



Gene- and pathway-level analyses of iCOGS variants highlight novel signaling pathways underlying familial breast cancer susceptibility

Christine Lonjou, Séverine Eon-marchais, Thérèse Truong, Marie-gabrielle Dondon, Mojgan Karimi, Yue Jiao, Francesca Damiola, Laure Barjhoux, Dorothée Le Gal, Juana Beauvallet, et al.

► To cite this version:

Christine Lonjou, Séverine Eon-marchais, Thérèse Truong, Marie-gabrielle Dondon, Mojgan Karimi, et al.. Gene- and pathway-level analyses of iCOGS variants highlight novel signaling pathways underlying familial breast cancer susceptibility. *International Journal of Cancer*, 2021, 148 (8), pp.1895-1909. 10.1002/ijc.33457 . hal-03345363

HAL Id: hal-03345363

<https://hal.science/hal-03345363>

Submitted on 17 Sep 2021
















HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Gene- and pathway-level analyses of iCOGS variants highlight novel signaling pathways underlying familial breast cancer susceptibility

Christine Lonjou^{1,2,3}  | Séverine Eon-Marchais^{1,2,3} | Thérèse Truong^{4,5}  |
 Marie-Gabrielle Dondon^{1,2,3}  | Mojgan Karimi^{4,5} | Yue Jiao^{1,2,3} |
 Francesca Damiola⁶  | Laure Barjhoux⁶ | Dorothee Le Gal^{1,2,3} |
 Juana Beauvallet^{1,2,3} | Noura Mebirouk^{1,2,3} | Eve Cavaciuti^{1,2,3} | Jean Chiesa⁷ |
 Anne Floquet⁸ | Séverine Audebert-Bellanger⁹ | Sophie Giraud¹⁰ |
 Thierry Frebourg¹¹ | Jean-Marc Limacher¹²  | Laurence Gladieff¹³  |
 Isabelle Mortemousque¹⁴ | Hélène Dreyfus^{15,16} | Sophie Lejeune-Dumoulin¹⁷ |
 Christine Lasset^{18,19,20} | Laurence Venat-Bouvet²¹  | Yves-Jean Bignon²²  |
 Pascal Pujol^{23,24}  | Christine M. Maugard^{25,26} | Elisabeth Luporsi²⁷ |
 Valérie Bonadona^{18,19,20} | Catherine Noguès^{28,29} | Pascaline Berthet³⁰ |
 Capucine Delnatte³¹ | Paul Gesta³² | Alain Lortholary³³ | Laurence Faivre^{34,35} |
 Bruno Buecher³⁶ | Olivier Caron³⁷  | Marion Gauthier-Villars³⁶ |
 Isabelle Coupier^{23,24} | Sylvie Mazoyer³⁸  | Luis-Cristobal Monraz^{1,2,3} |
 Maria Kondratova^{1,2,3} | Inna Kuperstein^{1,2,3}  | Pascal Guénel^{4,5}  |
 Emmanuel Barillot^{1,2,3}  | Dominique Stoppa-Lyonnet^{36,39}  | Nadine Andrieu^{1,2,3} |
 Fabienne Lesueur^{1,2,3} 

¹Inserm, U900, Institut Curie, Paris, France

²Mines ParisTech, Fontainebleau, France

³PSL Research University, Paris, France

⁴Université Paris-Saclay, UVSQ, Inserm, CESP, Villejuif, France

⁵Inserm U1018, CESP, Team Exposome and Heredity, Villejuif, France

⁶Department of BioPathology, Centre Léon Bérard, Lyon, France

⁷CHRU Hôpital Caremeau, Nîmes, France

⁸Institut Bergonié, Bordeaux, France

⁹Département de Génétique Médicale et Biologie de la Reproduction, CHU Brest, Hôpital Morvan, Brest, France

¹⁰Service de Génétique, Hospices Civils de Lyon, Groupement Hospitalier Est, Bron, France

Abbreviations: ACSN, Atlas of Cancer Signaling Network; AUC, area under the receiver-operator curve; BC, breast cancer; BCAC, Breast Cancer Association Consortium; CI, confidence interval; DAPPLE, Disease Association Protein-Protein Link Evaluator; ER, estrogen receptor; FDR, false-positive discovery rate; GWAS, genome-wide association study; KEGG, Kyoto Encyclopedia of Genes and Genomes; LD, linkage disequilibrium; MAF, minor allele frequency; OR, odds ratio; PPI, protein-protein interaction; PRS, polygenic risk score; QC, quality control; ROC, receiver-operating characteristic; SNP, single-nucleotide polymorphism.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2020 The Authors. *International Journal of Cancer* published by John Wiley & Sons Ltd on behalf of UICC.

- ¹¹Département de Génétique, Hôpital Universitaire de Rouen, Rouen, France
- ¹²Service d'Onco-Hématologie, Hôpital Pasteur, Colmar, France
- ¹³Service d'Oncologie Médicale, Institut Claudius Regaud—IUCT-Oncopole, Toulouse, France
- ¹⁴Service de Génétique, Hôpital Bretonneau, Tours, France
- ¹⁵Clinique Sainte Catherine, Avignon, France
- ¹⁶Département de Génétique, CHU de Grenoble, Hôpital Couple-Enfant, Grenoble, France
- ¹⁷Service de Génétique Clinique Guy Fontaine, CHU Lille, Lille, France
- ¹⁸Université Claude Bernard Lyon 1, Villeurbanne, France
- ¹⁹CNRS UMR 5558, Lyon, France
- ²⁰Centre Léon Bérard, Unité de Prévention et Epidémiologie Génétique, Lyon, France
- ²¹Service d'Oncologie Médicale, Hôpital Universitaire Dupuytren, Limoges, France
- ²²Département d'Oncogénétique, Université Clermont Auvergne, UMR INSERM, U1240, Centre Jean Perrin, Clermont Ferrand, France
- ²³Hôpital Arnaud de Villeneuve, CHU Montpellier, Service de Génétique Médicale et Oncogénétique, Montpellier, France
- ²⁴INSERM 896, CRCM Val d'Aurelle, Montpellier, France
- ²⁵Département d'Oncobiologie, LBBM, Hôpitaux Universitaires de Strasbourg, Génétique Oncologique Moléculaire, UF1422, Strasbourg, France
- ²⁶Hôpitaux Universitaires de Strasbourg, UF6948 Génétique Oncologique Clinique, Évaluation Familiale et Suivi, Strasbourg, France
- ²⁷ICL Alexis Vautrin, Unité d'Oncogénétique, Vandœuvre-lès-Nancy, France
- ²⁸Département d'Anticipation et de Suivi des Cancers, Oncogénétique Clinique, Institut Paoli-Calmettes, Marseille, France
- ²⁹Aix Marseille University, INSERM, IRD, SESSTIM, Marseille, France
- ³⁰Département de Biopathologie, Centre François Baclesse, Oncogénétique, Caen, France
- ³¹Institut de Cancérologie de l'Ouest, Unité d'Oncogénétique, Saint Herblain, France
- ³²CH Georges Renon, Service d'Oncogénétique Régional Poitou-Charentes, Niort, France
- ³³Centre Catherine de Sienne, Service d'Oncologie Médicale, Nantes, France
- ³⁴Institut GIMI, CHU de Dijon, Hôpital d'Enfants, Dijon, France
- ³⁵Oncogénétique, Centre de Lutte contre le Cancer Georges François Leclerc, Dijon, France
- ³⁶Institut Curie, Service de Génétique, Paris, France
- ³⁷Département de Médecine Oncologique, Gustave Roussy, Villejuif, France
- ³⁸Equipe GENDEV, Centre de Recherche en Neurosciences de Lyon, Inserm U1028, CNRS UMR5292, Université Lyon 1, Université St Etienne, Lyon, France
- ³⁹Inserm, U830, Université Paris-Descartes, Paris, France

Correspondence

Fabienne Lesueur, Inserm, U900, Institut Curie, Paris, France.
Email: fabienne.lesueur@curie.fr

Funding information

Institut National Du Cancer, Grant/Award Number: b2008-029/LL-LC; Ligue Nationale contre le Cancer, Grant/Award Numbers: PRE05/DSL, PRE07/DSL, PRE11/NA; Site de Recherche Intégrée sur le Cancer, Grant/Award Number: SiRIC Grant INCa-DGOS-4654

Abstract

Single-nucleotide polymorphisms (SNPs) in over 180 loci have been associated with breast cancer (BC) through genome-wide association studies involving mostly unselected population-based case-control series. Some of them modify BC risk of women carrying a *BRCA1* or *BRCA2* (*BRCA1/2*) mutation and may also explain BC risk variability in BC-prone families with no *BRCA1/2* mutation. Here, we assessed the contribution of SNPs of the iCOGS array in GENESIS consisting of BC cases with no *BRCA1/2* mutation and a sister with BC, and population controls. Genotyping data were available for 1281 index cases, 731 sisters with BC, 457 unaffected sisters and 1272 controls. In addition to the standard SNP-level analysis using index cases and controls, we performed pedigree-based association tests to capture transmission information in the sibships. We also performed gene- and pathway-level analyses to maximize the power to detect associations with lower-frequency SNPs or those with modest effect sizes. While SNP-level analyses identified 18 loci, gene-level analyses identified 112 genes. Furthermore, 31 Kyoto Encyclopedia of Genes and Genomes and 7 Atlas

of Cancer Signaling Network pathways were highlighted (false discovery rate of 5%). Using results from the “index case-control” analysis, we built pathway-derived polygenic risk scores (PRS) and assessed their performance in the population-based CECILE study and in a data set composed of GENESIS-affected sisters and CECILE controls. Although these PRS had poor predictive value in the general population, they performed better than a PRS built using our SNP-level findings, and we found that the joint effect of family history and PRS needs to be considered in risk prediction models.

