such genes (**Figure 3**), which represent 12.2% of the BCPGs identified in this work, a figure only slightly higher than expected (10%). These transcription factors target 252 genes (**Figure 3**), with nine also being BCPGs (*AHRR, BRCA2, CCND1, CDKN2B, ESR1, FOXP1, FTO, LPAR2, TERT*).

Figure 3

Breast cancer susceptibility regulons: targets of transcription factors whose germline variation is associated with breast cancer risk


Breast Cancer Susceptibility Genes Acting As Transcription Factors
ZEB1
GRHL2 CDH1
IL2 BCL6
ATP1A1 RCOR1
ZFPM2
SRY
ZMIZ1
MYC
AHR
ABCG2 CYP1A2
AHRR CYP1B1 MT2A
ARNT CYP2B6 MYC
BRCA1 FOS PCNA
CA9 IL13 RFC3
CCND1 IL1B UGT1A1
CCNG2 IL6 UGT1A6
CYP1A1 MFSD2A
BARX2
ESR1
CDA7L
MAOA
CREB5
ATP6V0E1 MX1
CFB RASGRP3
DGKG RDX
LGALS3BP STAT1
LY96 TGFBR3
MAPKAPK3 TNFRSF11B
CUX1
CDKN1B ECT2
CFTR FTO
CR2 RACGAP1
CYBB WNT5A
EBF1
ATF5
ESR1
ABCG2 CRHBP HTRA2 PTMA
AHR CTNNB1 JAK2 RARA
AMH CTSD JUN RET
AR CXCL12 JUNB RUNX2
AVP CYP19A1 KDR SERPINB9
BCL2 CYP1A1 KRT19 SERPINE1
BLM CYP1B1 MDM2 SP1
BRCA1 CYP2C19 MICB TAC3
BTG2 DCT MMP13 TERT
CCNA2 E2F1 MTA3 TFF1
CCND1 EGFR MYC TGFA
CD24 ESR1 NQO1 TP53
CDH1 ESRRA NR5A2 TYMS
CDK4 F12 NRF1 UGT1A4
CDKN1A FLT1 OXT UGT2B15
CDKN1B FOS PELP1 VEGFA
CEBPB FOXP1 PGR YWHAQ
CHAT GREB1 PLAC1 ZEB1
CRH HSPB1 PMAIP1 ZFHX3
ESRRG
NR0B1
PERM1
GRHL1
CYP11A1
DSG1
HNF4G
AFP
AKR1C4
CYP2D6
IRX3
MNX1
KLF12
PRL
TFAP2A
KLF5
BIRC5 CDKN2B FGFBP1 ITGB2
CCNB1 CXCR4 FOXO1 MMP9 NOTCH1
CCND1 FASN HIF1A MYH11
LMO4
BMP7 ESR1
BRCA1 MTA1
DLG4
LMX1B
CD2AP
COL4A1
COL4A4
NLRP2
NPHS2
TCF3
MAFF
OXTR
MBD1
ABCB1 MGMT
ALOX5 SETDB1
IGF2 SETDB1
IRF8
MDM4
TP53
MKL1
COL1A1 SRF
SCAI TAGLN
NCOA1
CDKN1A
EPO
NFIX
CDKN1A
CHI3L1
NR2F6
APOA2 CYP19A1
APOB LHCGR
APOC3 IL17
NR5A1
AMH CYP11A1 CYP11B2 CYP19A1 FSHR HSD3B2 LHB NOV PBX1 POU5F1 STAR
CITED2 CYP11B1 CYP17A1 CYP2B6 HSD3B1 INSL3 MC2R NR0B1 PNRC1 SRY STARD7
NRIP1
SLC7A1
ARNTL
CLOCK
CRY3
POU4F2
CCND1
HSP90AA1
RBPJ
CR2
IL6
NFKB1
PTEN
RUNX1
ALOX12 CSF1 IL3 MAD2L1 NCAM1 SLC22A4
BAALC CSF1R ITGA2B MECOM NF1 SPI1 TNFRSF18
BPI CSF2 ITGB2 MIR27A NFKB1 SYMPK IL2RA
CCL3 ELANE JUN MPL PF4 TIMP1 CTLA4
CCND2 FHL2 KLF4 MPO PIK3CD VEGFA RORC
CD34 GP1BA KLK3 MYH10 PRKCQ IL12
CD6 IL2 LGALS3 MYL9 RELA IFNG
SETBP1
PPP2R4
PTGER2
PTGS2
SNAI2
LPAR2 ITGB1 RUNX2
OCLN ITGB4 SOC9
BRCA2 BSG CXCL12
ITGA3 CDH1 TWIST1
SUGP1
IL1B
TCF7L2
ABCB1 GATA3 SNAI2
ACVRL1 GLCE STARD7
BIRC5 HECA STAT3
CTNNB1 KIBRA MYC
PTGS2
XRCC6
CD40
ERBB2
ZBED6
IGF2

Pathway analysis of the 252 targets demonstrated a significant enrichment in many cancer-related pathways, including those involved in the pathogenesis of different tumor types (mainly but not only carcinomas), cell cycle and apoptosis, multiple signaling pathways (such as p53, PI3K-Akt, Wnt, Hippo, Mapk, ErbB, HIF-1 and VEGF), hormone pathways (including sex hormones), immunity (with special regard to anti-viral immune response), and cell adhesion (**Table 2**).

Table 2: Pathway analysis of breast cancer predisposition genes (BCPGs) encoding transcription factors. Overlap: number of BCPGs over number of pathway genes. FDR: false discovery rate.

