Fine scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating *MAP3K1*.

Dylan M. Glubb,[1,146] Mel J. Maranian,[2,146] Kyriaki Michailidou,[3,146] Karen A. Pooley,[2,146] Kerstin B. Meyer,[4] Siddhartha Kar,[3] Saskia Carlebur,[4] Martin O'Reilly,[4] Joshua A. Betts,[1,5] Kristine M. Hillman,[1] Susanne Kaufmann,[1] Jonathan Beesley,[1] Sander Canisius[6], John L. Hopper,[7] Melissa C. Southey,[8] Helen Tsimiklis,[8] Carmel Apicella,[8] Marjanka K Schmidt,[6] Annegien Broeks,[6] Frans B Hogervorst,[6] C Ellen van der Schoot,[9] Kenneth Muir,[10,11] Artitaya Lophatananon,[10] Sarah Stewart-Brown,[10] Pornthep Siriwanarangsan,[12] Peter A. Fasching,[13,14] Matthias Ruebner,[13] Arif B. Ekici,[15] Matthias W. Beckmann,[13] Julian Peto,[16] Isabel dos-Santos-Silva,[16] Olivia Fletcher,[17] Nichola Johnson,[17] Paul D.P. Pharoah,[2,3] Manjeet K. Bolla[3], Qin Wang[3], Joe Dennis[3], Elinor J. Sawyer[18], Ian Tomlinson,[19] Michael J. Kerin,[20] Nicola Miller,[20] Barbara Burwinkel,[21] Frederik Marme,[21,22] Rongxi Yang,[21,23] Harald Surowy,[21,23] Pascal Guénel,[24,25] Thérèse Truong,[24,25] Florence Menegaux,[24,25] Marie Sanchez[24,25], Stig E. Bojesen[26-28], Børge G. Nordestgaard[26-28], Sune F Nielsen,[26,27] Henrik Flyger,[29] Anna González-Neira,[30] Javier Benitez,[31,32] M. Pilar Zamora,[33] Jose Ignacio Arias Perez,[34] Hoda Anton-Culver,[35] Susan L. Neuhausen,[36] Hermann Brenner,[37,38] Aida Karina Dieffenbach,[37,38] Volker Arndt,[37] Christa Stegmaier,[39] Alfons Meindl,[40] Rita K. Schmutzler,[41-44] Hiltrud.Brauch,[45,46] Yon-Dschun Ko,[47] Thomas Brüning,[48] The GENICA Network,[23,45-49] Heli Nevanlinna,[50] Taru A. Muranen,[50] Kristiina Aittomäki,[51] Carl Blomqvist,[52] Keitaro Matsuo,[53] Hidemi Ito,[54] Hiroji Iwata,[55] Hideo Tanaka,[54] Thilo Dörk,[56] Natalia V. Bogdanova,[57] Sonja Helbig,[56]

Annika Lindblom,[58] Sara Margolin,[59] Arto Mannermaa,[60-62] Vesa Kataja,[60-63] Veli-Matti Kosma,[60-62] Jaana M. Hartikainen,[60-62] kConFab Investigators,[64] Anna H. Wu,[65] Chiu-chen Tseng,[65] David Van Den Berg,[65] Daniel O. Stram,[65] Diether Lambrechts,[66,67] Hui Zhao,[66,67] Caroline Weltens,[68] Erik van Limbergen,[68] Jenny Chang-Claude,[69] Dieter Flesch-Janys,[70] Anja Rudolph,[69] Petra Seibold,[69] Paolo Radice,[71] Paolo Peterlongo,[72] Monica Barile[73] Fabio Capra,[72,74] Fergus J. Couch,[75] Janet E. Olson,[76] Emily Hallberg,[76] Celine Vachon,[76] Graham G. Giles[6,77] Roger L. Milne,[6,77] Catriona McLean,[78] Christopher A. Haiman,[65] Brian E. Henderson,[65] Fredrick Schumacher[65], Loic Le Marchand,[79] Jacques Simard,[80] Mark S. Goldberg,[81,82] France Labrèche,[83] Martine Dumont,[80] Soo Hwang Teo,[84,85] Cheng Har YIP,[85] Mee-Hoong See,[85] Belinda Cornes,[86] Ching-Yu CHENG,[86] M. Kamran Ikram,[86] Vessela Kristensen,[87-89] NBCS,[87,88,90-101] Wei Zheng,[102] Sandra L. Halverson,[102] Martha Shrubsole,[102] Jirong Long[102] Robert Winqvist,[103] Katri Pylkäs,[103] Arja Jukkola-Vuorinen,[104] Saila Kauppila,[105] Irene L. Andrulis,[106,107] Julia A. Knight,[108,109] Gord Glendon,[110] Sandrine Tchatchou,[110] Peter Devilee,[111] Robert A.E.M. Tollenaar,[111] Caroline Seynaeve,[112] Christi J. Van Asperen,[113] Montserrat García-Closas,[17,114] Jonine Figueroa,[115] Stephen J. Chanock,[115] Jolanta Lissowska,[116] Kamila Czene,[117] Daniel Klevebring,[117] Hatef Darabi,[117] Mikael Eriksson,[117] Maartje J. Hooning,[118] Antoinette Hollestelle,[118] John W.M. Martens,[118] J. Margriet Collée,[119] Per Hall,[117] Jingmei Li,[120] Keith Humphreys,[117] Xiao-Ou Shu,[102] Wei Lu,[121] Yu-Tang Gao,[122] Hui Cai,[102] Angela Cox,[123] Simon S. Cross,[124] Malcolm W. R. Reed,[123] William Blot,[102,125] Lisa B. Signorello,[102,125] Qiuyin Cai,[102] Mitul Shah,[2] Maya Ghoussaini,[2] Daehee Kang,[126-128] Ji-Yeob Choi,[127,128] Sue K. Park,[126-128] Dong-Young Noh,[129] Mikael Hartman,[130,131] Hui Miao,[130] Wei Yen Lim,[130] Anthony Tang,[132] Ute Hamann,[23] Diana Torres,[23,133] Anna Jakubowska,[134] Jan Lubinski,[134] Katarzyna

Jaworska,[134] Katarzyna Durda,[134] Suleeporn Sangrajrang,[135] Valerie Gaborieau,[136] Paul Brennan,[136] James McKay,[136] Curtis Olswold,[76] Susan Slager,[76] Amanda E. Toland,[137] Drakoulis Yannoukakos,[138] Chen-Yang Shen,[139-141] Pei-Ei Wu,[139,141] Jyh-Cherng Yu,[142] Ming-Feng Hou,[143] Anthony Swerdlow,[114,144] Alan Ashworth,[17] Nick Orr,[17] Michael Jones,[114] Guillermo Pita,[30] M. Rosario Alonso,[30] Nuria Álvarez,[30] Daniel Herrero,[30] Daniel C. Tessier,[145] Daniel Vincent,[145] Francois Bacot,[145] Craig Luccarini,[2] Caroline Baynes,[2] Shahana Ahmed,[2] Catherine S. Healey,[2] Melissa A. Brown,[5] Bruce A.J. Ponder,[4] Georgia Chenevix-Trench,[1] Deborah J. Thompson,[3] Stacey L. Edwards,[1,5] Douglas F. Easton,[2,3] Alison M. Dunning[2,147,*] and Juliet D. French[1,5,147, **]

1. Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4029, Australia.
2. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK.
3. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK.
4. CRUK Cambridge Institute and Department of Oncology, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK.
5. School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, QLD 4072, Australia
6. Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital, 1066 CX Amsterdam, The Netherlands.

7. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia.
8. Department of Pathology, The University of Melbourne, Melbourne, VIC 3010, Australia.
9. Sanquin Research, 1066 CX Amsterdam, The Netherlands.
10. Division of Health Sciences, Warwick Medical school, Warwick University, Coventry CV4 7AL, UK.
11. Institute of Population Health, University of Manchester, Manchester M13 9PL, UK.
12. Ministry of Public Health, Nonthaburi 11000, Thailand.
13. University Breast Center Franconia, Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, 91054 Erlangen, Germany.
14. David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA 90095, USA.
15. Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, 91054 Erlangen, Germany
16. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK.
17. Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London SW3 6JB, UK.

18. Research Oncology, Division of Cancer Studies, King's College London, Guy’s Hospital, London SE1 9RT, UK.
19. Wellcome Trust Centre for Human Genetics and Oxford Biomedical Research Centre, University of Oxford OX3 7BN, UK.
20. Clinical Science Institute, University Hospital Galway, Galway, Ireland.
21. Department of Obstetrics and Gynecology, University of Heidelberg, 69115 Heidelberg, Germany.
22. National Center for Tumor Diseases, University of Heidelberg, 69120 Heidelberg, Germany.
23. Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.
24. INSERM (National Institute of Health and Medical Research), CESP (Center for Research in Epidemiology and Population Health), U1018, Environmental Epidemiology of Cancer, 94807 Villejuif, France.
25. University Paris-Sud, UMRS 1018, 94807 Villejuif, France.
26. Copenhagen General Population Study, Herlev Hospital, Copenhagen University Hospital, 2730 Herlev, Denmark.
27. Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, 2730 Herlev, Denmark.
28. Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark.
29. Department of Breast Surgery, Herlev Hospital, Copenhagen University Hospital, 2730 Herlev, Denmark.
30. Human Genotyping-CEGEN Unit, Human Cancer Genetics Program, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain.

31. Centro de Investigación en Red de Enfermedades Raras (CIBERER), 46010 Valencia, Spain.
32. Human Genetics Group, Spanish National Cancer Centre (CNIO), and Biomedical Network on Rare Diseases (CIBERER), 28029 Madrid, Spain.
33. Servicio de Oncología Médica, Hospital Universitario La Paz, 28029 Madrid, Spain.
34. Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, 33012 Oviedo, Spain.
35. Department of Epidemiology, University of California Irvine, Irvine, CA 92697, USA.
36. Beckman Research Institute of City of Hope, Duarte, CA 91010, USA.
37. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.
38. German Cancer Consortium (DKTK), 69120 Heidelberg, Germany.
39. Saarland Cancer Registry, 66024 Saarbrücken, Germany.
40. Division of Gynaecology and Obstetrics, Technische Universität München, 81675 Munich, Germany
41. Division of Molecular Gyneco-Oncology, Department of Gynaecology and Obstetrics, University Hospital of Cologne, 50931 Cologne, Germany.
42. Centre of Familial Breast and Ovarian Cancer, Department of Gynaecology and Obstetrics and Centre for Integrated Oncology (CIO), Center for Molecular Medicine Cologne (CMMC), University Hospital of Cologne, 50937 Cologne, Germany.

43. Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50923 Cologne, Germany, on behalf of the German Consortium of Hereditary Breast and Ovarian Cancer (GC-HBOC).
44. Center for Integrated Oncology (CIO), Medical Faculty, University Hospital of Cologne, 50937 Cologne, Germany.
45. University of Tübingen, 72074 Tübingen, Germany.
46. Dr. Margarete Fischer-Bosch-Institute for Clinical Pharmacology, 70376 Stuttgart, Germany.
47. Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, 53113 Bonn, Germany.
48. Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), 44789 Bochum, Germany.
49. Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany.
50. Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, 00029 Helsinki, HUS, Finland.
51. Department of Clinical Genetics, Helsinki University Central Hospital, 00029 Helsinki, Finland.
52. Department of Oncology, University of Helsinki and Helsinki University Central Hospital, 00029 Helsinki, Finland.
53. Department of Preventive Medicine, Kyushu University Faculty of Medical Sciences, Fukuoka, Fukuoka 812-8582, Japan.
54. Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya 464-8681, Japan.

55. Department of Breast Oncology, Aichi Cancer Center Hospital, Nagoya 464-8681, Japan.

56. Gynaecology Research Unit, Hannover Medical School, 30625 Hannover, Germany.

57. Department of Radiation Oncology, Hannover Medical School, 30625 Hannover, Germany.

58. Department of Molecular Medicine and Surgery, Karolinska Institutet, 171 77 Stockholm, Sweden.

59. Department of Oncology - Pathology, Karolinska Institutet, 171 77 Stockholm, Sweden.

60. Cancer Center of Eastern Finland, University of Eastern Finland, Kuopio, Finland.

61. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, 70211 Kuopio, Finland.

62. School of Medicine, Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, 70211 Kuopio, Finland.

63. Cancer Center, Kuopio University Hospital, 70211 Kuopio, Finland.

64. Peter MacCallum Cancer Institute, East Melbourne, VIC 3002, Australia.

65. Department of Preventive Medicine, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, CA 90089, USA.

66. Laboratory for Translational Genetics, Department of Oncology, University of Leuven, 3000 Leuven, Belgium.

67. Vesalius Research Center, VIB, 3000 Leuven, Belgium.

68. University Hospital Gashuisberg, 3000 Leuven, Belgium.

69. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.

70. Department of Cancer Epidemiology/Clinical Cancer Registry and Institute for Medical Biometrics and Epidemiology, University Clinic Hamburg-Eppendorf, 20246 Hamburg, Germany.

71. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale Tumori (INT), 20133 Milan, Italy.

72. IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, 20139 Milan, Italy.

73. Division of Cancer Prevention and Genetics, Istituto Europeo di Oncologia, 20141 Milan, Italy.

74. Cogentech Cancer Genetic Test Laboratory, 20139 Milan, Italy.

75. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA.

76. Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA.

77. Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, VIC 3053, Australia.

78. Anatomical Pathology, The Alfred Hospital, Melbourne, VIC 3004, Australia.

79. University of Hawaii Cancer Centre Honolulu, HI 96813, USA.

80. Centre Hospitalier Universitaire de Québec Research Center and Laval University, Quebec City G1V 4G2, Canada.

81. Division of Clinical Epidemiology, McGill University Health Centre, Royal Victoria Hospital, Montreal, QC H3A 1A1, Canada.

82. Department of Medicine, McGill University, Montreal, QC H3A 1A1, Canada.

83. Département de médecine sociale et préventive, Département de santé environnementale et santé au travail, Université de Montréal, Montreal, QC H3A 3C2, Canada.

84. Cancer Research Initiatives Foundation, Sime Darby Medical Centre, 47500 Subang Jaya, Malaysia.

85. Breast Cancer Research Unit, University Malaya Cancer Research Institute, University Malaya Medical Centre (UMMC), 50603 Kuala Lumpur, Malaysia.

86. Singapore Eye Research Institute, National University of Singapore, 168751 Singapore.

87. Institute of Clinical Medicine, University of Oslo (UiO), 0450 Oslo, Norway.

88. Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Radiumhospitalet, 0310 Oslo, Norway.

89. Department of Clinical Molecular Biology (EpiGen), University of Oslo (UiO), 0450 Oslo, Norway.

90. Section of Oncology, Institute of Medicine, University of Bergen, N-5020 Bergen, Norway

91. Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, N-9038 Tromsø, Norway.

92. National Resource Centre for Long-term Studies after Cancer, Rikshospitalet-Radiumhospitalet Cancer Clinic, Montebello, 0310 Oslo, Norway.

93. Division of Cancer Medicine and Radiotherapy, Oslo University Hospital, Radiumhospitalet, 0310 Oslo, Norway.

94. Department of Surgery, Akershus University Hospital, 1478 Lørenskog, Norway.

95. Department of Radiology, Oslo University Hospital, Radiumhospitalet, 0310 Oslo, Norway.

96. Department of Pathology, Akershus University Hospital, 1478 Lørenskog, Norway.

97. Department of Oncology, Oslo University Hospital, Radiumhospitalet, 0310 Oslo, Norway.

98. Department of Oncology, Haukeland University Hospital, Bergen, 5021 Bergen, Norway.

99. Department of Community Medicine, Faculty of Health Sciences, University of Tromsø - The Arctic University of Norway, 9019 Tromsø, Norway.

100. Department of Clinical Molecular Biology (EpiGen), Institute of Clinical Medicine, Akershus University Hospital, University of Oslo, 1478 Lørenskog, Norway.

101. Department of Breast and Endocrine Surgery, Institute for Clinical Medicine, Ullevaal University Hospital, Oslo University Hospital, 0450 Oslo, Norway.

102. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center and Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37203, USA.

103. Laboratory of Cancer Genetics and Tumor Biology, Department of Clinical Chemistry and Biocenter Oulu, University of Oulu, NordLab Oulu/Oulu University Hospital, Oulu, Finland

104. Department of Oncology, Oulu University Hospital, University of Oulu, 90014 Oulu, Finland.

105. Department of Pathology, Oulu University Hospital, University of Oulu, 90014 Oulu, Finland.

106. Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, Canada.