KEYWORDS

familial breast cancer, single-nucleotide polymorphism, systems biology, association study

1 | INTRODUCTION

One of the strongest risk factors for the development of breast cancer (BC) is having a close relative affected with the disease. On the basis of the increased risk of BC in first-degree relatives of a woman with BC and segregation studies on BC cases in the families of affected women, it was estimated that 5% of these women carry a genetic predisposition factor transmitted according to a Mendelian dominant model.¹ Following the cloning of *BRCA1* and *BRCA2* (*BRCA1/2*) 25 years ago, diagnostic testing for pathogenic variants (or “mutation”) in these two major BC susceptibility genes involved in DNA damage response and DNA repair has been routine clinical practice in many developed countries. It has facilitated risk estimation and implementation of cancer prevention strategies and has now the potential to influence cancer therapy.^{2,3} More recently, other BC susceptibility genes have been identified essentially through resequencing of candidate genes investigated because of their direct or indirect functional link with *BRCA1* and *BRCA2* (*PALB2*, *ATM*, *CHEK2*, etc.),^{4–10} and BC risk associated with pathogenic variants in these genes ranges from elevated like *BRCA1/2* to moderate like *ATM*. In the meantime, common modest-risk single-nucleotide polymorphisms (SNPs) located in over 180 loci were detected by genome-wide association studies (GWAS). Combined as polygenic risk scores (PRS), these SNPs would explain about 10% of familial clustering.^{11–13} However, taking together all genetic variations involved in BC susceptibility, about 50% of the familial relative risk for BC remains unexplained. Given current data, it is very likely that the remaining familial aggregation of BC will be explained by many genetic alterations with a wide spectrum of associated risks, possibly in combination with other factors such as lifestyle or environmental-related factors.

Today, the search for new BC susceptibility variants seems to be in a dead end, where increasing the size of the studies has reached its limits and does not seem to bring new discoveries, and where alternative strategies must be developed. Here, we proposed to use multi-level approaches including single-variant, gene- and pathway-level analyses to maximize the power to detect modest effect sizes or lower-frequency BC predisposing variants,¹⁴ to explore the coherence

What's new?

Genetic studies have identified more than 180 single-nucleotide polymorphisms (SNPs) associated with breast cancer susceptibility, but these studies are reaching their limits. Here, the authors evaluated SNPs in the iCOGS genotyping array using a multilevel approach, including single variant, gene, and pathway analyses. They measured the contribution of the SNPs to breast cancer in patients who have a sister with breast cancer but do not carry a *BRCA1/2* mutation. They showed that a pathway-derived polygenic risk score performed poorly in the general population, and that the best predictive model must include family history.

in findings, and to get further insight into the underlying molecular mechanisms involved in BC susceptibility. In addition, we built new PRS for BC prediction based on pathway analyses and evaluated their performance.

2 | MATERIAL AND METHODS

2.1 | Study participants

The studied population consisted of women participating in GENESIS (GENE SISTers), a French resource for familial BC research.¹⁵ In brief, 1721 women affected with breast adenocarcinoma, not carrying a pathogenic variant in *BRCA1* and *BRCA2*, and having at least one sister with BC were enrolled in the study between 2007 and 2013 through the national network of cancer genetics clinics (<http://www.unicancer.fr/en/unicancer-group>). Affected sisters (N = 826), unaffected sisters (N = 599) and unrelated cancer-free friends or colleagues of index cases (controls) were also included (N = 1419). These latter were aged-matched (± 3 years) to cases at interview. Blood samples, clinical, familial and epidemiological data were collected for each

participant. Information about ethnic origin was self-reported by study subjects. Here we focused our analyses on subjects of European origin; those represented over 98% of the GENESIS population. After quality control (QC) procedures (see Section 2.3), we analyzed genotyping data from 1281 index cases, 731 affected sisters, 457 unaffected sisters and 1272 unrelated controls.

Validation of the pathway-specific PRS was performed in the CECILE population. CECILE is a population-based case-control study which was conducted in Côte d'Or and Ille-et-Vilaine, two administrative areas (*départements*) located in Eastern and Western parts of France, respectively.¹⁶ Cases were BC patients aged 25 to 75 years, with histologically confirmed invasive or in situ breast carcinoma diagnosed between 2005 and 2007. A total of 1232 incident BC cases and 1317 controls were enrolled in the study. Controls were selected from the general population among women living in the same areas with no personal history of BC. They were frequency-matched to the cases by 10-year age group and study area. A face-to-face interview with a trained nurse was conducted for all cases and controls. A standardized questionnaire was used to obtain information on hormonal and reproductive factors, personal medical history, family history of cancer. A blood sample was also collected during interview. iCOGS genotyping data were available for 1019 cases (of which 900 cases had invasive tumors and 119 had in situ tumor) and 999 controls.¹⁶ Demographic and clinical characteristics of GENESIS and CECILE women included in the analyses are presented in Table S1.

2.2 | Strategy

We performed data mining of SNPs on the iCOGS array^{17,18} in the GENESIS well-characterized population which includes familial BC cases with no *BRCA1/2* pathogenic variant, affected and unaffected sisters and cancer-free friends or colleagues serving as controls.¹⁵ We employed both unrelated case-control and pedigree-based designs at single-variant, gene and pathway levels. We also performed protein-protein interaction (PPI) analysis to identify genetic variation affecting common pathways and to compare results obtained with the different approaches.

We next assessed whether the cumulative effect of uncorrelated SNPs in genes of the identified BC-associated pathways, expressed as PRS, had predictive ability for BC by applying receiver-operating characteristic (ROC) analysis to the CECILE-independent data set involving unselected BC cases and controls from the French population.¹⁶ Finally, we also evaluated performance of the pathway-specific PRS in a data set composed of the GENESIS-affected sisters and CECILE controls. Figure S1 illustrates the study design.

2.3 | Genotyping and QC procedures

All study participants from GENESIS and CECILE were genotyped using the custom iCOGS array (Illumina Inc., San Diego, California)

targeting 211 155 SNPs throughout the genome. The array was designed in collaboration between the PRACTICAL, Breast Cancer Association Consortium (BCAC), Ovarian Cancer Association Consortium and Consortium of Investigators of Modifiers of *BRCA1/2* consortia. Genotyping of CECILE samples was performed in the context of studies conducted by BCAC, and these data contributed to the published GWAS.¹⁹ Detailed information about the design, genotyping and QC procedures for iCOGS can be found within the original publication.²⁰ Genotyping of GENESIS samples was performed subsequently at Genome Quebec and analyzed separately. In GENESIS, genotype calling was performed using Illumina GenomeStudio 2010 (Illumina Inc.). SNPs were excluded if genotyping rate was lower than 90%, or minor allele frequency (MAF) was <0.001 in the whole data set, or Hardy-Weinberg equilibrium was rejected ($P < .001$) in controls.

In order to identify potential duplicates and check for relatedness between study participants, kinship coefficients were calculated between all pairs of individuals with the *-genome -genome-full* command of PLINK²¹ using a subset of 81 057 independent SNPs (with $MAF \geq 0.07$ and $r^2 < .5$).

2.4 | SNP-level analysis

SNPs were first tested individually using PLINK version 1.7.²¹ Odds ratios (OR) were calculated for allelic model (a vs A). In the case-control analysis, reported P values are adjusted for age at diagnosis for cases and age at inclusion for controls. Multiple testing was taken into account by using Benjamini and Hochberg's procedure to compute the false-positive discovery rate (FDR), with a significance threshold of 0.05 (P_{FDR}).²² Family-based association tests were carried out using the "dfam" option of PLINK. This method implements the sib-transmission disequilibrium test and also allows for unrelated individuals to be included via a clustered analysis using the Cochran-Mantel-Haenszel method.²³

2.5 | Gene-level analysis

Gene-level analyses were performed using VEGAS2 (version 2, <https://vegas2.qimrberghofer.edu.au/vegas2v2>), a versatile gene-based test for GWAS,^{24,25} which performs gene-based tests based on association test from single-variant analyses and accounts for linkage disequilibrium (LD) between SNPs and the number of SNPs tested to avoid an increase in false-positive results due to genes with multiple, highly correlated markers. The method tests the evidence for association on a per-gene basis by summarizing either the full set of SNPs in the gene or a subset of the most significant SNPs. Here the 10% most significant SNPs in a gene were used. The results shown were obtained using GENESIS unrelated controls as reference data set for LD calculation. We considered a SNP to belong to a gene if it is located within 50 kb on either side of the gene's transcribed region, which we found to be a good balance between incorporating short-

range regulatory variants while maintaining the specificity of the result for a specific gene, as variants associated with neighboring genes can influence the test statistics for the gene of interest. VEGAS2 algorithm assigns SNPs to genes and calculates gene-based empirical association P values (P_{GENE}) while accounting for the LD structure within the gene.

SNPs were annotated using ANNOVAR.²⁶ Among the 197 182 analyzed iCOGS SNPs, 161 907 (82.1%) are located in the coding sequence of the genome or within 50 kb on either side of a gene's transcribed region according to the position information obtained from GENCODE Release 28 (<https://www.gencodegenes.org/releases/current.html>).

