Identification of novel susceptibility loci and genes for breast cancer risk: A transcriptome-wide association study of 229,000 women of European descent

Lang Wu[1,160], Wei Shi[2,160], Jirong Long[1], Xingyi Guo[1], Kyriaki Michailidou[3,4], Jonathan Beesley[2], Manjeet K. Bolla[3], Xiao-Ou Shu[1], Yingchang Lu[1], Qiuyin Cai[1], Fares Al-Ejeh[2], Esdy Rozali[2], Qin Wang[3], Joe Dennis[3], Bingshan Li[151], Chenjie Zeng[1], Helian Feng[5,6], Alexander Gusev[153, 154, 155], Richard T. Barfield[5], Irene L. Andrulis[7,8], Hoda Anton-Culver[9], Volker Arndt[10], Kristan J. Aronson[11], Paul L. Auer[12,13], Myrto Barrdahl[14], Caroline Baynes[15], Matthias W. Beckmann[16], Javier Benitez[17,18], Marina Bermisheva[19,20], Carl Blomqvist[21,159], Natalia V. Bogdanova[20,22,23], Stig E. Bojesen[24-26], Hiltrud Brauch[27-29], Hermann Brenner[10,29,30], Louise Brinton[31], Per Broberg[32], Sara Y. Brucker[33], Barbara Burwinkel[34,35], Trinidad Caldés[36], Federico Canzian[37], Brian D. Carter[38], J. Esteban Castelao[39], Jenny Chang-Claude[14,40], Xiaoqing Chen[2], Ting-Yuan David Cheng[41], Hans Christiansen[22], Christine L. Clarke[42], NBCS Collaborators[43-120,44-124,45], Margriet Collée[46], Sten Cornelissen[47], Fergus J. Couch[48], David Cox[49,50], Angela Cox[51], Simon S. Cross[52], Julie M. Cunningham[48], Kamila Czene[53], Mary B. Daly[54], Peter Devilee[55,56], Kimberly F. Doheny[57], Thilo Dörk[20], Isabel dos-Santos-Silva[58], Martine Dumont[59], Miriam Dwek[60], Diana M. Eccles[61], Ursula Eilber[14], A. Heather Eliassen[6,62], Christoph Engel[63], Mikael Eriksson[53], Laura Fachal[15], Peter A. Fasching[16,64], Jonine Figueroa[31,65], Dieter Flesch-Janys[66,67], Olivia Fletcher[68], Henrik Flyger[69], Lin Fritschi[70], Marike Gabrielson[53], Manuela Gago-Dominguez[71,72], Susan M. Gapstur[38], Montserrat García-Closas[31], Mia M. Gaudet[38], Maya Ghoussaini[15], Graham G. Giles[73,74], Mark S. Goldberg[75,76], David E. Goldgar[77], Anna González-Neira[17], Pascal Guénel[78], Eric Hahnen[79-81], Christopher A. Haiman[82], Niclas Håkansson[83], Per Hall[53], Emily Hallberg[84], Ute Hamann[85], Patricia Harrington[15], Alexander Hein[16], Belynda Hicks[86], Peter Hillemanns[20], Antoinette Hollestelle[87], Robert N. Hoover[31], John L. Hopper[74], Guanmengqian Huang[85], Keith Humphreys[53], David J. Hunter[6,158], Anna Jakubowska[88], Wolfgang Janni[89], Esther M. John[90-92], Nichola Johnson[68], Kristine Jones[86], Michael E. Jones[93], Audrey Jung[14], Rudolf Kaaks[14], Michael J. Kerin[94], Elza Khusnutdinova[19,95], Veli-Matti Kosma[96-98], Vessela N. Kristensen[99-101], Diether Lambrechts[102,103], Loic Le Marchand[104], Jingmei Li[157], Sara Lindström[5,105], Jolanta Lissowska[106], Wing-Yee Lo[27,28], Sibylle Loibl[107], Jan Lubinski[88], Craig Luccarini[15], Michael P. Lux[16], Robert J. MacInnis[73,74], Tom Maishman[61,108], Ivana Maleva Kostovska[20,109], Arto Mannermaa[96-98], JoAnn E. Manson[6,110], Sara Margolin[111], Dimitrios Mavroudis[112], Hanne Meijers-Heijboer[152], Alfons Meindl[113], Usha Menon[114], Jeffery Meyer[48], Anna Marie Mulligan[115,116], Susan L. Neuhausen[117], Heli Nevanlinna[118], Patrick Neven[119], Sune F. Nielsen[24,25], Børge G. Nordestgaard[24-26], Olufunmilayo I. Olopade[120], Janet E. Olson[84], Håkan Olsson[32], Paolo Peterlongo[121], Julian Peto[58], Dijana Plaseska-Karanfilska[109], Ross Prentice[12], Nadege Presneau[60], Katri Pylkäs[122,123], Brigitte Rack[89], Paolo Radice[125], Nazneen Rahman[126], Gad Rennert[127], Hedy S. Rennert[127], Valerie Rhenius[15], Atocha Romero[36,128], Jane Romm[57], Anja Rudolph[14], Emmanouil Saloustros[129], Dale P. Sandler[130], Elinor J. Sawyer[131], Marjanka K. Schmidt[47,132], Rita K. Schmutzler[79-81], Andreas Schneeweiss[34,133], Rodney J. Scott[134,135], Christopher Scott[84], Sheila Seal[126], Mitul Shah[15], Martha J. Shrubsole[1], Ann Smeets[119], Melissa C. Southey[136], John J. Spinelli[137,138], Jennifer Stone[139,140], Harald Surowy[34,35], Anthony J. Swerdlow[93,141], Rulla M. Tamimi[5,6,62], William Tapper[61], Jack A. Taylor[130,142], Mary Beth Terry[143], Daniel C. Tessier[144], Abigail Thomas[84], Kathrin Thöne[67], Rob A.E.M. Tollenaar[145], Diana Torres[85,146], Thérèse Truong[78], Michael Untch[147], Celine Vachon[84], David Van Den Berg[82], Daniel Vincent[144], Quinten Waisfisz[152], Clarice R. Weinberg[148], Camilla Wendt[111], Alice S. Whittemore[91,92], Hans Wildiers[119], Walter C.

Willett[6,62,156], Robert Winqvist[122,123], Alicja Wolk[83], Lucy Xia[82], Xiaohong R. Yang[31], Argyrios Ziogas[9], Elad Ziv[149], kConFab/AOCS Investigators[150], Alison M. Dunning[15], Paul D.P. Pharoah[3,15], Jacques Simard[59], Roger L. Milne[73,74], Stacey L. Edwards[2], Peter Kraft[5,6], Douglas F. Easton[3,15], Georgia Chenevix-Trench[2]*, Wei Zheng[1]*

***Corresponding Authors:** Wei Zheng, MD, PhD, Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, 2525 West End Ave, Suite 800, Nashville, Tennessee, 37203, USA. Email: wei.zheng@vanderbilt.edu and Georgia Chenevix-Trench, PhD, Cancer Division, QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston 4006, Australia. Email: Georgia.Trench@qimrberghofer.edu.au

Key words: eQTL, genetics, breast cancer, gene expression, GWAS, susceptibility

1. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA.
2. Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, Australia.
3. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
4. Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, Nicosia, Cyprus.
5. Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
6. Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA.
7. Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada.
8. Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada.
9. Department of Epidemiology, University of California Irvine, Irvine, CA, USA.
10. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
11. Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON, Canada.
12. Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.
13. Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI, USA.
14. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
15. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.
16. Department of Gynaecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen, Germany.

17. Human Cancer Genetics Program, Spanish National Cancer Research Centre, Madrid, Spain.
18. Centro de Investigación en Red de Enfermedades Raras (CIBERER), Valencia, Spain.
19. Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa, Russia.
20. Gynaecology Research Unit, Hannover Medical School, Hannover, Germany.
21. Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland.
22. Department of Radiation Oncology, Hannover Medical School, Hannover, Germany.
23. N.N. Alexandrov Research Institute of Oncology and Medical Radiology, Minsk, Belarus.
24. Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
25. Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
26. Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark.
27. Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, Germany.
28. University of Tübingen, Tübingen, Germany.
29. German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
30. Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.
31. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD, USA.
32. Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund, Sweden.
33. Department of Gynecology and Obstetrics, University of Tübingen, Tübingen, Germany.
34. Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg, Germany.
35. Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg, Germany.
36. Medical Oncology Department, CIBERONC Hospital Clínico San Carlos, Madrid, Spain.
37. Genomic Epidemiology Group, German Cancer Research Center (DKFZ), Heidelberg, Germany.
38. Epidemiology Research Program, American Cancer Society, Atlanta, GA, USA.
39. Oncology and Genetics Unit, Instituto de Investigacion Biomedica (IBI) Orense-Pontevedra-Vigo, Xerencia de Xestion Integrada de Vigo-SERGAS, Vigo, Spain.
40. University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
41. Department of Epidemiology, University of Florida, Gainesville, FL, USA.
42. Westmead Institute for Medical Research, University of Sydney, Sydney, Australia.
43. Department of Oncology, Haukeland University Hospital, Bergen, Norway.
44. National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital Radiumhospitalet, Oslo, Norway.
45. Oslo University Hospital, Oslo, Norway.

46. Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam, The Netherlands.
47. Division of Molecular Pathology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam, The Netherlands.
48. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN, USA.
49. Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK.
50. INSERM U1052, Cancer Research Center of Lyon, Lyon, France.
51. Sheffield Institute for Nucleic Acids, Department of Oncology and Metabolism, University of Sheffield, Sheffield, UK.
52. Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield, UK.
53. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden.
54. Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA, USA.
55. Department of Pathology, Leiden University Medical Center, Leiden, The Netherlands.
56. Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.
57. Center for Inherited Disease Research (CIDR), Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA.
58. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK.
59. Genomics Center, Centre Hospitalier Universitaire de Québec Research Center, Laval University, Québec City, QC, Canada.
60. Department of Biomedical Sciences, Faculty of Science and Technology, University of Westminster, London, UK.
61. Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton, UK.
62. Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
63. Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig, Germany.
64. David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA, USA.
65. Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh, UK.
66. Institute for Medical Biometrics and Epidemiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
67. Department of Cancer Epidemiology, Clinical Cancer Registry, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
68. The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London, UK.
69. Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev, Denmark.
70. School of Public Health, Curtin University, Perth, Australia.

71. Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago De Compostela, Spain.
72. Moores Cancer Center, University of California San Diego, La Jolla, CA, USA.
73. Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, Australia.
74. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia.
75. Department of Medicine, McGill University, Montréal, QC, Canada.
76. Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, QC, Canada.
77. Department of Dermatology, Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT, USA.
78. Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif, France.
79. Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne, Cologne, Germany.
80. Center for Integrated Oncology (CIO), University Hospital of Cologne, Cologne, Germany.
81. Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany.
82. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA.
83. Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden.
84. Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA.
85. Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg, Germany.
86. Cancer Genomics Research Laboratory, Leidos Biomedical Research, Frederick National Laboratory for Cancer Research, Frederick, MD, USA.
87. Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam, The Netherlands.
88. Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland.
89. Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm, Germany.
90. Department of Epidemiology, Cancer Prevention Institute of California, Fremont, CA, USA.
91. Department of Health Research and Policy - Epidemiology, Stanford University School of Medicine, Stanford, CA, USA.
92. Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA, USA.
93. Division of Genetics and Epidemiology, The Institute of Cancer Research, London, UK.
94. School of Medicine, National University of Ireland, Galway, Ireland.
95. Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa, Russia.
96. Translational Cancer Research Area, University of Eastern Finland, Kuopio, Finland.

97. Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio, Finland.
98. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio, Finland.
99. Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway.
100. Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway.
101. Department of Clinical Molecular Biology, Oslo University Hospital, University of Oslo, Oslo, Norway.
102. VIB KULeuven Center for Cancer Biology, VIB, Leuven, Belgium.
103. Laboratory for Translational Genetics, Department of Human Genetics, KU Leuven, Leuven, Belgium.
104. Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, USA.
105. Department of Epidemiology, University of Washington School of Public Health, Seattle, WA, USA.
106. Department of Cancer Epidemiology and Prevention, M. Sklodowska-Curie Institute - Oncology Center, Warsaw, Poland.
107. German Breast Group, GmbH, Neu Isenburg, Germany.
108. Southampton Clinical Trials Unit, Faculty of Medicine , University of Southampton, Southampton, UK.
109. Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov" , Macedonian Academy of Sciences and Arts, Skopje, Republic of Macedonia.
110. Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.
111. Department of Oncology - Pathology, Karolinska Institutet, Stockholm, Sweden.
112. Department of Medical Oncology, University Hospital of Heraklion, Heraklion, Greece.
113. Division of Gynaecology and Obstetrics, Technische Universität München, Munich, Germany.
114. Gynaecological Cancer Research Centre, Women’s Cancer, Institute for Women's Health, University College London, London, UK.
115. Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada.
116. Laboratory Medicine Program, University Health Network, Toronto, ON, Canada.
117. Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA, USA.
118. Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki, Finland.
119. Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium.
120. Center for Clinical Cancer Genetics and Global Health, The University of Chicago, Chicago, IL, USA.
121. IFOM, The FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan, Italy.
122. Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu, Finland.

123. Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu, Finland.
124. Department of Gynecology and Obstetrics, Ludwig-Maximilians University of Munich, Munich, Germany.
125. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS (Istituto Di Ricovero e Cura a Carattere Scientifico) Istituto Nazionale dei Tumori (INT), Milan, Italy.
126. Section of Cancer Genetics, The Institute of Cancer Research, London, UK.
127. Clalit National Cancer Control Center, Haifa, Israel.
128. Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid, Spain.
129. Hereditary Cancer Clinic, University Hospital of Heraklion, Heraklion, Greece.
130. Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA.
131. Research Oncology, Guy’s Hospital, King's College London, London, UK.
132. Division of Psychosocial Research and Epidemiology, The Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, The Netherlands.
133. National Center for Tumor Diseases, University of Heidelberg, Heidelberg, Germany.
134. Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, Australia.
135. Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Callaghan, Australia.
136. Department of Pathology, The University of Melbourne, Melbourne, Australia.
137. Cancer Control Research, BC Cancer Agency, Vancouver, BC, Canada.
138. School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada.
139. The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, Australia.
140. Department of Obstetrics and Gynaecology, University of Melbourne and the Royal Women's Hospital, Melbourne, Australia.
141. Division of Breast Cancer Research, The Institute of Cancer Research, London, UK.
142. Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA.
143. Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA.
144. McGill University and Génome Québec Innovation Centre, Montréal, QC, Canada.
145. Department of Surgery, Leiden University Medical Center, Leiden, The Netherlands.
146. Institute of Human Genetics, Pontificia Universidad Javeriana, Bogota, Colombia.
147. Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin, Germany.
148. Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC, USA.
149. Department of Medicine, Institute for Human Genetics, UCSF Helen Diller Family Comprehensive Cancer Center, University of California San Francisco, San Francisco, CA, USA.
150. Peter MacCallum Cancer Center, Melbourne, Australia.
151. Department of Molecular Physiology & Biophysics, Vanderbilt Genetics Institute, Vanderbilt University, Nashville, TN, USA.

152. Department of Clinical Genetics, VU University Medical Center, Amsterdam, The Netherlands.
153. Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA.
154. Department of Medicine, Harvard Medical School, Boston, MA.
155. Division of Genetics, Brigham and Women's Hospital, Boston, MA.
156. Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA.
157. Human Genetics, Genome Institute of Singapore, Singapore, Singapore.
158. Nuffield Department of Population Health, University of Oxford, Big Data Institute, Old Road Campus, Oxford OX3 7LF, UK.
159. Department of Oncology University of Örebro, Örebro, Sweden.
160. Lang Wu and Wei Shi are joint co-first authors.

Abstract:

Breast cancer risk variants identified in genome-wide association studies explain only a small fraction of familial relative risk, and genes responsible for these associations remain largely unknown. To identify novel risk loci and likely causal genes, we performed a transcriptome-wide association study evaluating associations of genetically predicted gene expression with breast cancer risk in 122,977 cases and 105,974 controls of European ancestry. We used data from 67 subjects included in the Genotype-Tissue Expression Project to establish genetic models to predict gene expression in breast tissue and evaluated model performance using data from 86 subjects included in The Cancer Genome Atlas. Of the 8,597 genes evaluated, significant associations were identified for 48 at a Bonferroni-corrected threshold of $P < 5.82\times10^{-6}$, including 14 genes at loci not yet reported for breast cancer risk. We silenced 13 genes and showed an effect for 11 on cell proliferation and/or colony forming efficiency. Our study provides new insights into breast cancer genetics and biology.

Breast cancer is the most commonly diagnosed malignancy among women in many countries[1]. Genetic factors play an important role in breast cancer etiology. Multiple high- and moderate-penetrance genes, including BRCA1, BRCA2, PALB2, CHEK2 and ATM, have been identified as contributors to familial breast cancer[2,3]. However, deleterious germline mutations in these genes are rare, thus accounting for only a small fraction of breast cancer cases in the general population[4,5]. Since 2007, genome-wide association studies (GWAS) have identified approximately 180 genetic loci harboring common, low-penetrance variants for breast cancer[6-13], but these more common variants explain less than 20% of familial relative risk[7].

A large proportion of disease-associated risk variants identified by GWAS are located in non-protein coding or intergenic regions and are not in linkage disequilibrium (LD) with any nonsynonymous coding single nucleotide polymorphisms (SNPs)[14]. Many of these susceptibility variants are located in gene regulatory elements[15,16], and it has therefore been hypothesized that most of the GWAS-identified associations may be driven by the regulatory function of risk variants on the expression levels of nearby genes. For breast cancer, recent studies have shown that GWAS-identified associations at 1p34, 1p36, 2q35, 5p12, 5p15.33, 5q11.2, 5q14, 6q25, 7q22, 9q31.2, 10q21.3, 10q26.13, 11p15, 11q13.3, 15q26.1, 19p13 and 19q13.31 are likely due to the effect of risk variants at these loci on regulating the expression of either nearby or more distal genes: CITED4, KLHDC7A, IGFBP5, FGF10/MRPS30, TERT, MAP3K1, ATP6AP1L, RMND1, RASA4/PRKRIP1, KLF4, NRBF2, FGFR2, PIDD1, CCND1, RCCD1, ABHD8, and ZNF404[7,9,10,13,17-22]. However, for the large majority of the GWAS-identified breast cancer risk loci, the genes responsible for the associations remain unknown.

Several recent studies have reported that regulatory variants may account for a large proportion of disease heritability not yet discovered through GWAS[23-25]. Many of these variants may have a small effect size, and thus are difficult to identify in individual SNP-based GWAS studies, even with a very large sample size. Applying gene-based approaches that aggregate the effects of multiple variants into a single testing unit may increase study power to identify novel disease-associated loci. Transcriptome-wide association studies (TWAS) systematically investigate across the transcriptome the association of genetically predicted gene expression with disease risk, providing an effective approach to identify novel susceptibility genes[26-29]. Instead of testing millions of SNPs in GWAS, TWAS evaluate the association of predicted expression for selected genes, thus greatly reducing the burden of multiple comparisons in statistical inference. Recently, Hoffman et al performed a TWAS including 15,440 cases and 31,159 controls and reported significant associations for five genes with breast cancer risk[30]. However, the sample size of that study was relatively small and several reported associations were not statistically significant after Bonferroni correction. Herein, we report results from a larger TWAS of breast cancer that used the MetaXcan method[26] to analyze summary statistics data from 122,977 cases and 105,974 controls of European descent from the Breast Cancer Association Consortium (BCAC).

Results

Gene expression prediction models

The overall study design is shown in **Supplementary Figure 1**. We used transcriptome and high-density genotyping data from 67 women of European descent included in the Genotype-Tissue Expression (GTEx) project to build genetic models to predict RNA expression levels for each of the genes expressed in normal breast tissues, by applying the elastic net method (α=0.5) with ten-fold cross-validation. Genetically regulated expression was estimated for each gene using variants within a 2 MB window flanking the respective gene boundaries, inclusive. SNPs with a minor allele frequency of at least 0.05 and included in the HapMap Phase 2 subset were used for model building. Of the models built for 12,696 genes, 9,109 showed a prediction performance (R^2) of at least 0.01 ($\geq$10% correlation between predicted and observed expression). For genes for which the expression could not be predicted well using this approach, we built models using only SNPs located in the promoter or enhancer regions, as predicted using three breast cell lines in the Roadmap Epigenomics Project/Encyclopedia of DNA Elements Project. This approach leverages information from functional genomics and reduces the number of variants for variable selection, and therefore potentially improving statistical power. This enabled us to build genetic models for additional 3,715 genes with $R^2 \geq 0.01$. **Supplementary Table 1** provides detailed information regarding the performance threshold and types of models built in this study. Overall, genes that were predicted with $R^2 \geq 0.01$ in GTEx data were also predicted well in The Cancer Genome Atlas (TCGA) tumor-adjacent normal tissue data (correlation coefficient of 0.55 for R^2 in two datasets; **Supplementary Figure 2**). Based on model performance in GTEx and TCGA, we prioritized 8,597 genes for analyses of the associations between predicted gene expression and breast cancer risk using the following criteria: 1) genes with a model prediction R^2 of at least 0.01 in the GTEx set (10% correlation) and a Spearman's correlation coefficient of $\geq$0.1 in the external validation experiment using TCGA data, 2) genes with a prediction R^2 of at least 0.09 (30% correlation) in the GTEx set regardless of their performance in the TCGA set, 3) genes with a prediction R^2 of at least 0.01 in the GTEx set (10% correlation) that could not be evaluated in the TCGA set because of a lack of data.