107. Department of Molecular Genetics, University of Toronto, Toronto, ON M5G 1X5, Canada

108. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada.

109. Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada.

110. ON Cancer Genetics Network, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada.

111. Department of Human Genetics and Department of Pathology, 2300 RC Leiden University Medical Center, Leiden, The Netherlands.

112. Family Cancer Clinic, Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands.

113. Department of Clinical Genetics, Leiden University Medical Center, 2300 RC Leiden, The Netherlands.

114. Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton SM2 5NG, UK.

115. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20892, USA.

116. Department of Cancer Epidemiology and Prevention, Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, 02-781 Warsaw, Poland.

117. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, 171 77 Stockholm, Sweden.

118. Department of Medical Oncology, Erasmus MC Cancer Institute, Rotterdam, The Netherlands.

119. Department of Clinical Genetics, Erasmus University Medical Center, 3008 AE Rotterdam, The Netherlands.

120. Human Genetics Division, Genome Institute of Singapore, 138672 Singapore.

121. Shanghai Center for Disease Control and Prevention, Changning, Shanghai 200336, China.

122. Department of Epidemiology, Shanghai Cancer Institute, Shanghai 200032, China.

123. Sheffield Cancer Research Centre, Department of Oncology, University of Sheffield, Sheffield S10 2RX, UK.

124. Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield S10 2RX, UK.

125. International Epidemiology Institute, Rockville, MD 20850, USA.

126. Department of Preventive Medicine, Seoul National University College of Medicine and Cancer Research Institute, Seoul National University, Seoul 110-799, Korea.

127. Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 151-742, Korea.

128. Cancer Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea.

129. Department of Surgery, Seoul National University, Bundang Hospital, Seongnam 110-744, Korea.

130. Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, 117597 Singapore.

131. Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, 119228 Singapore.

132. Division of General Surgery, National University Health System, 119228 Singapore.

133. Institute of Human Genetics, Pontificia Universidad Javeriana, Bogotá 11001000, Colombia.

134. Department of Genetics and Pathology, Pomeranian Medical University, 70-115 Szczecin, Poland.

135. National Cancer Institute, Bangkok 10400, Thailand.

136. International Agency for Research on Cancer, 69372 Lyon, France.

137. Department of Molecular Virology,Immunology and Medical Genetics, The Ohio State University, Columbus, OH 43210, USA.

138. Molecular Diagnostics Laboratory, IRRP, National Centre for Scientific Research "Demokritos", Aghia Paraskevi Attikis, Athens 15310, Greece.

139. Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan.

140. School of Public Health, China Medical University, Taichung 40402, Taiwan.

141. Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan.

142. Department of Surgery, Tri-Service General Hospital, Taipei 114, Taiwan.

143. Cancer Center and Department of Surgery, Kaohsiung Medical University Chung-Ho Memorial Hospital, Kaohsiung 807, Taiwan.

144. Division of Breast Cancer Research, Institute of Cancer Research, Sutton SM2 5NG, UK.

145. McGill University and Génome Québec Innovation Centre, Montréal, Québec H3A 0G1, Canada.

[146,147] These authors contributed equally and are listed in alphabetical order

*Correspondence: amd24@medschl.cam.ac.uk

**Correspondence: juliet.french@qimrberghofer.edu.au

Abstract

Genome Wide Association Studies (GWAS) revealed SNP rs889312 on 5q11.2 to be associated with breast cancer risk in women of European ancestry. In an attempt to identify the biologically relevant variants, we analysed 909 genetic variants across the 5q11.2 locus in 103,991 breast cancer cases and controls from 52 studies in the Breast Cancer Association Consortium. Multiple logistic regression analyses identified three independent risk signals: The strongest associations are with 15 correlated variants (iCHAV1) where the minor allele of best candidate, rs62355902, associates with significantly increased risks of both estrogen receptor-positive (ER+: OR=1.24, 95% CI 1.21-1.27; *P*-trend=5.7×10^{-44}) and estrogen receptor-negative tumors (ER-: OR=1.10, 95% CI 1.05-1.15; *P*-trend=3.0×10^{-4}). After adjustment for rs62355902, we found evidence for the association of a further 173 variants (iCHAV2) containing three subsets with a range of effects, of which the strongest is SNP rs113317823 (*P*-cond=1.61×10^{-5}); and five variants comprising iCHAV3 (lead rs11949391; ER+: OR=0.90, 95% CI 0.87-0.93; *P*-cond=$1.4x10^{-4}$). Twenty six percent of the prioritized candidate variants coincide with four putative regulatory elements that interact with the *MAP3K1* promoter through chromatin looping and affect *MAP3K1* promoter activity. Functional analysis indicates the cancer risk alleles of four candidates increase *MAP3K1* transcriptional activity: rs74345699 and rs62355900 (iCHAV1); rs16886397 (iCHAV2a); and rs17432750 (iCHAV3). Chromatin immunoprecipitation analysis revealed diminished GATA3 binding to the minor (cancer protective) allele of rs17432750, indicating a mechanism for its action. We propose that the cancer risk alleles act to increase *MAP3K1* expression *in vivo* and may promote breast cancer cell survival.

Introduction

One of the first Genome Wide Association Studies (GWAS) for breast cancer [MIM 114480] susceptibility identified a SNP at 5q11.2, rs889312, associated with risk of breast cancer in women of European ancestry.[1] In the most recent analyses by the Breast Cancer Association Consortium (BCAC), the minor allele of rs889312 was associated with a per-allele Odds Ratio (OR)=1.12 (95% Confidence Intervals (CI) 1.10-1.15; *P*-trend=1.8×10^{-26}).[2] The association was stronger for ER-positive disease (OR=1.14, 95% CI 1.11-1.17; *P*=$1.1\text{x}10^{-26}$ in the most recent BCAC analysis) but also seen for ER-negative (OR=1.06, 95% CI 1.03-1.10; *P*=0.0024) and triple negative disease (OR=1.11, 95% CI 1.02-1.20; *P*=0.016).[3] SNP rs889312 was also reported to be associated with an increased breast cancer risk in *BRCA2* [MIM 600185] mutation carriers.[4]

The GWAS SNP, rs889312, lies approximately 80kb centromeric of *MAP3K1* [MIM 600982], the gene encoding Mitogen-Activated Protein Kinase Kinase Kinase1, also known as MEK kinase 1 (MEKK1) – a stress-induced serine/threonine kinase with apparent dual functions: MEKK1 induces cell proliferation through a RAS/RAF/MEK/ERK signaling pathway[5] but, upon caspase cleavage, generates a fragment which has a pro-apoptotic function.[6,7] Furthermore, MEKK1 regulates transcription of key cancer related genes such as *MYC*[8] [MIM 190080], *TP53*[9] [MIM 191170] and *JUN*[10] [MIM 165160] through its signal transduction pathway. There is already evidence of a role for *MAP3K1* in breast cancer pathogenesis: *MAP3K1* driver mutations have been observed in luminal A and B type breast tumors[11] and *MAP3K1* expression has been associated with specific breast tumor subtypes.[12]

In this study, we have performed genetic epidemiological analyses on all common variants at 5q11.2, together with *in silico* and *in vitro* analyses of candidate causal variants, and identified strong candidates that we propose are functionally related to breast cancer risk. Specifically, we provide evidence that these associations are mediated through *MAP3K1*.

Materials and Methods

SNP selection and genotyping

We identified all SNPs from a 305kb interval on 5q11.2 (GRCh37 positions 55,983,657–56,288,810) with a minor allele frequency (MAF) >0.02 in Europeans using the March 2010 release of the 1000 Genomes Project.[13] SNPs with an Illumina designability score of >0.8, and r^2>0.1 with rs889312, together with a tagging set (r^2>0.9) for all other known SNPs in the interval, were selected for inclusion on the iCOGS custom array.[2] In total, 352 SNPs were selected, of which 300 passed post-genotyping quality control criteria.[2] To improve SNP coverage across the locus, 16 further SNPs, selected from the October 2010 release of the 1000 Genomes Project were genotyped in a subset of two BCAC studies (SEARCH and the combined Copenhagen studies CPGS and CCHS) using a Fluidigm array according to the manufacturer's instructions. These two datasets were used to impute all genotypes of other common variants in this interval, using IMPUTE2 and the January 2012 1000 Genomes Project release as reference.[14] All participants provided written informed consent. Study characteristics and iCOGS methodology have been previously reported.[2]

Statistical analyses

Associations with breast cancer risk in BCAC were evaluated by comparing SNP genotype frequencies in cases and controls using unconditional logistic regression. Analyses were adjusted by study and seven principal components.[2] The primary analysis fitted each SNP as an allelic dose and tested for association using a 1-degree-of-freedom trend test (*P*-trend) with associated odds ratio (OR) and 95% Confidence Intervals (95% CI). To identify independent risk signals, stepwise conditional analysis was performed in R with the function "step", including any variant with *P*-value<10^{-4} in the single SNP analysis to calculate the most parsimonious model using a penalty value of k=10. The null model included study and principal components. Haplotype analysis was performed in R using the package "haplo.stats", and the analyses adjusted for study and principal components.

Cell lines

The normal breast epithelial cell line Bre-80 was cultured as described previously.[15] The breast cancer cell lines MCF7, T-47D and MDA-MB231 were cultured in RPMI1640 medium supplemented with 10% fetal calf serum, antibiotics, sodium pyruvate and, in the case of MCF7 and T-47D cells, 10 μg/ml insulin.

Chromatin conformation capture (3C)

3C libraries were generated with EcoRI from the cell lines above as described previously.[16] 3C interactions were quantitated by real-time PCR (qPCR) with primers designed within the EcoRI restriction fragments spanning the 5q11.2 risk locus (**Table S1**). qPCR was performed as described previously[17] using at least two independent 3C libraries from each cell line with each experiment quantified in

duplicate. Two BAC clones (RP11-378G4 and RP11-1146C6) covering the 5q11.2 region were used to create an artificial library of ligation products in order to normalize for PCR efficiency. As an internal control, interaction frequencies were normalized to that of the EcoRI fragment immediately upstream of the promoter/bait fragment.

Plasmid construction

A *MAP3K1* promoter-driven luciferase reporter construct was generated by inserting a 1928 bp fragment containing the *MAP3K1* promoter and the transcription start site (GRch37 chr5:56109070-56110997) into the MluI and HindIII sites of pGL3-Basic. To assist cloning, AgeI and SbfI sites were inserted into the BamHI and SalI sites downstream of the luciferase gene. A 1575 bp PRE-A fragment, a 1765 bp PRE-B2 fragment, a 2357 bp PRE-B3 fragment, a 2203 bp PRE-C fragment and a 1519 bp PRE-D fragment were PCR generated using primers designed with AgeI and SbfI sites and cloned into the modified pGL3-*MAP3K1* promoter construct. PRE-B was too large (~7 kb) to be cloned in its entirety so three subregions containing iCHAV1 (PRE-B1), iCHAV2b (PRE-B2) and iCHAV3 (PRE-B3) variants were cloned separately. The minor alleles of individual SNPs were introduced into promoter and PRE sequences, containing the major alleles of any other causal candidate variants, by overlap extension PCR. All constructs were sequenced to confirm variant incorporation (AGRF, Brisbane, Australia). PCR primers are listed in **Table S2**. For the PRE-B1 construct, a 2129 bp region spanning chr5:56028968-56031097 (GRCh37) was synthesized with AgeI and SbfI sites incorporated at the 5' and 3' ends (GenScript, Piscataway, NJ, USA) to assist cloning into the *MAP3K1* promoter construct. The cloned regions are highlighted in **Figure 2B**.

Reporter assays and estrogen induction

Bre-80 and MCF7 cells were transfected with equimolar amounts of luciferase reporter plasmids and 50 ng of pRLTK transfection control plasmid with Lipofectamine 2000. The total amount of transfected DNA was kept constant at 600 ng for each construct by adding pUC19 as a carrier plasmid. Luciferase activity was measured 24 h post-transfection by the Dual-Glo Luciferase Assay System. To correct for differences in transfection efficiency or cell lysate preparation, *Firefly* luciferase activity was normalized to *Renilla* luciferase. For the assays under basal conditions, the activity of each construct was calculated relative to the construct containing the *MAP3K1* promoter alone, the activity of which was defined as 1.

For estrogen induction assays, we treated cells as described.[18] Briefly, 24 h after plating MCF7 cells into wells, medium was replaced with that containing 10 nM fulvestrant for 48 h, which was used to inhibit estrogen induced gene expression, thereby creating a baseline of expression for reporter assays. Cells were then incubated with fresh medium containing either 10 nM estrogen (17β-estradiol) or DMSO (as vehicle control) and transfected with reporter plasmids. Luciferase assays were performed as above after 24 h. Statistical significance was tested by repeated-measures ANOVA, using the Greenhouse-Geisser correction for non-sphericity, followed by Dunnett's multiple comparisons test in GraphPad Prism.

ChIP assays

Chromatin immunoprecipitation (ChIP) assays were carried out as previously

described.[19] GATA3 specific mouse monoclonal antibodies were obtained from Santa Cruz Biotech (sc268). Precipitate and input were used in qPCR using SYBR green master mix as described.[19] Primer sequences are listed in **Table S3**. Both rs17432750 primer sets gave identical enrichment and the identity of the larger fragment was confirmed by Sanger sequencing. All values obtained were normalized to input and enrichment is given relative to the negative *CCND1* [MIM 168461] control. Allele-specific ChIP was carried out by using a TaqMan SNP genotyping assay (Applied Biosystems) on the ChIP material. In the TaqMan assay two different fluorophores are each linked to a probe detecting the two different alleles. Amplification of each allele was followed using an Applied Biosystems Real Time PCR machine (7900HT) and the data analysed using the SDS software. The SDS software converts the raw data to fluorescence intensity for each allele and then plots the results as a scatter graph of allele X versus allele Y. We tested the accuracy of this assay by genotyping known mixtures of homozygous ZR751 and T47D (C/C), and MDA-MB0-468 (A/A) cell line DNA (**Figure S1**). For the allelic discrimination three independent experiments were carried out and gave similar results - a representative example is shown.

GATA3 siRNA knockdown for reporter assay

GATA3 (MIM 131320; L-003781-00) and non-targeting (D-001810-10-20) ON-TARGET*plus* SMARTpool siRNAs for were purchased from Thermo Scientific. For knockdown, Bre-80 cells were co-transfected with the relevant luciferase reporter plasmids and 100 nM of either *GATA3* or non-targeting siRNAs with Lipofectamine 2000. Luciferase assays were performed as described above after 24 h. To validate *GATA3* knockdown, qPCR was performed as described previously (**Figure S9**).[17]

Expression quantitative trait locus (eQTL) analysis

eQTL analysis was undertaken in two sample sets of adjacent normal breast samples from women of European decent: the first set contained 135 samples collected for the METABRIC study[20]; the second set was extracted from the TCGA breast cancer study[21] data and contained 56 samples. Matched gene expression (Illumina HT-12 v3 microarray for the METABRIC data; Agilent G4502A-07-3 microarray for the TCGA data) and germline SNP data that was either genotyped (Affymetrix SNP 6.0) or imputed (1,000 Genomes Project, March 2012 data using IMPUTE version 2.0) were used. Correlations between all imputed and genotyped variants at the 5q11.2 locus and expression levels of eight (METABRIC) or four (TCGA) genes present in the fine-mapped region were assessed using a linear regression model in which an additive effect on expression level was assumed for each copy of the rare allele. Calculations on the METABRIC data were carried out using the eMap library in R, and on the TCGA data using the SNPTEST software. [22]

Results

Genotyping of case-control studies

Three hundred SNPs at the 5q11.2 locus (GRch37 positions 55,983,657–56,288,810) were successfully genotyped using the iCOGs chip, in 103,991 breast cancer cases and controls from 52 BCAC studies, of which 41 (46,451 cases and 42,599 controls) were of European, nine were of Asian (6,269 cases and 6,624 controls) and two were of African-American ancestry (1,116 cases and 932 controls). Using these data, together with data on a further 16 SNPs, genotyped in two BCAC studies (SEARCH and the combined Copenhagen studies CPGS and CCHS), we

imputed genotypes for 909 variants with MAF>0.02 (out of a possible 911) and imputation r^2>0.3, using the January 2012 release of the 1000 Genomes Project as reference.

Potential breast cancer risk signals in European studies

Figure 1A shows the Manhattan plot of the 909 genotyped and imputed SNPs (r^2>0.3) for overall breast cancer risk in European studies. Genotype and association results for all 909 SNPs are presented in **Table S4**. Five hundred and forty one variants display association with overall breast cancer risk at *P*-trend<10^{-4} (**Table S5**). All associations are consistent with a log-additive model.