2.6 | PPI analysis

Analysis was performed using Disease Association Protein-Protein Link Evaluator (DAPPLE, version 19 and hg19 reference map) to investigate physical connections between proteins.²⁷ DAPPLE searches the InWeb database for PPI that have been reported in the literature and assigns a score reflecting the probability of being physically connected. The InWeb database compiles PPI data from numerous sources including Reactome, IntAct, the Molecular Interaction Database, the Biomolecular Interaction Network Database and Kyoto Encyclopedia of Genes and Genomes (KEGG).²⁷ DAPPLE is designed to analyze disease-associated SNPs or genes on the basis that disease-causing genetic variation is likely to affect common pathways that may be revealed by PPI.²⁷ Based on these interactions, DAPPLE forms networks of physical protein-protein connectivity where proteins are nodes connected by edges that represent interactions in the InWeb database. Here gene lists were provided as input. In DAPPLE, protein products of genes are scored based on their participation in direct or indirect networks. These scores are Bonferroni corrected for two tests if a protein participates in both direct and indirect networks (P_{corr}),²⁷ and the best score is assigned.

2.7 | Pathway-level analysis

Pathway-level analyses were performed using the set-based test implemented in PLINK.²¹ This test is a self-contained test which uses raw genotypes as input data; it calculates the average of test statistics as the pathway enrichment scores, using independent and significant (by preselected P value cutoff) SNPs in the pathway.²⁶ Here we considered as significant SNPs with empirical P value (P_{EMP}) ≤ 0.05 in the case-control and/or the family-based association test. For each pathway, independent SNPs are first identified ($r^2 < .5$), and from these an average statistic is calculated. The statistical significance of a pathway is computed using permutation, thereby efficiently correcting by the number of SNPs and the LD structure within the pathway. In order to account for the number of pathways tested, the FDR method was used.

In the present study, we used the reference biological pathway annotation databases KEGG,^{28,29} which is a collection of manually

curated pathway maps, and Atlas of Cancer Signaling Network (ACSN)³⁰ which describes tailored maps of molecular processes involved in cancer, to define the gene sets involving at least five genes.

With KEGG definitions (as of July 2018), a total of 319 curated biological pathways were tested. These pathways are organized into six maps (metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems and Human diseases) and 48 subgroups. For ACSN, we used version ACSN2.0 (release 2018) and tested 121 cancer modules (pathways). These modules are organized into 10 maps (adaptive immune response, angiogenesis, cell cycle and DNA repair, cell survival, EMT and cell senescence, fibroblasts, innate immune response, invasion and motility, regulated cell death, telomere maintenance).

2.8 | Pathway-derived PRS calculation and performance

To build each pathway-derived PRS, we considered all pathways with $P_{\text{EMP}} \leq .05$ in the case-control analysis and first selected the SNPs contributing to the associated pathway based on results of the PLINK set-based test. Then to create global PRS for KEGG (PRS_{KEGG}) and ACSN (PRS_{ACSN}), we combined the SNPs from the different selected pathways and applied the LD-driven clumping procedure from PLINK to exclude SNPs in strong LD ($r^2 \geq .8$). Pathway-derived PRS were calculated for each individual with the PLINK *-score* command using the following equation: $\text{PRS}_i = \sum_{n=1}^k \ln(\text{OR}_n) * C_{i,n}$, where i represents the individual whose score is calculated by summing over all SNPs n in the pathway ranging from the first SNP 1 to the last SNP k ; OR_n is the odds ratio of the risk allele for SNP _{n} obtained in the GENESIS case-control data set, and $C_{i,n}$ is the individual's count of risk alleles for SNP _{n} (0, 1 or 2). A higher PRS corresponds with having more risk alleles and thus, a higher amount of genetic risk for BC.

For each pathway-derived PRS, the ability of the model to discriminate between case and control individuals was evaluated by ROC curves, representing the sensitivity as a function of 1-specificity, using the R package "pROC," and the correlation between variables by Pearson's coefficient. The area under the receiver-operator curves (AUC), which is the probability that the predicted risk is higher for a case individual than for a control individual and ranges from 0.5 (equivalent to a coin toss) to 1.0 (perfect discrimination) was calculated for the different data sets.

3 | RESULTS

In the standard GENESIS case-control analyses, only index cases and unrelated controls were used, while all genotyped women, affected and unaffected, were used in the family-based analyses. The genomic control inflation factor³¹ which tests for population stratification, was close to 1 indicating the absence of population stratification in our data set (data not shown).

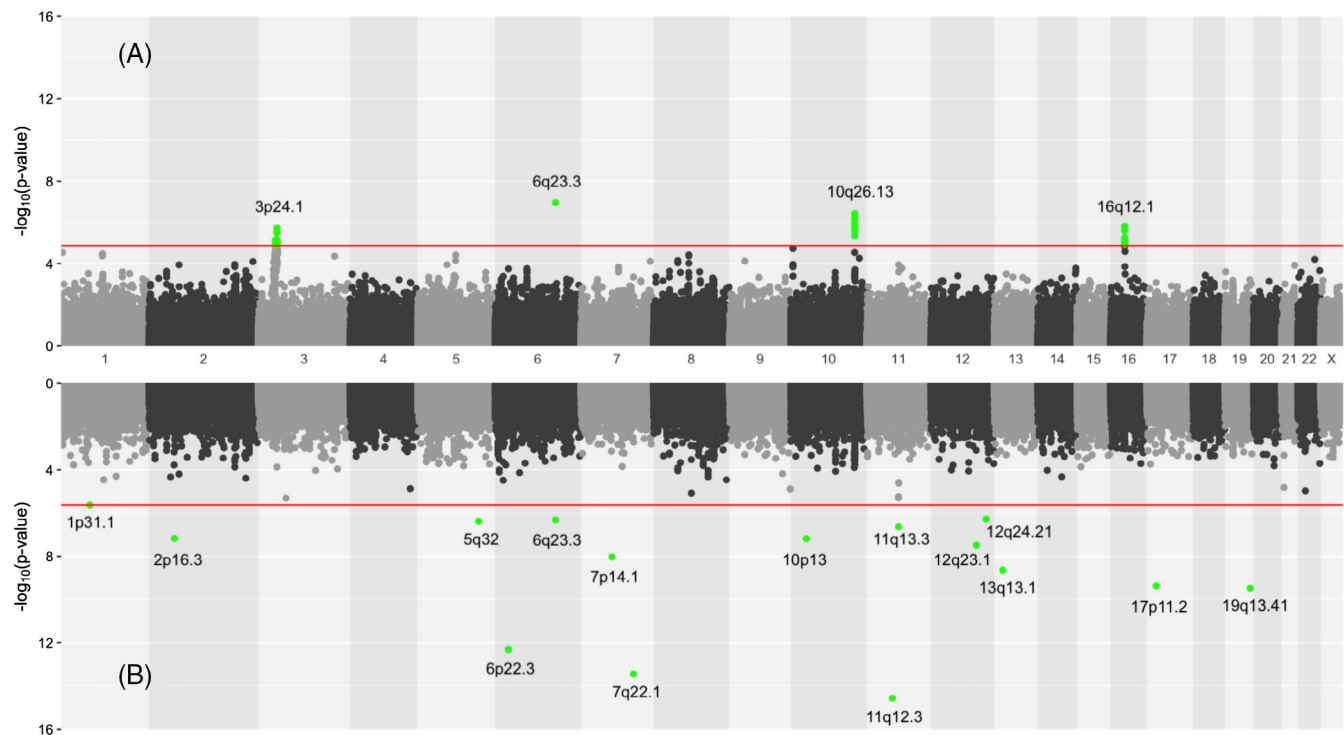


FIGURE 1 Miami plot of single-nucleotide polymorphism (SNP) association with breast cancer. A, Results of the case-control analysis. B, Results of the family-based association test. $-\log_{10} P$ values for SNP associations are plotted against the genomic coordinates (hg19). The red lines indicate the 10^{-5} threshold. Green points denote SNPs showing suggestive association with breast cancer

3.1 | SNP-level analysis

In the case-control analysis, no SNP reached the standard genome-wide significance P value threshold of 5×10^{-8} (Figure 1). However, after correction for multiple testing, SNPs at loci 3p24.1 (*NEK10/SLC4A7*), 6q23.3 (*ARFGEF3*) 10q26.13 (*FGFR2*) and 16q12.1 (*TOX3/CASC16*) were associated with BC risk with $P_{FDR} < .05$. Among these four loci, 3p24.1, 10q26.13 and 16q12.1 had been identified in the large-scale GWAS conducted by the BCAC¹⁹ while locus 6q23.3 was new. Results of the association test for the top SNP at each associated locus are presented in Table 1. This table also shows results of the family-based analysis. This latter analysis confirmed association with the new locus at 6q23.3, while the signal was not significant after correction for multiple testing at 3p24.1, 10q26.13 and 16q12.1 ($P_{FDR} < .05$). In addition, the family-based association test further identified significant SNPs at 14 loci for which mainly suggestive association was found in the case-control analysis (Figure 1). Among those, two SNPs were located within loci 11q13.3 and 12q24.21 that had been previously identified by the BCAC and top SNPs at 13 loci had MAF lower than 0.03 in GENESIS controls (Table S2). Summary statistics for the 72 SNPs with $P_{FDR} \leq .05$ in the standard case-control analysis or in the family-based analysis are provided in Table S3. We found that 17 genes located in 14 of the top 18 loci were probably biologically connected as the PPI networks formed by genes tagged by these SNPs had significant direct and/or indirect connectivity ($P_{corr} < .05$, based on 1000 network resampling) (Figure 2A and Table S4).