Pathways	**Overlap**	**FDR**	**Genes**
Pathways in cancer	51/397	<0.00001	RET;ITGB1;CDKN1A;SPI1;CDKN1B;ITGA2B;PTEN;PIK3CD;BRCA2;MECOM;CCND1;CDH1;MYC;HSP90AA1;ITGA3;WNT5A;ARNT;FOS;MMP9;AR;COL4A1;COL4A4;RARA;BIRC5;TP53;CSF1R;PTGER2;LPAR2;TGFA;CXCR4;KLK3;PTGS2;HIF1A;EGFR;FOXO1;RELA;RASGRP3;ERBB2;E2F1;JUN;CDKN2B;STAT1;STAT3;NFKB1;VEGFA;IL6;CXCL12;CDK4;MDM2;BCL2;CTNNB1
Prostate cancer	19/89	<0.00001	HSP90AA1;CDKN1A;CDKN1B;PTEN;TGFA;PIK3CD;KLK3;RELA;EGFR;FOXO1;NFKB1;AR;CCND1;ERBB2;BCL2;E2F1;MDM2;CTNNB1;TP53
Small cell lung cancer	18/86	<0.00001	ITGB1;CDKN2B;CDKN1B;ITGA3;ITGA2B;PTEN;PIK3CD;PTGS2;RELA;NFKB1;CCND1;CDK4;COL4A1;MYC;COL4A4;BCL2;E2F1;TP53
Non-small cell lung cancer	8/56	<0.00001	CCND1;CDK4;ERBB2;E2F1;TGFA;PIK3CD;TP53;EGFR
Pancreatic cancer	14/66	<0.00001	STAT1;STAT3;TGFA;PIK3CD;BRCA2;RELA;EGFR;NFKB1;VEGFA;CCND1;CDK4;ERBB2;E2F1;TP53
Bladder cancer	12/41	<0.00001	CDKN1A;CCND1;CDK4;CDH1;MYC;ERBB2;MDM2;E2F1;MMP9;TP53;EGFR;VEGFA
Chronic myeloid leukemia	12/73	<0.00001	CDKN1A;CDKN1B;MECOM;CCND1;CDK4;MYC;E2F1;MDM2;PIK3CD;TP53;RELA;NFKB1
Glioma	10/65	<0.00001	CDKN1A;CCND1;CDK4;E2F1;MDM2;PTEN;TGFA;PIK3CD;TP53;EGFR
Melanoma	10/71	<0.00001	CDKN1A;CCND1;CDK4;CDH1;E2F1;MDM2;PTEN;PIK3CD;TP53;EGFR
Endometrial cancer	9/52	<0.00001	CCND1;CDH1;MYC;ERBB2;PTEN;CTNNB1;PIK3CD;TP53;EGFR
Colorectal cancer	9/62	<0.00001	JUN;CCND1;MYC;BCL2;BIRC5;CTNNB1;PIK3CD;FOS;TP53
Renal cell carcinoma	6/66	0.00052	JUN;TGFA;ARNT;PIK3CD;HIF1A;VEGFA
Acute myeloid leukemia	8/57	<0.00001	SPI1;CCND1;MYC;STAT3;RARA;PIK3CD;RELA;NFKB1
PI3K-Akt signaling	37/341	<0.00001	ITGB1;CSF1R;CDKN1A;CDKN1B;FLT1;CSF1;EPO;ITGB4;ITGA2B;PTEN;LPAR2;PIK3CD;PRL;BRCA1;RELA;EGFR;CCND2;CCND1;YWHAQ;MYC;KDR;JAK2;HSP90AA1;ITGA3;NFKB1;IL2;VEGFA;COL1A1;IL3;IL6;COL4A1;CDK4;COL4A4;IL2RA;MDM2;BCL2;TP53
MAPK signaling	14/255	0.00001	JUN;SRF;HSPB1;FOS;RELA;EGFR;NFKB1;RASGRP3;MECOM;MAPKAPK3;MYC;IL1B;NF1;TP53
ErbB signaling	8/87	0.00004	JUN;CDKN1A;CDKN1B;MYC;ERBB2;TGFA;PIK3CD;EGFR
Ras signaling	11/227	0.00043	CSF1R;FLT1;CSF1;KDR;NF1;PIK3CD;RELA;EGFR;NFKB1;VEGFA;RASGRP3
p53 signaling	11/69	<0.00001	CDKN1A;CCNB1;CCND2;CCND1;CDK4;CCNG2;SERPINE1;MDM2;PTEN;PMAIP1;TP53
FoxO signaling	15/133	<0.00001	CDKN2B;CDKN1A;CDKN1B;STAT3;PTEN;PIK3CD;EGFR;FOXO1;CCNB1;IL6;CCND2;CCND1;BCL6;CCNG2;MDM2
cAMP signaling	13/199	<0.00001	OXTR;JUN;PTGER2;PIK3CD;FOS;ATP1A1;RELA;NFKB1;MC2R;FSHR;AMH;MYL9;CFTR
Rap1 signaling	14/211	<0.00001	ITGB1;CSF1R;FLT1;CSF1;ITGA2B;ITGB2;LPAR2;PIK3CD;EGFR;VEGFA;RASGRP3;CDH1;