In a forward stepwise conditional analysis, using the SNPs listed in the **Table S5**, the best model included three SNPs: (1) rs62355902 – conditional *P*-value (*P*-cond)=$8.6x10^{-26}$; (2) rs113317823 - *P*-cond=$2.8x10^{-5}$ and (3) rs11949391 - *P*-cond=$9.7x10^{-5}$. No significant evidence for heterogeneity was observed among odds ratios (ORs) for these SNPs among studies of European ancestry (minimum observed *P*-het=0.14 and maximum I^2 = 19.3% for SNP rs11949391 - **Figure S2**). Each SNP remaining in the conditional analysis model indicates the existence of a separate genetic risk signal (previously defined[23] as an Independent set of Correlated Highly Associated Variants – iCHAV), each of which will contain at least one directly causal variant.

The most significantly associated variant overall, is rs62355902 (OR (per minor allele) =1.21, 95% CI 1.19-1.24; *P*-trend=9.5×10^{-49}). This is the most significant of 15 strongly correlated SNPs (r^2>0.93) lying within a 50Kb interval (GRCh37 positions

56,003,831–56,053,745; **Table 1** and marked in red in **Table S5**), which we have designated as iCHAV1. These iCHAV1 SNPs all have likelihood ratios of <100:1 relative to the best candidate SNP, rs62355902, and thus cannot be excluded from further analysis, remaining as strong candidate causal variants based on epidemiological evidence. After conditioning on iCHAV1 top SNP, rs62355902, the most strongly associated variant is SNP rs113317823: OR (per minor allele)=1.22, 95% CI 1.18-1.26; *P*-trend=7.0x10^{-25}; conditional OR=1.12, 95% CI 1.05-1.20; *P*-cond=2.8x10^{-5}. One hundred and seventy two variants have likelihood ratios of <100:1 relative to rs113317823 and these constitute iCHAV2 (highlighted in shades of blue and yellow **Table S5**). SNP rs113317823 is partially correlated with iCHAV1 top candidate rs62355902 (r^2=0.19) and this adds complexity to the subsequent analysis – this is explored in more detail by the haplotype analysis described below.

After adjustment for the top iCHAV1 (rs62355902) and iCHAV2 (rs113317823) SNPs, the best remaining significantly associated SNP is rs11949391 (iCHAV3): OR (per minor allele)=0.91, 95% CI 0.89-0.94; *P*-trend=9.4×10^{-12}; and conditional OR=0.95, 95% CI 0.92-0.98; *P*-cond=9.7×10^{-5}. Four other SNPs have likelihood ratios of <100:1 relative to rs11949391 and are highly correlated with rs11949391 (r^2≥0.95) but not with either of the top iCHAV1 or iCHAV2 SNPs (r^2<0.04, marked in yellow in **Table S5,** listed in **Table 1)** and these are thus candidate causal variants for iCHAV3.

Effects on Estrogen Receptor-positive (ER+) and negative (ER-) tumor subtypes

Candidate causal SNPs in iCHAV1 are associated with risks of both ER+ and ER- disease (**Table 1**). However the OR is greater for ER+ (rs62355902 OR=1.24, 95% CI 1-21-1.27) than ER- disease (OR=1.10, 95% CI 1.05-1.15; *P*-difference=$1.5x10^{-5}$; **Table 1**). By contrast, the minor alleles of candidate causal SNPs in iCHAV3 are protective against ER+ tumor development (rs11949391 OR=0.90, 95% CI 0.87-0.93; *P*=$1.9x10^{-10}$) but have no apparent effect on ER- tumor risk (OR=1.01, 95% CI 0.96-1.06; *P*=0.66; *P*-difference=$6.5x10^{-5}$; **Table 1**). The lead SNP in iCHAV2, rs113317823, remains significantly associated with ER+ tumor risk (*P*-cond=$9.7x10^{-5}$, after adjustment for rs62355902) but not with ER- tumor risk (*P*-cond=0.099), but the difference in the OR by ER subtype is only borderline significant (*P*-difference=0.02; **Table 1**).

Effects of haplotypes on breast cancer risk

While iCHAVs 1 and 3 represent sets of highly correlated SNPs (r^2>0.93) with the lead SNP), the set of 173 SNPs, labeled iCHAV2, includes three subsets, defined according to their correlations with rs113317823 (iCHAV2) and rs62355902 (iCHAV1): these subsets are iCHAV2a - lead SNP rs113317823 - representing 90 SNPs correlated with rs113317823 (r^2>0.53) and with iCHAV1 SNP rs62355902 (r^2=0.19-0.29), marked in mid blue **Table S5**; iCHAV2b - lead SNP rs62355899 representing 66 SNPs, which are independent of rs113317823 (r^2≤0.01) but correlated with rs62355902 (r^2=0.59-0.62) (conditional OR=0.90, 95% CI 0.86-0.95; *P*-cond=$3.0x10^{-5}$), marked in **Table S5** in teal; and i**CHAV2c** - lead SNP rs7721581 - representing 17 SNPs that are modestly correlated with rs113317823 (r^2=0.14-0.16) but independent of rs62355902 (r^2≤0.01) (conditional OR=0.96, 95% CI 0.93-0.98; *P*-cond=$4.8x10^{-4}$) marked in **Table S5** in pale blue.

To further clarify these association signals, we performed haplotype analyses, based on the lead SNPs from iCHAVs 1, 2a, 2b, 2c and 3. These five SNPs define seven common haplotypes (**Table 2**). Two of these: **h5**, which carries the risk alleles of iCHAV1, iCHAV2a and iCHAV2b; and **h6**, which carries the risk alleles of iCHAV1 and iCHAV2b, are strongly associated with risk, with the risk being higher for haplotype **h6 (***P*=1.66x10^{-29}**)**. These results are consistent with the observation that SNPs in iCHAV1 are the most strongly associated with risk. They are also consistent with a model in which either a SNP in iCHAV2a or iCHAV2c (in combination with iCHAV1) increases risk, or a SNP in iCHAV2b reduces risk. These hypotheses are difficult to distinguish since the iCHAV1 risk allele never occurs alone – it occurs in combination with either iCHAV2a/iCHAV2c or with iCHAV2b, but not both. Some support for the iCHAV1+iCHAV2a hypothesis is provided by the fact that haplotypes **h3** and **h4**, which carry the risk alleles for iCHAV2a but not iCHAV1, though rare, are associated with an increased risk. Evidence against the iCHAV1+iCHAV2c hypothesis is provided by the fact that **h1***,* which carries the risk allele for iCHAV2c alone, is not associated with an increased risk. These observations are consistent with the regression analyses, in which iCHAV2c is less likely than iCHAV2a or iCHAV2b SNPs to harbour a causal variant, (likelihood ratio ~30:1, after adjustment for iCHAV1). Haplotype **h2**, which carries the minor allele of iCHAV3 SNP rs11949391, was associated with a reduced ER+ (though not ER-) breast cancer risk, consistent with the effect of the iCHAV3 SNP in the regression analysis.

We conclude that, at least, one of the 90 SNPs in iCHAV2a (positions 56,001,002-56,270,717) or one of the 66 SNPs in iCHAV2b (positions 55,998,085-56,183,743)

are causally related to risk, together with a variant in iCHAV1 and a variant in iCHAV3.

Risk associations in Asian and African-American studies

We tested all genotyped and imputed SNPs with MAF>0.02 and imputation r^2>0.3 in the nine Asian studies (6,269 cases and 6,624 controls; SNPs n=1045) and the two African-American studies (1,116 cases and 932 controls; SNPs n=1601) for association with overall breast cancer risk. None reached genome-wide levels of significance (P<5x10^{-8}) but the lead SNPs of each iCHAV display compatible effects in all three ethnic groups. This is most apparent for iCHAV2a SNP rs113317823: (European–unadjusted) OR=1.22, 95% CI 1.18-1.26; P-trend=7.0×10^{-25}; (Asian) OR=1.19, 95% CI 1.11-1.27; P-trend=1.4×10^{-5}; (African-American) OR=1.04, 95% CI 0.77-1.31; P-trend=0.78, **Table S6**.

Identification of putative regulatory elements (PREs) that interact with the *MAP3K1* promoter

After the epidemiological analyses, 15 iCHAV1, 90 iCHAV2a, 66 iCHAV2b and 5 iCHAV3 variants remain as strong causal candidates (**Figure 1A** and **Table S7**), whereas the data provide weaker support for iCHAV2c being causal. As iCHAV2a and 2b were comprised of a large number of variants, we prioritized these for functional analysis using a threshold of P-cond<1x10^{-4} and focused on the remaining 30 iCHAV2a and 10 iCHAV2b variants, in addition to the iCHAV1 and iCHAV3 candidates (**Table 1**). Next we examined whether these 62 iCHAV variants coincide with PREs that may affect gene expression.

Using publicly available ENCODE chromatin immunoprecipitation sequencing (ChIP-seq) data from MCF7 and HMEC cells, we identified regions marked by histone modifications associated with transcriptional enhancers (mono- or di-methylation of H3 lysine 4 (H3K4Me1 and H3K4Me2) and acetylation of H3 lysine 27 (H3K27Ac)) or bound by ERα, FOXA1 or GATA3, transcription factors known to play a role in breast cancer. Next, we mined RNA polymerase II ChIA-PET (chromatin-interaction analysis with paired-end tag sequencing) data, previously generated in MCF7 cells,[24,25] and identified multiple long-range chromatin interactions between discrete regions of the iCHAV loci and the promoter of *MAP3K1* (**Figure S3**). Consequently, we performed chromosome conformation capture (3C) experiments to analyze interactions between the *MAP3K1* promoter and these regions within 5q11.2. Using 3C in a normal mammary epithelial cell line, Bre-80, we found several regions that interact with the *MAP3K1* promoter (**Figure 2C**). Similar 3C profiles were observed in two ER+ cell lines (MCF7 and T-47D) and in ER– MDA-MB231 breast cancer cells. (**Figure S4**).

PREs were defined as the loci encompassing functional elements identified from the ENCODE data within a *MAP3K1* promoter interacting region. This analysis revealed four PREs (A-D) that coincide with the iCHAV candidates prioritized for functional analyses (**Figure 2**). Consistently, ENCODE ChIA-PET data demonstrates that all four PREs interact with the *MAP3K1* promoter (**Figure S3A**) but not with the promoters of other nearby genes in MCF7 cells (**Figure S3B**). It should be noted that all four PREs contain iCHAV2a or 2b variants (*P*-cond>$1x10^{-4}$) that were not prioritized for functional analyses but could not be excluded as causal candidates after the log likelihood testing (**Figure 2A**). Furthermore, additional PREs, some

containing such iCHAV2a or 2b variants, are apparent at this locus (**Figure 2**). Twenty six percent of the 62 iCHAV variants, prioritized for functional analyses, are coincident with a PRE. This is an enrichment of the 16% of SNPs at this locus with MAF in the range of the prioritized iCHAV variants (dbSNP 138, MAF=0.04-0.18, accessed through the UCSC Genome Browser) that are located in a PRE.

The risk alleles of iCHAV1 candidates rs74345699 and rs62355900 further induce PRE-C enhancer activity after estrogen stimulation

For functional analysis of iCHAV1, we focused on seven of the 15 candidate causal variants, coincident with a PRE (PREs A-C; **Figure 2**). We first examined the effects of these PREs on *MAP3K1* promoter activity by cloning the relevant genomic regions into luciferase reporter gene constructs containing the *MAP3K1* promoter. All three of the PREs had effects on *MAP3K1* promoter activity: PRE-A acted as a silencer and reduced promoter activity by 62% (P=0.006) in Bre-80 cells (**Figure 3A**) with a similar but non-significant trend (P=0.056) in MCF7 cells (**Figure S5A**). A subregion of PRE-B, termed PRE-B1, containing two candidate variants acted as an enhancer in Bre-80 cells, increasing promoter activity by 42% (P=0.047; **Figure 3B**), but had no significant effect in MCF7 cells – indicating that PRE-B1 may function in specific breast cell types (**Figure S5B**). PRE-C acted as an enhancer and increased *MAP3K1* promoter activity by 90% (P=0.034) in MCF7 (**Figure 3C**) and 77% (P=0.034) in Bre-80 (**Figure S5D**) cells. Introduction of the iCHAV1 minor alleles into the PRE-A, PRE-B1 and PRE-C reference constructs did not detectably alter *MAP3K1* promoter activity (**Figure 3A-C; S5A-B and D**).

ChIP-seq data from the ENCODE project indicate that at least two different transcription factors implicated in estrogen signaling, FOXA1 and ER-α, bind within PRE-B1 and PRE-C, respectively (**Figure 2B**). We thus tested whether the iCHAV1 candidates within PRE-B1 and PRE-C confer estrogen dependent effects on *MAP3K1* promoter activity. We first confirmed that *MAP3K1* expression is up-regulated by estrogen stimulation (**Figure S6**). Then, using reporter assays, we examined the effects of estrogen induction on PRE-B1 and PRE-C by measuring the changes in *MAP3K1* promoter activity between estrogen stimulated and unstimulated cells. We showed that the PRE-C enhancer containing the protective (major) alleles (**Figure 3D**), but not PRE-B1 (**Figure S5C**), has a significant induction in activity, compared with the promoter construct (72%, *P*=0.012), after estrogen stimulation. Compared to this reference PRE-C enhancer, induction was 23% (*P*=0.032) and 15% (*P*=0.011) greater in PRE-C enhancers containing the risk (minor) alleles of iCHAV1 candidate SNPs, rs74345699 and rs62355900, respectively (**Figure 3D**). Of note, neither of these candidates significantly affected *MAP3K1* promoter activity in the absence of estrogen.

The risk allele of iCHAV2a candidate rs16886397 in PRE-D enhances *MAP3K1* promoter activity

Of the 30 candidate causal variants for iCHAV2a at *P*-cond<$1x10^{-4}$, one variant, rs77371588, coincided with a PRE (**Figure 2**) and we thus prioritized this SNP for functional analysis. Using luciferase reporter assays, we demonstrated that the reference PRE-D acted as an enhancer of the *MAP3K1* promoter in Bre-80 cells, increasing *MAP3K1* promoter activity by 69% (*P*=0.013; **Figure S7A**). The PRE-D enhancer containing the risk (minor) allele of rs77371588 had 23% greater enhancer

activity than the reference PRE-D, but this effect did not reach statistical significance (*P*=0.103; **Figure S7A**). By contrast the same PRE-D reference construct did not impact *MAP3K1* promoter activity in MCF7 cells nor did the introduction of the risk allele of rs77371588 into PRE-D significantly alter its activity (**Figure 3E**). As we had generated the PRE-D construct it was straightforward to test rs16886397, an additional iCHAV2a causal candidate located in PRE-D that did not reach the threshold for the functional prioritization (*P*-cond<$1x10^{-4}$) but did pass the likelihood ratio threshold of <100:1 for defining the causal iCHAV candidates. In MCF7 cells, the construct containing the minor (risk) allele of rs16886397 had 21% (*P*=0.049) greater *MAP3K1* promoter activity than the reference PRE-D (**Figure 3E**) and, thus, rs16886397 appears to confer enhancer activity on PRE-D. In Bre-80 cells, in contrast, the minor allele of rs16886397 had no effect on PRE-D activity (**Figure S7A**).

The minor allele of iCHAV2b candidate, rs62355881, increases PRE-B2 enhancer activity

Of the ten iCHAV2b candidates at *P*-cond<$1x10^{-4}$, three variants were coincident with PRE-B and two flanked the boundaries of PRE-C (**Figure 2**). We prioritized the three variants in PRE-B as these were the most compelling functional candidates due to their central location in several functional elements (**Figure 2B**). Using reporter assays, we demonstrated that the reference PRE-B2 construct (a subregion of PRE-B containing the iCHAV2b variants) acted as enhancer and increased the activity of the *MAP3K1* promoter by 152% (*P*=0.032) and 143% (*P*=0.048) in MCF7 and Bre-80 cells, respectively (**Figure 3F**). The introduction of the minor (potentially protective) allele of the iCHAV2b candidate rs62355881 into the reference PRE-B2

construct led to a 29% (*P*=0.017) increase in the enhancer activity of PRE-B2 in MCF7 cells and the haplotype construct containing the minor alleles of all three iCHAV2b candidates demonstrated a similar effect (*P*=0.030; **Figure 3F**). However, these effects were not seen in Bre-80 cells, indicating another cell type specific effect, nor did the other iCHAV2b candidates have any effect on PRE-B2 activity in either cell line (**Figures 3F and S7B**).