3.2 | Gene-level analysis

To better characterize the molecular and cellular mechanisms involved in the pathogenesis of BC, we next focused on the subset of iCOGS SNPs tagging 32 444 genes, coding RNA, pseudogenes, miRNA or lncRNA (Table S5). The gene-level analysis identified 112 genes with $P_{GENE} \leq .001$ either in the case-control analysis or in the family-based analysis (Table S6). Among the top 112 genes, only 8 are located in a region highlighted previously in the SNP-level analysis (Table S4), demonstrating the advantage of the gene-level analysis to highlight new candidates. Of the 112 genes, 30 are directly or indirectly connected in a PPI network (Figure 2B). To get a more general overview of the interconnections between genes identified in the gene-level analysis and candidate genes at loci identified in the SNP-level analysis, we also constructed a PPI network combining results obtained with the two approaches. The final network including 41 genes is shown in Figure 2C and the DAPPLE score for each gene, reflecting its participation in the network (P_{corr}) is provided in Table S4.

Furthermore, we interrogated whether iCOGS SNPs in or nearby the 112 top ranked genes were acting as cis-eQTLs using independent mRNA expression data from 1092 breast invasive carcinomas from the Cancer Genome Atlas available through PanCanQTL project.³² We found cis-eQTLs for *AKNAD1* (1p13.3), *GPSM2* (1p13.3), *NEK10* (3p24.1), *SLC4A7* (3p24.1), *CDC25A* (3p21.31), *ADCY5* (3q21.1), *ALDH5A1* (6p22.3), *HSPA14* (10p13), *PLCE1* (10q23.33), *UCP3* (11q13.4), *TOX3* (16q12.1), *MAP2K3* (17p11.2) and *FBXO7* (22q12.3),

TABLE 1 Breast cancer associated loci identified in the SNP-based analysis

Locus	Gene or region containing the top SNP ^a	# sigSNPs ^b (CC/Fam.)	Best SNP	Nucleotide change ^c (strand)	Effect allele frequency in controls	Case-control analysis		Family-based analysis	
						OR ^d (95% CI)	P value	P _{FDR}	P _{FDR}
1p31.1	ADGRL4 (=LTD1)	0/1	rs17102586	T > C (+)	0.03	1.55 (1.14, 2.12)	6.0 × 10 ⁻³	.67	2.3 × 10 ⁻⁶
2p16.3	MSH6, FBXO11, RPL36AP15	0/1	rs2020912	A > G (-)	0.007	2.20 (1.20, 4.03)	1.0 × 10 ⁻²	.75	6.7 × 10 ⁻⁸
3p24.1	NEK10, SLC4A7	29/0	rs9828914	C > G (+)	0.36	0.73 (0.65, 0.83)	1.9 × 10 ⁻⁶	.03	1.4 × 10 ⁻⁴
5q32	JAKMIP2, SPINK1	0/1	rs7735394	A > C (-)	0.002	3.80 (1.22, 11.9)	2.0 × 10 ⁻²	.80	4.2 × 10 ⁻⁷
6p22.3	ALDH5A1, KIAA0319	0/1	rs7764860	A > G (-)	0.01	2.28 (1.48, 3.51)	1.8 × 10 ⁻⁴	.25	4.9 × 10 ⁻¹³
6q23.3	ARFGEF3 (=KIAA1244)	1/1	rs203136	A > C (-)	0.33	1.38 (1.23, 1.56)	1.1 × 10 ⁻⁷	.02	4.8 × 10 ⁻⁷
7p14.1	ELMO1	0/1	rs17170951	T > C (-)	0.003	1.94 (0.79, 4.80)	1.5 × 10 ⁻¹	.94	9.5 × 10 ⁻⁹
7q22.1	CYP3A7, CYP3A7-CYP3A51P, CYP3A4	0/1	rs2687117	T > C (+)	0.002	6.70 (2.32, 19.4)	4.5 × 10 ⁻⁴	.40	3.6 × 10 ⁻¹⁴
10p13	SUV39H2, DCLRE1C	0/1	rs1062884	G > T (+)	0.0004	13.1 (1.65, 104)	1.5 × 10 ⁻²	.77	6.5 × 10 ⁻⁸
10q26.13	FGFR2	15/0	rs2981579	A > G (+)	0.44	1.35 (1.20, 1.52)	3.9 × 10 ⁻⁷	.03	1.8 × 10 ⁻⁴
11q12.3	Intergenic ^e (FTH1, INCENP)	0/1	rs1024123	A > G (+)	0.0004	24.2 (3.24, 180)	2.0 × 10 ⁻³	.54	2.7 × 10 ⁻¹⁵
11q13.3	Intergenic ^e (MYEOV, CCND1)	0/1	rs662169	A > G (+)	0.12	1.39 (1.18, 1.65)	1.1 × 10 ⁻⁴	.18	2.3 × 10 ⁻⁷
12q23.1	NR1H4	0/1	rs11110398	A > G (-)	0.0004	17.2 (2.17, 135)	7.0 × 10 ⁻³	.70	3.3 × 10 ⁻⁸
12q24.21	Intergenic ^e (TBX3, MED13L)	0/1	rs74710455	A > G (-)	0.003	3.68 (1.53, 8.81)	4.0 × 10 ⁻³	.62	5.3 × 10 ⁻⁷
13q13.1	PDS5B	0/1	rs17077706	A > G (-)	0	—	—	—	2.3 × 10 ⁻⁹
16q12.1	TOX3, CASC16	13/0	rs45465998	T > C (+)	0.25	1.38 (1.21, 1.57)	1.5 × 10 ⁻⁶	.03	9.4 × 10 ⁻⁴
17p11.2	MAP2K3	0/1	rs2885765	A > G (+)	0.01	12.1 (2.80, 52)	8.3 × 10 ⁻⁴	.51	4.2 × 10 ⁻¹⁰
19q13.41	SIGLEC22P, CD33, SIGLEC11, LINC01872	0/1	rs117239811	G > C (-)	0	—	—	—	3.3 × 10 ⁻¹⁰

Abbreviations: CI, confidence interval; FDR, false-positive discovery rate; MAF, minor allele frequency; OR, odds ratio; SNP, single-nucleotide polymorphism.

^aAccording to GENCODE definition. A SNP is linked to a gene if it is located within 50 kb on either side of the gene's transcribed region.

^bSNPs with a P_{FDR} ≤ .05 in the case-control (CC) analysis and in the family-based (Fam.) analysis.

^cEffect allele is underlined.

^dOdds ratio of the logistic regression when adjusting for age at diagnosis for cases and age at interview for controls.

^eThe closest genes on either side of the top SNP are indicated in brackets.

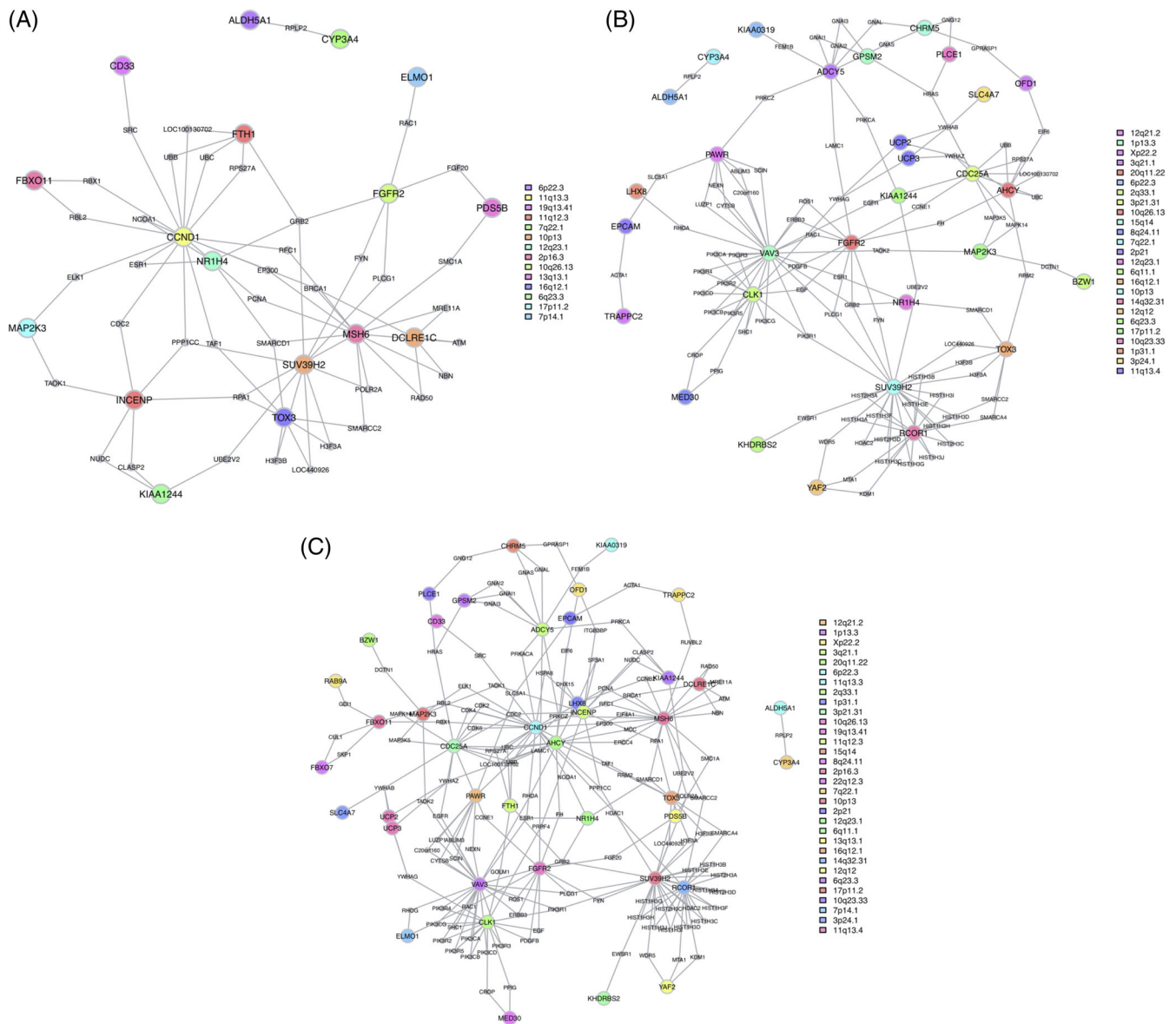


FIGURE 2 Physical interactions among proteins encoded by genes associated with breast cancer or genes in the associated intervals. A, Protein-protein interaction (PPI) network obtained with genes located within the 18 loci from the single-nucleotide polymorphism (SNP)-level analysis. B, PPI network obtained with the 112 top genes from the gene-level analysis. C, PPI network obtained with the input gene list combining input lists from Figures 2A and 2B

indicating that causal SNPs at the associated loci could alter the regulation of the expression of these 13 genes that therefore represent good candidates to prioritize for further functional biological studies (Table S7).