Association analyses of predicted gene expression with breast cancer risk

Using the MetaXcan method[26], we performed association analyses to evaluate predicted gene expression and breast cancer risk using the meta-analysis summary statistics of individual genetic variants generated for 122,977 breast cancer cases and 105,974 controls of European ancestry included in BCAC. For the majority of the tested genes, most of the SNPs selected for prediction models were used for the association analyses (e.g., ≥95% predicting SNPs used for 83.8% of the tested genes, and ≥80% predicting SNPs used for 95.6% of the tested genes). Lambda 1,000 ($\lambda_{1,000}$), a standardized estimate of the genomic inflation scaling to a study of 1,000 cases and 1,000 controls, was 1.004 in our study (Quantile-quantile (QQ) plot presented in **Supplementary Figure 3 (A)**). Of the 8,597 genes evaluated in this study, we identified 179 genes whose predicted expression was associated with breast cancer risk at $P<1.05\times10^{-3}$, a FDR-corrected significance level (**Figure 1**, **Supplementary Table 2**). Of these, 48 showed a significant association at the Bonferroni-corrected threshold of $P\leq5.82\times10^{-6}$ (**Figure 1**, **Tables 1-3**), including 14 genes located at 11 loci that are 500 kb away from any of the risk variants identified in previous GWAS of breast cancer risk (**Table 1**). An association between lower predicted expression and increased breast cancer risk was detected for LRRC3B (3p24.1), SPATA18 (4q12), UBD (6p22.1), MIR31HG (9p21.3), RIC8A (11p15.5), B3GNT1 (11q13.2), GALNT16 (14q24.1) and MAN2C1 and CTD-2323K18.1 (15q24.2). Conversely, an association between higher predicted expression and increased breast cancer risk was identified for ZSWIM5 (1p34.1), KLHDC10 (7q32.2), RP11-867G23.10 (11q13.2), RP11-218M22.1 (12p13.33) and

PLEKHD1 (14q24.1). The remaining 34 significantly associated genes are all located at breast cancer susceptibility loci identified in previous GWAS (**Tables 2-3**). Among them, 23 have not yet been previously implicated as genes responsible for association signals with breast cancer risk identified at these loci through expression quantitative trait loci (eQTL) and/or functional studies, and do not harbor GWAS or fine-mapping identified risk variants (**Table 2**), while the other eleven (KLHDC7A[7], ALS2CR12[31], CASP8[31,32], ATG10[9], SNX32[33], STXBP4[34,35] , ZNF404[8], ATP6AP1L[9], RMND1[17], L3MBTL3[6], and RCCD1[10]) had been reported as potential causal genes at breast cancer susceptibility loci or harbor GWAS or fine-mapping identified risk variants (**Table 3**). Except for RP11-73O6.3 and L3MBTL3, there was no evidence of heterogeneity in the gene-expression association (I^2<0.2) across the iCOGS, OncoArray, and GWAS datasets included in our analyses (**Supplementary Table 3**). Overall, through our agnostic search, we identified 37 novel susceptibility genes for breast cancer, including 21 protein-coding genes, 15 long non-coding RNAs (lncRNAs) and a processed transcript, and confirmed eleven genes known to potentially play a role in breast cancer susceptibility.

To determine whether the associations between predicted gene expression and breast cancer risk were independent of the association signals identified in previous GWAS, we performed conditional analyses adjusting for the GWAS-identified risk SNPs closest to the TWAS-identified gene (**Supplementary Table 4**)[36]. We found that the associations for 11 genes (LRRC3B, SPATA18, KLHDC10, MIR31HG, RIC8A, B3GNT1, RP11-218M22.1, MAN2C1, CTD-2323K18.1 **(Table 1)**, ALK, CTD-3051D23.1 **(Table 2)**) remained statistically significant at $P<5.82\times10^{-6}$ (**Tables 1-3**). This suggests the expression of these genes may be associated with breast cancer risk independent of the GWAS-identified risk variant(s). For nine of the genes

(SPATA18, KLHDC10, MIR31HG, RIC8A, RP11-218M22.1, MAN2C1, CTD-2323K18.1 **(Table 1)**, ALK, and CTD-3051D23.1 **(Table 2)**), the significance level of the association remained essentially unchanged, suggesting these associations may be entirely independent of GWAS-identified association signals.

Of the 131 genes showing a significant association at P values between 5.82×10^{-6} and 1.05×10^{-3} (significant after FDR-correction but not Bonferroni-correction), 38 are located at GWAS-identified breast cancer risk loci (± 500 kb of the index SNPs) (**Table 4**). Except for RP11-400F19.8, there was no evidence of heterogeneity in TWAS association ($I^2<0.2$) across the iCOGS, OncoArray, and GWAS studies (**Supplementary Table 3**). After adjusting for the index SNPs, breast cancer associations for MTHFD1L, PVT1, RP11-123K19.1, FES, RP11-400F19.8, CTD-2538G9.5, and CTD-3216D2.5 remained significant at $p \leq 1.05\times10^{-3}$, again suggesting that the association of these genes with breast cancer risk may be independent of the GWAS-identified association signals (**Table 4**).

For 41 of the 48 associated genes that reached the Bonferroni-corrected significant level, we obtained individual-level data from subjects included in the iCOGS (n=84,740) and OncoArray (n=112,133) datasets, which was 86% of the subjects included in the analysis using summary statistics (**Supplementary Table 5**). The results from the analysis using individual-level data were very similar to those described above using MetaXcan analyses (Pearson correlation of z-scores was 0.991 for iCOGS data and 0.994 for OncoArray data), although not all associations reached the Bonferroni-corrected significant level, possibly due to a smaller sample size (**Supplementary Table 5**). Conditional analyses using individual level data also revealed consistent results compared with analyses using summary data. We found that for several genes within the same genomic region, their predicted expression levels were correlated with each other (**Tables 1-3**). The associations between predicted expression of PLEKHD1 and ZSWIM5 and breast cancer risk were largely influenced by their corresponding closest risk variants identified in GWAS, although these risk variants are >500 kb away from these genes (**Table 1**). There were significant correlation of rs999737 and rs1707302 with genetically predicted expression of PLEKHD1 (r = -0.47 in the OncoArray dataset and -0.48 in the iCOGS dataset) and ZSWIM5 (r = 0.50 in the OncoArray dataset and 0.51 in the iCOGS dataset), respectively.

INQUISIT algorithm scores for the identified genes

For the 48 associated genes after Bonferroni correction, we assessed their integrated expression quantitative trait and in silico prediction of GWAS target **(**INQUISIT) scores[7] to assess whether there are other lines of evidence beyond the scope of eQTL for supporting our TWAS-identified genes as candidate target genes at GWAS-identified loci. The detailed methodology for INQUISIT scores have been described elsewhere[7]. In brief, a score for each gene-SNP pair is calculated across categories representing potential regulatory mechanisms - distal or proximal gene regulation (promoter). Features contributing to the score are based on functionally important genomic annotations such as chromatin interactions, transcription factor binding, and eQTLs. Compared with evidence from eQTL only, INQUISIT scores incorporate additional lines of evidence, including distal regulations. The INQUISIT scores for our identified genes are shown in **Supplementary Table 6**. Except for UBD with a very low score in the distal regulation category (0.05), none of the genes at novel loci (**Table 1**) showed evidence to be potential target genes for any of the GWAS-identified breast cancer susceptibility loci. This is interesting and within the expectation since these genes may represent novel association signals. There was evidence suggesting that RP11-439A17.7, NUDT17, ANKRD34A, BTN3A2, AP006621.6, RPLP2, LRRC37A2, LRRC37A, KANSL1-AS1, CRHR1 and HAPLN4 listed in Table 2, and all eleven genes listed in Table 3, may be target genes for risk variants identified in GWAS at these loci (**Supplementary Table 6**). For NUDT17, ANKRD34A, RPLP2, LRRC37A2, LRRC37A, KANSL1-AS1, CRHR1, HAPLN4, KLHDC7A, ALS2CR12, CASP8, ATG10, ATP6AP1L, L3MBTL3, RMND1, SNX32, RCCD1, STXBP4 and ZNF404, the INQUISIT scores were not derived only from eQTL data, providing orthogonal support for these loci. For these loci, the associations of candidate causal SNPs with breast cancer risk may be mediated through these genes. This is in general consistent with the findings from the conditional analyses described above.

Pathway enrichment analyses

Ingenuity Pathway Analysis (IPA)[37] suggested potential enrichment of cancer-related functions for the significantly associated protein-coding genes identified in this study (**Supplementary Table 7**). The top canonical pathways identified in these analyses included apoptosis related pathways (Granzyme B signaling (p=0.024) and cytotoxic T lymphocyte-mediated apoptosis of target cells (p=0.046)), immune system pathway (inflammasome pathway (p=0.030)), and tumoricidal function of hepatic natural killer cells (p=0.036). The identified pathways are largely consistent with findings in previous studies[7]. For the significantly associated lncRNAs identified in this study, pathway analysis of their highly co-expressed protein-coding genes also revealed potential over-representation of cancer related functions (**Supplementary Table 7**).

Knockdown of predicted risk-associated genes in breast cells

To assess the function of genes whose high levels of predicted expression were associated with increased breast cancer risk, we selected 13 genes for knockdown experiments in breast cells: ZSWIM5, KLHDC10, RP11-218M22.1 and PLEKHD1 (**Table 1**), UBLCP1, AP006621.6, RP11-467J12.4, CTD-3032H12.1 and RP11-15A1.7 (**Table 2**), and ALS2CR12, RMND1, STXBP4 and ZNF404 (**Table 3**). As negative controls, we selected B2M, ARHGDIA and ZAP70 using the following criteria: 1) at least 2 MB from any known breast cancer risk locus; 2) not an essential gene in breast cancer[38,39]; and 3) not predicted to be a target gene in INQUISIT. In addition, as positive controls, we included in the experiments PIDD1 (**Table 4**)[7], NRBF2[20] and ABHD8[22], which have been functionally validated as the target genes at breast cancer risk loci. We performed quantitative PCR (qPCR) on a panel of three 'normal' mammary epithelial and 15 breast cancer cell lines to analyze their expression level (**Supplementary Figure 4 and Supplementary Table 8**). All 19 genes were expressed in the normal mammary epithelial line 184A1[40] and the luminal breast cancer cell lines, MCF7 and T47D, so we used these cell lines for the proliferation assay, and MCF7 for the colony formation assay[41]. We also evaluated SNX32, ALK and BTN3A2 by qPCR, but they were not expressed in T47D and MCF7 cells; therefore they were not evaluated further. It was difficult to design siRNAs against RP11-867G23.1 and RP11-53O19.1 because they both have multiple transcripts with limited, GC-rich regions in common. We did not include RPLP2 because it is already known to be an essential gene for breast cancer survival[42]. Knockdown of the 19 tested genes was achieved by small short interfering RNA (siRNA) (**Supplementary Table 9**) and the knockdown efficiency was calculated in 184A1, MCF7 and T47D for each siRNA pair. Robust knockdown of the gene of interests (GOI) was validated by qPCR with the majority of the siRNAs (**Supplementary Figure 5**).