The protective allele of iCHAV3 candidate, rs17432750, reduces PRE-B3 enhancer activity and GATA3 binding

Of the five iCHAV3 candidates, we focused on two variants coincident with PRE-B3 (**Figure 2B**), for functional analysis. Using reporter assays, we demonstrated that the reference PRE-B3 construct (a third subregion of PRE-B), containing the risk (major) iCHAV3 alleles, increased *MAP3K1* promoter activity by 166% (*P*=0.041; **Figure 4A**) and 110% (*P*=0.035; **Figure S8A**) in Bre-80 and MCF7 cells, respectively. Reversing the orientation of PRE-B3 in the reporter gene construct had no effect on its activity in either cell line (**Figure S8B and C**), indicating that it acts as a typical enhancer. Next, we introduced the protective (minor) alleles of the iCHAV3 candidates into the reference PRE-B3 construct. The protective A allele of rs17432750 had a repressive effect and reduced PRE-B3 enhancer activity by 43% (*P*=0.024; **Figure 4A**) in Bre-80 cells. The same allele had a similar but non-significant effect (*P*=0.150) in MCF7 cells (**Figure S8A**). By contrast the protective allele of the second iCHAV3 candidate, rs11956804, had no significant effect on enhancer activity in either cell line (**Figures 4A and S8A**). The haplotype construct containing both iCHAV3 variants also had no significant effect on enhancer activity either (**Figures 4A and**

S8A), suggesting the possibility of an interaction between the two iCHAV3 minor alleles in this construct.

We observed from MCF7 ChIP-seq data that a region containing rs17432750 binds the GATA3 transcription factor (**Figure 2B**) and that the sequence around rs17432750 shows homology to the GATA3 position weight matrix (**Figure 4B**). Using GATA3 ChIP-assays, followed by qPCR detection, we confirmed that there was a consistent 2-fold enrichment of the sequence surrounding this SNP in precipitated DNA, when compared to a *CCND1* negative control (**Figure 4C**). We also tested the allele-specificity of GATA3 binding using a Taqman genotyping assay for rs17432750 on ChIP samples from MCF7 cells. The allelic discrimination plot of these data showed an enrichment of the risk (major) C allele in the GATA3 ChIP samples (**Figure 4D**). The ratio of the two alleles in three independent ChIP experiments indicated a 3.7-fold greater GATA3 binding to the risk C allele, compared to the protective A allele, in MCF7 cells.

To determine whether differential GATA3 transcription factor binding may explain the effects of rs17432750 in the reporter assays, we knocked down *GATA3* using siRNA and found the enhancer activity of the reference PRE-B3 construct, containing the risk C allele, was reduced by 33% in Bre-80 cells (P=0.001; **Figure 4E**). *GATA3* knockdown had no effect on the construct containing the protective A allele of (rs17432750) or the *MAP3K1* promoter alone (**Figure 4E**). Diminished GATA3 binding to the protective A allele thus appears to be responsible for the decrease in PRE-B3 enhancer activity observed under basal conditions (**Figure 4A**).

Expression quantitative trait (eQTL) analyses

Given the findings of these functional studies, an obvious hypothesis is that the candidate causal variants in the iCHAVs would be associated with differences in expression of *MAP3K1,* and possibly other local genes in normal breast cells. We therefore explored potential eQTL associations of all locus SNPs with genes lying within ~1 Mb of the locus in 135 normal breast tissue samples from the METABRIC study and 56 further normal breast samples from TCGA. Summary results for representative SNPs from the three iCHAVs are presented in **Table S8.** None of the iCHAV-representative SNPs show detectable differences in *MAP3K1* expression in this dataset. This locus has a positive control eQTL: SNP rs832402 is the most strongly associated SNP with *SETD9* expression in both METABRIC (P=5.93x10^{-9}) and TCGA (P=1.96x10^{-7}), but it is not a strong candidate breast cancer risk SNP (P-cond=1.46x10^{-3}). The positive control SNP is in iCHAV2c (correlated with lead SNP, rs7721581 at r^2=0.74) and SNP rs7721581 is consequently also associated, although less significantly, with *SETD9* expression (METABRIC P=4.38x10^{-8}; TCGA P=2.78x10^{-4}). However, as none of the other representative iCHAV SNPs are associated with *SETD9* expression, it appears unlikely that detectable *SETD9* expression differences in normal breast cells are the underlying cause of breast cancer generated by the candidate functional variants we have identified.

Discussion

In this fine scale mapping study, we have found clear evidence for at least three independent breast cancer risk variants in European women: SNPs in iCHAV1 and iCHAV2, each have the greatest effects on breast cancer in the unadjusted analysis, with minor alleles conferring increased risks of 25-30% for ER+ and ~10% for ER-

tumor-development, while the minor alleles of SNPs in iCHAV3 have a protective effect of ~10% against ER+ breast cancer, although no apparent effect on ER- tumor risk (**Table 1**). The originally detected GWAS tag SNP, rs889312, is most correlated with iCHAV1 (r^2=0.5) and can be excluded from causality within iCHAV1 with a likelihood ratio of >10^{21}:1. Within iCHAV2, we have additionally found evidence of three subsets of variants with a range of effects: iCHAVs 2a, 2b and 2c. The epidemiological analyses suggest iCHAV2c is least likely to be causally related to risk and we conclude that at least one of the variants in iCHAV2a or iCHAV2b is functional. It should be noted that the correlations between candidate causal SNPs in iCHAVs 1 and 2 have added a level of complexity to this analysis and its interpretation that we have not recognized in previous fine-scale mapping studies of breast cancer risk loci.

We have separately identified at least four PREs within the 5q11.2 locus that contain iCHAV candidate variants and interact with the *MAP3K1* promoter in normal and cancer mammary epithelial cells. 3C analysis indicates that there are long range chromatin interactions between these PREs and the *MAP3K1* promoter, while interactions between the PREs and the promoters of other nearby genes are not evident from available ChIA-PET studies.[24,25] Although we cannot rule out interactions between the iCHAVs and the promoters of other genes in the region such as *SETD9* or *MIER3*, a proposed candidate gene for this risk locus,[26] we propose that *MAP3K1* is the likely target gene of the 5q11.2 breast cancer susceptibility locus. Consistent with this proposition and our analyses, Corradin *et al*. identified an enhancer (chr5:56052477-56053943) coincident with the element we

have termed PRE-C that is predicted to regulate *MAP3K1*, based on correlation of cell-type specific H3K4me1 modification and *MAP3K1* expression data.[27]

Our reporter assays indicate that PREs B, C and D act as enhancers while PRE-A is a silencer of the *MAP3K1* promoter. Having identified these regulatory elements we have investigated whether the iCHAV candidate causal variants, within these PREs, detectably modify their regulatory activity. We found that the risk alleles of two iCHAV1 candidates, rs74345699 and rs62355900, in the PRE-C enhancer, act to further induce *MAP3K1* promoter activity in breast cancer cells under estrogen stimulation; the risk allele of iCHAV2a candidate, rs16886397, in PRE-D confers enhancer activity on this PRE for *MAP3K1* promoter activity in breast cancer cells; the potentially protective allele of iCHAV2b candidate, rs62355881, in the PRE-B2 enhancer, increases *MAP3K1* promoter activity in breast cancer cells; while the protective allele of iCHAV3 SNP rs17432750 diminishes the enhancer activity of PRE-B3 for the *MAP3K1* promoter in these cells.

Due to experimental constraints, we have only been able to examine the functions of a minority of the iCHAV causal candidates and, thus, we cannot exclude the presence of more functional variants across the recognizable iCHAVs. Of the candidates we have examined in reporter gene assays, four support a hypothesis that alleles, which increase *MAP3K1* expression, also increase breast cancer risk. Currently, the allelic effects of a fifth candidate, iCHAV2b SNP rs62355881, may be inconsistent with this hypothesis as the conditional analysis suggests iCHAV2b has a protective effect. However, we do not have clear epidemiological evidence that iCHAV2b has an individual effect on risk: iCHAV2b is only observed in the presence

of iCHAV1, which in combination on the same haplotype act to increase breast cancer risk, and we could not test such iCHAV haplotype effects in reporter gene assays, given the number of candidates and the size of the region they encompass.

Consistent with our hypothesis that candidate causal risk alleles act by increasing transcriptional activation of *MAP3K1*, Godde et al. have recently demonstrated that up-regulation of MAP kinase activity in mouse mammary basal progenitor cells is associated with ductal hyperplasia and accelerated tumor progression.[28] This hypothesis is also supported by one known function of *MAP3K1* protein (MEKK1): knockdown of *MAP3K1* in human breast cancer cells reduces tumor invasiveness and progression in a mouse model.[29] Furthermore, studies have shown that MEKK1 has an anti-apoptotic effect and enhances cancer cell survival,[30,31] although, upon stress stimulus, caspase cleavage of the protein generates a fragment that plays a pro-apoptotic role.[6,7] These dual functions of MEKK1 suggest it has a complex role in cell fate decisions. In this context, it is intriguing that somatic *MAP3K1* driver mutations, found in tumor sequencing studies, are mostly truncating and are predicted to disrupt MEKK1 signaling,[11] since inactivation of the kinase domain at the protein terminus reduces apoptotic responses in cells exposed to stress.[7] Thus, it appears that germline cancer risk alleles act by increasing *MAP3K1* expression, but once a tumor has developed, somatic mutations drive cancer progression by disrupting MEKK1 signaling within the tumor. Dysregulation of *MAP3K1* expression or MEKK1 function may thus promote tumorigenesis by perturbing a balance between cell apoptosis and survival. Similar conflicting effects of germline risk variants and somatic mutations have been observed at other breast cancer risk loci. For example, *CCND1* is frequently amplified in breast tumors and here

overexpression appears to play an important role in breast cancer pathogenesis,[32] even though germline breast cancer risk alleles at 11q13 reduce *CCND1* transcriptional activity.[17] Amplification of *TERT* [MIM 187270] is also common within breast and other tumors[33] but, again, germline breast cancer risk alleles reduce *TERT* transcriptional activity.[34] These observations therefore challenge the notion that variants at loci such as 5q11.2 act in the same manner as somatic tumor driver mutations to confer germline risk of tumor development.

iCHAV1 spans multiple PREs (PREs 1, 2 and 3) and is consistent with a recent proposal that genetic susceptibility to common diseases can be explained by multiple enhancer variants in linkage disequilibrium, each with modest effects on gene expression, which cooperatively act to alter gene expression.[27] There is also some suggestion from the haplotype data that iCHAV2a, which is in linkage disequilibrium (r^2=0.19-0.29) with iCHAV1, may in combination with iCHAV1 have a cooperative effect on risk. The modest effect sizes observed in the reporter assays for iCHAV1 and 2a variants may be a consequence of the fact that, due to size limitations of reporter gene constructs, we could not examine these variants in combination.

Some SNP effects in our reporter assays were suggestive of cell-line and stimulus-dependence, highlighting the importance of cellular and environmental context when assessing SNP functionality.[23] Similarly, we defined our PREs on the basis of chromatin modification and conformation states, both of which can change during development or in response to stimulus. Additionally, the effects of some SNPs may not have been observed because plasmid reporter gene constructs do not reflect the native genomic context nor the chromatin or methylation state of genomic DNA. The

chromatin state of transiently transfected DNA, for example, is considered to be more open and disorganized than the corresponding chromatin in the native genomic context and may not have the repressive chromatin structures, found in genomic DNA, that inhibit binding of ubiquitous transcription factors.[35] A disorganized chromatin state may explain the inconsistent effect of rs17432750 in the PRE-B3 constructs (**Figure 4A**). The abrogation of the effect of rs17432750 in the haplotype construct suggests some interaction or cooperative effect of the minor alleles of rs17432750 and rs11956804 to enhance promoter activity. This effect may be possible due to the more open and permissive chromatin structures associated with transfected plasmid DNA.

It is noteworthy that, in available normal breast tissue, the top candidate causal variants show no association with *MAP3K1* expression, although iCHAV2c variants are associated with significant differences in *SETD9* mRNA levels. Similarly, normal breast tissue eQTL studies with strong candidate causal variants at the 11q13,[17]10q26,[15] 2q35[36] and 5p15.33[34] breast cancer loci have indicated that available normal breast samples may be inappropriate for these studies. Sample sizes are large enough to detect significant eQTLs at these loci but those detected do not appear to drive breast cancer risk. It is possible that tissue heterogeneity, developmental stage or stimulus-dependent effects prevent the detection of risk driving eQTLs in currently available normal breast samples. Indeed, the finding that up-regulation of MAP kinase activity in mammary progenitor cells is associated with mammary tumorigenesis[28] suggests that increased *MAP3K1* expression in specific breast cell populations, possibly at a specific point in time, could drive breast cancer risk.

Transcription factor binding studies indicate that the C (risk) allele of rs17432750 preferentially binds GATA3 over the A (protective) allele. Increased binding of the C allele by GATA3 appears to explain the activity of the PRE-B3 enhancer in which it is located (**Figure 4E**). These findings suggest that SNP rs17432750 is a strong causal candidate for the protective effect of iCHAV3. The transcription factor GATA3 has multiple regulatory roles and can affect histone modifications associated with enhancers, and the binding of other transcription factors related to breast cancer such as ER-α and FOXA1.[37] We have previously identified GATA3 to be a mediator of breast cancer risk across multiple loci[38] and specifically at the 11q13 locus.[17]

In conclusion, we have found evidence for the existence of at least three breast cancer risk iCHAVs that partially coincide with four *MAP3K1* gene regulatory elements at 5q11.2. Genetic epidemiological studies within BCAC reduced the catalogue of potentially causal variants from 909 to 193 candidates within five iCHAVs, of which at least three must be functional. Functional studies on candidates that lie within the identified regulatory elements have shown that the effects of strong candidate cancer risk alleles, in iCHAVs 1, 2a and 3, are compatible with the hypothesis that they act via increased expression of *MAP3K1*. Moreover, the function of *MAP3K1* protein (MEKK1) suggests that increased expression may alter the balance between apoptosis and cell survival in breast cancer cells, thus explaining the risks conferred by the candidate alleles.

Supplemental Data

Supplemental Data include Acknowledgements, nine figures and eight tables.

Web Resources

1,000 Genomes Project: http://www.1000genomes.org/

R: http://www.bios.unc.edu/~weisun/software/eMap

UCSC Genome Browser: http://genome.ucsc.edu

Online Mendelian Inheritance in Man: http://www.omim.org

References

1. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447, 1087-1093.

2. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al. (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 45, 353-361, 361e351-352.

3. Broeks, A., Schmidt, M.K., Sherman, M.E., Couch, F.J., Hopper, J.L., Dite, G.S., Apicella, C., Smith, L.D., Hammet, F., Southey, M.C., et al. (2011). Low penetrance breast cancer susceptibility loci are associated with specific breast tumor subtypes: findings from the Breast Cancer Association Consortium. Hum Mol Genet 20, 3289-3303.

4. Antoniou, A.C., Spurdle, A.B., Sinilnikova, O.M., Healey, S., Pooley, K.A., Schmutzler, R.K., Versmold, B., Engel, C., Meindl, A., Arnold, N., et al. (2008). Common breast cancer-predisposition alleles are associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers. Am J Hum Genet 82, 937-948.

5. Keshet, Y., and Seger, R. (2010). The MAP kinase signaling cascades: a system of hundreds of components regulates a diverse array of physiological functions. Methods Mol Biol 661, 3-38.

6. Schlesinger, T.K., Bonvin, C., Jarpe, M.B., Fanger, G.R., Cardinaux, J.R., Johnson, G.L., and Widmann, C. (2002). Apoptosis stimulated by the 91-kDa caspase cleavage MEKK1 fragment requires translocation to soluble cellular compartments. J Biol Chem 277, 10283-10291.

7. Widmann, C., Gerwins, P., Johnson, N.L., Jarpe, M.B., and Johnson, G.L. (1998). MEK kinase 1, a substrate for DEVD-directed caspases, is involved in genotoxin-induced apoptosis. Mol Cell Biol 18, 2416-2429.

8. Alarcon-Vargas, D., Tansey, W.P., and Ronai, Z. (2002). Regulation of c-myc stability by selective stress conditions and by MEKK1 requires aa 127-189 of c-myc. Oncogene 21, 4384-4391.

9. Fuchs, S.Y., Adler, V., Pincus, M.R., and Ronai, Z. (1998). MEKK1/JNK signaling stabilizes and activates p53. Proc Natl Acad Sci U S A 95, 10541-10546.