3.3 | Pathway-level analysis

To discover novel sets of variants with related functions which could help explain the observed data, we performed pathway-level analyses using summary association statistics from the single-SNP analysis. The reference biological pathway annotation databases KEGG^{28,29} and ACSN³⁰ were used to define the gene sets.

In the case-control analysis, 31 KEGG pathways were identified with $P_{EMP} \leq .05$ and 4 with $P_{FDR} \leq .05$ out of the 319 pathways tagged by iCOGS SNPs (Table 2). Top pathways were involved in endocytosis, signaling pathways regulating pluripotency of stem cells, regulation of actin cytoskeleton, cell growth/death (p53 signaling pathway, apoptosis), and pathways altered in prostate and gastric cancers. Using ACSN2.0 annotation, suggestive association was found for 13 out of 121 tested modules ($P_{EMP} \leq .05$) but no modules were significantly associated with BC after FDR correction (Table 3). Top ACSN modules were all involved in cell survival (WNT noncanonical, PI3K/AKT/MTOR, MAPK, extracellular matrix).

With the family-based association test, corresponding numbers were 63 KEGG pathways with $P_{EMP} \leq .05$ (of which 29 with

TABLE 2 KEGG pathways associated with breast cancer susceptibility, with empirical P value (P_{EMP}) $\leq .05$ in the case-control study^a

KEGG group	KEGG ID	Pathway definition	Genes from the 112 top genes list involved in the pathway	#SNP ^b	#sigSNP ^c	#Gene ^d	P_{EMP}	P_{FDR}	#sigGene ^e
Cell growth death	hsa04115	p53 signaling pathway		687	29	15	4.00×10^{-4}	.04	3
Cell growth death	hsa04210	Apoptosis		1153	63	38	1.20×10^{-3}	.06	9
Cell growth death	hsa04215	Apoptosis multiple species		334	13	8	4.70×10^{-3}	.15	1
Cell growth death	hsa04217	Necroptosis		935	45	32	4.40×10^{-3}	.15	7
Cell motility	hsa04810	Regulation of actin cytoskeleton (–)	CHRM5, FGFR2, VAV3	1707	130	63	1.10×10^{-3}	.06	16
Cellular community eukaryotes	hsa04550	Signaling pathways regulating pluripotency of stem cells	FGFR2	1189	95	34	4.00×10^{-4}	.04	8
Transport catabolism	hsa04144	Endocytosis (–)	FGFR2	1826	140	55	3.00×10^{-4}	.04	19
Cancers	hsa05200	Pathways in cancer	ADCY5, FGFR2	5553	334	71	1.28×10^{-2}	.29	31
Cancers	hsa05215	Prostate cancer (–)	FGFR2	1391	82	29	5.00×10^{-4}	.04	5
Cancers	hsa05226	Gastric cancer (–)	FGFR2	1766	119	52	1.60×10^{-3}	.07	16
Cancers	hsa05230	Central carbon metabolism in cancer (–)	FGFR2	686	63	21	3.20×10^{-3}	.13	3
Drug resistance	hsa01521	EGFR tyrosine kinase inhibitor resistance (–)	FGFR2	1016	73	27	9.00×10^{-3}	.24	5
Drug resistance	hsa01524	Platinum drug resistance		644	31	19	1.40×10^{-2}	.3	2
Endocrine metabolic diseases	hsa04932	Nonalcoholic fatty liver disease (NAFLD)		890	50	36	2.70×10^{-2}	.39	9
Infectious diseases	hsa05134	Legionellosis	CLK1	320	28	15	4.84×10^{-2}	.51	7
Infectious diseases	hsa05145	Toxoplasmosis	MAP2K3	975	59	44	1.68×10^{-2}	.33	9
Infectious diseases	hsa05164	Influenza A	MAP2K3	1186	80	54	2.90×10^{-2}	.39	10
Infectious diseases	hsa05165	Human papillomavirus infection		3298	185	66	3.16×10^{-2}	.39	18
Infectious diseases	hsa05168	Herpes simplex infection		1079	72	52	4.97×10^{-2}	.51	9
Infectious diseases	hsa05169	Epstein-Barr virus infection (–)	MAP2K3	1423	83	60	2.99×10^{-2}	.39	15
Neurodegenerative diseases	hsa05016	Huntington's disease (–)	RCOR1	1223	50	36	2.30×10^{-2}	.39	7
Signal transduction	hsa04010	MAPK signaling pathway (–)	FGFR2, MAP2K3	2760	171	72	7.70×10^{-3}	.22	17
Signal transduction	hsa04014	Ras signaling pathway (–)	FGFR2	2229	144	71	9.90×10^{-3}	.24	12
Signal transduction	hsa04151	PI3K-Akt signaling pathway (–)	FGFR2	3726	224	63	3.17×10^{-2}	.39	15
Signal transduction	hsa04340	Hedgehog signaling pathway		378	10	9	2.91×10^{-2}	.39	2
Signal transduction	hsa04668	TNF signaling pathway	MAP2K3	1033	46	30	2.97×10^{-2}	.39	4
Glycan biosynthesis	hsa00601	Glycosphingolipid biosynthesis lacto and neolacto series		175	9	6	4.22×10^{-2}	.48	2
Glycan biosynthesis	hsa00603	Glycosphingolipid biosynthesis globo and isoglobo series		118	6	4	4.72×10^{-2}	.51	1
Lipid metabolism	hsa00062	Fatty acid elongation		132	6	4	4.17×10^{-2}	.48	2
Immune system	hsa04622	RIG-I-like receptor signaling pathway		389	25	15	2.16×10^{-2}	.39	7
Immune system	hsa04657	IL-17 signaling pathway		588	27	20	2.85×10^{-2}	.39	6

Note: The corrected P values (P_{FDR}) are also provided. (–) indicates that the pathway is not significant anymore in the case-control analysis after excluding SNPs in the 112 top genes from the VEGAS2 analysis.

Abbreviations: FDR, false-positive discovery rate; KEGG, Kyoto Encyclopedia of Genes and Genomes; SNP, single-nucleotide polymorphism.

^aResults of the gene set analysis performed with PLINK²¹ using SNP P values obtained in the GENESIS case-control set. Reported P values are adjusted for age at diagnosis for cases and age at inclusion for controls.

^bNumber of iCOGS SNPs in the pathway.

^cNumber of SNPs contributing to the pathway with $P \leq .05$ in the SNP-level analysis.

^dNumber of genes linked to the contributing SNPs in the pathway.

^eNumber of genes with $P \leq .05$ in the gene-level analysis.

TABLE 3 ACSN pathways associated with breast cancer susceptibility, with empirical P value (P_{EMP}) $\leq .05$ in the case-control study^a

Cancer hallmark	Map	Pathway definition	Genes from the 112 top genes list involved in the pathway	#SNP ^b	#sigSNP ^c	#Gene ^d	P_{EMP}	P_{FDR}	#sigGene ^e
Activating invasion and metastasis	EMT senescence	EMT regulators (–)	FGFR2	5628	358	73	1.10×10^{-2}	.22	38
Activating invasion and metastasis	Cell survival; EMT senescence	ECM (–)	FGFR2	1920	148	47	2.00×10^{-3}	.06	9
Activating invasion and metastasis	Invasion motility	Invasion motility (–)	FGFR2	1640	116	70	2.70×10^{-2}	.30	20
Avoiding immune destruction	Innate immune response	Markers NK		49	1	1	1.90×10^{-2}	.26	0
Evading growth suppressors	Cell survival	WNT noncanonical	ADCY5, FGFR2	3669	238	70	9.99×10^{-4}	.06	34
Evading growth suppressors	Cell survival	MAPK (–)	FGFR2, MAP2K3	2195	136	60	2.00×10^{-3}	.06	11
Evading growth suppressors	Cell survival	PI3K AKT MTOR (–)	FGFR2	2738	167	58	9.99×10^{-4}	.06	17
Genome instability and mutation	Cell cycle and DNA repair	NER		366	24	15	4.00×10^{-2}	.37	4
Resisting cell death	Regulated cell death	Caspases	MAP2K3	1254	61	38	1.30×10^{-2}	.22	9
Resisting cell death	Regulated cell death	TRAIL response		184	20	10	1.20×10^{-2}	.22	2
Resisting cell death	Regulated cell death	FAS response		130	11	7	1.60×10^{-2}	.24	2
Tumor promoting inflammation	Fibroblasts	Matrix regulation		215	9	6	2.30×10^{-2}	.28	0
Tumor promoting inflammation	Adaptive immune response	TCR signaling (–)	MAP2K3, VAV3	1655	90	56	3.10×10^{-2}	.31	13

Note: The corrected P values (P_{FDR}) are also provided. (–) indicates that the pathway is not significant anymore in the case-control analysis after excluding the 112 top genes from the VEGAS2 analysis. Abbreviations: ACSN, Atlas of Cancer Signaling Network; ECM, extracellular matrix; EMT, epithelial mesenchymal transition; FDR, false-positive discovery rate; NER, nucleotide excision repair; SNP, single-nucleotide polymorphism.