To evaluate the survival and proliferation ability of cells following gene interruption, we used an IncuCyte to quantify cell proliferation in real time and quantified the corrected proliferation of cells with knocking down of GOI in comparison to that of cells with non-target control (NTC) siRNA). As expected, knockdown of the three negative control genes (B2M, ARHGDIA and ZAP70) did not significantly change cell proliferation in any of the three cell lines (**Figure 2**A, **Supplementary Figure 6)**. However, with the exception of *UBLCP1, RMND1* and *STXBP4,* knockdown of all other genes (11 TWAS-identified genes along with two known genes, *ABHD8* and *NRBF2*) resulted in significantly decreased cell proliferation in 184A1 normal breast cells, with *KLHDC10, PLEKHD1, RP11-218M22.1, AP006621.6, ZNF404, RP11-467J12.4, CTD-3032H12.1* and *STXBP4* showing a similar effect in one or both cancer cell lines. Down-regulation of three lncRNAs (*RP11-218M22.1*, *RP11-467J12.4* and *CTD-3032H12.1)* resulted in significant reduction in cell proliferation in all three cell lines. We also evaluated the effect of inhibition of these genes on colony forming ability in MCF7 cells. Knockdown of the three negative control genes did not significantly affect colony forming efficiency (CFE). By contrast, knockdown of PIDD1, RP11-15A1.7, RP11-218M22.1, AP006621.6, ZNF404, RP11-467J12.4 and CTD-3032H12.1 resulted in significantly decreased colony forming efficiency in MCF7 cells compared to the NTC (**Figure 2B, Supplementary Figure 7).**

Discussion

This is the largest study to systematically evaluate associations of genetically predicted gene expression across the human transcriptome with breast cancer risk. We identified 179 genes showing a significant association at the FDR-corrected significance level. Of these, 48 showed a significant association at the Bonferroni-corrected threshold, including 14 genes at genomic loci that have not previously been implicated for breast cancer risk. Of the 34 genes we identified that are located at known risk loci, 23 have not previously been shown to be the targets of GWAS-identified risk SNPs at corresponding loci and not harbor any risk SNPs. Our study provides substantial new information to improve the understanding of genetics and etiology for breast cancer, the most common malignancy among women in most countries.

It is possible that TWAS-identified genes may be associated with breast cancer risk through their correlation with disease causal genes. To determine the potential functional significance of TWAS-identified genes and provide evidence for causal inference, we knocked down 13 genes for which high predicted levels of expression were associated with an increased breast cancer risk, in one normal and two breast cancer cell lines, and measured the effect on proliferation and colony forming efficiency. Although there was some variation between cell lines, knockdown of 11 of the 13 genes showed an effect in at least one cell line, particularly on proliferation in 184A1 normal breast cells; the effects were strongest and most consistent for the lncRNAs, RP11-218M22.1, RP11-467J12.4 and CTD-3032H12.1. The observation of a more consistent effect in the normal breast cell line compared with the cancer cell lines is not surprising as cancer cell lines have increased capacity to handle gene interference through mutations which enhance cell survival. Rewiring of pathways and compensatory mechanisms is a hallmark of cancer. Knockdown of PIDD1, NRBF2 and ABHD8¸ for which breast cancer risk associated haplotypes have been shown to be associated with increased expression in reporter assays[7,20,22], affected either proliferation or colony forming efficiency, supporting the results from this study. Knockdown of UBLCP1 and RMND1 did not affect proliferation or colony formation but they could mediate breast cancer risk through other mechanisms.

Some of the genes with strong functional evidence from our study have been reported to have important roles in carcinogenesis. For example, RP11-467J12.4 (PR-lncRNA-1) is a p53-regulated lncRNA that modulates gene expression in response to DNA damage downstream of p53[43]. STXBP4 encodes Syntaxin binding protein 4, a scaffold protein that can stabilise and prevent degradation of an isoform of p63, a member of the p53 tumor suppressor family[44]. KLHDC10 encodes a member of the Kelch superfamily that can activate apoptosis signal-regulating kinase 1, contributing to oxidative stress-induced cell death[45]. Notably, another member of this superfamily, KLHDC7A, has recently been identified as the target gene at the 1p36 breast cancer risk locus[7].

SNX32, ALK and BTN3A2 are also likely susceptibility genes for breast cancer risk. However, their low or absent expression in our chosen breast cell lines prevented further functional analysis. SNX32 (Sorting Nexin 32) is not well characterized, but ALK (Anaplastic lymphoma kinase) copy number gain and overexpression have been reported in aggressive and metastatic breast cancers[46]. Therapeutic targeting of ALK rearrangement has significantly improved survival in advanced ALK-positive lung cancer[47], making it an attractive target for breast and other cancers. BTN3A2 is a member of the B7/butyrophilin-like group of Ig superfamily receptors modulating the function of T-lymphocytes. While the exact role of BTN3A2 remains unknown, over-expression of this gene in epithelial ovarian cancer is associated with higher infiltrating immune cells and a better prognosis[48].

Our analyses identified multiple genes with reduced expression levels associated with increased breast cancer risk. Among them, LRRC3B and CASP8 are putative tumor suppressors in multiple cancers, including breast cancer. Leucine-rich repeat-containing 3B (LRRC3B) is a putative LRR-containing transmembrane protein, which is frequently inactivated via promoter hypermethylation leading to inhibition of cancer cell growth, proliferation, and invasion[49]. CASP8 encodes a member of the cysteine-aspartic acid protease family, which play a central role in cell apoptosis. Previous studies have suggested that caspase-8 may act as a tumor suppressor in certain types of lung cancer and neuroblastoma, although this function has not yet been demonstrated in breast cancer. Notably, several large association studies have identified SNPs at the 2q33/CASP8 locus associated with increased breast cancer risk[31,50]. Consistent with our data, eQTL analyses showed that the risk alleles for breast cancer were associated with reduced CASP8 mRNA levels in both peripheral blood lymphocytes and normal breast tissue[31].

For seven of the genes listed in Tables 1 and 2, we found some evidence from studies using tumor tissues, in vitro or in vivo experiments linking them to cancer risk (**Supplementary Table 10)**, although their association with breast cancer has not been previously demonstrated in human studies. For five of them, including LRRC3B, SPATA18, RIC8A, ALK and CRHR1, previous in vitro and in vivo experiments and human tissue studies showed a consistent direction of the association as demonstrated in our studies. For two other genes (UBD and MIR31HG), however, results from previous studies were inconsistent, reporting both potential promoting and inhibiting effects on breast cancer development. Future studies are needed to evaluate functions of these genes.

We included a large number of cases and controls in this study, providing strong statistical power for the association analysis. This large sample size enabled us to identify a large number of candidate breast cancer susceptibility genes, much larger than the number identified in a TWAS study with a sample size of about 20% of ours[30]. The previous study included subjects of different races, which could affect the results as linkage disequilibrium (LD) patterns differ by races. Of the five genes reported in that smaller TWAS that showed a suggestive association with breast cancer risk, the association for the RCCD1 gene was replicated in our study **(Table 3)**. The other four genes (ANKLE1, DHODH, ACAP1 and LRRC25) were not evaluated in our study because of unsatisfactory performance of our breast specific models for these genes which were built using the GTEx reference dataset including only female European descendants. In our study, the expression prediction model for ANKLE1 has a marginal performance in predicting gene expression (R^2=0.013 in the GTEx). The model, however, did not perform well in the TCGA data. For ACAP1 and LRRC25, previous results for suggestive associations were based on blood tissue models.

A substantial proportion of SNPs included in the OncoArray and iCOGS were selected from breast cancer GWAS and fine-mapping analyses, and thus these arrays were enriched for association signals with breast cancer risk. As a result, the overall λ value for the BCAC association analyses of individual variants is 1.26 after adjusting for population stratifications (QQ plot in **Supplementary Figure 3 (B)**)[7]. The λ value for the associations of the ~257,000

SNPs included in the gene expression prediction models of the 8,597 genes tested in our association analysis is 1.40 (QQ plot in **Supplementary Figure 3 (C)**). This higher λ value is perhaps expected because of a potential further enrichment of breast cancer associated signals in the set of SNPs selected to predict gene expression. There could be additional gain of power (and thus a higher λ value) in TWAS as it aggregates the effect of multiple SNPs to predict gene expression and use genes as the unit for association analyses. The lambda (λ) for our associated analyses of 8,597 genes was 1.51 (QQ plot presented in **Supplementary Figure 3 (A)**) likely due to the potential enrichment and power gain discussed above as well as our large sample size, and the highly polygenic nature of the disease[7,51]. Interestingly, high λ values were also found in recent large studies of other polygenic traits, such as body mass index (BMI) ($\lambda = 1.99$) and height ($\lambda = 2.7$)[52,53]. The $\lambda_{1,000}$, a standardized estimate of the genomic inflation scaling to a study of 1,000 cases and 1,000 controls, is 1.004 in our study.

The statistical power of our study is very large to detect associations for genes with a relatively high cis-heritability (h^2) (**Supplementary Figure 8)**. For example, our study has 80% statistical power to detect an association with breast cancer risk at $P<5.82\times10^{-6}$ with an OR of 1.07 or higher per one standard deviation increase (or decrease) in the expression level of genes with an h^2 of 0.1 or higher. One limitation of our study is the small sample size for building gene expression prediction models, which may have affected the precision of model parameter estimates. The prediction performance (R^2) for several of the genes identified in our study was not optimal, and thus additional research is needed to confirm our findings. We expect that models built with a larger sample size (and thus with more stable estimates of model parameters) will identify additional association signals. We used samples from women of European origin in model building, given differences in gene expression patterns between males and females and in genetic architecture across ethnicities[54]. We also used gene expression data of tumor-adjacent normal tissue samples from European descendants in TCGA as an external validation step to prioritize genes for association analyses. Given potential somatic alterations in tumor-adjacent normal tissues, we retained all models showing a prediction performance (R^2) of at least 0.09 in GTEx, regardless of their performance in TCGA. Not all genes have a significant hereditary component in expression regulation, and thus these genes could not be investigated in our study. For example, previous studies have provided strong evidence to support a significant role of the TERT, ESR1, CCND1, IGFBP5, TET2 and MRPS30 genes in the etiology of breast cancer. However, expression of these genes cannot be predicted well using the data from female European descendants included in the GTEx and thus they were not included in our association analyses. **Supplementary Table 11** summarizes the performance of prediction models and association results for breast cancer target genes reported previously at GWAS-identified loci.

In summary, our study has identified multiple gene candidates that can be further functionally characterized. By evaluating the associations of predicted gene expression levels with breast cancer risk, we provided evidence for the direction of the association for the identified genes. The silencing experiments we performed suggest that many of the genes identified by TWAS are likely to mediate risk of breast cancer by affecting proliferation or colony forming efficiency, two of the hallmarks of cancer. Further investigation of genes identified in our study will provide additional insight into the biology and genetics of breast cancer.

Methods

Building of gene expression prediction models

We used transcriptome and high-density genotyping data from the Genotype-Tissue Expression (GTEx) study to establish prediction models for genes expressed in normal breast tissues. Details of the GTEx have been described elsewhere[55]. Genomic DNA samples obtained from study subjects included in the GTEx were genotyped using Illumina OMNI 5M or 2.5M SNP Array and RNA samples from 51 tissue sites were sequenced to generate transcriptome profiling data. Genotype data were processed according to the GTEx protocol (http://www.gtexportal.org/home/documentationPage). SNPs with a call rate $< 98\%$, with differential missingness between the two array experiments (5M/2.5M Arrays), with Hardy-Weinberg equilibrium p-value $< 10^{-6}$ (among subjects of European ancestry), or showing batch effects were excluded. One Klinefelter individual, three related individuals, and a chromosome 17 trisomy individual were also excluded. The genotype data were imputed to the Haplotype Reference Consortium reference panel[56] using Minimac3 for imputation and SHAPEIT for prephasing[57,58]. SNPs with high imputation quality ($r^2 \geq 0.8$), minor allele frequency (MAF) ≥ 0.05, and included in the HapMap Phase 2 version, were used to build expression prediction models. For gene expression data, we used Reads Per Kilobase per Million (RPKM) units from RNA-SeQC[59]. Genes with a median expression level of 0 RPKM across samples were removed, and the RPKM values of each gene were log2 transformed. We performed quantile normalization to bring the expression profile of each sample to the same scale, and performed inverse quantile normalization for each gene to map each set of expression values to a standard normal. We adjusted for the top ten principal components (PCs) derived from genotype data and the top 15 probabilistic estimation of expression residuals (PEER) factors to correct for batch effects and experimental confounders in model building[60]. Genetic and transcriptome data from 67 female subjects of European descent without a prior breast cancer diagnosis were used to build gene expression prediction models for this study.