			KDR;CTNNB1
Chemokine signaling	10/187	0.00039	CXCL12;STAT1;STAT3;CCL3;PIK3CD;CXCR4;JAK2;RELA;NFKB1;PF4
Jak-STAT signaling	19/158	<0.00001	CDKN1A;CSF2;EPO;STAT1;IL13;STAT3;MPL;PIK3CD;PRL;IL2;IL3;IL6;CCND2;CCND1; IFNG;MYC;IL2RA;BCL2;JAK2
NF-kappa B signaling	9/93	0.00001	CD40;CXCL12;IL1B;BCL2;LY96;PRKCQ;PTGS2;RELA;NFKB1
Hippo signaling	15/153	<0.00001	WNT5A;SERPINE1;ITGB2;AFP;BMP7;CCND2;CCND1;YWHAQ;CDH1;DLG4;MYC;BIRC5; SNAI2;CTNNB1;AMH
Wnt signaling	7/142	0.00496	JUN;CCND2;CCND1;MYC;WNT5A;CTNNB1;TP53
HIF-1 signaling	19/103	<0.00001	CDKN1A;FLT1;CDKN1B;EPO;SERPINE1;STAT3;CYBB;ARNT;PIK3CD;HIF1A;RELA;EGFR; NFKB1;VEGFA;IL6;IFNG;ERBB2;BCL2;TIMP1
VEGF signaling	6/61	0.00034	MAPKAPK3;KDR;HSPB1;PIK3CD;PTGS2;VEGFA
TNF signaling	13/110	<0.00001	JUN;CEBPB;CSF2;CSF1;PIK3CD;FOS;PTGS2;MMP9;RELA;NFKB1;IL6;IL1B;JUNB
Apoptosis	12/140	<0.00001	IL3;JUN;BCL2;BIRC5;HTRA2;PMAIP1;PIK3CD;FOS;TP53;CTSD;RELA;NFKB1
Cell cycle	15/124	<0.00001	CDKN2B;CDKN1A;PCNA;CDKN1B;CCNA2;CCNB1;CCND2;CCND1;YWHAQ;CDK4;MYC; E2F1;MDM2;TP53;MAD2L1
Estrogen signaling	8/99	0.00012	HSP90AA1;JUN;SP1;PIK3CD;FOS;ESR1;MMP9;EGFR
Steroid hormone biosynthesis	15/58	<0.00001	HSD3B2;UGT1A1;HSD3B1;UGT2B15;AKR1C4;CYP19A1;CYP17A1;CYP11B2;CYP11A1; CYP11B1;CYP1A2;CYP1A1; CYP1B1;UGT1A4;UGT1A6
Prolactin signaling	14/72	<0.00001	LHCGR;STAT1;STAT3;LHB;PIK3CD;PRL;FOS;ESR1;RELA;NFKB1;CYP17A1;CCND2;CCND1; JAK2
Thyroid hormone signaling	12/118	<0.00001	NOTCH1;CCND1;STAT1;MYC;MDM2;CTNNB1;PIK3CD;ATP1A1;TP53;ESR1;HIF1A;FOXO1
Oxytocin signaling pathway	10/158	0.00011	OXTR;JUN;CDKN1A;CCND1;PIK3CD;OXT;FOS;PTGS2;MYL9;EGFR
T cell receptor signaling	11/104	<0.00001	JUN;CSF2;IFNG;CDK4;CTLA4;PIK3CD;PRKCQ;FOS;RELA;NFKB1;IL2
Toll-like receptor signaling	11/106	<0.00001	JUN;IL6;CD40;STAT1;IL1B;CCL3;LY96;PIK3CD;FOS;RELA;NFKB1
Cytokine-cytokine receptor interaction	26/265	<0.00001	CSF1R;CD40;CSF2;FLT1;CSF1;EPO;MPL;CXCR4;TNFRSF11B;PRL;EGFR;KDR;CCL3;AMH; TNFRSF18;IL13;BMP7;IL2; VEGFA;IL3;IL6;CXCL12;IFNG;IL1B;IL2RA;PF4
Leukocyte transendothelial migration	10/118	<0.00001	ITGB1;OCLN;CXCL12;ITGB2;CYBB;CTNNB1;PIK3CD;CXCR4;MYL9;MMP9
Regulation of actin cytoskeleton	10/214	0.00107	ITGB1;ITGA3;ITGB4;RDX;ITGA2B;ITGB2;PIK3CD;MYL9;MYH10;EGFR
Hepatitis B	22/146	<0.00001	JUN;CDKN1A;PCNA;CDKN1B;STAT1;STAT3;PTEN;PIK3CD;FOS;MMP9;RELA;NFKB1; CCNA2;IL6;CCND1;YWHAQ;CDK4;MYC;E2F1;BCL2;BIRC5;TP53
HTLV-I infection	27/258	<0.00001	CD40;CDKN1A;SPI1;PCNA;CSF2;SRF;ITGB2;PIK3CD;RELA;CCND2;TERT;CCND1;MYC; E2F1;JUN;CDKN2B;WNT5A;FOS;NFKB1;IL2;IL6;CDK4;IL2RA;CTNNB1;TCF3;TP53; MAD2L1
Measles	19/136	<0.00001	CDKN1B;STAT1;MX1;IL13;STAT3;PIK3CD;RELA;NFKB1;IL2;IL6;CCND2;CCND1;IFNG; CDK4;IL1B;IL2RA;PRKCQ;JAK2;TP53
Herpes simplex infection	12/185	0.00001	JUN;IL6;IFNG;STAT1;IL1B;FOS;JAK2;TP53;CLOCK;RELA;NFKB1;ARNTL
Epstein-Barr virus infection	19/202	<0.00001	JUN;CD40;CDKN1A;CR2;SPI1;CDKN1B;STAT3;HSPB1;PIK3CD;RELA;NFKB1;CCNA2; IFNG;YWHAQ;MYC;MDM2;BCL2;TP53;PTMA
Influenza A	10/175	0.00023	JUN;IL6;IFNG;STAT1;IL1B;MX1;PIK3CD;JAK2;RELA;NFKB1
Hepatitis C	9/133	0.00015	OCLN;CDKN1A;STAT1;STAT3;PIK3CD;TP53;RELA;EGFR;NFKB1
Viral carcinogenesis	17/205	<0.00001	JUN;CDKN2B;CDKN1A;CDKN1B;SRF;STAT3;PIK3CD;RELA;NFKB1;CCNA2;CCND2; CCND1;YWHAQ;CDK4;MDM2;PMAIP1;TP53
Chemical carcinogenesis	10/82	<0.00001	UGT1A1;CYP1A2;UGT2B15;CYP1A1;CYP1B1;ARNT;UGT1A4;CYP2C19;PTGS2;UGT1A6
Drug metabolism_cytochrome P450	9/69	<0.00001	CYP2B6;UGT1A1;MAOA;CYP2D6;CYP1A2;UGT2B15;UGT1A4;CYP2C19;UGT1A6
Xenobiotics metabolism_cytochrome P450	9/73	<0.00001	CYP2B6;UGT1A1;CYP2D6;CYP1A2;UGT2B15;CYP1A1;CYP1B1;UGT1A4;UGT1A6
Insulin resistance	9/109	0.00003	IL6;PPP2R4;STAT3;PTEN;PIK3CD;PRKCQ;RELA;FOXO1;NFKB1
Choline metabolism in cancer	8/101	0.00013	DGKG;SLC22A4;JUN;SP1;PIK3CD;FOS;HIF1A;EGFR
Retinol metabolism	7/65	0.00005	CYP2B6;UGT1A1;CYP1A2;UGT2B15;CYP1A1;UGT1A4;UGT1A6
Central carbon metabolism in cancer	8/67	<0.00001	RET;MYC;ERBB2;PTEN;PIK3CD;TP53;HIF1A;EGFR
Focal adhesion	20/202	<0.00001	ITGB1;JUN;FLT1;ITGA3;ITGB4;ITGA2B;PTEN;PIK3CD;EGFR;VEGFA;COL1A1;CCND2;