^aResults of the gene set analysis performed with PLINK (Purcell et al., 2007) using SNP P -values obtained in the GENESIS case-control set. Reported P -values are adjusted for age at diagnosis for cases and age at inclusion for controls.

^bNumber of iCOGS SNPs in the pathway.

^cNumber of SNPs contributing to the pathway with $P \leq .05$ in the SNP-level analysis.

^dNumber of genes linked to the contributing SNPs in the pathway.

^eNumber of genes with $P \leq .05$ in the gene-level analysis.

$P_{FDR} \leq .05$), and 23 ACSN modules with $P_{EMP} \leq .05$ (of which 7 with $P_{FDR} \leq .05$). Results of these analyses are shown in Tables S8 and S9.

When reiterating the case-control and family-based association tests after excluding SNPs tagging the top 112 genes from the gene-level analysis, we found that association signals for a number of KEGG and ACSN pathways were driven by genes *FGFR2*, *MAP2K3*, *ADCY5* and *CYP3A4* (Tables 2, 3, S8 and S9) which are part of the above described 41-gene PPI network (Figure 2C). Hence, the pathway-level approach support findings of the SNP- and gene-level analyses and further identified new sets of functionally related genes pathways, such as genes involved in the KEGG definitions “p53 signaling pathway,” “apoptosis,” and “platinum drug resistance” and genes involved in ACSN modules of the innate immune response (“markers of the myeloid-derived suppressor cells [MDSC]” and “antigen presentation” modules), opening up new venues to explore in experimental studies.

3.4 | Pathway-derived PRS

The combined effect of SNPs related to genes in identified pathways was expressed as pathway-derived PRS. These PRS were built using summary statistics from the SNP-based analysis conducted in the GENESIS “index case-control” data set. A total of 672 SNPs linked to the 31 top KEGG pathways and 473 SNPs linked to the 13 top ACSN modules ($P_{EMP} \leq .05$) were used to build two PRS, named $PRS_{KEGG-672}$ and $PRS_{ACSN-473}$, respectively. We also built a PRS by restricting the SNP selection to the 211 SNPs linked to the 4 KEGG-associated pathways with $P_{FDR} \leq .05$ ($PRS_{KEGG-211}$) and for comparison, a 4-SNP-derived PRS (PRS_{4-SNPs}) corresponding to a polygene including only the four SNPs associated with BC ($P_{FDR} \leq .05$) in the classical single-SNP-level analysis conducted in the “index case-control” set (Table 1). The complete list of SNPs used for each PRS is provided in Table S10.

Association of these PRS with BC and their performance were assessed in two validation sets: the CECILE population (set I), which includes 1019 BC cases and 999 controls also genotyped with the iCOGS array,^{18,33} and set II which includes the affected sisters of 731 GENESIS index cases and the 999 CECILE controls. Because affected sisters in set II are not genetically independent from the cases of our discovery set (GENESIS index cases), we first evaluated the degree of correlation between the $PRS_{KEGG-672}$ of the siblings. We found that the Pearson correlation between sisters was 0.46 when considering the 675 affected sib pairs for whom genotyping data were available for both the index case and the sister, and 0.48 when considering the 448 sib pairs for whom genotyping data was available for both the index case and an unaffected sister. This suggests that the PRS correlation between two sisters is independent from BC status.

Table 4 shows the associations between $PRS_{KEGG-672}$, $PRS_{KEGG-211}$, $PRS_{ACSN-473}$ and PRS_{4-SNPs} quintiles in the different validation sets. In set I, women with a $PRS_{KEGG-211}$ in the highest quintile had a significant increased risk of BC as compared to women in the middle quintile used as reference (OR = 1.33). This risk was even higher when restricting the

analysis to CECILE cases with at least one first-degree relative affected with BC (OR = 1.84).

In set II, for each of the tested PRS we observed that women in the lowest quintile had a reduced risk of BC (OR from 0.52 to 0.68), and those in the highest quintile had an increased risk of BC (OR from 1.82 to 2.47) as compared to women in the middle quintile.

Overall, we found that these PRS had very little discriminative capacity within CECILE (AUC ranging from 0.53 to 0.55), but they performed slightly better to discriminate BC cases with a first degree relative affected with BC (AUC ranging from 0.55 to 0.58). Interestingly, we found that performance of the pathway-derived PRS was improved in Set II, with $PRS_{KEGG-672}$ representing the best predictor of BC risk (AUC = 0.66, 95% confidence interval [CI]: 0.63, 0.68; Table 4).

Moreover, the difference in BC risk between women in the lowest quintile and women in the highest quintile was bigger for $PRS_{KEGG-672}$ and $PRS_{ACSN-473}$ than for PRS_{4-SNPs} in set II, showing that our system biology-based strategy to identify genes and SNPs to prioritize leads to relevant SNP selections in the high-risk population. Overall, women in the highest quintile of each pathway-derived PRS were at higher risk of BC than women in the highest quintile of PRS_{4-SNPs} (Table 4).

4 | DISCUSSION

For a better understanding of the genetic basis underlying familial BC unexplained by *BRCA1/2* pathogenic variants, we performed data mining of GENESIS GWAS data using prior biological knowledge on gene function, under the assumption that BC in high-risk families could be caused by the joint effects of alterations in multiple functionally related genes.^{34–36} Since pathway-based methods strongly reduce the number of association tests, such approaches may substantially increase the power to identify new genetic variation compared to the classical GWAS approach where a large number of markers are individually tested for association and stringent significance thresholds are applied.³⁷ However, an important limitation of employing a gene- or pathway-based approach is the omission of intergenic regions. In the present study, we assigned variants that lie within 50 kb on either side of a gene's coding sequence boundaries to compute its association *P* value. With this gene definition, 17.9% of the iCOGS SNP were not linked to a gene. This choice might have therefore ignored distantly located risk variants associated to key genes; however, we chose to use this SNP selection criterion to strike a balance between inclusions of possible *cis*-regulatory variants and maintaining specificity of a gene.

One strength of the GENESIS population is that in addition to the index cases of *BRCA1/2* negative families, affected and unaffected sisters had been also genotyped allowing to apply beyond to a classical case-control study design, a family-based GWAS approach which could be a more potent way of identifying rare variants involved in BC susceptibility.^{14,38} Indeed, among the 18 identified loci, 14 of them were found at $P_{FDR} < .05$ only in the family-based approach (Table 1),

TABLE 4 Association of pathway-derived polygenic risk scores with breast cancer in the validation sets

Set I										Set II					
		All CECILE cases (N = 1019) (%)				CECILE cases with family history of BC ^b (N = 176) (%)				GENESIS affected sisters (N = 731) (%)					
Quintile	Range	CECILE controls (N = 999) (%)	OR ^a	95% CI	P value	OR ^a	95% CI	P value	OR ^a	95% CI	P value	OR ^a	95% CI	P value	
PRS _{KEGG-672}															
Q1	≤ -0.0026	200 (0.2)	0.84	0.63, 1.12	.23	0.78	0.45, 1.36	.39	0.56	0.39, 0.82	.002				
Q2	-0.0026 to -0.0012	200 (0.2)	0.97	0.73, 1.28	.82	0.81	0.47, 1.41	.46	0.75	0.52, 1.06	.10				
Q3	-0.0012 to 0	200 (0.2)	Ref	—	—	Ref	—	—	Ref	—	—				
Q4	0 to 0.0015	200 (0.2)	1.14	0.87, 1.50	.34	1.28	0.77, 2.12	.34	1.46	1.07, 2.00	.02				
Q5	> 0.0015	199 (0.2)	1.25	0.95, 1.64	.11	1.63	1.01, 2.64	.05	2.32	1.71, 3.13	<.001				
AUC (95% CI): 0.54 (0.52, 0.57)															
AUC (95% CI): 0.58 (0.54, 0.63)															
PRS _{KEGG-211}															
Q1	≤ -0.0063	200 (0.2)	0.89	0.67, 1.18	.42	0.83	0.47, 1.47	.52	0.68	0.48-0.95	.03				
Q2	-0.0063 to -0.0036	200 (0.2)	0.83	0.63, 1.11	.21	1.10	0.64, 1.89	.73	0.72	0.52-1.02	.06				
Q3	-0.0036 to -0.0011	200 (0.2)	Ref	—	—	Ref	—	—	Ref	—	—				
Q4	-0.0011 to 0.0018	200 (0.2)	1.16	0.88, 1.52	.30	1.31	0.78, 2.21	.31	1.19	0.87-1.62	.28				
Q5	> 0.0018	199 (0.2)	1.33	1.02, 1.75	.04	1.84	1.13, 3.02	.02	1.82	1.35-2.46	<.001				
AUC (95% CI): 0.55 (0.52, 0.57)															
AUC (95% CI): 0.57 (0.52, 0.62)															
PRS _{ACSN-473}															
Q1	≤ -0.0041	200 (0.2)	0.90	0.68, 1.20	.47	0.81	0.47, 1.42	.47	0.52	0.35, 0.76	.001				
Q2	-0.0041 to -0.0022	200 (0.2)	1.16	0.88, 1.52	.31	1.03	0.61, 1.75	.90	0.93	0.66, 1.31	.67				
Q3	-0.0022 to -8e-04	200 (0.2)	Ref	—	—	Ref	—	—	Ref	—	—				
Q4	-8e-04 to 9e-04	200 (0.2)	1.05	0.80, 1.39	.72	1.25	0.76, 2.08	.38	1.61	1.17, 2.20	.003				
Q5	> 9e-04	199 (0.2)	1.32	1.00, 1.73	.05	1.42	0.86, 2.32	.17	2.47	1.82, 3.34	<.001				
AUC (95% CI): 0.53 (0.51, 0.56)															
AUC (95% CI): 0.56 (0.40, 0.63)															
PRS _{4-SNPs^c}															
Q1	≤ 6e-04	206 (0.21)	0.74	0.55, 1.00	.05	0.57	0.32, 1.00	.05	0.59	0.42, 0.83	.003				
Q2	6e-04 to 0.0392	233 (0.23)	1.06	0.80, 1.40	.70	0.80	0.48, 1.32	.37	0.82	0.60, 1.13	.23				
Q3	0.0392-0.0767	166 (0.17)	Ref	—	—	Ref	—	—	Ref	—	—				
Q4	0.0767-0.0903	195 (0.20)	0.99	0.74, 1.32	.94	0.95	0.57, 1.58	.85	1.08	0.79, 1.49	.62				
Q5	≥0.0903	199 (0.20)	1.13	0.85, 1.50	.40	1.03	0.63, 1.70	.90	1.56	1.16, 2.12	.004				
AUC (95% CI): 0.53 (0.50, 0.56)															
AUC (95% CI): 0.55 (0.51, 0.59)															
AUC (95% CI): 0.59 (0.56, 0.62)															