We built an expression prediction model for each gene by using the elastic net method as implemented in the glmnet R package, with α=0.5, as recommended by Gamazon et al[27]. The genetically regulated expression for each gene was estimated by including variants within a 2 MB window flanking the respective gene boundaries, inclusive. Expression prediction models were built for protein coding genes, long non-coding RNAs (lncRNAs), microRNAs (miRNAs), processed transcripts, immunoglobulin genes, and T cell receptor genes, according to categories described in the Gencode V19 annotation file (http://www.gencodegenes.org/releases/19.html). Pseudogenes were not included in the present study because of potential concerns of inaccurate calling[61]. Ten-fold cross-validation was used to validate the models internally. Prediction R^2 values (the square of the correlation between predicted and observed expression) were generated to estimate the prediction performance of each of the gene prediction models established.

For genes that cannot be predicted well using the above approach, we built models using only SNPs located in predicted promoter or enhancer regions in breast cell lines. This approach reduces the number of variants for model building, and thus potentially improves model accuracy, by increasing the ratio of sample size to effective degrees of freedom. SNP-level annotation data in three breast cell lines, namely, Breast Myoepithelial Primary Cells (E027), Breast variant Human Mammary Epithelial Cells (vHMEC) (E028), and HMEC Mammary Epithelial Primary Cells (E119) in the Roadmap Epigenomics Project/Encyclopedia of DNA Elements Project[16], were downloaded from http://archive.broadinstitute.org/mammals/haploreg/data/ (Version 4.0, assessed on December 6, 2016). SNPs in regions classified as promoters (TssA, TssAFlnk), enhancers (Enh, EnhG), or regions with both promoter and enhancer signatures (ExFlnk) according to the core 15 chromatin state model[16] in at least one of the cell lines were retained as input SNPs for model building.

Evaluating performance of gene expression prediction models using The Cancer Genome Atlas (TCGA) data

To assess further the validity of the models, we performed external validation using data generated in tumor-adjacent normal breast tissue samples obtained from 86 European-ancestry female breast cancer patients included in the TCGA. Genotype data were imputed using the same approach as described for GTEx data. Expression data were processed and normalized using a similar approach as described above. The predicted expression level for each gene was calculated using the model established using GTEx data and then compared with the observed level of that gene using the Spearman's correlation.

Evaluating statistical power for association tests

We conducted a simulation analysis to assess the power of our TWAS analysis. Specifically, we set the number of cases and controls to be 122,977 and 105,974, respectively, and generated the gene expression levels from the empirical distribution of predicted gene expression levels in the BCAC. We calculated statistical power at $P<5.82\times10^{-6}$ (the significance level used in our TWAS) according to cis-heritability (h^2) which we aim to capture using gene expression prediction models (R^2). The results based on 1000 replicates are summarized in **Supplementary Figure 8**. Based on the power calculation, our TWAS analysis has 80% power to detect a minimum odds ratio of 1.11, 1.07, 1.05, 1.04, or 1.03 for breast cancer risk per one standard deviation increase (or decrease) in the expression level of a gene whose cis-heritability is 5%, 10%, 20%, 40%, or 60%, respectively.

Association analyses of predicted gene expression with breast cancer risk

We used the following criteria to select genes for the association analysis: 1) with a model prediction R^2 of ≥ 0.01 in GTEx and a Spearman's correlation coefficient of ≥ 0.1 in TCGA, 2) with a prediction R^2 of ≥ 0.09 in GTEx regardless of the performance in TCGA, 3) with a prediction R^2 of ≥ 0.01 in GTEx but unable to be evaluated in TCGA. The second group of genes was selected because some gene expression levels might have changed in TCGA tumor-adjacent normal tissues, and thus it is anticipated that some genes may show low prediction performance in TCGA data due to the influence of tumor growth[62,63]. Overall, a total of 8,597 genes met the criteria and were evaluated for their expression-trait associations.

To identify novel breast cancer susceptibility loci and genes, the MetaXcan method, as described elsewhere, was used for the association analyses[26]. Briefly, the formula:

$$Z_g \approx \sum_{l \in \mathrm{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)}$$

was used to estimate the Z-score of the association between predicted expression and breast cancer risk. Here w_{lg} is the weight of SNP l for predicting the expression of gene g, $\hat{\beta}_l$and $\mathrm{se}(\hat{\beta}_l)$ are the GWAS association regression coefficient and its standard error for SNP l, and $\hat{\sigma}_l$ and $\hat{\sigma}_g$ are the estimated variances of SNP l and the predicted expression of gene g respectively. Therefore, the weights for predicting gene expression, GWAS summary statistics results, and correlations between model predicting SNPs are the input variables for the MetaXcan analyses. For this study we estimated correlations between SNPs included in the prediction models using the phase 3, 1000 Genomes Project data focusing on European population.

For the association analysis, we used the summary statistics data of genetic variants associated with breast cancer risk generated in 122,977 breast cancer patients and 105,974 controls of European ancestry from the Breast Cancer Association Consortium (BCAC). The details of the BCAC have been described elsewhere[7,9,13,64,65]. Briefly, 46,785 breast cancer cases and 42,892 controls of European ancestry were genotyped using a custom Illumina iSelect genotyping array (iCOGS) containing ~211,155 variants. A further 61,282 cases and 45,494 controls of European ancestry were genotyped using the OncoArray including 570,000 SNPs (http://epi.grants.cancer.gov/oncoarray/). Also included in this analysis were data from nine GWAS studies including 14,910 breast cancer cases and 17,588 controls of European ancestry. Genotype data from iCOGS, OncoArray and GWAS were imputed using the October 2014 release of the 1000 Genomes Project data as reference. Genetic association results for breast cancer risk were combined using inverse variance fixed effect meta-analyses[7]. For our study, only SNPs with imputation $r^2 \geq 0.3$ were used. All participating BCAC studies were approved by their appropriate ethics review boards. This study was approved by the BCAC Data Access Coordination Committee.

Lambda 1,000 ($\lambda_{1,000}$) was calculated to represent a standardized estimate of the genomic inflation scaling to a study of 1,000 cases and 1,000 controls, using the following formula: $\lambda_{1,000}=1+(\lambda_{obs}-1) \times (1/n_{cases}+1/n_{controls})/(1/1,000_{cases}+1/1,000_{controls})$[66,67]. We used a Bonferroni corrected p threshold of 5.82×10^{-6} (0.05/8,597) to determine a statistically significant association for the primary analyses. To identify additional gene candidates at previously identified susceptibility loci, we also used a false discovery rate (FDR) corrected p threshold of 1.05×10^{-3} (FDR $\leq$ 0.05) to determine a significant association. Associated genes with an expression of >0.1 RPKM in less than 10 individuals in GTEx data were excluded as the corresponding prediction models may not be stable.

To determine whether the predicted expression-trait associations were independent of the top signals identified in previous GWAS, we performed GCTA-COJO analyses developed by Yang et al[36] to calculate association betas and standard errors of variants with breast cancer risk after adjusting for the index SNPs of interest. We then re-ran the MetaXcan analyses using the association statistics after conditioning on the index SNPs. This information was used to determine whether the detected expression-trait associations remained significant after adjusting for the index SNPs.

For 41 identified associated genes at the Bonferroni-corrected threshold, we also performed analyses using individual level data in iCOGS (n=84,740) and OncoArray (n=112,133) datasets. We generated predicted gene expression using predicting SNPs, and then assessed the association between predicted gene expression and breast cancer risk adjusting for study and nine principal components in iCOGS dataset, and country and the first ten principal components in OncoArray dataset. Conditional analyses adjusting for index SNPs were performed to assess potential influence of reported index SNPs on the association between predicted gene expression and breast cancer risk. Furthermore, we evaluated whether the predicted expression levels of genes within a same genomic region were correlated with each other by using the OncoArray data.

INQUISIT algorithm scores for TWAS-identified genes

To evaluate whether there are additional lines of evidence supporting the identified genes as putative target genes of GWAS identified risk SNPs beyond the scope of eQTL, we assessed their INQUISIT algorithm scores, which have been described elsewhere[7]. Briefly, this approach evaluates chromatin interactions between distal and proximal regulatory transcription-factor binding sites and the promoters at the risk regions using Hi-C data generated in HMECs[68] and Chromatin Interaction Analysis by Paired End Tag (ChiA-PET) in MCF7 cells. This could detect genome-wide interactions brought about by, or associated with, CCCTC-binding factor (CTCF), DNA polymerase II (POL2), and Estrogen Receptor (ER), all involved in transcriptional regulation[68]. Annotation of predicted target genes used the Integrated Method for Predicting Enhancer Targets (IM-PET)[69], the Predicting Specific Tissue Interactions of Genes and Enhancers (PreSTIGE) algorithm[70], Hnisz[71] and FANTOM[72]. Features contributing to the scores are based on functionally important genomic annotations such as chromatin interactions, transcription factor binding, and eQTLs. The detailed information for the INQUISIT pipeline and scoring strategy has been included in a previous publication[7]. In brief, besides assigning integral points according to different features, we also set up-weighting and down-weighting criteria according to breast cancer driver genes, topologically associated domain (TAD) boundaries, and gene expression levels in relevant breast cell lines. Scores in the distal regulation category range from 0-7, and in the promoter category from 0-4. A score of "none" represents that no evidence was found for regulation of the corresponding gene.

Functional enrichment analysis using Ingenuity Pathway Analysis (IPA)

We performed functional enrichment analysis for the identified protein-coding genes reaching Bonferroni corrected association threshold. To assess potential functionality of the identified lncRNAs, we examined their co-expressed protein-coding genes determined using expression data of normal breast tissue of European females in GTEx. Spearman's correlations between protein-coding genes and identified lncRNAs of ≥ 0.4 or ≤ -0.4 were used to indicate a high co-expression. Canonical pathways, top associated diseases and biofunctions, and top networks associated with genes of interest were estimated using IPA software[37].

Gene expression in breast cell lines

Total RNA was isolated from 18 cell lines (**Supplementary Table 8**) using the RNeasy Mini Kit (Qiagen). cDNA was synthesized using the SuperScript III (Invitrogen) and amplified using the Platinum SYBR Green qPCR SuperMix-UDG cocktail (Invitrogen). Two or three primer pairs were used for each gene and the mRNA levels for each sample was measured in technical triplicates for each primer set. The primer sequences are listed in **Supplementary Table 12**. Experiments were performed using an ABI ViiA(TM) 7 System (Applied Biosystems), and data processing was performed using ABI QuantStudio™ Software V1.1 (Applied Biosystems). The average of Ct from all the primer pairs for each gene was used to calculate ΔC_T. The relative quantitation of each mRNA normalizing to that in 184A1 was performed using the comparative Ct method ($\Delta\Delta C_T$) and summarized in **Supplementary Figure 4**.