			CCND1;COL4A1;COL4A4;ERBB2;KDR;BCL2;CTNNB1;MYL9
Cell adhesion molecules (CAMs)	9/142	0.00024	ITGB1;OCLN;CD40;CD6;CDH1;ITGB2;CTLA4;NCAM1;CD34
Adherens junction	5/74	0.00540	CDH1;ERBB2;SNAI2;CTNNB1;EGFR
Tight junction	9/139	0.00021	OCLN;CDK4;PTEN;SYMPK;CTNNB1;MYH11;PRKCQ;MYL9;MYH10
ECM-receptor interaction	8/82	0.00003	COL1A1;ITGB1;ITGA3;COL4A1;ITGB4;COL4A4;ITGA2B;GP1BA
Transcriptional misregulation in cancer	27/180	<0.00001	CSF1R;CD40;CEBPB;CDKN1A;CSF2;SPI1;FLT1;CDKN1B;MPO;RELA;FOXO1;CCND2;MYC;ELANE;MMP9;RUNX2;NFKB1;PBX1;IL3;IL6;ZEB1;BCL6;SP1;MDM2;RARA;TCF3;TP53
MicroRNAs in cancer	26/297	<0.00001	CDKN1A;CDKN1B;NOTCH1;ABCB1;PTEN;BRCA1;PTGS2;SLC7A1;EGFR;CCND2;CCND1;MYC;ERBB2;E2F1;CYP1B1;RDX;STAT3;MIR27A;MMP9;NFKB1;VEGFA;FOXP1;ZEB1;MDM2;BCL2;TP53
Signaling pathways of stem cells	11/142	<0.00001	ZFHX3;SETDB1;MYC;WNT5A;STAT3;CTNNB1;PIK3CD;TCF3;KLF4;JAK2;POU5F1
Proteoglycans in cancer	20/203	<0.00001	ITGB1;CDKN1A;RDX;STAT3;WNT5A;IGF2;TWIST1;PIK3CD;HIF1A;ESR1;MMP9;EGFR;VEGFA;CCND1;MYC;ERBB2;KDR;MDM2;CTNNB1;TP53

Network analysis revealed a very high degree of connectivity across these target genes (observed edges: 3105; expected: 1104; PPI enrichment p-value: <10E-16); influence analysis showed that the top ten most influential proteins largely overlapped with those identified in the primary analysis (8/10), with ESR1 being the second ranking molecule (**Suppementary Table 5**).

Data were available for 238 SNPs linked to 275 genes which also allowed us to perform a subgroup analysis dedicated to estrogen receptor negative breast cancer (only four SNPs were available for estrogen receptor positive cases). Pathway and network analysis yielded results very similar to those obtained in the primary analysis where all breast cancer cases (both receptor positive and negative) were included (data not shown), likely because of the high degree of overlapping between the SNPs (and consequently of genes) of the two series.

SNPs shared with other tumors

Finally, we assessed whether some breast cancer risk associated SNPs are shared with other malignancies, a phenomenon known as pleiotropy (65). Querying the GWAS Catalog, we found 37 breast cancer risk SNPs shared with other eight tumor types (details are reported in **Table 3**): ovarian carcinoma (n=7), prostate carcinoma (n=4), lung carcinoma (n=2), thyroid carcinoma (n=1), esophageal carcinoma (n=1), renal cell carcinoma (n=1), cutaneous melanoma (n=1), glioma/glioblastoma (n=1) and a tumor miscellany mainly including ovarian, prostate and lung carcinoma (n=28). In two cases, the breast cancer susceptibility locus was shared with other three tumor types: one SNP (rs13016963) was located in chromosome 2q33.1 (sharing tumors: prostate and esophageal carcinomas, and cutaneous melanoma), the other SNP (rs10069690) in chromosome 5p15.33 (sharing tumors: ovarian and thyroid carcinomas, and glioma/glioblastoma).