Abbreviations: AUC, area under the receiver-operator curve; CI, confidence interval; FDR, false-positive discovery rate; OR, odds ratio; PRS, polygenic risk scores; SNP, single-nucleotide polymorphism.
^aAdjusted for age at diagnosis for cases and age at interview for controls.
^bCECILE cases with at least one first-degree relative affected with breast cancer at inclusion.
^cPRS built using SNPs with $P_{FDR} \leq .05$ in the case-control analysis.

and remarkably, risk alleles at these loci were quite rare in our control population and were associated with a relatively high size effect ($OR > 2$) in the case-control analysis. Moreover, in the family-based association test, 6 of the 18 associated loci contain genes found to be associated with BC risk in the gene-level analysis (*SPINK1* at 5q32, *ALDH5A1* and *KIAA0319* at 6p22.3, *CYP3A7*, *CYP3A51P* and *CYP3A4* at 7q22.1, *NR1H4* at 12q23.1, *MAP2K3* at 17p11.2), supporting that this approach can help identifying rare risk alleles for familial BC that could be missed applying a classical case-control association study design. However, we acknowledge that the new associations obtained in the family-based analyses only should be interpreted with caution as no additional set with genotyping data was available to replicate them. Moreover, the top SNPs at these associated loci have a $MAF < 0.01$ in the control population, and six of them were not reported in the BCAC meta-analysis.¹⁹ Conversely, the family-based approach failed to identify common SNPs at the well-known BC susceptibility 3p24.1 (*NEK10*, *SLC4A7*), 10q26 (*FGFR2*) and 16q12.1 (*TOX3*, *CASC16*).

The gene-level analyses identified additional signals among the several loci that demonstrated suggestive but nonsignificant association peaks in our single-SNP analyses, but for which no individual variant had achieved significance. Indeed, among the top 112 genes, 91 were located within 17 new loci (Table S4). Although none of the proteins encoded by the 112 top genes had known experimentally validated direct biological connections, 30 second-order neighbors were identified, that is, two proteins from the input were connected to each other via a common interactor protein (Figure 2B). Furthermore, the final PPI network built with proteins encoded by genes at known or novel potential BC susceptibility loci involved 41 proteins with indirect connectivity (Figure 2C). This suggests that although proteins encoded by genes in the associated intervals do not interact directly with each other, they may represent converging hubs of BC-relevant protein networks.

The limitation to using PPI data or pathway data from curated databases such as KEGG and ACSN is that proteins for which no high-confidence interactions exist will be left out of the analysis. As such, our analysis is limited to proteins present in the databases. On the other hand, pathway-level analyses using such databases allow to confidently highlight relevant biological pathways and may help to identify the best candidate in these pathways for therapeutic intervention. For instance, targeting p53 signaling pathway and apoptosis pathway as described in the KEGG “cell growth death” group or in modules “WNT noncanonical,” “MAPK” and “EMT regulators” from the ACSN cell survival map might also have clinical implications for finding additional drug targets. Similarly, gene products involved in the “Toll-like receptor signaling pathway” and “chemokine signaling pathway” from the KEGG immune system group or in ACSN modules “TH1” (adaptive immune response map), “Antigen presentation” and “markers MDSC” (innate immune response map) may be good candidates to target.

In addition to the identification of potential drug targets, these observations may also be used for prevention. Under the assumption that proteins interacting with multiple associated pathway

members and encoded elsewhere in the genome themselves carry an excess of association to BC, we built weighted pathway-derived PRS and explored the potential for using them as predictors for BC in the general population and in a population with familial predisposition. We found that each of the pathway-derived PRS had very little discriminative capacity within the general population, which may be due to overfitting of the model. This could be explained by the number of pathways (and of SNPs) which are considered to build the PRS. PRS_{KEGG-672} and PRS_{ACSN-473} were built using 672 and 473 SNPs, respectively, considering SNPs of pathways associated with BC with $P_{EMP} \leq .05$, while PRS_{KEGG-211} was built restricting the number of pathways to those after applying FDR correction ($P_{FDR} \leq .05$). However, the three pathway-derived PRS performed better in CECILE than the 4-SNPs PRS constructed using significant SNPs of the single-marker analysis, and they also performed better to discriminate affected women with a family history of BC. Besides, the performance of our pathway-derived PRS in the high-risk GENESIS population is close to that of the PRS recently published by BCAC based on 313 BC associated SNPs developed on a data set comprising over 170 000 subjects of European ancestry from 69 studies.³⁹

GENESIS cases are from HBOC families who received genetic counseling and who were tested negative for *BRCA1/2* mutations. Despite ascertainment of GENESIS families, investigated cases were not specifically early onset cases (mean age at BC diagnosis was 50.6 years in GENESIS index cases and 54.4 years in CECILE cases; Table S1). Eighty-four percent of GENESIS index cases with verified pathology data and 85% of CECILE cases have developed estrogen receptor positive (ER+) breast tumors. Therefore, the GENESIS population has a tumor type's distribution more comparable to that of the general population than has the *BRCA1/2* carriers' population (*BRCA1* mutation carriers developing mainly ER– tumors). In order to get as much power as possible and because cases with an ER– tumor were few, we chose to build pathway-derived PRS for our entire population. It is also important to note that only six SNPs of our pathway-derived PRS are included or are strongly correlated with a SNP of the 313-SNP PRS developed by Mavaddat et al.³⁹ (Table S10). Hence our strategy to select SNPs to be included in PRS for prediction of BC might pave the way for future research in subpopulations for which the classical approach will never be powerful enough.

To conclude with, our findings also further underline the need for developing new strategies to analyze family-based genetic data, as well as methodological approaches to identify altered biological mechanisms due to genetic variants and nongenetic factors which both may underline the predisposition. Analyzing genome-wide data through gene sets defined by functional pathways offers the potential of greater power discovery and natural connections to biological mechanisms. The identification of new BC susceptibility genes and biological mechanisms in which they are involved may help formulate new hypotheses or substantiate existing hypotheses regarding BC etiology,⁴⁰ and genes that rank high in these pathways can serve as candidates for further genetic and functional studies. In turn, this may open new therapeutic avenues. Furthermore, our data confirm that

strategies employed to construct population specific PRS need to be improved and that the joined effect of these PRS and family history needs to be considered in risk prediction models to improve surveillance and medical management of women at higher risk.