10. Clarke, N., Arenzana, N., Hai, T., Minden, A., and Prywes, R. (1998). Epidermal growth factor induction of the c-jun promoter by a Rac pathway. Mol Cell Biol 18, 1065-1073.

11. Stephens, P.J., Tarpey, P.S., Davies, H., Van Loo, P., Greenman, C., Wedge, D.C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G.R., et al. (2012). The

landscape of cancer genes and mutational processes in breast cancer. Nature 486, 400-404.

12. Nordgard, S.H., Johansen, F.E., Alnaes, G.I., Naume, B., Borresen-Dale, A.L., and Kristensen, V.N. (2007). Genes harbouring susceptibility SNPs are differentially expressed in the breast cancer subtypes. Breast Cancer Res 9, 113.

13. Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.

14. Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.

15. Meyer, K.B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S.L., French, J.D., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., et al. (2013). Fine-Scale Mapping of the FGFR2 Breast Cancer Risk Locus: Putative Functional Variants Differentially Bind FOXA1 and E2F1. Am J Hum Genet 93, 1046-1060.

16. Tan-Wong, S.M., French, J.D., Proudfoot, N.J., and Brown, M.A. (2008). Dynamic interactions between the promoter and terminator regions of the mammalian BRCA1 gene. Proc Natl Acad Sci U S A 105, 5160-5165.

17. French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al. (2013).

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.

18. Prall, O.W., Sarcevic, B., Musgrove, E.A., Watts, C.K., and Sutherland, R.L. (1997). Estrogen-induced activation of Cdk4 and Cdk2 during G1-S phase progression is accompanied by increased cyclin D1 expression and decreased cyclin-dependent kinase inhibitor association with cyclin E-Cdk2. J Biol Chem 272, 10882-10894.
19. Schmidt, D., Wilson, M.D., Spyrou, C., Brown, G.D., Hadfield, J., and Odom, D.T. (2009). ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. Methods 48, 240-248.
20. Curtis, C., Shah, S.P., Chin, S.F., Turashvili, G., Rueda, O.M., Dunning, M.J., Speed, D., Lynch, A.G., Samarajiwa, S., Yuan, Y., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature 486, 346-352.
21. Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. Nature 490, 61-70.
22. Marchini, J., and Howie, B. (2010). Genotype imputation for genome-wide association studies. Nature reviews Genetics 11, 499-511.
23. Edwards, S.L., Beesley, J., French, J.D., and Dunning, A.M. (2013). Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet 93, 779-797.
24. Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 148, 84-98.

25. Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462, 58-64.

26. denDekker, A.D., Xu, X., Vaughn, M.D., Puckett, A.H., Gardner, L.L., Lambring, C.J., Deschenes, L., and Samuelson, D.J. (2012). Rat Mcs1b is concordant to the genome-wide association-identified breast cancer risk locus at human 5q11.2 and MIER3 is a candidate cancer susceptibility gene. Cancer Res 72, 6002-6012.

27. Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal-Lari, R., Lupien, M., Markowitz, S., and Scacheri, P.C. (2013). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. Genome Res.

28. Godde, N.J., Sheridan, J.M., Smith, L.K., Pearson, H.B., Britt, K.L., Galea, R.C., Yates, L.L., Visvader, J.E., and Humbert, P.O. (2014). Scribble modulates the MAPK/Fra1 pathway to disrupt luminal and ductal integrity and suppress tumour formation in the mammary gland. PLoS Genet 10, e1004323.

29. Cuevas, B.D., Winter-Vann, A.M., Johnson, N.L., and Johnson, G.L. (2006). MEKK1 controls matrix degradation and tumor cell dissemination during metastasis of polyoma middle-T driven mammary cancer. Oncogene 25, 4998-5010.

30. Hirano, T., Shino, Y., Saito, T., Komoda, F., Okutomi, Y., Takeda, A., Ishihara, T., Yamaguchi, T., Saisho, H., and Shirasawa, H. (2002). Dominant negative MEKK1 inhibits survival of pancreatic cancer cells. Oncogene 21, 5923-5928.

31. Nawata, R., Yujiri, T., Nakamura, Y., Ariyoshi, K., Takahashi, T., Sato, Y., Oka, Y., and Tanizawa, Y. (2003). MEK kinase 1 mediates the antiapoptotic effect

of the Bcr-Abl oncogene through NF-kappaB activation. Oncogene 22, 7774-7780.

32. Arnold, A., and Papanikolaou, A. (2005). Cyclin D1 in breast cancer pathogenesis. J Clin Oncol 23, 4215-4224.

33. Cao, Y., Bryan, T.M., and Reddel, R.R. (2008). Increased copy number of the TERT and TERC telomerase subunit genes in cancer cells. Cancer Sci 99, 1092-1099.

34. Bojesen, S.E., Pooley, K.A., Johnatty, S.E., Beesley, J., Michailidou, K., Tyrer, J.P., Edwards, S.L., Pickett, H.A., Shen, H.C., Smart, C.E., et al. (2013). Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet 45, 371-384, 384e371-372.

35. Smith, C.L., and Hager, G.L. (1997). Transcriptional regulation of mammalian genes in vivo. A tale of two templates. J Biol Chem 272, 27493-27496.

36. Ghoussaini, M., Edwards, S.L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., Kar, S., Hillman, K.M., Kaufmann, S., Glubb, D.M., et al. (2014). Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. Nat Commun 4, 4999.

37. Theodorou, V., Stark, R., Menon, S., and Carroll, J.S. (2013). GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res 23, 12-22.

38. Fletcher, M.N.C., Castro, M.A.A., Wang, X., de Santiago, I., O'Reilly, M., Chin, S.-F., Rueda, O.M., Caldas, C., Ponder, B.A.J., Markowetz, F., et al. (2013). Master regulators of FGFR2 signalling and breast cancer risk. Nat Commun 4, 2464.

Figure Titles and Legends

Figure 1. Genetic mapping and chromatin state of the 5q11.2 locus. (A) Manhattan plot of overall breast cancer risk in Europeans at the 5q11.2 locus. SNPs are plotted based on their chromosomal position on the x axis and *P*-values (log_{10} values) for association. The span of the iCHAVs in terms of chromosomal location and *P*-value is displayed using shading and candidate causal variants from the iCHAV are colored black. The lead SNPs from each iCHAV, the original GWAS tag SNP (rs889312) and the three genes present in the region are shown. The dotted line intersects the y-axis at $P=10^{-8}$ and indicates conventional genome-wide significance. (B) The chromatin state of the 5q11.2 locus in human mammary epithelial cells (HMECs) is shown using ENCODE ChIP-seq data from H3K4me3, H3K4me2, H3k4me1 and DNaseI studies accessed from the UCSC Genome Browser (http://genome.ucsc.edu). Transcription factor (TF) binding from ENCODE ChIP-seq studies of 161 TFs in 91 cell lines is also displayed.

Figure 2. Candidate causal variants are located in PREs that interact with the *MAP3K1* promoter. (A) The candidate causal variants associated with breast cancer risk from iCHAVs 1, 2a, 2b and 3 were mapped to PREs at the 5q11.2 locus. (B) PREs (highlighted) were identified using ChIP-seq data (H3K4Me1 studies in 7 ENCODE cell lines (GM12878, H1-hESC, HSMM, HUVEC, K562, NHEK and NHLF); H3K4Me1 and H3K4Me2 in MCF7 and HMEC cells; transcription factors ER-α, FOXA1 and GATA3 in MCF7 cells) accessed from the UCSC Genome Browser. Regions cloned into reporter gene constructs are also shown. (C) 3C analysis of

interactions between EcoRI fragments at the 5q11.2 locus, encompassing the PREs coincident with candidate causal variants, and the *MAP3K1* promoter in Bre-80 cells (error bars represent standard deviation and a representative graph is shown).

Figure 3. Risk (minor) alleles of iCHAV1 and iCHAV2a SNPs enhance *MAP3K1* promoter activity in luciferase reporter assays. PRE-A, PRE-B, PRE-C and PRE-D regions containing the major allelic variants of iCHAV1, iCHAV2a and 2b SNPs were cloned downstream of a *MAP3K1* promoter-driven luciferase construct (Promoter) to create reference (Ref) PRE constructs. Minor allelic variants of the iCHAV1, iCHAV2a and 2b SNPs were engineered into the constructs and are designated by the rs ID of the corresponding SNP. Constructs containing minor allelic haplotypes (Haplotype) were also generated. Cells were transiently transfected with each of these constructs and assayed for luciferase activity after 24 h. Panels (A-B) show results from assays of PRE-A and PRE-B1, respectively, in Bre-80 cells. Panels (C, E and F) show results from assays of PRE-C, PRE-D and PRE-B2 constructs, respectively, in MCF7 cells under basal conditions. Panel (D) shows results after estrogen induction of MCF7 cells. For each reporter construct in this assay, the luciferase activity of estrogen treated cells was normalized to the activity of the corresponding vehicle treated cells. Error bars denote standard error of the mean (SEM) from three experiments performed with triplicates. *P*-values were determined by repeated-measures ANOVA followed by Dunnett's multiple comparisons test (**P*<0.05 and ***P*<0.01).

Figure 4. The protective (minor) allele of iCHAV3 SNP rs17432750 demonstrates diminished PRE-B3 enhancer activity in luciferase reporter assays and reduced GATA3 binding in ChIP analysis. (A) The PRE-B3 region containing the major allelic variants of iCHAV3 SNPs was cloned downstream of a *MAP3K1* promoter-driven luciferase construct (Promoter) to create reference (Ref) PRE constructs. Minor allelic variants of iCHAV3 SNPs were engineered into the constructs which are designated by the rs ID of the corresponding SNP. A construct containing the minor allelic haplotypes (Haplotype) was also generated. Bre-80 cells were transiently transfected with each of these constructs and assayed for luciferase activity after 24 h. Error bars denote SEM from three experiments performed in triplicate. *P*-values were determined by repeated-measures ANOVA followed by Dunnett's multiple comparisons test (**P*<0.05). (B) Position Weight Matrix of GATA3 is shown relative to the negative strand of the sequences surrounding rs17432750. (C) GATA3 ChIP assays demonstrate enrichment of rs17432750 relative to the *CCND1* negative control. A GATA3 site from the ER-α enhancer was included as positive control. Results from two biological repeats are shown and error bars denote SD of three technical repeats. (D) Genotyping of rs17432750 in MCF7 genomic DNA versus MCF7 GATA3-ChIP DNA. Homozygous cell lines ZR751 (C/C), T47D (C/C), MDA-MB-468 (A/A) and no template controls (NTC) were included as reference for the assay. The risk (major) C allele is preferentially precipitated in the ChIP experiment. (E) Luciferase assay in Bre-80 cells showing the effect of *GATA3* siRNA silencing on the activity of the *MAP3K1* promoter alone (Promoter) and with PRE-B3 constructs containing the C allele (ref PRE-B3) and protective A (minor) allele rs17432750 (rs17432750). Error bars denote SEM from three experiments performed in triplicate. *P*-values were determined by two-way repeated-measures

ANOVA followed either by Sidak's multiple comparisons test (to analyze the effect of *GATA3* knockdown within constructs) or Dunnett's multiple comparison test (to analyze differences in activity between constructs). (**P*<0.05, ***P*<0.01 and ****P*<0.001). The level of *GATA3* knockdown is shown in **Figure S9**.

Tables

iCHAV	SNP	Chr Position (GRch37)	Alleles (major/ minor)	MAF	r^2 with lead iCHAV1 SNP	Imp r^2	Overall breast cancer risk				ER-positive breast cancer risk				ER-negative breast cancer risk				ER-positive /negative
							OR	(95%CI)	*P*-trend	*P*-cond	OR	(95%CI)	*P*-trend	*P*-cond	OR	(95%CI)	*P*-trend	*P*-cond	*P*-diff
1	rs10461612	56003831	G/C	0.18	0.96	0.97	1.21	(1.18-1.23)	1.08E-47		1.23	(1.20-1.26)	1.71E-43		1.09	(1.04-1.14)	4.13E-04		1.47E-05
1	rs7709971	56007339	G/A	0.18	0.96	0.97	1.21	(1.18-1.23)	1.05E-47		1.24	(1.21-1.27)	1.33E-43		1.09	(1.04-1.14)	4.86E-04		1.17E-05
1	rs7714232	56011357	A/T	0.18	0.94	0.99	1.21	(1.18-1.23)	4.76E-48		1.23	(1.20-1.26)	5.09E-44		1.09	(1.04-1.14)	5.64E-04		8.03E-06
1	rs12653202	56016918	A/C	0.18	0.94	1	1.21	(1.18-1.23)	2.17E-47		1.23	(1.20-1.26)	7.65E-42		1.09	(1.04-1.14)	5.04E-04		2.68E-05
1	rs59270457	56017887	T/G	0.18	0.96	1	1.21	(1.18-1.23)	1.22E-47		1.23	(1.20-1.26)	1.02E-42		1.09	(1.04-1.14)	4.48E-04		1.76E-05
1	rs61055995	56019064	A/T	0.18	0.96	1	1.21	(1.18-1.23)	1.16E-47		1.23	(1.20-1.26)	7.36-E43		1.09	(1.04-1.14)	4.70E-04		1.62E-05
1	rs16886165	56023083	T/G	0.18	0.96	1	1.21	(1.18-1.23)	1.14E-47		1.23	(1.20-1.26)	6.22E-43		1.09	(1.04-1.14)	4.56E-04		1.55E-05
1	rs16886181	56029243	T/C	0.18	0.96	1	1.21	(1.18-1.23)	4.80E-48		1.23	(1.20-1.26)	2.87E-43		1.10	(1.05-1.15)	2.69E-04		2.26E-05
1	chr5:56030827:I	56030827	T/TA	0.18	0.97	0.99	1.20	(1.18-1.23)	2.33E-47		1.23	(1.20-1.26)	1.27E-42		1.09	(1.04-1.14)	5.82E-04		1.24E-05
1	rs66893416	56051596	G/A	0.18	0.99	0.97	1.21	(1.18-1.23)	1.19E-47		1.23	(1.20-1.26)	4.60E-43		1.09	(1.04-1.14)	4.42E-04		1.05E-05
1	rs62355900	56052695	T/C	0.18	0.98	1	1.20	(1.18-1.23)	4.26E-47		1.23	(1.20-1.26)	8.93E-43		1.09	(1.04-1.14)	5.84E-04		1.21E-05
1	rs74345699	56053479	C/T	0.18	0.99	0.99	1.20	(1.18-1.23)	5.08E-47		1.23	(1.20-1.26)	1.57E-42		1.09	(1.04-1.14)	7.01E-04		1.05E-05
1	rs62355901	56053535	T/C	0.18	0.98	0.93	1.21	(1.18-1.24)	3.90E-47		1.24	(1.20-1.27)	5.41E-42		1.10	(1.04-1.15)	5.20E-04		1.49E-05
1	rs62355902 [a]	56053723	A/T	0.18	-	0.95	1.21	(1.19-1.24)	9.50E-49		1.24	(1.21-1.27)	5.71E-44		1.10	(1.05-1.15)	3.02E-04		1.47E-05
1	rs149188233	56053745	A/AAAAC	0.18	0.99	0.98	1.20	(1.18-1.23)	3.92E-47		1.23	(1.20-1.26)	1.00E-42		1.09	(1.04-1.14)	6.48E-04		1.13E-05
2a	rs77961606	56040643	T/C	0.05	0.28	0.98	1.27	(1.22-1.31)	1.03E-28	9.21E-05	1.29	(1.25-1.34)	3.42E-25	4.55E-04	1.09	(1.01-1.18)	3.59E-02	7.58E-01	1.04E-04
2a	rs16886272	56067434	A/G	0.07	0.19	0.95	1.21	(1.17-1.25)	8.89E-24	4.14E-05	1.25	(1.20-1.29)	3.41E-23	1.01E-05	1.07	(0.99-1.14)	8.47E-02	8.08E-01	5.42E-05
2a	rs113317823[a]	56087883	T/C	0.08	0.19	0.72	1.22	(1.18-1.26)	7.00E-25	1.61E-05	1.24	(1.20-1.29)	2.02E-21	9.74E-05	1.12	(1.05-1.20)	2.56E-03	9.90E-02	1.93E-02
2a	rs77371588	56134560	G/T	0.05	0.26	0.98	1.27	(1.22-1.31)	1.36E-27	9.03E-05	1.29	(1.25-1.34)	2.01E-24	2.68E-04	1.11	(1.02-1.19)	2.11E-02	5.37E-01	3.32E-04
2a	rs60590641	56141155	G/A	0.05	0.26	0.98	1.27	(1.22-1.31)	1.16E-27	7.93E-05	1.30	(1.25-1.35)	1.08E-24	2.01E-04	1.11	(1.02-1.19)	2.10E-02	5.33E-01	3.05E-04
2a	rs61154548	56224720	C/T	0.05	0.24	0.97	1.27	(1.23-1.32)	4.32E-27	2.76E-05	1.30	(1.25-1.36)	4.03E-24	9.06E-05	1.10	(1.01-1.19)	3.65E-02	6.19E-01	3.37E-04
2a	chr5:56237882:D	56237882	CAGTTAAGTTTA/C	0.04	0.23	0.93	1.28	(1.24-1.33)	1.82E-25	8.94E-05	1.32	(1.26-1.37)	3.49E-23	1.57E-04	1.10	(1.01-1.20)	3.80E-02	6.02E-01	5.14E-04
2a	rs79565352	56251753	G/A	0.06	0.20	1	1.23	(1.19-1.27)	7.45E-24	6.46E-05	1.25	(1.20-1.29)	8.52E-20	7.69E-04	1.11	(1.03-1.19)	1.42E-02	3.04E-01	2.20E-03
2a	rs74571895	56254772	A/G	0.05	0.20	0.99	1.24	(1.20-1.28)	1.71E-24	3.62E-05	1.26	(1.21-1.30)	1.54E-20	3.99E-04	1.11	(1.03-1.19)	1.16E-02	2.70E-01	2.10E-03
2a	rs78743305	56256217	G/A	0.05	0.20	0.99	1.24	(1.20-1.28)	1.41E-24	3.28E-05	1.26	(1.21-1.30)	1.20E-20	3.54E-04	1.11	(1.03-1.19)	1.11E-02	2.64E-01	2.03E-03
2a	rs7726354	56256483	C/T	0.06	0.20	1	1.23	(1.19-1.27)	4.51E-24	5.13E-05	1.25	(1.20-1.30)	4.61E-20	5.90E-04	1.11	(1.03-1.19)	1.30E-02	2.88E-01	2.11E-03