These shared SNPs were associated with 34 genes: when we input these BCPGs into a pathway analysis, enrichment in apoptosis and cancer-related pathways was observed (**Supplementary Table 6**). Upon network analysis, the connectivity was very low (observed edges: 2; expected edges: 1; PPI enrichment P-value:

0.386). Adding 10 first-shell interactors showed the enrichment in cancer-related pathways as well as ribosome machinery and degenerative disease pathways (**Supplementary Table 6**).

Table 3: Breast cancer risk associated single nucleotide polymorphisms (SNPs) shared with other malignancies

Cancer	Chromosome	SNP	Genes
Glioma/Glioblastoma	5p15.33	rs10069690	TERT
Ovarian carcinoma	5p15.33	rs10069690	TERT
Thyroid carcinoma	5p15.33	rs10069690	TERT
Miscellany	9p21.3	rs1011970	CDKN2B
Prostate carcinoma	9p21.3	rs1011970	CDKN2B
Miscellany	19p13.11	rs10419397	ABHD8;ANKLE1;BABAM1
Ovarian carcinoma	19p13.11	rs10419397	ABHD8;ANKLE1;BABAM1
Miscellany	14q24.1	rs10483813	RAD51B
Miscellany	10q26.13	rs1078806	FGFR2
Miscellany	10q26.13	rs11200014	FGFR2
Lung carcinoma	13q13.1	rs11571833	BRCA2
Miscellany	14q24.1	rs11844632	RAD51B
Miscellany	20q11.22	rs11907546	CHMP4B
Miscellany	10q26.13	rs1219648	FGFR2
Esophageal carcinoma	2q33.1	rs13016963	CASP8;ALS2CR12
Melanoma (cutaneous)	2q33.1	rs13016963	CASP8;ALS2CR12
Prostate carcinoma	2q33.1	rs13016963	CASP8;ALS2CR12
Miscellany	3p24.1	rs1352941	NEK10;SLC4A7
Miscellany	19p13	rs1469713	MAU2;SUGP1;NDUFA13;GATAD2A;CILP2;TM6SF2
Prostate carcinoma	5p15.33	rs2242652	TERT
Ovarian carcinoma	19p13.11	rs2363956	ABHD8;ANKLE1;MRPL34;OCEL1
Miscellany	3p24.1	rs2590265	NEK10
Miscellany	10q26.13	rs2912780	FGFR2
Miscellany	10q26.13	rs2981575	FGFR2
Miscellany	10q26.13	rs2981582	FGFR2
Miscellany	10q26.13	rs3135718	FGFR2
Prostate carcinoma	1q32.1	rs4245739	MDM4;PIK3C2B
Miscellany	19p13.11	rs4808075	ANO8;ABHD8;ANKLE1;BABAM1;OCEL1
Ovarian carcinoma	19p13.11	rs4808075	ANO8;ABHD8;ANKLE1;BABAM1;OCEL1
Miscellany	3p24.1	rs481519	NEK10
Miscellany	3p24.1	rs571978	NEK10
Miscellany	3p24.1	rs580057	NEK10
Miscellany	5q11.2	rs59957907	C5orf67;MAP3K1

Miscellany	19p13.11	rs61494113	ABHD8;ANKLE1;OCEL1
Ovarian carcinoma	19p13.11	rs61494113	ABHD8;ANKLE1;OCEL1
Miscellany	14q24.1	rs61986943	RAD51B
Ovarian carcinoma	9p34	rs635634	SURF6;ABO
Miscellany	5q11.2	rs6450401	MAP3K1;SETD9
Miscellany	5q11.2	rs6890270	MIER3;SETD9
Miscellany	14q24.1	rs71423318	RAD51B
Renal cell carcinoma	11q22.3	rs74911261	KDELC2
Miscellany	5q11.2	rs7709971	C5orf67;MAP3K1
Miscellany	5q11.2	rs7714232	C5orf67;MAP3K1
Lung carcinoma	5p15.33	rs7726159	TERT
Miscellany	5p15.33	rs7726159	TERT
Ovarian carcinoma	19p13.11	rs8170	PLVAP;NR2F6;BABAM1;MRPL34;USHBP1;ABHD8;ANKLE1

Discussion

We reported on the first knowledge-base dedicated on GWAS-based evidence linking common germline variants to the risk of breast cancer. The information on breast cancer risk associated SNPs forms a knowledge-base which will be made publicly available at our cancer dedicated website (www.mmmp.org (66)) and will be annually updated.

Following the principles of integrative analysis of genomic data, we combined genome-wide information from different sources (e.g., high-throughput genotyping experiments, eQTL analysis, LD analysis, and so on) to make the most of the available evidence (8,9). This is of particular relevance because most SNPs do not have a direct functional effect, indeed a large proportion of associated SNPs are not in the coding regions of genes, and thus additional information is needed to link them to a gene. Then, we used these data to make tentative inferences on the pathways (and most influential molecules within them) whose variation can affect the risk of developing breast cancer.

Data from almost 400,000 women affected with breast cancer showed that 281 SNPs are significantly associated with the risk of this disease, which reduced to 233 risk loci when linkage disequilibrium was taken into account. These findings add new information to the already existing recent literature reviews on this subject, which report up to 172 common variants linked to breast cancer susceptibility (3,4,6,67-71). These SNPs are estimated to account 15-20% of the genetic component of disease risk (3,72), which clearly implies that much more work is needed to fully elucidate the molecular basis of breast cancer predisposition. It has been argued that future GWAS will not lead to the discovery of many more risk variants (3). This appears especially true in terms of rare variants (that is, variants with a MAF <1%) (3,72), as GWAS studies are designed to identify only common polymorphisms (MAF >1%) through a tagging strategy (tested tag-SNPs are in high linkage disequilibrium with non-tested SNPs). Moving forward, massively parallel sequencing technology (a.k.a. next generation sequencing [NGS], which can directly interrogate every genomic position) could provide investigators with the right tool to overcome the challenging hurdle of interrogating rarer variants which may affect risk, thus adding essential information to this field of investigation (73).