ACKNOWLEDGMENTS

We are most grateful to all subjects who so willingly participated in the GENESIS study. We would like to thank Juliette Coignard for helpful discussions regarding the analysis of the iCOGS data. We wish to pay a tribute to Olga M. Sinilnikova, who was one of the initiators and principal investigators of GENESIS and who died prematurely on June 30, 2014. We thank all the GENESIS collaborating cancer clinics (Clinique Sainte Catherine, Avignon: H. Dreyfus; Hôpital Saint Jacques, Besançon: M-A. Collonge-Rame; Institut Bergonié, Bordeaux: M. Longy, A. Floquet, E. Barouk-Simonet; CHU, Brest: S. Audebert; Centre François Baclesse, Caen: P. Berthet; Hôpital Dieu, Chambéry: S. Fert-Ferrer; Centre Jean Perrin, Clermont-Ferrand: Y-J. Bignon; Hôpital Pasteur, Colmar: J-M. Limacher; Hôpital d'Enfants CHU—Centre Georges François Leclerc, Dijon: L. Faivre-Olivier; CHU, Fort de France: O. Bera; CHU Albert Michallon, Grenoble: D. Leroux; Hôpital Flaubert, Le Havre: V. Layet; Centre Oscar Lambret, Lille: P. Vennin†, C. Adenis; Hôpital Jeanne de Flandre, Lille: S. Lejeune-Dumoulin, S. Manouvrier-Hanu; CHRU Dupuytren, Limoges: L. Venat-Bouvet; Centre Léon Bérard, Lyon: C. Lasset, V. Bonadona; Hôpital Edouard Herriot, Lyon: S. Giraud; Institut Paoli-Calmettes, Marseille: F. Eisinger, L. Huiart; Centre Val d'Aurelle—Paul Lamarque, Montpellier: I. Coupier; CHU Arnaud de Villeneuve, Montpellier: I. Coupier, P. Pujol; Centre René Gauducheau, Nantes: C. Delnatte; Centre Catherine de Sienne, Nantes: A. Lortholary; Centre Antoine Lacassagne, Nice: M. Frénay, V. Mari; Hôpital Caremeau, Nîmes: J. Chiesa; Réseau Oncogénétique Poitou Charente, Niort: P. Gesta; Institut Curie, Paris: D. Stoppa-Lyonnet, M. Gauthier-Villars, B. Buecher, A. de Pauw, C. Abadie, M. Belotti; Hôpital Saint-Louis, Paris: O. Cohen-Haguenauer; Centre Viggo-Petersen, Paris: F. Cornélis; Hôpital Tenon, Paris: A. Fajac; GH Pitié Salpêtrière et Hôpital Beaujon, Paris: C. Colas, F. Soubrier, P. Hammel, A. Fajac; Institut Jean Godinot, Reims: C. Penet, T. D. Nguyen; Polyclinique Courlancy, Reims: L. Demange†, C. Penet; Centre Eugène Marquis, Rennes: C. Dugast†; Centre Henri Becquerel, Rouen: A. Chevrier, T. Frebourg, J. Tinat, I. Tennevet, A. Rossi; Hôpital René Huguenin/Institut Curie, Saint Cloud: C. Noguès, L. Demange†, E. Mouret-Fourme; CHU, Saint-Etienne: F. Prieur; Centre Paul Strauss, Strasbourg: J-P. Fricker, H. Schuster; Hôpital Civil, Strasbourg: O. Caron, C. Maugard; Institut Claudius Regaud, Toulouse: L. Gladieff, V. Feille; Hôpital Bretonneau, Tours: I. Mortemousque; Centre Alexis Vautrin, Vandoeuvre-les-Nancy: E. Luporsi; Hôpital de Bravois, Vandoeuvre-les-Nancy: P. Jonveaux; Gustave Roussy, Villejuif: A. Chompret†, O. Caron). †Deceased prematurely. Financial support for GENESIS was provided by the Ligue Nationale contre le Cancer (grants PRE05/DSL, PRE07/DSL, PRE11/NA), the French National Institute of Cancer (INCa grant no b2008-029/LL-LC) and the comprehensive cancer center SiRIC (Site de Recherche Intégrée sur le Cancer: Grant INCa-DGOS-4654).

CONFLICT OF INTEREST

Dr P. Pujol is a consultant for AstraZeneca, Pfizer, Roche, MSD, Exact Sciences, Abbvie, OncoDNA, Takeda and Novartis. He received research funding from AstraZeneca, Pfizer and Novartis. The other authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

Written informed consent for the present study was obtained from all participants from GENESIS and CECILE. The two studies were approved by the appropriate Advisory Committees on the Treatment of Health Research Information (Comité Consultatif de Protection des Personnes dans la Recherche Biomédicale [CCPPRB] Ile-de-France III for GENESIS and CCPPRB Kremlin-Bicêtre for CECILE) and by the National Data Protection authority.

ORCID

Christine Lonjou  <https://orcid.org/0000-0003-1226-6992>

Thérèse Truong  <https://orcid.org/0000-0002-2943-6786>

Marie-Gabrielle Dondon  <https://orcid.org/0000-0001-6016-8524>

Francesca Damiola  <https://orcid.org/0000-0002-0238-1252>

Jean-Marc Limacher  <https://orcid.org/0000-0002-7723-5508>

Laurence Gladieff  <https://orcid.org/0000-0002-6980-9719>

Laurence Venat-Bouvet  <https://orcid.org/0000-0002-0716-2550>

Yves-Jean Bignon  <https://orcid.org/0000-0001-9378-6353>

Pascal Pujol  <https://orcid.org/0000-0001-8315-4715>

Olivier Caron  <https://orcid.org/0000-0001-8934-2071>

Sylvie Mazoyer  <https://orcid.org/0000-0002-2135-0160>

Inna Kuperstein  <https://orcid.org/0000-0001-8086-8915>

Pascal Guénel  <https://orcid.org/0000-0002-8359-518X>

Emmanuel Barillot  <https://orcid.org/0000-0003-2724-2002>

Dominique Stoppa-Lyonnet  <https://orcid.org/0000-0002-5438-8309>

Fabienne Lesueur  <https://orcid.org/0000-0001-7404-4549>

REFERENCES

1. Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet.* 2003;72:1117-1130.
2. Domchek SM, Weber BL. Clinical management of BRCA1 and BRCA2 mutation carriers. *Oncogene.* 2006;25:5825-5831.
3. Farmer H, McCabe N, Lord CJ, et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature.* 2005;434:917-921.
4. Rahman N, Seal S, Thompson D, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet.* 2007;39:165-167.
5. Renwick A, Thompson D, Seal S, et al. ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat Genet.* 2006;38:873-875.

6. Tavtigian SV, Oefner PJ, Babikyan D, et al. Rare, evolutionarily unlikely missense substitutions in ATM confer increased risk of breast cancer. *Am J Hum Genet.* 2009;85:427-446.
7. Teo ZL, Park DJ, Provenzano E, et al. Prevalence of PALB2 mutations in Australasian multiple-case breast cancer families. *Breast Cancer Res.* 2013;15:R17.
8. Le Calvez-Kelm F, Lesueur F, Damiola F, et al. Rare, evolutionarily unlikely missense substitutions in CHEK2 contribute to breast cancer susceptibility: results from a breast cancer family registry (CFR) case-control mutation screening study. *Breast Cancer Res.* 2011; 13:R6.
9. Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations in PALB2. *N Engl J Med.* 2014;371:497-506.
10. Girard E, Eon-Marchais S, Olaso R, et al. Familial breast cancer and DNA repair genes: insights into known and novel susceptibility genes from the GENESIS study, and implications for multigene panel testing. *Int J Cancer.* 2018;144:1962-1974.
11. Easton DF, Pooley KA, Dunning AM, et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature.* 2007; 447:1087-1093.
12. 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;39:870-874.
13. 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;39:865-869.
14. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009;461:747-753.
15. Sinilnikova OM, Dondon MG, Eon-Marchais S, et al. GENESIS: a French national resource to study the missing heritability of breast cancer. *BMC Cancer.* 2016;16:13.
16. Truong T, Liqueur B, Menegaux F, et al. Breast cancer risk, nightwork, and circadian clock gene polymorphisms. *Endocr Relat Cancer.* 2014; 21:629-638.
17. Bahcall OG. iCOGS collection provides a collaborative model. Foreword. *Nat Genet.* 2013;45:343.
18. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet.* 2013;45:353-361.61e1-2.
19. Milne RL, Kuchenbaecker KB, Michailidou K, et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat Genet.* 2017;49:1767-1778.
20. Eeles RA, Olama AA, Benlloch S, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet.* 2013;45:385-391.91e1-2.
21. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007;81:559-575.
22. Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res.* 2001; 125:279-284.
23. Agresti A. *Categorical Data Analysis.* New York, NY: Wiley J; 1990: 100-102.
24. Liu JZ, McRae AF, Nyholt DR, et al. A versatile gene-based test for genome-wide association studies. *Am J Hum Genet.* 2010;87: 139-145.
25. Mishra A, Macgregor S. VEGAS2: software for more flexible gene-based testing. *Twin Res Hum Genet.* 2015;18:86-91.
26. Wang K, Li M, Hakonarson H. Analysing biological pathways in genome-wide association studies. *Nat Rev Genet.* 2010;11: 843-854.
27. Rossin EJ, Lage K, Raychaudhuri S, et al. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* 2011;7:e1001273.
28. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000;28:27-30.
29. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2017;45:D353-D361.
30. Kuperstein I, Bonnet E, Nguyen HA, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google maps. *Oncogenesis.* 2015;4:e160.
31. Yang J, Weedon MN, Purcell S, et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet.* 2011;19:807-812.
32. Gong J, Mei S, Liu C, et al. PanCanQTL: systematic identification of cis-eQTLs and trans-eQTLs in 33 cancer types. *Nucleic Acids Res.* 2018;46:D971-D976.
33. Menegaux F, Truong T, Anger A, et al. Night work and breast cancer: a population-based case-control study in France (the CECILE study). *Int J Cancer.* 2013;132:924-931.
34. Zhang K, Cui S, Chang S, Zhang L, Wang J. I-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res.* 2010;38:W90-W95.
35. Gui H, Li M, Sham PC, Cherny SS. Comparisons of seven algorithms for pathway analysis using the WTCCC Crohn's disease dataset. *BMC Res Notes.* 2011;4:386.
36. Ramanan VK, Shen L, Moore JH, Saykin AJ. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet.* 2012;28:323-332.
37. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet.* 2010;86:929-942.
38. Ott J, Kamatani Y, Lathrop M. Family-based designs for genome-wide association studies. *Nat Rev Genet.* 2011;12:465-474.
39. Mavaddat N, Michailidou K, Dennis J, et al. Polygenic risk scores for prediction of breast cancer and breast cancer subtypes. *Am J Hum Genet.* 2019;104:21-34.
40. Aterido A, Julia A, Ferrandiz C, et al. Genome-wide pathway analysis identifies genetic pathways associated with psoriasis. *J Invest Dermatol.* 2016;136:593-602.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lonjou C, Eon-Marchais S, Truong T, et al. Gene- and pathway-level analyses of iCOGS variants highlight novel signaling pathways underlying familial breast cancer susceptibility. *Int. J. Cancer.* 2021;148:1895–1909. <https://doi.org/10.1002/ijc.33457>