2a	rs79760198	56262639	T/C	0.05	0.23	0.98	1.27	(1.23-1.31)	1.39E-26	2.71E-05	1.30	(1.25-1.35)	3.04E-23	1.34E-04	1.11	(1.03-1.20)	1.74E-02	3.91E-01	1.23E-03
2a	rs79041328	56264375	A/G	0.05	0.23	0.99	1.27	(1.23-1.31)	1.48E-26	2.70E-05	1.30	(1.24-1.35)	3.42E-23	1.38E-04	1.11	(1.03-1.20)	1.60E-02	3.71E-01	1.42E-03
2a	rs6893174	56265025	C/T	0.05	0.22	1	1.26	(1.22-1.30)	8.59E-26	4.75E-05	1.28	(1.23-1.33)	2.83E-22	2.77E-04	1.11	(1.02-1.19)	1.88E-02	4.02E-01	1.56E-03
2a	rs113325879	56266142	G/A	0.05	0.23	0.99	1.27	(1.23-1.31)	1.30E-26	2.53E-05	1.30	(1.25-1.35)	3.73E-23	1.45E-04	1.11	(1.03-1.20)	1.52E-02	3.59E-01	1.47E-03
2a	rs142258027	56266278	G/A	0.05	0.22	0.96	1.27	(1.22-1.31)	8.33E-26	4.57E-05	1.29	(1.24-1.34)	1.83E-22	2.34E-04	1.11	(1.02-1.20)	1.97E-02	4.12E-01	1.46E-03
2a	rs113173541	56266689	C/A	0.05	0.23	0.98	1.27	(1.23-1.31)	1.34E-26	2.57E-05	1.30	(1.25-1.35)	4.05E-23	1.51E-04	1.11	(1.03-1.20)	1.52E-02	3.60E-01	1.49E-03
2a	rs78925509	56267141	A/G	0.05	0.23	0.98	1.27	(1.23-1.31)	1.41E-26	2.63E-05	1.30	(1.25-1.35)	4.38E-23	1.56E-04	1.11	(1.03-1.20)	1.55E-02	3.63E-01	1.49E-03
2a	rs73122135	56267308	G/A	0.05	0.23	0.96	1.27	(1.22-1.31)	1.93E-26	3.14E-05	1.29	(1.24-1.35)	4.03E-23	1.63E-04	1.11	(1.02-1.19)	1.98E-02	4.30E-01	1.02E-03
2a	rs80089016	56267444	G/A	0.05	0.23	0.98	1.27	(1.23-1.32)	1.43E-26	2.64E-05	1.30	(1.25-1.35)	4.61E-23	1.59E-04	1.11	(1.03-1.20)	1.54E-02	3.62E-01	1.52E-03
2a	rs73122138	56267949	G/A	0.05	0.23	0.96	1.26	(1.22-1.31)	6.61E-26	5.47E-05	1.29	(1.24-1.34)	1.54E-22	3.19E-04	1.11	(1.02-1.19)	2.27E-02	4.66E-01	1.17E-03
2a	rs80310238	56267984	C/A	0.05	0.23	0.98	1.27	(1.23-1.32)	1.45E-26	2.67E-05	1.30	(1.25-1.35)	5.00E-23	1.66E-04	1.11	(1.03-1.20)	1.55E-02	3.64E-01	1.53E-03
2a	rs77706078	56268686	G/A	0.05	0.23	0.98	1.27	(1.23-1.32)	1.40E-26	2.63E-05	1.30	(1.25-1.35)	5.05E-23	1.67E-04	1.11	(1.03-1.20)	1.52E-02	3.59E-01	1.57E-03
2a	chr5:56268884:I	56268884	C/CTGG GAGGT	0.04	0.22	0.96	1.28	(1.24-1.33)	8.44E-27	1.76E-05	1.31	(1.25-1.36)	3.05E-23	1.16E-04	1.12	(1.03-1.21)	1.15E-02	3.00E-01	2.05E-03
2a	rs2408651	56268991	T/C	0.05	0.23	0.95	1.26	(1.22-1.31)	1.08E-25	6.53E-05	1.29	(1.24-1.34)	2.10E-22	3.50E-04	1.11	(1.02-1.19)	2.19E-02	4.57E-01	1.19E-03
2a	rs79459889	56269286	G/A	0.05	0.23	0.98	1.27	(1.23-1.31)	2.09E-26	3.08E-05	1.29	(1.24-1.35)	8.65E-23	2.06E-04	1.12	(1.03-1.20)	1.43E-02	3.49E-01	1.74E-03
2a	rs111773762	56269336	T/C	0.05	0.23	0.98	1.27	(1.23-1.31)	2.53E-26	3.33E-05	1.29	(1.24-1.35)	7.04E-23	1.87E-04	1.11	(1.03-1.20)	1.50E-02	3.58E-01	1.57E-03
2a	rs80097053	56269492	C/T	0.05	0.23	0.98	1.27	(1.23-1.31)	2.50E-26	3.31E-05	1.29	(1.24-1.35)	7.03E-23	1.87E-04	1.12	(1.03-1.20)	1.49E-02	3.56E-01	1.58E-03
2a	rs59536253	56269512	A/G	0.05	0.23	0.96	1.27	(1.22-1.31)	2.84E-26	3.73E-05	1.29	(1.24-1.34)	8.40E-23	2.26E-04	1.11	(1.03-1.20)	1.64E-02	3.88E-01	1.40E-03
2a	rs112032073	56270717	G/T	0.05	0.23	0.98	1.27	(1.23-1.31)	2.40E-26	3.27E-05	1.30	(1.24-1.35)	7.36E-23	1.92E-04	1.12	(1.03-1.20)	1.41E-02	3.44E-01	1.69E-03
2b	rs12659430	56032619	A/G	0.13	0.62	1.00	1.15	(1.12-1.18)	1.73E-19	4.42E-05	1.17	(1.13-1.20)	3.95E-18	1.81E-03	1.08	(1.02-1.14)	8.95E-03	6.26E-01	1.96E-02
2b	rs62355881	56033093	T/C	0.13	0.62	1.00	1.15	(1.12-1.18)	1.61E-19	4.68E-05	1.17	(1.13-1.20)	3.62E-18	1.92E-03	1.08	(1.02-1.14)	8.95E-03	6.26E-01	1.93E-02
2b	rs2113084	56033671	T/C	0.13	0.62	1.00	1.15	(1.12-1.18)	1.97E-19	4.02E-05	1.17	(1.13-1.20)	4.80E-18	1.62E-03	1.08	(1.02-1.14)	8.92E-03	6.27E-01	2.02E-02
2b	rs62355882	56045110	C/T	0.13	0.60	1.00	1.15	(1.12-1.18)	3.69E-19	7.81E-05	1.16	(1.13-1.20)	2.64E-17	1.06E-03	1.08	(1.02-1.14)	1.22E-02	5.60E-01	2.42E-02
2b	rs62355899[a]	56050465	A/G	0.13	0.62	1.00	1.15	(1.12-1.18)	2.98E-19	3.04E-05	1.17	(1.13-1.20)	5.52E-18	2.04E-03	1.08	(1.02-1.14)	1.11E-02	5.42E-01	1.86E-02
2b	rs10513090	56054065	T/C	0.13	0.62	1.00	1.15	(1.12-1.18)	1.91E-19	3.67E-05	1.17	(1.13-1.20)	2.92E-18	2.79E-03	1.08	(1.02-1.13)	1.43E-02	4.46E-01	1.32E-02
2b	rs6895844	56057940	G/A	0.13	0.62	1.00	1.15	(1.12-1.18)	7.44E-20	7.50E-05	1.17	(1.13-1.20)	1.17E-18	4.66E-03	1.07	(1.02-1.13)	1.63E-02	3.99E-01	1.08E-02
2b	rs6882255	56058171	T/A	0.13	0.63	0.97	1.15	(1.12-1.18)	3.03E-20	9.13E-05	1.17	(1.14-1.21)	8.02E-19	5.26E-03	1.08	(1.02-1.13)	1.47E-02	4.15E-01	9.86E-03
2b	rs12654584	56058415	T/G	0.13	0.62	1.00	1.15	(1.12-1.18)	5.90E-20	8.57E-05	1.17	(1.13-1.20)	8.79E-19	5.31E-03	1.07	(1.02-1.13)	1.61E-02	4.01E-01	1.01E-02
2b	rs12654176	56058487	A/G	0.13	0.62	1.00	1.15	(1.12-1.18)	4.29E-20	9.99E-05	1.17	(1.14-1.20)	8.04E-19	5.66E-03	1.08	(1.02-1.13)	1.33E-02	4.64E-01	1.19E-02
3	rs17432750	56031822	C/A	0.16	0.04	0.98	0.92	(0.89-0.94)	4.55E-11	1.61E-04	0.90	(0.87-0.93)	9.03E-11	1.48E-04	1.01	(0.96-1.06)	6.53E-01	2.44E-01	5.13E-05

3	rs199945768	56033233	CG/C	0.16	0.04	0.92	0.91	(0.89-0.94)	6.35E-11	2.66E-04	0.90	(0.87-0.93)	1.90E-10	3.09E-04	1.01	(0.96-1.06)	6.61E-01	2.49E-01	6.53E-05
3	rs11956804	56033233	C/A	0.16	0.04	0.94	0.91	(0.89-0.94)	7.39E-11	1.86E-04	0.90	(0.87-0.93)	1.90E-10	2.05E-04	1.01	(0.96-1.06)	6.61E-01	2.22E-01	7.44E-05
3	rs11949391[a]	56045081	T/C	0.16	0.04	1	0.91	(0.89-0.94)	9.36E-12	5.57E-05	0.90	(0.87-0.93)	1.00E-10	1.44E-04	1.01	(0.96-1.06)	8.14E-01	3.50E-01	1.27E-04
3	rs6884514	56062614	C/T	0.16	0.04	1	0.92	(0.89-0.95)	6.04E-10	5.80E-04	0.91	(0.88-0.94)	1.90E-09	6.74E-04	1.01	(0.96-1.06)	6.15E-01	2.30E-01	1.13E-04

Table 1. iCHAVs with the strongest effects on breast cancer risk in European studies

Single SNP risk estimates for the top candidates in each iCHAV with overall breast cancer risk and subtypes by estrogen receptor status. Results are given as ORs with 95% CI (using the minor variant alleles as the reference), per-allele *P*-trend and *P*-cond, (*P*-cond being conditional on the iCHAV1 lead SNP rs62355902). The ER-positive/negative *P*-diff is from a case-only analysis, comparing the effect sizes in the subtypes by estrogen receptor status. For iCHAVs 2a and 2b, only SNPs with *P*-cond<$1x10^{-4}$ for overall breast cancer risk are shown here - full results are listed in **Table S5** and all other iCHAV2a and 2b candidate variants are listed in **Table S7**.

[a]The best candidate SNP in the iCHAV.

Table 2. Breast cancer risk in Europeans by haplotypes of five iCHAV representative SNPs

							Overall Breast cancer			ER+ Breast cancer			ER- Breast cancer		
Haplotype	**iCHAV1[a]**	**iCHAV2a[b]**	**iCHAV2b[c]**	**iCHAV2c[d]**	**iCHAV3[e]**	**Frequency**	**OR**	**(95%CI)**	***P*-value**	**OR**	**(95%CI)**	***P*-value**	**OR**	**(95%CI)**	***P*-value**
h_0[f]	1	1	1	1	1	0.52	-	-	-	-	-	-	-	-	-
h1	1	1	1	2	1	0.12	1.01	(0.99-1.05)	4.50E-01	1.03	(0.99-1.07)	1.12E-01	1.04	(0.98-1.11)	2.08E-01
h2	1	1	1	1	2	0.15	0.95	(0.92-0.97)	2.08E-04	0.95	(0.92-0.99)	5.23E-03	1.04	(0.98-1.10)	1.67E-01
h3	1	2	1	1	1	0.02	1.08	(0.99-1.17)	7.57E-02	1.13	(1.02-1.24)	1.68E-02	1.08	(0.92-1.27)	3.45E-01
h4	1	2	1	2	1	0.01	1.05	(0.99-1.17)	3.81E-01	0.99	(0.87-1.13)	8.75E-01	1.17	(0.96-1.42)	1.12E-01
h5	2	1	2	1	1	0.11	1.16	(1.12-1.20)	2.00E-20	1.19	(1.14-1.23)	9.37E-20	1.10	(1.04-1.18)	1.82E-03
h6	2	2	1	2	1	0.05	1.31	(1.25-1.37)	1.66E-29	1.34	(1.27-1.42)	3.74E-27	1.15	(1.05-1.25)	3.64E-03

[a] rs62355902, representing 15 SNPs
[b] rs113317823,representing 90 SNPs which have $r^2 \geq 0.53$ with rs113317823 and r^2 in [0.19-0.29] with rs62355902
[c] rs62355899, representing 66 SNPs which have $r^2 \leq 0.01$ with rs113317823 and r^2 in [0.59-0.62] with rs62355902
[d] rs7721581, representing 17 SNPs which have r^2 in [0.14-0.16] with rs113317823 and $r^2 \leq 0.01$ with rs62355902
[e] rs11949391, representing 5 SNPs
[f] h_1-h_6 are compared to h_0 (the reference haplotype carrying the major alleles of all 5 SNPs).
'1' represents major alleles and '2' minor alleles in each SNP