The data collected in our knowledge-base can be used to build polygenic predictive models and thus help optimize breast cancer secondary prevention programs (i.e. early detection by mammographic screening) by selecting women at higher risk (74-77). So far, such models have yielded generally unsatisfactory results, as their accuracy remains too low to be clinically implemented. This could be due to the fact that the complex genetic architecture of sporadic breast cancer predisposition remains still to be fully elucidated, as well as to the lack of information on gene-environment interactions (78,79). Nevertheless, the systematic collection of variants associated with breast cancer risk, along with information on their functional effect (as proposed in our knowledge-base) is the first step to build more effective predictive tools.

We utilized the collected information to generate tentative mechanistic hypotheses on the pathways whose perturbation (as determined by germline variation of the corresponding genes) affect breast cancer susceptibility. Some studies have already investigated the role of the variation of a single pathway across the results of multiple GWAS or the variation of multiple pathways within a single GWAS in the determinism of breast carcinogenesis (80,81). However, to the best of our knowledge, this is the first time that the comprehensive collection of variants linked to breast cancer risk by means of all available GWAS (and their meta-analyses) has been employed to systematically explore the cell pathways potentially involved in breast cancer development. Our gene set enrichment analysis led to the identification of multiple pathways well known to be involved in cancer development in general (such as apoptosis, cell cycle, and signal transduction) and breast cancer in particular (such as steroid hormone pathways). As regards the latter, the estrogen receptor pathway was confirmed to play a pivotal role in the carcinogenesis of a hormone dependent neoplasm such as breast carcinoma (82), within this frame, the gene encoding the estrogen receptor alpha (ESR1) was a key influencer in the generated networks of BCPGs (see **Figure 1** and **Supplementary Table 5**). This finding might be of relevance with regard to breast cancer chemoprevention, which aims to reduce disease incidence by the administration of anti-estrogen drugs such as selective estrogen receptor modifiers (e.g., tamoxifen) (83). The selection of women who could most benefit from these risk reducing medications might be improved by genetic testing based on polymorphisms that affect breast cancer risk (84).

Another interesting piece of information yielded from data analysis is the high degree of overlap between network-guided gene set enrichment primary analysis and the pathway analysis performed with the targets of BCSGs acting as transcription factors (see **Table 1** and **Table 2**). This finding supports the hypothesis that most of the biological effect of the SNPs linked to breast cancer risk might actually be mediated by regulons governed by the transcription factors associated with those SNPs. Notably, our data confirm the results of a recent publication where investigators have identified a breast cancer risk regulatory network comprising some of the transcription factors we identified as BCPGs (85).

Besides well known cancer-related pathways (such as apoptosis, signal transduction and so on), our gene set enrichment analysis showed that germline variation of other pathways might be of particular relevance for breast cancer susceptibility, such as those involved in anti-viral immunity, degenerative diseases as well as peroxisome and ribosome activity (see **Table 1** and **Table 2**). Actually, peroxisomes are known to be linked to carcinogenesis through their production of reactive oxygen species (86), which in turn can initiate tumor development by causing DNA damage. Of special interest is also the case of genes encoding ribosome proteins, which were repeatedly enriched in our pathway and network analyses of the whole series, as well as in the analysis of pleiotropic SNPs. Indeed, it has recently been suggested that ribosome derangement may play a significant role in both development and progression of different tumor types (87,88), including breast cancer (89).

In conclusion, we present the first knowledge-base dedicated to sporadic breast cancer predisposition variants. This wealth of information can inform future studies aimed to dissect the molecular epidemiology and the molecular basis of this disease.

References

1) DeSantis CE, Bray F, Ferlay J, et al. International Variation in Female Breast Cancer Incidence and Mortality Rates. Cancer Epidemiol Biomarkers Prev 2015;24(10):1495-506.
2) Senft D, Leiserson MDM, Ruppin E, et al. Precision Oncology: The Road Ahead. Trends Mol Med 2017;23(10):874-898.
3) Skol AD, Sasaki MM, Onel K. The genetics of breast cancer risk in the post-genome era: thoughts on study design to move past BRCA and towards clinical relevance. Breast Cancer Res 2016;18(1):99.
4) Ghoussaini M, Pharoah PDP, Easton DF. Inherited genetic susceptibility to breast cancer: the beginning of the end or the end of the beginning? Am J Pathol 2013;183(4):1038-1051.
5) Zhang B, Beeghly-Fadiel A, Long J, et al. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. Lancet Oncol 2011;12(5):477-88.
6) Fachal L, Dunning AM. From candidate gene studies to GWAS and post-GWAS analyses in breast cancer. Curr Opin Genet Dev 2015;30:32-41.
7) MacArthur J, Bowler E, Cerezo M, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 2017;45(D1):D896-D901.
8) Kristensen VN, Lingjaerde OC, Russnes HG, et al. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 2014;14:299-313.
9) Thingholm LB, Andersen L, Makalic E, et al. Strategies for Integrated Analysis of Genetic, Epigenetic, and Gene Expression Variation in Cancer: Addressing the Challenges. Front Genet 2016;7:2.
10) Mooney MA, Nigg JT, McWeeney SK, et al. Functional and genomic context in pathway analysis of GWAS data. Trends Genet 2014;30(9):390-400.
11) Creixell P, Reimand J, Haider S, et al. Pathway and network analysis of cancer genomes. Nat Methods 2015;12:615-621.
12) The GTEx Consortium. Genetic effects on gene expression across human tissues. Nature 2017;550:204-213.