Short interfering RNA (siRNA) silencing

MCF7 and T47D cells were reverse-transfected with siRNAs targeting genes of interest (GOI) or a non-targeting control siRNA (consi; Shanghai Genepharma) with RNAiMAX (Invitrogen) according to the manufacturer's protocol. Verification of siRNA knockdown of gene expression by qPCR was performed 36 hours after transfection.

Proliferation and colony formation assays

For proliferation assays, MCF7 and T47D cells were trypsinized at 16 hours post-transfection and seeded into 24 well plates to achieve ~10% confluency. Phase-contrast images were collected with IncuCyte ZOOM (Essen Bioscience) for seven days. Duplicate samples were assessed for each GOI siRNA transfected cells along with non-target control si (NTCsi) treated cells in the same plate. 184A1 cells were reverse-transfected in 96 well plates to achieve 50% confluence at 8 hours after transfection. Two independent experiments were carried out for all siRNAs in all three cell lines. Each cell proliferation time-course was normalized to the baseline confluency and analyzed in GraphPad Prism. The area under the curve was calculated for each concentration (n=4) and used to calculate corrected proliferation (Corrected proliferation % = 100 +/- (relative proliferation in indicated siRNA - proliferation in NTC siRNA) / knockdown efficiency ("+" if the GOI promotes proliferation and "-" if it inhibits proliferation)). For each gene, results from two siRNAs in two independent experiments were averaged and summarized in **Figure 2** and **Supplementary Figure 6**. For colony formation assays; the same number of GOI siRNA transfected MCF7 cells was seeded in 6 well plates at 16 hours after transfection to assay colony forming efficiency at two weeks. All siRNA-treated cells were seeded in duplicate. Colonies (defined to consist of at least 50 cells) were fixed with methanol, stained with crystal violet (0.5% w/v), scanned and counted using ImageJ as batch analysis by a self-defined plug-in

Macro. Correct CFE % = 100 +/- (relative CFE in indicated siRNA - CFE in NTC siRNA) / knockdown efficiency ("+" if the GOI promotes CF and "-" if it inhibits CF). For each gene, results from two siRNAs in two independent experiments were averaged and summarized in **Figure 2** and **Supplementary Figure 7**.

Data availability

The GTEx data are publicly available via dbGaP (www.ncbi.nlm.nih.gov/gap; dbGaP Study Accession: phs000424.v6.p1). TCGA data are publicly available via National Cancer Institute's Genomic Data Commons Data Portal (https://gdc.cancer.gov/). Most of the BCAC data used in this study are or will be publicly available via dbGAP. Data from some BCAC studies are not publicly available due to restraints imposed by the ethics committees of individual studies; requests for further data can be made to the BCAC (http://bcac.ccge.medschl.cam.ac.uk/) Data Access Coordination Committee.

Code availability

The computer codes used in our study are available upon reasonable request.

Acknowledgements

The authors thank Jing He, Wanqing Wen, Ayush Giri, and Todd Edwards of Vanderbilt Epidemiology Center and Rao Tao of Department of Biostatistics, Vanderbilt University Medical Center for their help with the data analysis of this study. The authors also would like to thank all the individuals for their participation in the parent studies and all the researchers, clinicians, technicians and administrative staff for their contribution to the studies. We are also grateful to

Hae Kyung Im of University of Chicago for her help. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. This project at Vanderbilt University Medical Center was supported in part by grants R01CA158473 and R01CA148677 from the U.S. National Institutes of Health as well as funds from Anne Potter Wilson endowment. Dr. Wu was supported by the Vanderbilt Molecular and Genetic Epidemiology of Cancer (MAGEC) training program funded by the US National Cancer Institute grant R25 CA160056 (PI: X.-O. Shu). Genotyping of the OncoArray was principally funded from three sources: the PERSPECTIVE project, funded from the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the Ministère de *l'Économie, de la Science et de l'Innovation du Québec* through Genome Québec, and the Quebec Breast Cancer Foundation; the NCI Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative and Discovery, Biology and Risk of Inherited Variants in Breast Cancer (DRIVE) project (NIH Grants U19 CA148065 and X01HG007492); and Cancer Research UK (C1287/A10118 and. C1287/A16563). **BCAC** is funded by Cancer Research UK [C1287/A16563], by the European Community's Seventh Framework Programme under grant agreement 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union's Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec – grant # PSR-SIIRI-701. Combining the GWAS data was supported in part by The National Institute of Health (NIH) Cancer Post-Cancer GWAS initiative grant U19 CA 148065 (DRIVE, part of the GAME-

ON initiative). A full description of funding and acknowledgments for BCAC studies are included in the Acknowledgments for BCAC studies section of the **Supplementary Material**.

Author Contributions

W.Z. and J.L. conceived the study. L.W. contributed to the study design, and performed statistical analyses. L.W., W.Z. and G.C.-T. wrote the manuscript with significant contributions from W.S., J.L., X.G., and S.L.E.. W.S. performed the *in vitro* experiments. G.C.-T. directed the *in vitro* experiments. X.G. contributed to the model building and pathway analyses. J.B. contributed to the bioinformatics analyses. F.A.-E., E.R., and S.L.E. contributed to the *in vitro* experiments. Y. L. and C. Z. contributed to the model building. K.M., M.K.B., X.-O.S., Q.W., J.D., B.L., C.Z., H.F., A.G., R.T.B., A.M.D., P.D.P.P., J.S., R.L.M., P.K., and D.F.E, contributed to manuscript revision, statistical analyses and/or BCAC data management. I.L.A., H.A.-C., V.A., K.J.A., P.L.A., M. Barrdahl, C.B., M.W.B., J.B., M. Bermisheva, C.B., N.V.B., S.E.B., H. Brauch, H. Brenner, L.B., P.B., S.Y.B., B.B., Q.C., T.C., F.C., B.D.C., J.E.C., J.C.-C., X.C., T.-Y.D.C., H.C., C.L.C., NBCS Collaborators, M.C., S.C., F.J.C., D.C., A.C., S.S.C., J.M.C., K.C., M.B.D., P.D., K.F.D., T.D., I.d.S.S., M. Dumont, M. Dwek, D.M.E., U.E., H.E., C.E., M.E., L.F., P.A.F., J.F., D.F.-J., O.F., H.F., L.F., M. Gabrielson, M.G.-D., S.M.G., M.G.-C., M.M.G., M. Ghoussaini, G.G.G., M.S.G., D.E.G., A.G.-N., P.G., E. Hahnen, C.A.H., N.H., P. Hall, E. Hallberg, U.H., P. Harrington, A. Hein, B.H., P. Hillemanns, A. Hollestelle, R.N.H., J.L.H., G.H., K.H., D.J.H., A.J., W.J., E.M.J., N.J., K.J., M.E.J., A. Jung, R.K., M.J.K., E.K., V.-M.K., V.N.K., D.L., L.L.M., J. Li, S.L., J. Lissowska, W.-Y.L., S.Loibl, J.L., C.L., M.P.L., R.J.M., T.M., I.M.K., A. Mannermaa, J.E.M., S.M., D.M., H.M.-H., A. Meindl, U.M., J.M., A.M.M., S.L.N., H.N., P.N., S.F.N., B.G.N., O.I.O., J.E.O., H.O., P.P., J.P., D.P.-K., R.P., N.P., K.P.,

B.R., P.R., N.R., G.R., H.S.R., V.R., A. Romero, J.R., A. Rudolph, E.S., D.P.S, E.J.S., M.K.S., R.K.S., A.S., R.J.S., C. Scott, S.S., M.S., M.J.S., A.S., M.C.S., J.J.S., J.S., H.S., A.J.S., R.T., W.T., J.A.T., M.B.T., D.C.T., A.T., K.T., R.A.E.M.T., D.T., T.T., M.U., C.V., D.V.D.B., D.V., Q.W., C.R.W., C.W., A.S.W., H.W., W.C.W., R.W., A.W., L.X., X.R.Y., A.Z., E.Z., kConFab/AOCS Investigators contributed to the collection of the data and biological samples for the original BCAC studies. All authors have reviewed and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

1. Kamangar, F., Dores, G.M. & Anderson, W.F. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. J Clin Oncol **24**, 2137-50 (2006).
2. Beggs, A.D. & Hodgson, S.V. Genomics and breast cancer: the different levels of inherited susceptibility. Eur J Hum Genet **17**, 855-6 (2009).
3. Southey, M.C. et al. PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. J Med Genet (2016).
4. Nathanson, K.L., Wooster, R. & Weber, B.L. Breast cancer genetics: what we know and what we need. Nat Med **7**, 552-6 (2001).
5. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. Br J Cancer **83**, 1301-8 (2000).
6. Milne, R.L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat Genet **49**, 1767-1778 (2017).
7. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. Nature **551**, 92-94 (2017).
8. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet **45**, 353-61, 361e1-2 (2013).
9. Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet **47**, 373-80 (2015).
10. Cai, Q. et al. Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. Nat Genet **46**, 886-90 (2014).
11. Zheng, W. et al. Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. Hum Mol Genet **22**, 2539-50 (2013).
12. Zhang, B., Beeghly-Fadiel, A., Long, J. & Zheng, W. Genetic variants associated with breast-cancer risk: comprehensive research synopsis, meta-analysis, and epidemiological evidence. Lancet Oncol **12**, 477-88 (2011).
13. French, J.D. et al. Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am J Hum Genet **92**, 489-503 (2013).
14. Hindorff, L.A. et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A **106**, 9362-7 (2009).
15. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. Nature **489**, 57-74 (2012).
16. Roadmap Epigenomics, C. et al. Integrative analysis of 111 reference human epigenomes. Nature **518**, 317-30 (2015).
17. Dunning, A.M. et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet **48**, 374-86 (2016).
18. Ghoussaini, M. et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nat Commun **4**, 4999 (2014).

19. Li, Q. et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell **152**, 633-41 (2013).
20. Darabi, H. et al. Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. Am J Hum Genet **97**, 22-34 (2015).
21. Glubb, D.M. et al. Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. Am J Hum Genet **96**, 5-20 (2015).
22. Lawrenson, K. et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast-ovarian cancer susceptibility locus. Nat Commun **7**, 12675 (2016).
23. Lee, D. et al. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet **47**, 955-61 (2015).
24. Finucane, H.K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet **47**, 1228-35 (2015).
25. Gusev, A. et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. Am J Hum Genet **95**, 535-52 (2014).
26. Barbeira, A.N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. bioRxiv (2017).
27. Gamazon, E.R. et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet **47**, 1091-8 (2015).
28. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet **48**, 245-52 (2016).
29. Zhu, Z. et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nat Genet **48**, 481-7 (2016).
30. Hoffman, J.D. et al. Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. PLoS Genet **13**, e1006690 (2017).
31. Lin, W.Y. et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. Hum Mol Genet **24**, 285-98 (2015).
32. Camp, N.J. et al. Discordant Haplotype Sequencing Identifies Functional Variants at the 2q33 Breast Cancer Risk Locus. Cancer Res **76**, 1916-25 (2016).
33. Li, Q. et al. Expression QTL-based analyses reveal candidate causal genes and loci across five tumor types. Hum Mol Genet **23**, 5294-302 (2014).
34. Caswell, J.L. et al. Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors. Hum Mol Genet **24**, 7421-31 (2015).
35. Darabi, H. et al. Fine scale mapping of the 17q22 breast cancer locus using dense SNPs, genotyped within the Collaborative Oncological Gene-Environment Study (COGs). Sci Rep **6**, 32512 (2016).
36. Yang, J. et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet **44**, 369-75, S1-3 (2012).
37. Kramer, A., Green, J., Pollard, J., Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. Bioinformatics **30**, 523-30 (2014).
38. Koh, J.L. et al. COLT-Cancer: functional genetic screening resource for essential genes in human cancer cell lines. Nucleic Acids Res **40**, D957-63 (2012).
39. Marcotte, R. et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov **2**, 172-89 (2012).