Figure 1
Click here to download high resolution image

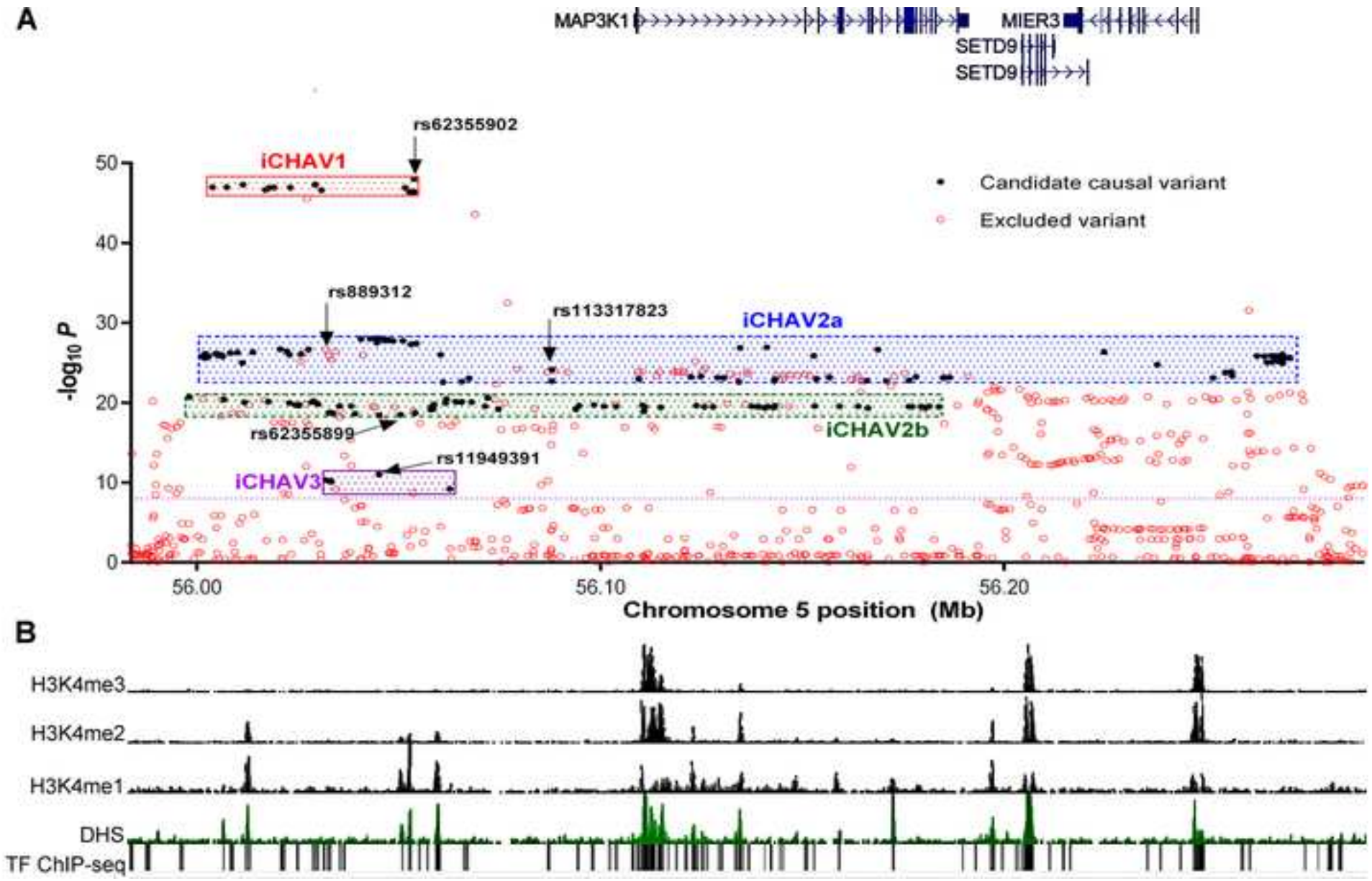


Figure 2

Click here to download high resolution image

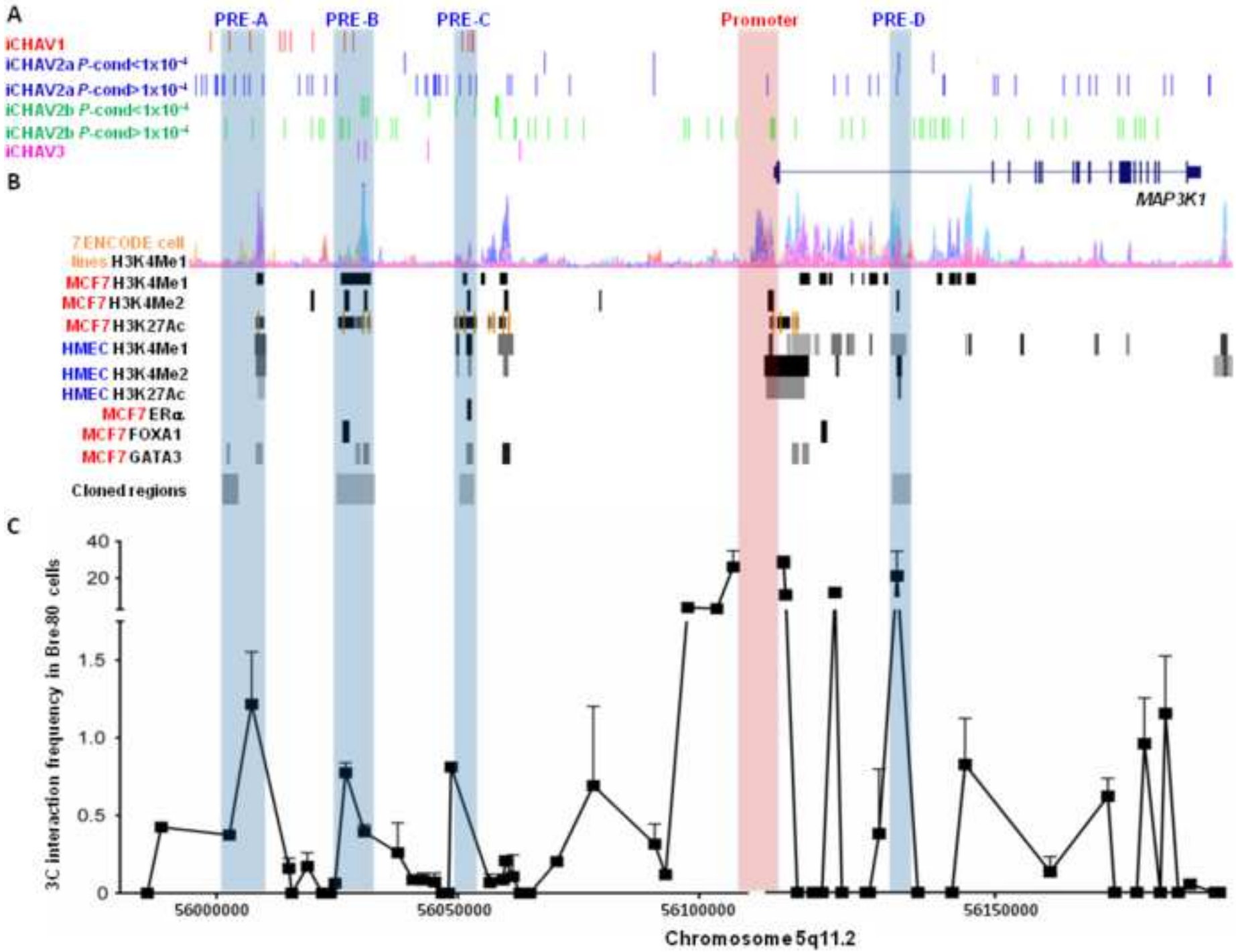


Figure 3

Click here to download high resolution image

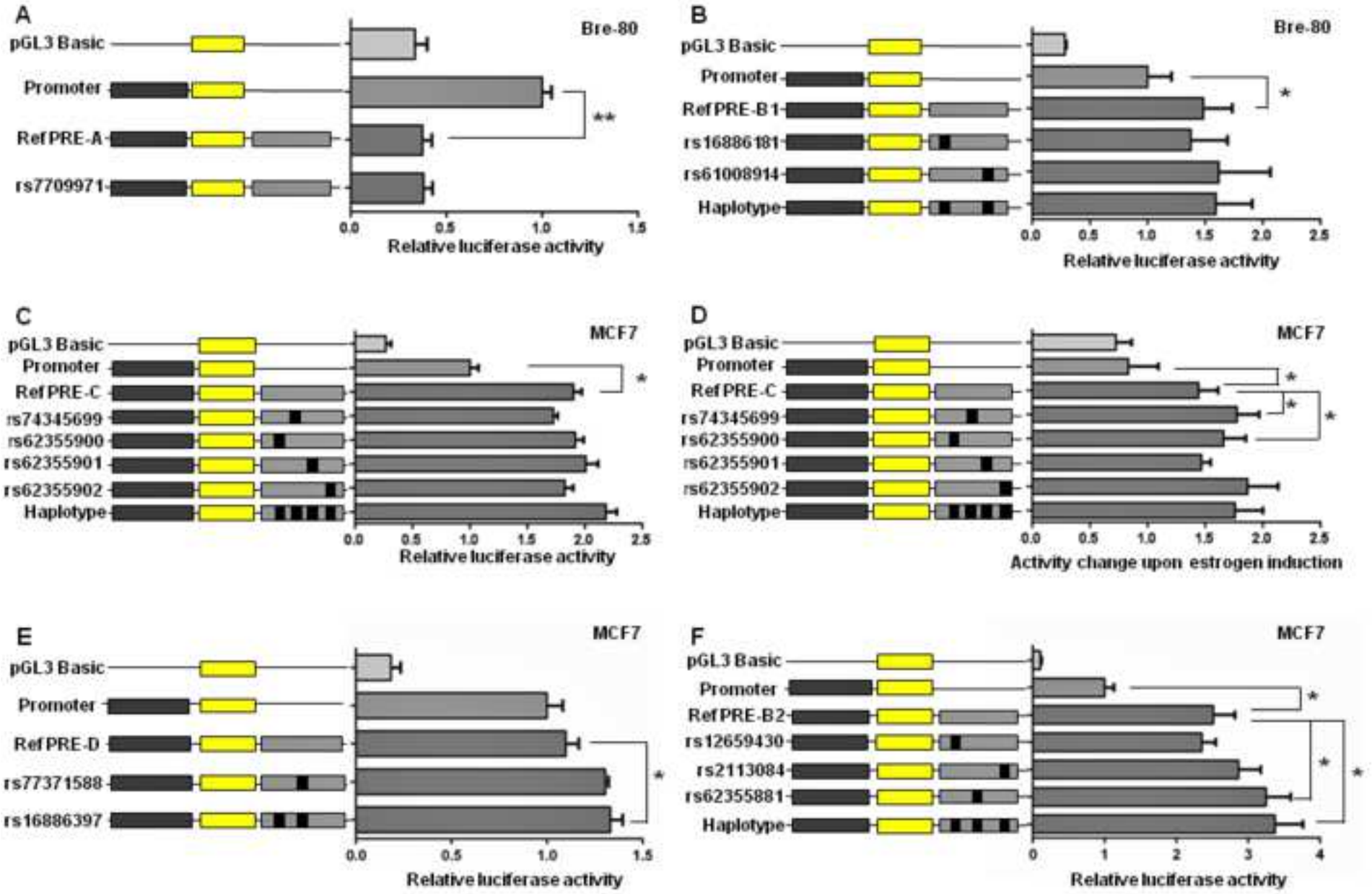
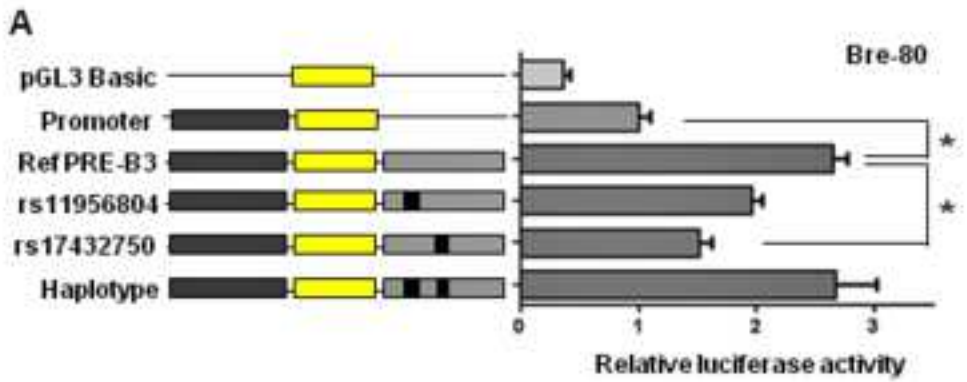
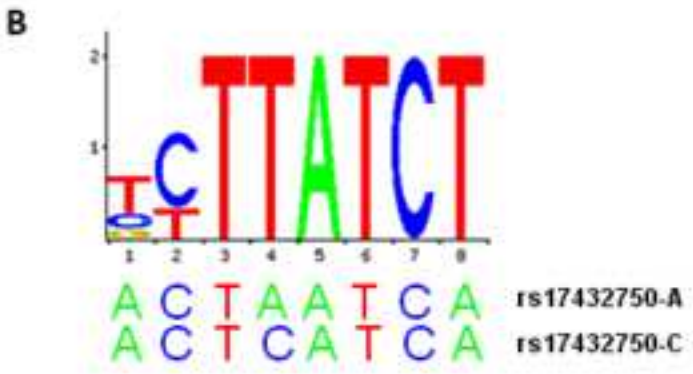
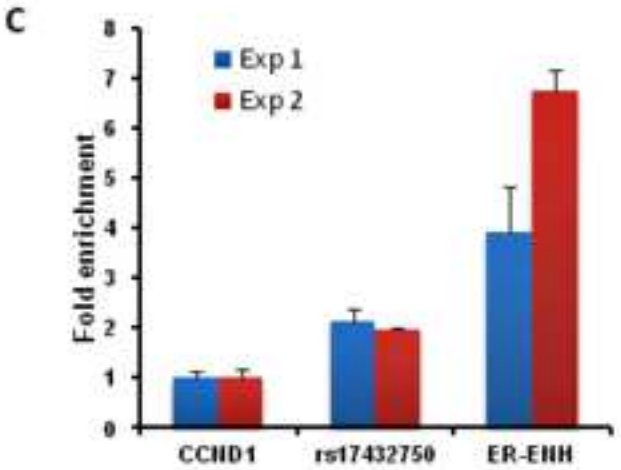
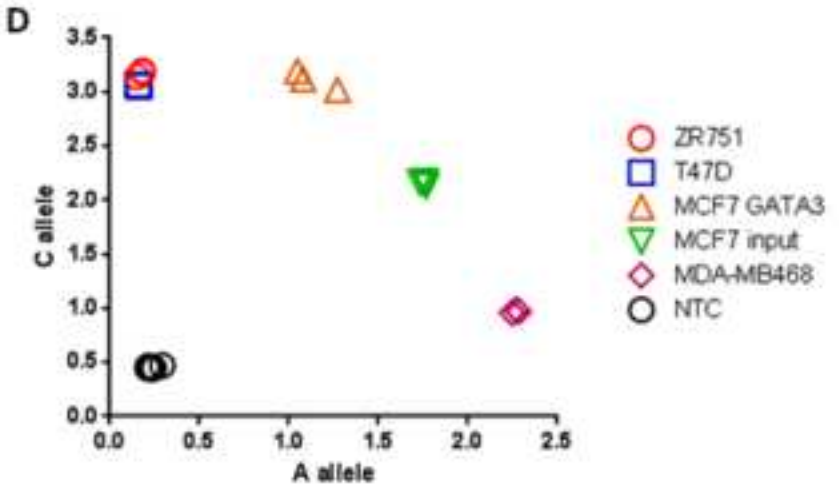


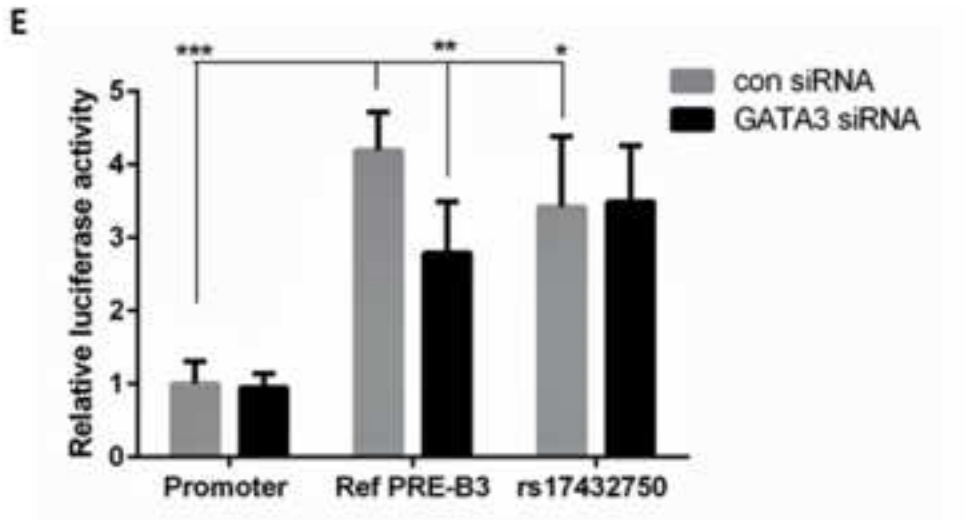
Figure 4
Click here to download high resolution image