13) Ernst J, Kheradpour P, Mikkelsen TS, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature 2011;473(7345):43-9.

14) Bernstein BE, Stamatoyannopoulos JA, Costello JF, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010;28(10):1045-8.

15) Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res 2012;22(9):1790-7.

16) Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res 2012;40(Database issue):D930-4.

17) Machiela MJ, Chanock SJ. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. Bioinformatics 2015;31(21):3555-7.

18) Kuleshov MV, Jones MR, Rouillard AD, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 2016;44(W1):W90-7.

19) Rivals I, Personnaz L, Taing L, et al. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007;23(4):401-7.

20) Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100(16):9440-5.

21) Szklarczyk D, Morris JH, Cook H, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res 2017;45:D362-68.

22) He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS Genet 2006;2(6):e88.

23) Bean DM, Heimbach J, Ficorella L, et al. EsyN: network building, sharing and publishing. PLoS One 2014;9(9):e106035.

24) Morone F, Makse HA. Influence maximization in complex networks through optimal percolation. Nature 2015;524(7563):65-8.

25) UniProt Consortium. UniProt: a hub for protein information. Nucleic Acids Res 2015;43(Database issue):D204-12.

26) Han H, Cho JW, Lee S, et al. TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Res 2018;46(D1):D380-D386.

27) Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 2007;447(7148):1087-93.

28) Hunter DJ, Kraft P, Jacobs KB, et al. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007;39(7):870-4.

29) Stacey SN, Manolescu A, Sulem P, et al. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007;39(7):865-9.

30) Gold B, Kirchhoff T, Stefanov S, et al. Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. Proc Natl Acad Sci U S A 2008;105(11):4340-5.

31) Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 2009;41(3):324-8.

32) Thomas G, Jacobs KB, Kraft P, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet 2009;41(5):579-84.

33) Turnbull C, Ahmed S, Morrison J, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010;42(6):504-7.

34) Long J, Cai Q, Shu XO, et al. Identification of a functional genetic variant at 16q12.1 for breast cancer risk: results from the Asia Breast Cancer Consortium. PLoS Genet 2010;6(6):e1001002.

35) Antoniou AC, Wang X, Fredericksen ZS, et al. A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population. Nat Genet 2010;42(10):885-92.

36) Li J, Humphreys K, Heikkinen T, et al. A combined analysis of genome-wide association studies in breast cancer. Breast Cancer Res Treat 2011;126(3):717-27.

37) Gaudet MM, Kirchhoff T, Green T, et al. Common genetic variants and modification of penetrance of BRCA2-associated breast cancer. PLoS Genet 2010;6(10):e1001183.

38) Fletcher O, Johnson N, Orr N, et al. Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst 2011;103(5):425-35.

39) Cai Q, Long J, Lu W, et al. Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. Hum Mol Genet 2011;20(24):4991-9.

40) Haiman CA, Chen GK, Vachon CM, et al. A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer. Nat Genet 2011;43(12):1210-4.

41) Long J, Cai Q, Sung H, et al. Genome-wide association study in east Asians identifies novel susceptibility loci for breast cancer. PLoS Genet 2012;8(2):e1002532.

42) Kim HC, Lee JY, Sung H, et al. A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study. Breast Cancer Res 2012;14(2):R56.

43) Elgazzar S, Zembutsu H, Takahashi A, et al. A genome-wide association study identifies a genetic variant in the SIAH2 locus associated with hormonal receptor-positive breast cancer in Japanese. J Hum Genet 2012;57(12):766-71.

44) Siddiq A, Couch FJ, Chen GK, et al. A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. Hum Mol Genet 2012;21(24):5373-84.

45) Orr N, Lemnrau A, Cooke R, et al. Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk. Nat Genet 2012;44(11):1182-4.

46) Rinella ES, Shao Y, Yackowski L, et al. Genetic variants associated with breast cancer risk for Ashkenazi Jewish women with strong family histories but no identifiable BRCA1/2 mutation. Hum Genet 2013;132(5):523-36.

47) Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013;45(4):353-61.

48) Garcia-Closas M, Couch FJ, Lindstrom S, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet 2013;45(4):392-8.

49) Couch FJ, Wang X, McGuffog L, et al. Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet 2013;9(3):e1003212.

50) Low SK, Takahashi A, Ashikawa K, et al. Genome-wide association study of breast cancer in the Japanese population. PLoS One 2013;8(10):e76463.

51) Purrington KS, Slager S, Eccles D, et al. Genome-wide association study identifies 25 known breast cancer susceptibility loci as risk factors for triple-negative breast cancer. Carcinogenesis 2014;35(5):1012-9.

52) Ahsan H, Halpern J, Kibriya MG, et al. A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. Cancer Epidemiol Biomarkers Prev 2014;23(4):658-69.

53) Sawyer E, Roylance R, Petridis C, et al. Genetic predisposition to in situ and invasive lobular carcinoma of the breast. PLoS Genet 2014;10(4):e1004285.

54) Cai Q, Zhang B, Sung H, et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. Nat Genet 2014;46(8):886-90.

55) Fejerman L, Ahmadiyeh N, Hu D, et al. Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25. Nat Commun 2014;5:5260.

56) Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 2015;47(4):373-80.

57) Palomba G, Loi A, Porcu E, et al. Genome-wide association study of susceptibility loci for breast cancer in Sardinian population. BMC Cancer 2015;15:383.

58) Couch FJ, Kuchenbaecker KB, Michailidou K, et al. Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat Commun 2016;7:11375.