40. Walen, K.H. & Stampfer, M.R. Chromosome analyses of human mammary epithelial cells at stages of chemical-induced transformation progression to immortality. Cancer Genet Cytogenet **37**, 249-61 (1989).
41. Treszezamsky, A.D. et al. BRCA1- and BRCA2-deficient cells are sensitive to etoposide-induced DNA double-strand breaks via topoisomerase II. Cancer Res **67**, 7078-81 (2007).
42. Marcotte, R. et al. Essential gene profiles in breast, pancreatic, and ovarian cancer cells. Cancer Discov **2**, 172-189 (2012).
43. Sanchez, Y. et al. Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. Nat Commun **5**, 5812 (2014).
44. Li, Y., Peart, M.J. & Prives, C. Stxbp4 regulates DeltaNp63 stability by suppression of RACK1-dependent degradation. Mol Cell Biol **29**, 3953-63 (2009).
45. Sekine, Y. et al. The Kelch repeat protein KLHDC10 regulates oxidative stress-induced ASK1 activation by suppressing PP5. Mol Cell **48**, 692-704 (2012).
46. Kim, M.H. et al. Anaplastic lymphoma kinase gene copy number gain in inflammatory breast cancer (IBC): prevalence, clinicopathologic features and prognostic implication. PLoS One **10**, e0120320 (2015).
47. Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer. N Engl J Med **373**, 1582 (2015).
48. Le Page, C. et al. BTN3A2 expression in epithelial ovarian cancer is associated with higher tumor infiltrating T cells and a better prognosis. PLoS One **7**, e38541 (2012).
49. Kan, L. et al. LRRC3B is downregulated in non-small-cell lung cancer and inhibits cancer cell proliferation and invasion. Tumour Biol **37**, 1113-20 (2016).
50. Cox, A. et al. A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet **39**, 352-8 (2007).
51. Yang, J. et al. Genomic inflation factors under polygenic inheritance. Eur J Hum Genet **19**, 807-12 (2011).
52. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. Nature **542**, 186-190 (2017).
53. Turcot, V. et al. Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity. Nat Genet **50**, 26-41 (2018).
54. Mele, M. et al. Human genomics. The human transcriptome across tissues and individuals. Science **348**, 660-5 (2015).
55. Consortium, G.T. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science **348**, 648-60 (2015).
56. McCarthy, S. et al. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet **48**, 1279-83 (2016).
57. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. Nat Methods **9**, 179-81 (2012).
58. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet **5**, e1000529 (2009).
59. DeLuca, D.S. et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics **28**, 1530-2 (2012).

60. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc **7**, 500-7 (2012).
61. Guo, X., Lin, M., Rockowitz, S., Lachman, H.M. & Zheng, D. Characterization of human pseudogene-derived non-coding RNAs for functional potential. PLoS One **9**, e93972 (2014).
62. Casbas-Hernandez, P. et al. Tumor intrinsic subtype is reflected in cancer-adjacent tissue. Cancer Epidemiol Biomarkers Prev **24**, 406-14 (2015).
63. Huang, X., Stern, D.F. & Zhao, H. Transcriptional Profiles from Paired Normal Samples Offer Complementary Information on Cancer Patient Survival--Evidence from TCGA Pan-Cancer Data. Sci Rep **6**, 20567 (2016).
64. Ghoussaini, M. et al. Genome-wide association analysis identifies three new breast cancer susceptibility loci. Nat Genet **44**, 312-8 (2012).
65. Garcia-Closas, M. et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. Nat Genet **45**, 392-8, 398e1-2 (2013).
66. Devlin, B. & Roeder, K. Genomic control for association studies. Biometrics **55**, 997-1004 (1999).
67. Freedman, M.L. et al. Assessing the impact of population stratification on genetic association studies. Nat Genet **36**, 388-93 (2004).
68. Rao, S.S. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell **159**, 1665-80 (2014).
69. He, B., Chen, C., Teng, L. & Tan, K. Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A **111**, E2191-9 (2014).
70. Corradin, O. et al. Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res **24**, 1-13 (2014).
71. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. Cell **155**, 934-47 (2013).
72. Consortium, F. et al. A promoter-level mammalian expression atlas. Nature **507**, 462-70 (2014).

Figure Legends

Figure 1. Manhattan plot of association results from the breast cancer transcriptome-wide association study. The red line represents $P = 5.82 \times 10^{-6}$. The blue line represents $P = 1.00 \times 10^{-3}$.

Figure 2. Heat maps of proliferation and colony formation efficiency in breast cells. (A) 184A1, MCF7 or T47D cells were transfected with indicated siRNAs over seven days and phase-contrast images collected using an IncuCyte ZOOM. Each cell proliferation time-course was normalized to the baseline confluency and analyzed using GraphPad Prism. Corrected proliferation % = 100 +/- (relative proliferation in indicated siRNA - proliferation in control siRNA (consi))/knockdown efficiency. **(B)** MCF7 cells were transfected with indicated siRNAs, then reseeded after 16 hours for colony formation (CF) assay. At day 14, colonies were fixed with methanol, stained with crystal violet, scanned and batch analyzed by ImageJ. Corrected CF efficiency (CFE) % = 100 +/- (relative CFE in indicated siRNA - CFE in control siRNA (consi))/knockdown efficiency. Error bars, SD (N=2). P-values were determined by one-way ANOVA followed by Dunnett's multiple comparisons test: *P-value < 0.05. NTC: non-target control.

Table 1. Fourteen expression-trait associations for genes located at genomic loci at least 500 kb away from any GWAS-identified breast cancer risk variants

Region	Gene[a]	Type[b]	Z score	P value[c]	R^2[c]	Closest risk SNP[d]	Distance to the closest risk SNP (kb)	P value after adjusting for adjacent risk SNPs[e]
1p34.1	**ZSWIM5**	Protein	5.26	1.43×10^{-7}	0.17	rs1707302	829	0.006
3p24.1	LRRC3B	Protein	-9.57	1.11×10^{-21}	0.17	rs653465	591	1.60×10^{-6}
4q12	SPATA18	Protein	-4.62	3.86×10^{-6}	0.11	rs6815814	14,101	3.98×10^{-6}
6p22.1	UBD	Protein	-4.87	1.10×10^{-6}	0.13	rs9257408	597	0.94
7q32.2	**KLHDC10**	Protein	5.21	1.92×10^{-7}	0.14	rs4593472	892	2.90×10^{-7}
9p21.3	MIR31HG	lncRNA	-5.02	5.22×10^{-7}	0.12	rs1011970	502	1.23×10^{-7}
11p15.5	RIC8A	Protein	-5.27	1.40×10^{-7}	0.15	rs6597981	588	4.95×10^{-6}
11q13.2	B3GNT1	Protein	-5.85	4.88×10^{-9}	0.09	rs3903072	530	3.50×10^{-6}
11q13.2	RP11-867G23.10	transcript	4.71	2.49×10^{-6}	0.03	rs3903072	594	2.61×10^{-4}
12p13.33	**RP11-218M22.1**	lncRNA	5.02	5.27×10^{-7}	0.19	rs12422552	13,641	5.17×10^{-7}
14q24.1	GALNT16	Protein	-8.27	1.38×10^{-16}	0.04	rs999737	691	8.57×10^{-4}
14q24.1	**PLEKHD1**	Protein	7.50	6.55×10^{-14}	0.02	rs999737	917	0.12
15q24.2	MAN2C1 [f]	Protein	-5.32	1.02×10^{-7}	0.39	rs2290203	15,851	9.56×10^{-8}
15q24.2	CTD-2323K18.1 [f]	lncRNA	-4.65	3.27×10^{-6}	0.07	rs2290203	15,619	3.16×10^{-6}

[a] Genes that were siRNA-silenced for functional assays are bolded; SNPs used to predict gene expression are listed in the Supplementary Table 13

[b] Protein: protein coding genes; lncRNA: long non-coding RNAs; transcript: processed transcript

[c] P value: derived from association analyses; associations with $p \leq 5.82 \times 10^{-6}$ considered statistically significant based on Bonferroni correction of 8,597 tests (0.05/8,597); R^2: prediction performance (R^2) derived using GTEx data.

[d] Risk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes are presented in the **Supplementary Table 4**

[e] Use of COJO method[36]

[f] Predicted expression of MAN2C1 and CTD-2323K18.1 was correlated (spearman R=0.76)

Table 2. Twenty-three expression-trait associations for genes located at genomic loci within 500 kb of any previous GWAS-identified breast cancer risk variants but not yet implicated as target genes of risk variants[#]

Region	Gene[a]	Type[b]	Z score	P value[c]	R^2[c]	Closest risk SNP[d]	Distance to the closest risk SNP (kb)	P value after adjusting for adjacent risk SNPs[e]
1p11.2	RP11-439A17.7	lncRNA	-5.34	9.07×10^{-8}	0.22	rs11249433	442	0.02
1q21.1	NUDT17	Protein	-6.27	3.58×10^{-10}	0.01	rs12405132	56	0.08
1q21.1	ANKRD34A	Protein	-5.05	4.42×10^{-7}	0.01	rs12405132	169	4.28×10^{-5}
2p23.1-2p23.2	ALK	Protein	4.67	3.06×10^{-6}	0.06	rs4577244	295	2.70×10^{-6}
3p21.31	PRSS46	Protein	-5.83	5.68×10^{-9}	0.13	rs6796502	89	0.002
3q12.2	RP11-114I8.4	lncRNA	-5.84	5.19×10^{-9}	0.02	rs9833888	356	0.09
5p12	RP11-53O19.1	lncRNA	10.38	2.94×10^{-25}	0.03	rs10941679	39	7.46×10^{-4}
5q33.3	**UBLCP1**	Protein	5.93	3.04×10^{-9}	0.07	rs1432679	446	0.37
5q33.3	RP11-32D16.1	lncRNA	-5.41	6.37×10^{-8}	0.09	rs1432679	283	1.32×10^{-4}
6p22.2	BTN3A2	Protein	4.61	3.97×10^{-6}	0.28	rs71557345	229	0.72
6q23.1	RP11-73O6.3 [f]	lncRNA	-6.61	3.74×10^{-11}	0.11	rs6569648	105	0.41
11p15.5	**AP006621.6** [g]	lncRNA	5.61	2.01×10^{-8}	0.34	rs6597981	21	0.52
11p15.5	RPLP2 [g]	Protein	4.64	3.46×10^{-6}	0.27	rs6597981	7	0.51
14q32.33	CTD-3051D23.1	lncRNA	-5.06	4.21×10^{-7}	0.05	rs10623258	97	7.05×10^{-7}
16q12.2	**RP11-467J12.4**	lncRNA	8.04	9.02×10^{-16}	0.23	rs3112612	434	0.79
16q12.2	**CTD-3032H12.1**	lncRNA	4.92	8.58×10^{-7}	0.03	rs28539243	290	0.006
17q21.31	LRRC37A [g]	Protein	-5.89	3.85×10^{-9}	0.43	rs2532263	118	0.79
17q21.31	KANSL1-AS1 [g]	lncRNA	-5.58	2.44×10^{-8}	0.62	rs2532263	18	0.95
17q21.31	CRHR1 [g]	Protein	-5.29	1.22×10^{-7}	0.22	rs2532263	339	0.99
17q21.31	LINC00671	lncRNA	-5.85	4.95×10^{-9}	0.07	rs72826962	190	0.26
17q21.31	LRRC37A2	Protein	-5.77	7.93×10^{-9}	0.46	rs2532263	336	0.93
19p13.11	HAPLN4	Protein	-7.13	9.88×10^{-13}	0.02	rs2965183	172	0.22
19q13.31	**RP11-15A1.7** [h]	lncRNA	5.45	5.06×10^{-8}	0.02	rs3760982	215	0.28