Supplementary Data

Supplemental Acknowledgements

Manuscript writing group: DMG, MJM, KM, KAP, KBM, AMD, JDF. Locus SNP selection: MG, ED, AMD. iCOGS genotyping, calling and QC: MJM, SEB, SFN, AG-N, MRA, DH, JB, DCT, DV, FB, FR, SA, CL, CB, DC, JC, JD, CSH, JS, AMD, GC-T, DFE. Imputation: KM, DFE. Statistical analyses and programming: KM, KAP, DB, ED, JT, SK, DFE. Functional analysis and bioinformatics: DMG, SLE, KBM, SC, MO, MJM JAB, KMH, SK, JB and JDF. COGS coordination: PH, DFE, JB, AMD, BCAC coordination: DFE, GC-T, PDP. BCAC data management: MKB, QW. All other authors provided participant samples and phenotype information, and read and approved the manuscript. BCAC thanks all the individuals who took part in these studies and all the researchers, clinicians, technicians and administrative staff who have enabled this work to be carried out. BCAC is funded by Cancer Research UK [C1287/A10118, C1287/A12014] and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). Meetings of the BCAC have been funded by the European Union COST programme (BM0606). This study would not have been possible without the contributions of the following: Andrew Berchuck (OCAC), Rosalind A. Eeles, Ali Amin Al Olama, Zsofia Kote-Jarai, Sara Benlloch (PRACTICAL), Antonis Antoniou, Lesley McGuffog, and Ken Offit (CIMBA), Andrew Lee, Ed Dicks, and the staff of the Centre for Genetic Epidemiology Laboratory, the staff of the CNIO genotyping unit, Sylvie LaBoissière, Frederic Robidoux and the staff of the McGill University and Génome Québec Innovation Centre, the staff of the Copenhagen DNA laboratory, and Julie M. Cunningham, Sharon A. Windebank, Christopher A. Hilker, Jeffrey Meyer and the staff of Mayo Clinic Genotyping Core Facility. 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 - grant #CRN-87521, and the Ministry of Economic Development, Innovation and Export Trade of Quebec – grant # PSR-SIIRI-701. This study makes use of data generated by the Molecular Taxonomy of Breast Cancer International Consortium. Funding for the project was provided by Cancer Research UK and the British Columbia Cancer Agency Branch. The QIMR Berghofer group was supported by project grants from the National Health and Medical Research Council of Australia (1021731, 1058415). The work by KBM, MO, SC and BJAP is supported by Cancer Research UK and the Breast Cancer Research Foundation. DFE is a Principal Research Fellow of CR-UK. GCT is an NHMRC Senior Principal Research Fellow. SLE and JDF are supported by Fellowships from the National Breast Cancer Foundation (NBCF) Australia. The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

BCAC investigating groups:
The Australian Breast Cancer Family Study (ABCFS) would like to thank Maggie Angelakos, Judi Maskiell, Gillian Dite. ABCFS was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Australia Fellow and a Victorian Breast Cancer Research Consortium Group Leader. M.C.S. is a NHMRC Senior Research Fellow and a Victorian Breast Cancer Research Consortium Group Leader. The ABCS study would like to thank Sten Cornelissen, Richard van Hien, Linde Braaf, Senno Verhoef, Laura van 't Veer, Emiel Rutgers. The **ABCS** study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]; BBMRI-NL, which is a Research Infrastructure financed by the Dutch government (NWO 184.021.007); and the Dutch National Genomics Initiative. The ACP study wishes to thank the participants in the Thai Breast Cancer study. Special Thanks also go to the Thai Ministry of Public Health

(MOPH), doctors and nurses who helped with the data collection process. Finally, the study would like to thank Dr Prat Boonyawongviroj, the former Permanent Secretary of MOPH and Dr Pornthep Siriwanarungsan, the Department Director-Generalof Disease Control who have supported the study throughout. The ACP study is funded by the Breast Cancer Research Trust, UK. The BBCC study would like to thank Matthias Rübner, Alexander Hein, Michael Schneider. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCS would like to thank Eileen Williams, Elaine Ryder-Mills, Kara Sargus. The **BBCS** is funded by Cancer Research UK and Breakthrough Breast Cancer and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). The BIGGS would like to thank Niall McInerney, Gabrielle Colleran, Andrew Rowan, Angela Jones. ES is supported by NIHR Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London, United Kingdom. IT is supported by the Oxford Biomedical Research Centre. The BSUCH would like to thank Peter Bugert, Medical Faculty Mannheim. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). The CECILE study would like to thank Claire Mulot and Pierre Laurent-Puig (Biobank Saints-Pères Paris), Pierre Kerbrat (Centre Eugène Marquis, Rennes), Patrick Arveux (Centre Georges François Leclerc, DijonThe **CECILE** study was funded by Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Ligue contre le Cancer Grand Ouest, Agence Nationale de Sécurité Sanitaire (ANSES), Agence Nationale de la Recherche (ANR). The CGPS study would like to thank Staff and participants of the Copenhagen General Population Study. It would also like to thank Dorthe Uldall Andersen, Maria Birna Arnadottir, Anne Bank and Dorthe Kjeldgård Hansen for the excellent technical assistance. The Danish Breast Cancer Group (DBCG) is acknowledged for the tumor information. The CGPS was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council and Herlev Hospital. The CNIO-BCS would like to thank Guillermo Pita, Charo Alonso, Daniel Herrero, Nuria Álvarez, Pilar Zamora, Primitiva Menendez, the Human Genotyping-CEGEN Unit (CNIO). The CNIO-BCS was supported by the Genome Spain Foundation, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI081120). The Human Genotyping-CEGEN Unit (CNIO) is supported by the Instituto de Salud Carlos III. The CTS study would like to thank the CTS Steering Committee: Leslie Bernstein, Susan Neuhausen, James Lacey, Sophia Wang, Huiyan Ma, Yani Lu, Jessica Clague DeHart at the Beckman Research Institute of City of Hope, Dennis Deapen, Rich Pinder, Eunjung Lee, and Fred Schumacher at the University of Southern California, Pam Horn-Ross, Peggy Reynolds, Christina Clarke Dur and David Nelson at the Cancer Prevention Institute of California, and Hoda Anton-Culver, Argyrios Ziogas, and Hannah Park at the University of California Irvine.. The **CTS** was initially supported by the California Breast Cancer Act of 1993 and the California Breast Cancer Research Fund (contract 97-10500) and is currently funded through the National Institutes of Health (R01 CA77398). Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885. HAC receives support from the Lon V Smith Foundation (LVS39420). The ESTHER study would like to thank Hartwig Ziegler, Sonja Wolf, Volker Hermann. The ESTHER study was supported by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe). The GENICA Network would like to thank Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart, and University of Tübingen, Germany; [HB, Wing-Yee Lo, Christina Justenhoven], Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany [YDK, Christian Baisch], Institute of Pathology, University of Bonn, Germany [Hans-Peter Fischer], Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany [Ute Hamann], Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, Germany [TB, Beate Pesch, Sylvia Rabstein, Anne Lotz]; and Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, Germany [Volker Harth]. The **GENICA** was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, the Institute for Prevention and

Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, as well as the Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany. The HEBCS would like to thank Kirsimari Aaltonen, Karl von Smitten, Sofia Khan, Tuomas Heikkinen, Irja Erkkilä. The HEBCS was financially supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (266528), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation. The **HERPACC** was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports, Culture and Technology of Japan, by a Grant-in-Aid for the Third Term Comprehensive 10-Year Strategy for Cancer Control from Ministry Health, Labour and Welfare of Japan, by Health and Labour Sciences Research Grants for Research on Applying Health Technology from Ministry Health, Labour and Welfare of Japan and by National Cancer Center Research and Development Fund. The HMBCS would like to thank Peter Hillemanns, Hans Christiansen and Johann H. Karstens. . The **HMBCS** was supported by a grant from the Friends of Hannover Medical School and by the Rudolf Bartling Foundation. Financial support for **KARBAC** was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and and Bert von Kantzows foundation. The KBCP would like to thank Eija Myöhänen, Helena Kemiläinen. The **KBCP** was financially supported by the special Government Funding (EVO) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations, and by the strategic funding of the University of Eastern Finland. The kConFab/AOCS study would like to thank Heather Thorne, Eveline Niedermayr, all the kConFab research nurses and staff, the heads and staff of the Family Cancer Clinics, and the Clinical Follow Up Study (which has received funding from the NHMRC, the National Breast Cancer Foundation, Cancer Australia, and the National Institute of Health (USA)) for their contributions to this resource, and the many families who contribute to kConFab**kConFab** is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. **LAABC** is supported by grants (1RB-0287, 3PB-0102, 5PB-0018, 10PB-0098) from the California Breast Cancer Research Program. Incident breast cancer cases were collected by the USC Cancer Surveillance Program (CSP) which is supported under subcontract by the California Department of Health. The CSP is also part of the National Cancer Institute's Division of Cancer Prevention and Control Surveillance, Epidemiology, and End Results Program, under contract number N01CN25403. The LMBC would like to thank Gilian Peuteman, Dominiek Smeets, Thomas Van Brussel and Kathleen Corthouts. LMBC is supported by the 'Stichting tegen Kanker' (232-2008 and 196-2010). Diether Lambrechts is supported by the FWO and the KULPFV/10/016-SymBioSysII. The MARIE study would like to Judith Heinz, Nadia Obi, Alina Vrieling, Sabine Behrens, Ursula Eilber, Muhabbet Celik, Til Olchers and Stefan Nickels. The **MARIE** study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419], the Hamburg Cancer Society, the German Cancer Research Center and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. The MBCSG would like to thank Siranoush Manoukian, Bernard Peissel, Giulietta Scuvera and Daniela Zaffaroni of the Fondazione IRCCS Istituto Nazionale dei Tumori (INT); Bernardo Bonanni and Irene Feroce of the Istituto Europeo di Oncologia (IEO) and Loris Bernard and the personnel of the Cogentech Cancer Genetic Test Laboratory. **MBCSG** is supported by grants from the Italian Association for Cancer Research (AIRC) and by funds from the Italian citizens who allocated the 5/1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale Tumori, according to Italian laws (INT-Institutional strategic projects “5x1000”). The **MCBCS** was supported by the NIH grants CA128978, CA116167, CA176785 an NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation and a generous gift from the David F. and Margaret T. Grohne Family Foundation and the Ting Tsung and Wei Fong Chao Foundation. **MCCS** cohort recruitment in the was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. The MEC was support by NIH grants CA63464, CA54281, CA098758 and CA132839. The MTLGEBCS would like to thank Martine Tranchant (CHU de Québec Research Center), Marie-France Valois, Annie Turgeon and Lea Heguy (McGill University Health Center, Royal Victoria Hospital; McGill University) for DNA extraction, sample management and skillful technical assistance. J.S. is Chairholder of

the Canada Research Chair in Oncogenetics. The work of **MTLGEBCS** was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701. The MYBRCA study would like to thank Phuah Sze Yee, Peter Kang, Kang In Nee, Kavitta Sivanandan, Shivaani Mariapun, Yoon Sook-Yee, Daphne Lee, Teh Yew Ching and Nur Aishah Mohd Taib for DNA Extraction and patient recruitment. **MYBRCA** is funded by research grants from the Malaysian Ministry of Science, Technology and Innovation (MOSTI), Malaysian Ministry of Higher Education (UM.C/HlR/MOHE/06) and Cancer Research Initiatives Foundation (CARIF). Additional controls were recruited by the Singapore Eye Research Institute, which was supported by a grant from the Biomedical Research Council (BMRC08/1/35/19/550), Singapore and the National medical Research Council, Singapore (NMRC/CG/SERI/2010). The NBCS was supported by grants from the Norwegian Research council, 155218/V40, 175240/S10 to ALBD, FUGE-NFR 181600/V11 to VNK and a Swizz Bridge Award to ALBD. The NBHS_TN study would like to thank participants and research staff for their contributions and commitment to this study. The NBHS was supported by NIH grant R01CA100374. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The OBCS study would like to thank Meeri Otsukka, Kari Mononen, Mervi Grip. The **OBCS** was supported by research grants from the Finnish Cancer Foundation, the Academy of Finland (grant number 250083, 122715 and Center of Excellence grant number 251314), the Finnish Cancer Foundation, the Sigrid Juselius Foundation, the University of Oulu, the University of Oulu Support Foundation and the special Governmental EVO funds for Oulu University Hospital-based research activities. The OFBCR study would like to thank Teresa Selander, Nayana Weerasooriya. The Ontario Familial Breast Cancer Registry (OFBCR) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ORIGO study would like to thank E. Krol-Warmerdam, and J. Blom for patient accrual, administering questionnaires, and managing clinical information. The LUMC survival data were retrieved from the Leiden hospital-based cancer registry system (ONCDOC) with the help of Dr. J. Molenaar. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS study would like to thank Louise Brinton, Mark Sherman, Neonila Szeszenia-Dabrowska, Beata Peplonska, Witold Zatonski, Pei Chao, Michael Stagner. The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. The pKARMA study would like to thank The Swedish Medical Research Counsel. The pKARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. The RBCS study would like to thank Petra Bos, Jannet Blom, Ellen Crepin, Elisabeth Huijskens, Annette Heemskerk, the Erasmus MC Family Cancer Clinic. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318). The SASBAC would like to thank The Swedish Medical Research Counsel. The SASBAC study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. The SBCGS would like to thank participants and research staff for their contributions and commitment to this study. The SBCGS was supported primarily by NIH grants R01CA64277, R01CA148667, and R37CA70867. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The scientific development and funding of this project were, in part, supported by the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network U19 CA148065. The SBCS would like to thank Sue Higham, Helen Cramp, Sabapathy Balasubramanian, Ian Borck and Dan Connley. The **SBCS** was supported by Yorkshire Cancer Research S295, S299, S305PA and Sheffield Exeprimental Cancer Medicine Centre. The **SCCS** is supported by a grant from the National Institutes of Health (R01 CA092447). Data on SCCS cancer cases used in this publication were provided by the Alabama Statewide Cancer Registry; Kentucky Cancer Registry, Lexington, KY; Tennessee Department of Health, Office of Cancer Surveillance; Florida Cancer Data System; North Carolina Central Cancer Registry, North Carolina Division of Public Health; Georgia Comprehensive Cancer Registry; Louisiana Tumor Registry; Mississippi Cancer Registry; South Carolina Central Cancer Registry; Virginia Department of Health, Virginia Cancer Registry; Arkansas

Department of Health, Cancer Registry, 4815 W. Markham, Little Rock, AR 72205. The Arkansas Central Cancer Registry is fully funded by a grant from National Program of Cancer Registries, Centers for Disease Control and Prevention (CDC). Data on SCCS cancer cases from Mississippi were collected by the Mississippi Cancer Registry which participates in the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or the Mississippi Cancer Registry.The SEARCH study would like to thank The SEARCH and EPIC teams. **SEARCH** is funded by a programme grant from Cancer Research UK [C490/A10124] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. **SEBCS** was supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0000347). **SGBCC** is funded by the National Medical Research Council start-up Grant and National University Cancer Institute Singapore (NCIS) Centre Grant. The study would like to thank the participants and research coordinator Kimberley Chua. Additional controls were recruited by the Singapore Consortium of Cohort Studies-Multi-ethnic cohort (SCCS-MEC), which was funded by the Biomedical Research Council, grant number: 05/1/21/19/425. The SKKDKFZS study thanks all study participants, clinicians, family doctors, researchers and technicians for their contributions and commitment to this study. SKKDKFZS is supported by the DKFZ. The SZBCS was supported by Grant PBZ_KBN_122/P05/2004. The TBCS was funded by The National Cancer Institute Thailand. The **TNBCC** was supported by: a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation, the Stefanie Spielman Breast Cancer fund and the OSU Comprehensive Cancer Center, the Hellenic Cooperative Oncology Group research grant (HR R_BG/04) and the Greek General Secretary for Research and Technology (GSRT) Program, Research Excellence II, the European Union (European Social Fund – ESF), and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - ARISTEIA. The **TWBCS** is supported by the Taiwan Biobank project of the Institute of Biomedical Sciences, Academia Sinica, Taiwan. The UKBGS study would like to thank Breakthrough Breast Cancer and the Institute of Cancer Research for support and funding of the Breakthrough Generations Study, and the study participants, study staff, and the doctors, nurses and other health care providers and health information sources who have contributed to the study. We acknowledge NHS funding to the Royal Marsden/ICR NIHR Biomedical Research Centre. The UKBGS is funded by Breakthrough Breast Cancer and the Institute of Cancer Research (ICR), London. ICR acknowledges NHS funding to the NIHR Biomedical Research Centre.