59) Fehringer G, Kraft P, Pharoah PD, et al. Cross-Cancer Genome-Wide Analysis of Lung, Ovary, Breast, Prostate, and Colorectal Cancer Reveals Novel Pleiotropic Associations. Cancer Res 2016;76(17):5103-14.

60) Han MR, Long J, Choi JY, et al. Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci. Hum Mol Genet 2016;25(15):3361-3371.

61) Kar SP, Beesley J, Amin Al Olama A, et al. Genome-Wide Meta-Analyses of Breast, Ovarian, and Prostate Cancer Association Studies Identify Multiple New Susceptibility Loci Shared by at Least Two Cancer Types. Cancer Discov 2016;6(9):1052-67.

62) Huo D, Feng Y, Haddad S, et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. Hum Mol Genet 2016;25(21):4835-4846.

63) Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. Nature 2017;551(7678):92-94.

64) Milne RL, Kuchenbaecker KB, Michailidou K, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet 2017;49(12):1767-1778.

65) Wu YH, Graff RE, Passarelli MN, et al. Identification of Pleiotropic Cancer Susceptibility Variants from Genome-Wide Association Studies Reveals Functional Characteristics. Cancer Epidemiol Biomarkers Prev 2018;27(1):75-85.

66) Mocellin S, Rossi CR. The melanoma molecular map project. Melanoma Res 2008;18(3):163-5.

67) Peng S, Lü B, Ruan W, et al. Genetic polymorphisms and breast cancer risk: evidence from meta-analyses, pooled analyses, and genome-wide association studies. Breast Cancer Res Treat 2011;127(2):309-24.

68) Maxwell KN, Nathanson KL. Common breast cancer risk variants in the post-COGS era: a comprehensive review. Breast Cancer Res 2013;15(6):212.

69) Sapkota Y. Germline DNA variations in breast cancer predisposition and prognosis: a systematic review of the literature. Cytogenet Genome Res 2014;144(2):77-91.

70) Walsh MF, Nathanson KL, Couch FJ, et al. Genomic Biomarkers for Breast Cancer Risk. Adv Exp Med Biol 2016;882:1-32.

71) Lilyquist J, Ruddy KJ, Vachon CM, et al. Common Genetic Variation and Breast Cancer Risk - Past, present, and future. Cancer Epidemiol Biomarkers Prev 2018; [Epub ahead of print]

72) Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A 2014;111(4):E455-64.

73) Chandler MR, Bilgili EP, Merner ND. A Review of Whole-Exome Sequencing Efforts Toward Hereditary Breast Cancer Susceptibility Gene Discovery. Hum Mutat 2016;37(9):835-46.

74) Evans DG, Brentnall A, Byers H, et al. The impact of a panel of 18 SNPs on breast cancer risk in women attending a UK familial screening clinic: a case-control study. J Med Genet 2017;54(2):111-113.

75) Shieh Y, Hu D, Ma L, et al. Breast cancer risk prediction using a clinical risk model and polygenic risk score. Breast Cancer Res Treat 2016;159(3):513-25.

76) Mavaddat N, Pharoah PD, Michailidou K, et al. Prediction of breast cancer risk based on profiling with common genetic variants. J Natl Cancer Inst 2015;107(5).

77) Cuzick J, Brentnall AR, Segal C, et al. Impact of a Panel of 88 Single Nucleotide Polymorphisms on the Risk of Breast Cancer in High-Risk Women: Results From Two Randomized Tamoxifen Prevention Trials. J Clin Oncol 2017;35(7):743-750.

78) Rudolph A, Song M, Brook MN, et al. Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. Int J Epidemiol 2018; [Epub ahead of print]

79) Ito H, Matsuo K. Molecular epidemiology, and possible real-world applications in breast cancer. Breast Cancer 2016;23(1):33-8.

80) Lee YH, Kim JH, Song GG. Genome-wide pathway analysis of breast cancer. Tumour Biol 2014;35(8):7699-705.

81) Scarbrough PM, Weber RP, Iversen ES, et al. A Cross-Cancer Genetic Association Analysis of the DNA Repair and DNA Damage Signaling Pathways for Lung, Ovary, Prostate, Breast, and Colorectal Cancer. Cancer Epidemiol Biomarkers Prev 2016;25(1):193-200.

82) Li T, Zhao J, Yang J, et al. A Meta-Analysis of the Association between ESR1 Genetic Variants and the Risk of Breast Cancer. PLoS One 2016;11(4):e0153314.

83) Mocellin S, Pilati P, Briarava M, et al. Breast Cancer Chemoprevention: A Network Meta-Analysis of Randomized Controlled Trials. J Natl Cancer Inst 2015;108(2). pii: djv318.

84) Ziv E, Tice JA, Sprague B, et al. Using Breast Cancer Risk Associated Polymorphisms to Identify Women for Breast Cancer Chemoprevention. PLoS One 2017;12(1):e0168601.

85) Castro MA, de Santiago I, Campbell TM, et al. Regulators of genetic risk of breast cancer identified by integrative network analysis. Nat Genet 2016;48(1):12-21.

86) Gào X, Schöttker B. Reduction-oxidation pathways involved in cancer development: a systematic review of literature reviews. Oncotarget 2017;8(31):51888-51906.

87) Pelletier J, Thomas G, Volarević S. Ribosome biogenesis in cancer: new players and therapeutic avenues. Nat Rev Cancer 2018;18(1):51-63.

88) Bustelo XR, Dosil M. Ribosome biogenesis and cancer: basic and translational challenges. Curr Opin Genet Dev 2017;48:22-29.

89) Belin S, Beghin A, Solano-Gonzàlez E, et al. Dysregulation of ribosome biogenesis and translational capacity is associated with tumor progression of human breast cancer cells. PLoS One 2009;4(9):e7147.