[#] not yet reported from eQTL and/or functional studies as target genes of GWAS-identified risk variants and not harbor GWAS or fine-mapping identified risk variants

[a] Genes that were siRNA-silenced for functional assays are bolded; SNPs used to predict gene expression are listed in the Supplementary Table 13

[b] Protein: protein coding genes; lncRNA: long non-coding RNAs

[c] P value: nominal P value from association analysis; the threshold after Bonferroni correction of 8,597 tests ($0.05/8,597=5.82\times10^{-6}$) was used; R^2: prediction performance (R^2) derived using GTEx data

[d] Risk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes are presented in the **Supplementary Table 4**

[e] Use of COJO method[36]; all index SNPs in the corresponding region were adjusted in the conditional analyses

[f] Predicted expression of RP11-73O6.3 and L3MBTL3 was correlated (spearman R=0.88)

[g] Predicted expression of AP006621.6 and RPLP2 was correlated; predicted expression of LRRC37A, KANSL1-AS1, and CRHR1 was correlated (spearman R>0.1)

[h] Predicted expression of RP11-15A1.7 and ZNF404 was correlated (spearman R=0.64)

Table 3. Eleven expression-trait associations for genes previously reported as potential target genes of GWAS-identified breast cancer risk variants or genes harboring risk variants

Region	**Gene**[a]	**Type**[b]	**Z score**	**P value**[c]	**R^2**[c]	**Closest risk SNP**[d]	**Distance to the closest risk SNP (kb)**	**P value after adjusting for adjacent risk SNPs**[e]	**Association direction reported previously**[f]	**Reference**
1p36.13	KLHDC7A	Protein	-5.67	1.40×10^{-8}	0.04	rs2992756	0.085	0.06	-	[7]
2q33.1	**ALS2CR12**	Protein	6.70	2.11×10^{-11}	0.10	rs1830298	intron of the gene	0.17	NA	[31]
2q33.1	CASP8	Protein	-8.05	8.51×10^{-16}	0.22	rs3769821	intron of the gene	0.16	-	[31,32]
5q14.1	ATG10	Protein	-6.65	2.85×10^{-11}	0.51	rs7707921	intron of the gene	0.21	NA	[9]
5q14.2	ATP6AP1L	Protein	-4.98	6.32×10^{-7}	0.63	rs7707921	37	0.98	NA	[9]
6q23.1	L3MBTL3[g]	Protein	-6.69	2.27×10^{-11}	0.10	rs6569648	208	0.44	NA	[6]
6q25.1	**RMND1**	Protein	4.76	1.95×10^{-6}	0.13	rs3757322	169	1.11×10^{-4}	mixed	[17]
11q13.1	SNX32	Protein	4.70	2.60×10^{-6}	0.19	rs3903072	18	0.17	NA	[33]
15q26.1	RCCD1	Protein	-7.18	7.23×10^{-13}	0.13	rs2290203	6	1.66×10^{-4}	-	[10]
17q22	**STXBP4**	Protein	6.69	2.21×10^{-11}	0.03	rs6504950	intron of the gene	0.90	+ in GTEx	[34,35]
19q13.31	**ZNF404**[h]	Protein	7.42	1.15×10^{-13}	0.15	rs3760982	90	0.005	NA	[8]

[a] Genes that were siRNA silenced for functional assays are bolded; SNPs used to predict gene expression are listed in the Supplementary Table 13

[b] Protein: protein coding genes; lncRNA: long non-coding RNAs; NA: not available

[c] P value: nominal P value from association analysis; the threshold after Bonferroni correction of 8,597 tests ($0.05/8{,}597=5.82\times10^{-6}$) was used; R^2: prediction performance (R^2) derived using GTEx data .

[d] Risk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes are presented in the **Supplementary Table 4**

[e] Use of COJO method[36]; all index SNPs in the corresponding region were adjusted for the conditional analyses

[f] -: inverse association; +: positive association; mixed: both inverse and positive associations reported; NA: not available

[g] Predicted expression of L3MBTL3 and RP11-73O6.3 was correlated (spearman R=0.88)

[h] Predicted expression of ZNF404 and RP11-15A1.7 was correlated (spearman R=0.64)

Table 4. Genes at GWAS-identified breast cancer risk loci (± 500kb of the index SNPs) whose predicted expression levels were associated with breast cancer risk at p-values between 5.82×10^{-6} and 1.05×10^{-3} (FDR corrected p-value≤0.05)

Region	**Gene**	**Type**[a]	**Z score**	**P value**[b]	R^{2b}	**Closest risk SNP**[c]	**Distance to the closest risk SNP (kb)**	**P value after adjusting for adjacent risk SNPs**[d]
1p34.1	UQCRH	Protein	-3.90	9.51×10^{-5}	0.12	rs1707302	168	0.06
1p22.3	LMO4	Protein	-3.76	1.73×10^{-4}	0.09	rs12118297	15	0.002
2p23.3	DNAJC27-AS1	lncRNA	3.84	1.24×10^{-4}	0.03	rs6725517	65	0.13
4p14	KLHL5	Protein	3.52	4.35×10^{-4}	0.13	rs6815814	230	0.03
5q11.2	AC008391.1	miRNA	-4.03	5.60×10^{-5}	0.13	rs16886113	242	0.76
6p22.1	HCG14	lncRNA	-3.47	5.19×10^{-4}	0.11	rs9257408	61	0.03
6p22.2	TRNAI2	miRNA	-3.71	2.09×10^{-4}	0.02	rs71557345	307	0.007
6q25.1	MTHFD1L	Protein	3.85	1.17×10^{-4}	0.10	rs3757318	491	2.36×10^{-4}
8q24.21	PVT1	transcript	3.85	1.20×10^{-4}	0.03	rs11780156	81	1.09×10^{-4}
9q33.3	RP11-123K19.1	lncRNA	-4.10	4.05×10^{-5}	0.05	rs10760444	20	1.26×10^{-4}
10q25.2	RP11-57H14.3	lncRNA	3.42	6.16×10^{-4}	0.08	rs7904519	108	0.002
10q26.13	RP11-500G22.2	lncRNA	4.48	7.54×10^{-6}	0.15	rs2981582	336	0.91
11p15.5	PTDSS2	Protein	-3.47	5.16×10^{-4}	0.04	rs6597981	312	0.02
11p15.5	AP006621.5	Protein	4.35	1.37×10^{-5}	0.51	rs6597981	19	0.01
11p15.5	PIDD1	Protein	4.24	2.28×10^{-5}	0.45	rs6597981	intron of the gene	0.12
11p15.5	MRPL23-AS1	lncRNA	-3.86	1.12×10^{-4}	0.10	rs3817198	95	0.06
11q13.1-11q13.2	PACS1	Protein	-3.59	3.36×10^{-4}	0.06	rs3903072	255	0.001
12p11.22	RP11-860B13.1	lncRNA	3.46	5.42×10^{-4}	0.17	rs10771399	221	0.86
13q22.1	KLF5	Protein	-4.08	4.44×10^{-5}	0.22	rs6562760	306	NA
14q24.1	CTD-2566J3.1	lncRNA	-3.84	1.22×10^{-4}	0.04	rs2588809	64	0.55
14q32.33	C14orf79	Protein	4.37	1.22×10^{-5}	0.11	rs10623258	240	0.91
15q26.1	FES	Protein	4.37	1.26×10^{-5}	0.21	rs2290203	73	3.04×10^{-6}
16q12.2	BBS2	Protein	3.97	7.23×10^{-5}	0.26	rs2432539	80	0.36
16q12.2	CRNDE	lncRNA	3.28	1.05×10^{-3}	0.02	rs28539243	271	0.69
16q24.2	RP11-482M8.1	lncRNA	3.32	9.16×10^{-4}	0.02	rs4496150	441	0.19

17q11.2	GOSR1	Protein	3.79	1.51×10^{-4}	0.10	rs146699004	376	0.04
17q21.2	ATP6V0A1	Protein	3.61	3.02×10^{-4}	0.03	rs72826962	162	0.01
17q21.2	RP11-400F19.8	transcript	-3.96	7.65×10^{-5}	0.01	rs72826962	122	6.62×10^{-4}
17q21.31	RP11-105N13.4	transcript	-4.51	6.46×10^{-6}	0.02	rs2532263	359	NA
17q25.3	CBX8	Protein	4.38	1.16×10^{-5}	0.05	rs745570	6	0.99
19p13.11	CTD-2538G9.5	lncRNA	3.56	3.76×10^{-4}	0.01	rs8170	432	4.38×10^{-4}
19p13.11	HOMER3	Protein	-3.87	1.08×10^{-4}	0.10	rs4808801	469	0.18
20q11.22	CTD-3216D2.5	lncRNA	4.03	5.60×10^{-5}	0.16	rs2284378	281	9.24×10^{-4}
22q13.1	TRIOBP	Protein	3.34	8.34×10^{-4}	0.07	rs738321	396	0.003
22q13.1	RP5-1039K5.13	lncRNA	3.73	1.93×10^{-4}	0.01	rs738321	99	0.053
22q13.1	CBY1	Protein	3.91	9.34×10^{-5}	0.05	chr22:39359355	289	0.06
22q13.1	APOBEC3A	Protein	-4.11	3.98×10^{-5}	0.07	chr22:39359355	0.2	0.02
22q13.2	RP1-85F18.6	lncRNA	3.52	4.28×10^{-4}	0.12	rs73161324	460	0.72

[a] Protein: protein coding genes; lncRNA: long non-coding RNAs; transcript: processed transcript

[b]P value: nominal P value from association analysis; R^2: prediction performance derived using GTEx data.

[c] Risk SNPs identified in previous GWAS or fine-mapping studies. The risk SNP closest to the gene is presented. A full list of all risk SNPs, and their distances to the genes are presented in the **Supplementary Table 4**

[d] Use of COJO method[36]; all index SNPs in the corresponding region were adjusted for the conditional analyses